Visualizing Language in Hollywood Screenplays

Joseph Rafla

Department of Languages, Literatures, and Cultures

Digital Humanities

McGill University, Montréal

July 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Arts

Table of Contents

Abstract

The language of the film screen is ubiquitous. Hollywood's wide release feature films, like *The Godfather* (1972) or *Wonder Woman* (2017), are experienced by millions of people all over the world. However, before a movie makes it on screen, it has to make it as a screenplay, the blue print of a movie. With films costing 10's of millions of dollars to produce, it helps to have the best blueprint possible. Increasingly, computers are capable of linguistic analysis at scale. "It is the work of humanities scholars to trace patterns in language, in narratives, in art, and other cultural artifacts" (Armoza 7). In my research, I look at the language of one specific type of scene which is common to all screenplays. It's the scene that changes everything: when Alice drinks the potion, shrinks to thumb size, and jumps through the keyhole, when Don Corleone gets gunned down, or when Forrest Gump runs so hard he breaks his leg braces. These scenes change what the story is about. I use computational methods to look at a small collection of such events, and attempt to visualize their sentiment language as a way of seeing patterns in a screenplay, a process, which is helpful in the development of screenplays. The computational techniques I use, from corpus linguistics, are not specific to screenplays as they can be applied to any type of text. The computation's outputs led me to the making of the Sentiment Pulse Visualization, or SentiPulse Viz, a data driven visualization which 'maps' the changes in the sentiment represented in text of events. The SentiPulse viz uses, as a visual model, the graph of Pulsar CP 1919, the first recorded black hole, which was discovered in a Cambridge observatory in 1963.

Résumé

Le langage de l'écran de film est omniprésent. Les longs métrages hollywoodiens, comme The Godfather (1972) ou Wonder Woman (2017), sont vécus par des millions de personnes dans le monde entier. Cependant, avant qu'un film ne soit affiché à l'écran, il doit être transformé en scénario, l'empreinte bleue d'un film. Avec des films coûtant des dizaines de millions de dollars à produire, il est utile d'avoir le meilleur schéma possible. Les ordinateurs sont de plus en plus capables d'analyses linguistiques à grande échelle. «Les érudits en sciences humaines ont pour tâche de tracer des modèles linguistiques, narratifs, artistiques et autres artefacts culturels» (Armoza 7). Dans mes recherches, je regarde le langage d'un type spécifique de scène qui est commun à tous les scénarios. C'est la scène qui change tout: quand Alice boit la potion, réduit à la taille du pouce et saute à travers le trou de la serrure, lorsque Don Corleone est abattu ou lorsque Forrest Gump s'exécute tellement, il se casse les attelles. Ce sont des scènes qui changent le sujet de l'histoire. J'utilise des méthodes de calcul pour examiner une petite collection d'événements de ce type et tenter de visualiser leur langage des sentiments comme un moyen de voir les motifs d'un scénario, processus utile pour le développement de scénarios. Les techniques de calcul que j'utilise, de la linguistique de corpus, ne sont pas spécifiques aux scénarios car elles peuvent être appliquées à tout type de texte. Les résultats du calcul m'ont conduit à la réalisation de la visualisation de sentiment, ou SentiPulse Viz, une visualisation pilotée par les données qui "mappe" les changements du sentiment représentés dans le texte. Le SentiPulse utilise, comme modèle visuel, le graphique de Pulsar CP 1919, le premier trou noir enregistré, découvert dans un observatoire de Cambridge en 1963.

Acknowledgments

I would like to acknowledge the following people for the generous assistance and support they provided me in doing the research and writing of this thesis:

Thesis Supervisor: Stéfan Sinclair, Associate Professor, McGill University. Professor Sinclair's patience and generosity in allowing me to pursue my research findings is very much appreciated, as was his direction in using D3 and Observable's platform. "The Art of Literary Text Analysis" was an invaluable resource.

Text analysis: Andrew Piper, Professor, McGill University. Professor Piper's classes in text analysis covered a range of concepts and techniques based on corpus linguistics, which have aided me in navigating the computational analysis of the corpus. His book *Enumerations*, has been helpful in understanding how to make interpretations of data.

Web-based site design: Cecily Raynor, Assistant Professor, McGill University, through her class presentations and lectures showed me a range of diverse website designs, data visualization, and their components, which has helped in the design of my visualizations.

Code tutorials and assistance, Marc Cataford, Phd. Computer Engineering, McGill University, helped me adapt Python code from NLTK, and helped me understand key computing concepts.

Michael Vastola, Web Engineer, BS in Computer Engineering, Tufts University, gave me tutorials on the use of D3, data driven documents, and their application on the Observable platform, and helped me sort out the code used on the Observable platform.

Introduction

As a screenplay editor in the Canadian film industry, I was assigned the job of understanding and reporting on the construction of hundreds of fledgling screenplays, which in turn necessitated an analysis of the screenplay's narrative structure.

Based on years of close reading screenplays, I began to ask questions such as, how common are certain events in screenplays, are all events equally important to the narrative's shape, could events have a particular linguistic fingerprint, i.e. do they share identifiable linguistic features which set them apart from the rest of the screenplay? Furthermore, can salient linguistic patterns, if they exist, be made explicit through a visual representation?

"It is the work of humanities scholars to trace patterns in language, in narratives, in art, and other cultural artifacts" (Armoza 7). Jonathon Armoza's careful and considered research into topic modelling has been a guiding example of how a humanist who uses computational methods can share their findings in a highly accessible fashion. As he points out, the advent of digital technology has allowed scholars to examine cultural artifacts, in this case screenplays, where "there is some description of those objects of study that lay beyond our faculties" (Armoza 7). The ability to search for patterns in language such as parts of speech, sentiment markers, clusters, etc. is possible in my thesis due to recent advances in open source, pre-existing code that is available to create web-based visualizations through such sites as observablehq.com. Observable's diverse library of data visualizations was instrumental in creating the visualizations in chapter 4.

The research-in-progress of Andrew Piper and Eva Portelance, summarized in their poster, "Understanding Narratives: Computational approaches to detecting narrative frames" was

my first introduction to narrative computational analysis. Piper writes that, “Understanding narrative structure at a large scale remains a challenging problem within the field of cultural analytics and computational linguistics” (Piper, Poster). Piper and Portelance’s narrative epistemology is based on the work of Gerald Prince, who defines a narrative thus from Piper’s poster, “...narrative is essentially a mode of verbal presentation and involves the LINGUISTIC recounting or telling of EVENTS.’ -Dictionary of Narratology, by Gerald Prince, 2003” (Piper, Poster). The telling of events is a core function of narrative according to Gerald Prince, Mieke Bal, and Robert McKee, among other narratologists. Prince’s definition above is meant to include all events in all narratives. So, if a lead character brushes her teeth, and then goes on to assassinate the prime minister, Prince’s definition categorizes both equally as events. However, the impact of these two events on the narrative are not equal; one is decisive in shaping the narrative, while the other is often incidental. Events that change what the story is about is a key feature of the kinds of events I will analyse.

William Labov, a sociolinguist, identifies an event as the raison d’être of the narrative. The event in Labov’s analysis defines the reason why the narrative is being recounted, as well as its “tellability”, i.e. why it matters, or its noteworthiness (Labov 367). Event therefore is a central concept, not only in narrative theory, but as we shall see, in several other humanities disciplines.

However, it is only recently that the event has become a more central concern in the study of narrative theory, “The concept of event has become prominent in recent work on narratology” (Huhn 80). In the writing of screenplays, the event is the core unit through which the screenwriter propels the narrative (McKee 33).

Further research led to “The Living Handbook of Narratology”, an open access publication, whose editor-in-chief is Peter Hühn of the Interdisciplinary Centre for Narratology, University of Hamburg. The Hamburg scholars take Prince’s definition of event, which they call a Type I event, applicable in general to all changes of state in all narrative, and differentiate it from a Type II event, a specific type of change of state, which must occur at least once for the writing to be categorized as narrative. It is on the basis of this distinction between Type I and Type II events that I was able to clearly and consistently identify and tag events and nonevents in my corpus of screenplays.

The purpose of the computational analysis of events and nonevents, the content of chapter three, is to search for salient linguistic patterns which I then use in chapter four to make visualizations that differentiate events from nonevents. The computational methods, based on corpus linguistics, are fairly straight forward: most frequent terms, clustering, parts of speech, and sentiment analysis. I chose these because I am familiar with them through text analysis courses.

The work of “Scientific American’s” chief illustrator, Jen Christensen, set the stage for my research into spatialization and the use of metaphor in visually representing abstract linguistic data. My goal in the final chapter is to use an existing data visualization platform, ObservableHQ, to create a visualization of sentiment data in events/nonevents using landscape as a metaphorical representation: the SentiPulse Visualization.

Chapter 1

1.0 Introducing the Narrative Event

Chapter 1 deals with the following two questions: 1. what is an event and 2. what is its purpose in the narrative of a screenplay? By posing the questions in this way I am emphasizing two critical aspects. The first is the operational definition of a narrative event. The goal of such a definition is to provide a clear and concise set of criteria for choosing the scenes that will inform this study's corpus. The second question addresses the centrality of the event in the meaning and construction of a screenplay narrative. However, I will start by defining what type of screen narratives I will address. The following definition of the arch-plot is from Robert McKee's *Story:*

> Classical design means a story built around an active protagonist who struggles against primarily external forces of antagonism to pursue his or her desire, through continuous time, within a consistent and causally fictional reality, to a closed ending of absolute irreversible change. (Mckee 45)

The choice of the arch-plot screenplay as the subject of this thesis came down to choosing the most language-rich type of screenplay. That is, the type of screenplay where narrative descriptions are explicit, detailed, and complete, as is dialogue. The justification for choosing this type of screenplay is discussed in section 2.0 "The Films and their Data".

In French structuralist narrative theory (1966-1980), the narrator is a constitutive element of the narrative (Schmid 17)."[The Narrator] symbolizes the epistemological view familiar to us since Kant that we do not grasp the world as it is in itself, but as it passed through the medium of a contemplative mind" (Schimd 18).

A turn away from the narrator as a constitutive element of narrative occurred in poststructuralist narratology (1980-1990), which was "dominated by…a widening of narratology's scope beyond literary narrative" (Meister LHN). From the poststructuralist perspective, the constitutive elements of narrative are temporality, and change of state (Schmid 18).

The narrative event, as defined by Miek Bal: "is a transition from one state to another state" (Bal 5). Of the two constitutive elements, temporality and change of state, it is the second, a change of state, that is of interest in this study, because it will form the operational definition which I will use to identify the data for the computational analysis that is to follow. Change of state implies temporality because change is a process, and of necessity takes place within a timeframe (Bal 7). That a story has a beginning, a middle, and an end, implies the temporality which harkens back to Aristotle's definition of Greek tragedy.

In Aristotelean terms, the type of event that I researched is closest to the "peripeteia", or the moment of reversal, the point at which the story's trajectory is altered; the peripeteia is often associated with the climax event after which the story moves to its denouement (Brittanica).

1.1 Aristotle

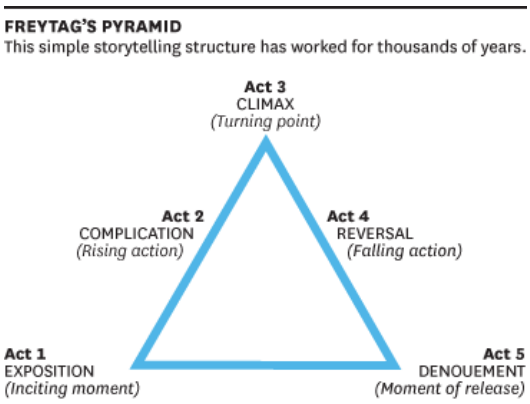
According to Aristotle, "Tragedy is an imitation of an action that is complete, and whole…A whole is that which has a beginning, middle, and an end" (Aristotle 14).
J.C. Meister points to Aristotle's *Poetics* as a "precursor" to the narratological modelling of narrative, "Aristotle's Poetics presented a second criterion that has remained fundamental for the understanding of narrative: the distinction between the totality of events taking place in a depicted world and the *de facto* narrated plot or *muthos*. He pointed out that the latter is always a

construct presenting a subset of events, chosen and arranged according to aesthetic considerations" (Meister LHN, emphasis mine). This is a clear definition of event as a constitutive element of drama.

Aristotle's definition is relevant in that he is addressing dramatic structure through events, "Tragedy is an imitation not only of a complete action, but of events inspiring fear or pity" (Aristotle 18). McKee writes that screenplays contain 40-60 events, "for a typical film, the writer will choose forty to sixty Story Events or, as they're commonly know, scenes" (McKee 35). Although events are often contained in a scene, they can be a part of and not the whole scene.

I analyze what are commonly called major story events (McKee 35); the type of events that change the story's trajectory, or change what the story is about: end of act turn, inciting incident, climax, etc. In Aristotle's terminology such scenes are more broadly called peripeteia: "In classical tragedy (and hence in other forms of drama, fiction, etc.): a point in the plot at which a sudden reversal occurs" (OED peripeteia). The clear, consistent identification of such scenes is foundational to the study of the corpus I will examine.

The earliest films, such as Georges Méliès' *Le Voyage Dans la Lune* (1902), based on two narrative texts: Jules Verne's *From the Earth to the Moon* and H. G. Wells' *The First Men in the Moon,* amount to staged dramas captured on film (Méliès). From the beginning the film narrative has drawn on the novel's narrative in terms of its structure and other story elements.



credit (Nexus MediaWorks)

Figure 1. Freytag's visual representation of Aristotle's narrative structure of Greek tragedy.

It is not an exaggeration to say that the modern screenplay narrative structure has its roots firmly planted in antiquity. Freytag's Pyramid (Fig. 1), the visual representation of Aristotle's design of Greek tragedy, has at each of its points an event, which changes the direction of the plot. Another way to say this is that these critical events change what the story is about. What the above diagram labels inciting moment is generally referred to in screenplay terminology as the inciting incident (McKee 181), "In most cases, the inciting incident is a single event that either happens directly to the protagonist or is caused by the protagonist" (McKee 190). The inciting incident propels and gives life, is its raison d'être in Labov's epistemology, to the ensuing narrative.

For example in the film *"The Godfather"* (1972), the inciting incident occurs on page twenty-four of the screenplay (IMSDb). Don Corleone rejects Sollozo's request to help protect his fledgling drug racket, which is considered an act of aggression by his fellow mafiosi. This in turn leads to the event that sets the story on a new course: Don Corleone is gunned down, which changes what the story is about; the story is no longer about a happy, prosperous mafia family, but changes its direction to become a story about revenge and succession. In terms of screenplay

structure, it is the turn that ends the first act, and sets Michael on his course to become the new Godfather. The end of act turn is an example of the kind of event whose language this study seeks to examine .

In contrast to structuralist narrative theory, which came to be called “narratology”, a term originally coined by Tzevan Todorov in *The Grammar of the Decameron* (1969), poststructuralism focused on the content of the narrative rather than the narrator-as-authority (Meister LHN). The foundation of Todorov’s work was laid in Vladimir Propp’s *Morphology of the Folktale* (1928).

1.2 Event in Post-structural Narrative Theory

Further to Bal’s definition of event as change of state, she identifies event, not as an atomistic narrative element, but as the building block of the content of narrative, “That with which the narrative text is concerned, the contents it conveys to its readers, is a series of connected events caused or experienced by actors presented in a specific manner” (Bal 8), In addition, according to Wolf Schmid, “The minimal condition of narrativity is that at least one change of state must be represented” (Schmid 19). But even that is not entirely accurate. Take for example Samuel Beckett’s *Waiting for Godot,* a drama in which no story-turning event ever occurs, only the anticipation of such a possible event occurs. However, the raison d’être of the play is the anticipated event, one event which never takes place.

Event as change of state is a broad, inclusive definition. Narrative theorists such as Bal's or Prince’s seek to define event in terms which apply to all narratives, whether in the form of a novel, play, or film. Here for example is Prince’s definition of event, “A change of STATE

manifested in DISCOURSE by a PROCESS STATEMENT in the mode of Do or Happen" (PRINCE Dictionary 28). However, in seeking to operationalize this definition, i.e. for the purposes of identifying what is an event and what is a non-event in screenplays, it is too broad: sneezing is an event, sleeping is an event, crossing the street is an event. Such a definition may indeed be universal, but in terms of analyzing the language of change-inducing events in screenplays it is too broad. This study seeks to understand events that are of significance to the story's trajectory, and so not every change of state is of interest. The definition of event needs to go beyond the singularly linguistic.

The event, as defined in poststructuralist narratology, through the works of Gerald Prince, Mieke Bal, Peter Hühn, and Wolf Schimd among others, will be examined, and an amalgam of their varied definitions will be crafted to allow for the clear and consistent gathering of the data that will power the computational analysis to follow.

A precise, operational definition of narrative event is crucial in the process of choosing the corpus, because of the huge variety of how events are represented in screenplays, and crucially because the choice of events will form the data for computational analysis; if the data is not precisely, coherently, and consistently defined, its computation is likely to produce incoherent outputs.

Both Bal's and Prince's definitions result in a potentially enormous number of events in any given narrative. Bal identifies three criteria to narrow the scope of an event's definition. The first is change as expressed by a phrase's verb. She distinguishes between "John is ill" and "John falls ill", where the first phrase describes a condition while in the second the verb "falls" implies a process of change (Bal 155). Her two other limiting criteria are "choice"

and “confrontation”. Briefly, choice is based on the idea that an event is either “functional” or “non-functional”. A functional event presents the actor with a choice, one that is either realized or not realized. Her final criteria is “confrontation”, which is defined as a linguistic argument: “Every phase of the fabula — every functional event — has three components: two actors and one action; stated in the logical terms, two arguments and one predicate” (Bal 157). Bal does provide sample sentences, but they are not examples of actual narratives and are therefore examples without context. Bal acknowledges this when she makes the point that it is only in context, “only in a series that events become meaningful for the further development of the fabula” (Bal 156). In this quote Bal points to an aspect of event that bears further inquiry when she indicates that only in a series do events become meaningful for the trajectory of the narrative.

A screenplay is by definition a series of connected events; however, in screenplays there are singular events that completely alter the development of the fabula, e.g. an end of act turn such as the attempted assassination of Don Corleone in *The Godfather* (1972). This type of event stands alone, in the sense that it represents the actions that lead to a change in what the story is about. Without this kind of event the story does not turn, does not develop further. In other words an event, one event, can and often does, as Ball writes, “become meaningful for the further development of the fabula” in the types of screenplays I am studying.

While Bal’s definition narrows the concept of event as a series that is *“meaningful for the development of the plot”* [emphasis mine], it does not help us model it, because it does not specify an interpretation that takes into account the intradeigetic context of the kind of event I’m concerned with, i.e. does not address how the event alters the trajectory of the story.

> The linguistic form in which the information is embodied can be an indication, but it is not always decisive. Furthermore, the general assumption that every event is indicated by a verb of action does not work either. It is, of course, possible to restate every event so that a verb of action appears in the sentence…This provides a convenient means of making explicit any implicit relationships between facts, and can lead to a preliminary selection of events (Bal 156).

Here Bal is defining event in linguistic terms, but in the previous quote she indicates a heuristic process as well, when she says that events are meaningful for the development of the plot. She addresses the linguistic form in which the information of an event is represented, and talks about how events become meaningful for the story's trajectory. Determining meaning is a heuristic process.

Bal's definition of event, as well as Prince's definition is so broad that all events are lumped together. For the purposes of understanding narrative structure, these accurate and inclusive definitions are of limited use in this study. A different type of filter is required.

What is the difference between defining a term such as 'event', and making it operational? For example, Bal and Prince define 'event' quite thoroughly, but their definitions are made to be as broad as possible; they are meant to capture every verb-induced event in every type of narrative. To make it operational, a clear set of criteria is required, which is something a definition does not fully address.

Because not all events are equal in terms of their impact on the narrative, one that drives the story's premise is what might be called a primary event. In screenplay parlance, they are turning point events, pinch point events, inciting incident events, or resolution events (Mckee 181, 206, 208). Thus far, the critical aspect that is missing from the definition of event is how to make its definition operable in a computational analysis. The research in computational narratology by J.C. Meister, whose lectures I attended at the Digital Humanities 2017 conference

at McGill University, opened the door to the clear and accessible work of the Hamburg narratology scholars.

1.3 The Hamburg School

The recent scholarly and computational work of the "Hamburg Narratology Research Group (http://www.lhn.uni-hamburg.de) which includes scholars such as Peter Hühn, Jan Christoph Miester, and Wolf Schimd, among others, provides a concept of event that proves usable in establishing a model upon which to cull data from the screenplays. The following section draws on their work as it pertains to this study.

When I came across the Hamburg Narratology Research Group's distinction between different types of events, I found scholars who were concerned with a focus on events as linguistic constructs, but also events as plot altering elements of story. This had a significant impact on my research. Given the clear and precise criteria laid out by the Hamburg scholars, I have been able to gather the kind of data I want to study using a precise and consistent categorization in the context of a computational analysis.

1.4 The Event that Makes a Difference

In his "Event and Eventfulness", narratologist Peter Hühn makes a distinction between two types of event. He calls them Type I and Type II events, "Event I involves all kinds of change of state…" and is a familiar restatement of Prince's linguistic definition of event, "A change of state manifested in discourse by a process of statement in the mode of 'do' or 'happen'" (*Dictionary* Glossary). In Hühn's words, "In language, a Type I event is expressed by the difference of predicates" (Hühn 160). It is a linguistic definition that is broadly applicable,

objective and verifiable, "…a type of narration that can be described linguistically and manifests itself in predicates that express changes…" (Hühn 159). An example of a study which uses Type I events, is "Understanding Narrative: Computational approaches to detecting narrative frames", by Andrew Piper and Eva Portelance. This ongoing study looks at scene changes as a "disruption in the sequence of events" (Piper and Portelance). Because the study is focused on scene changes it does not directly analyze events, but uses change of state, i.e. their disruption as part of what defines a scene change. Event here is of the Type I, and is linguistically based: "The relationship between and AGENT of an event (VERB) and its THEME (object)." (Piper and Portelance). This is a use of Type I event as a component signal in the understanding of what Piper and others call "narrative frames".

Hühn identifies a second type of event which he calls a Type II event. He starts by contrasting the two types of event: "The two types of event correspond to broad and narrow definitions of narrativity, respectively: narration as the relation of changes of any kind [Type I] and narration as the representation of changes with certain qualities [Type II]" (Hühn 159). Type I events refer to all kinds of change of state, whereas Type II events refers to a change of state which fulfills a set of conditions, " e.g., of being a decisive, unpredictable turn in the narrated happenings, a deviation from the normal, expected course of things, as is implied by event in everyday language" (Hühn 160). Furthermore, unlike Type I, identifying Type II events is a matter of interpretation. The focus on Type II events in my study is in part motivated by the goal of assisting the development of fully realized screenplays. It is Type II events that often have the greatest impact, produce the greatest audience engagement, and provide a structure that often defines a screenplay in terms of its plot (McKee 44).

1.5 Type II Event: the Criteria

More specifically "The event has to be defined as a change of state that fulfills certain conditions" (Schmid 24). Schmid first outlines what he calls basic requirements of an event:

1) Factual or Real: in the context of the story (i.e. narrative content) the event cannot be imagined, or desired, or dreamed. "However, the real acts of wishing, imaging, or dreaming can qualify as [Type II] events" (Schmid 24).
2) Resultative: minimally, the event must have a beginning and an end within the story's beginning and end. So an event cannot be entered in medias res, nor can its ending be implied. That is, for the purposes of this thesis, an event needs to be stated explicitly and fully. Screenwriters use "FADE IN", or "CUT", and other camera directions in the beginning, middle, or end of an event to allude to the action instead of stating it in language. This is a time-honoured cinematic writer's technique, which when used makes the event unsuitable for linguistic analysis. These two criteria, real and resultative, are common to all Type II events.

A Type II event, as defined by Hühn et al., admits of gradation; it is more or less eventful depending to what degree it fulfills the following five criteria established by Wolf Schimd. The first two criteria, relevance and surprise, are considered the minimum criteria to identify a Type II event. (It is described in the next paragraph.) However, in defining event for the purpose of this thesis, the event must meet all of the following criteria and in addition it must alter the trajectory of the story. This last criteria is one that I have added to further narrow and specify the type of event I am tagging for computation.

The following is a list of the relevant criteria to be used in tagging an event:

1. Relevance: the change of state must be relevant to the narrative, "Eventfulness increases in conjunction with the degree to which the change of state is felt to be an essential part of the narrative world in which it occurs. For example: the following two scenes from *The Godfather* (1972) differ in the degree of their eventfulness, i.e. in the degree to which they are relevant and essential to the story. On page thirty-two Don Corleone is gunned down. This is a highly relevant event, it changes what the story is about, it is in fact essential to the story. Whereas on page thirty-seven Tom is kidnapped by Solozzo and told that the Don has been killed, and furthermore that Tom is to convince Sonny not to go to war. Tom is kidnapped, a change of state which is relevant to the narrative. How essential is it? Is it as essential as the assassination of the Don? In regards to the overall narrative's plot structure, Tom's kidnapping is relatively minor. Sollozo could have delivered the message that he wants to negotiate, not go to war, by other means. However, Tom's kidnapping is not as essential to the narrative as the gunning down of the Don, which changes what the story is about. While Tom's kidnapping, as an event, is relevant to the narrative, it is not essential. It does not alter the story's direction nor change its shape, it does not alter what the story is about, that is, you could remove Tom's kidnapping and the story's trajectory remains intact; therefore, it is not as eventful as the attempted assassination. The evaluation of relevance and essentiality are interpretive exercises, but that does not mean that any interpretation goes. It is not a subjective interpretation, but rather one that is intradeigetic to the narrative, especially when it is a character from within the story who is interpreting the event, in contrast to a reader's interpretation. These interpretations are fundamentally human choices, which at this time can not be automated. (No, not subjective; without the Type II I'm

discussing, the narrative falls apart. In a complex story there could be room for interpretation as to which of the events turns the story, but an event, a change of state, at least one, will turn the story.)

I am making a distinction between two criteria, "relevance" and "essentialness", which Schmid seems to collapse into one criterion. Although they are related, I do not think that they are necessarily interchangeable concepts. An event may be relevant without being essential. I am interested in events that are essential to the narrative, i.e., an event, which if removed, results in the given narrative collapsing.

2. Unpredictability: "Eventfulness increases in proportion to the extent to which a change of state deviates from the doxa of the narrative (i.e. what is generally expected in the narrative world.)" (Schmid 26). To illustrate this feature I will again refer to *The Godfather* (1972). From page one to page twenty-four of the screenplay, the writer exposes us to the Corleone family, its various members, and to the world they inhabit. It is a world full of joy, mischief , and violence, representing the violence of the shadow world of the gangster: "My father made him an offer he couldn't refuse." (Godfather 13), or the famous nightmarish scene when the character of the film producer, Woltz, wakes to find his prized stallion's head in his blood-soaked bed (Godfather 22): the consequence of refusing the Don's request. Up to that point the narrative world portrayed Don Corleone as a benevolent caretaker of the community. The horse head scene deviates from the doxa of the first act in that it turns sharply from the first act's portrayal of the Don as largely benevolent, violent yes, but just. This doxa is established in the opening scene of the story, when the Don refuses Bonasera

(Godfather 2): the Don refuses to murder the boys, as Bonasera asks, but agrees only to administer a punishment equal to the crime, a severe beating. The decapitation of the horse, as an event, is therefore a surprise that sweeps away the idea that the Don and his family are benevolent, loving caretakers of their friends and community. "A highly eventful change is paradoxical in the literal sense of the word: it is not what we expect" (Schmid 26). The ability of a reader to predict an event based on their understanding of narrative/literary patterns is not what is being referred to here, rather it is whether an event deviates from expectations within the given narrative world, the doxa.

Essentialness and surprise are the two sine qua non criteria that "underlie the continuum of eventfulness" (Schmid 26); "A change of state must meet both of these requirements to a minimum degree" (Schmid 27), the more essential and surprising the event the higher its degree of eventfulness.

The following three criteria, also used in determining an event's eventfulness, provide a refinement to understanding the degree of eventfulness, they are according to Schimd less crucial than the above criteria:

3. Persistence: "The eventfulness of a change of state increases with its consequences for the thought and action of the affected subject in the framework of the narrated world" (Schmid 27). This criteria can be quite difficult to pin down. Much depends on whether or not the style of writing is ambiguous. For example, in the novella *Death in Venice* by Thomas Mann, adapted by Mann to the screen for Luchino Visconti, the protagonist, Gustav von Aschenback, is struck by the beauty of a young Polish boy, Tadzio. The initial attraction to Tadzio is only aesthetic, but as the story develops, it becomes an infatuation. The

consequences of the thought and action of the event, the protagonist's first sight of Tadzio, forms the spine of the story and in this case becomes very persistent and therefore highly eventful. Another way to say it is, to what extent does an event give rise to ideas and actions and to what extent do they drive/influence the character? In the above cited example, an appreciative interest turns into an obsession, which clearly drives the protagonist to consider a breach of community law. So, persistence is a prominent feature of this story's inciting incident.

A contrasting example of the persistence of an event's effect on a character's actions and thoughts can be found in *The Godfather* (1972), when Michael Corleone promises to conform to his beloved's desire to continue to be socially legitimate, to not become a mafioso like his father and brothers:

MICHAEL
Luca Brasi held a gun to his head, and my father assured him that either his brains or his signature would be on the contract.
(then)
That's a true story.
(then)
That's my family, Kay. It's not me.

Michael sides with Kay's values against his family's values. The persistence of this thought, this promise to her to be legitimate, evaporates into thin air once his father is gunned down, a Type II event, which gives rise to the idea that he must avenge his father by embracing his family's values, a conviction which in turn takes root and persistently drives Michael's actions and thoughts towards revenge and succession at the story's conclusion.

4. Irreversibility: “Eventfulness increases with the irreversibility of the new condition which arises from a change of state” (Schmid 28). What is done cannot be undone is one way to understand this criterion . For example, in *Erin Brokovich* (2002), the inciting incident is the contamination of the groundwater of the town of Hinkley by highly toxic industrial chemical waste. As an event, the wells of Hinkley cannot be un-poisoned, the toxic waste ravages the people of Hinkley: their cancers, organ failures, and their bleeding to death cannot be undone. These events are highly eventful precisely because, among other criteria, they are absolutely irreversible. The poisoning of the wells of Hinkley, as an event, happens before the story starts and continues for the duration of the story.

5. Non-iterativity: “Repeated transformations, even if they are both relevant and unpredictable, represent at best a low level of eventfulness” (Schmid 29). The repeated use of an event becomes a trope, a way for the writer to comment on the story, or the world of the story. For example in *Groundhog Day* (1993), the protagonist wakes up at the same hour of the same day repeatedly. In *Edge of Tomorrow* (2014), the protagonists are repeatedly sent back to the same moment of conflict in order to learn about the enemy and implement a counter-attack. In both cases the iterative part of the structure is by definition predictable, and so in Schmid’s schema it qualifies for a lower order of eventfulness.

In conclusion, the criteria used in choosing an event for this study are:

1) Real: The change must be factual in the context of the story.

2) Resultative: minimally, the event must have a beginning and an end within the story's beginning and end.
3) Unpredictable: the extent to which an event deviates from the doxa of the story.
4) Irreversible: What is done cannot be undone.
5) Essential: The change must be essential, not just relevant, to the development of the narrative.
6) Persistence: the degree to which a thought or action influences the narrative.
7) Non-iterative: the event is not repeated.

The first two criteria, real and resultative, are necessary to all Type II events in Schmid's schema. In choosing the event corpus, I use all seven of Schmid's criteria, with one change: instead of "relevant", I use "essential" as a criteria. However, in any given event, some of these features will be stronger than others.

1.6 Type I & II Event: Compare and Contrast

What features do type I and II events have in common? They both represent a change of state, they are temporally based, they are action oriented, and they are both expressed in language (in screenplays). However Type II events are what Hühn calls a hermeneutic category; they require interpretation. They are "an interpretation and context-dependent type of narration that implies changes of a special kind" (Hühn 160). The reader or story character can and often does provide the interpretation of such events. Type I events, as in the Piper/Portelance study, are objective and do not require interpretation, they are a function of linguistics.

Both event types are changes of state, but for our purposes while every narrative event is a change of state, not every change of state is a narrative event (Schmid 24). Every type I event is a change of state, not every change of state is a Type II event, far from it.

Hühn claims that a Type II event contributes to a particular type of narrative's "raison d'etre, or tellability" (Hühn 160). "A Type I event is a defining feature inherent in every kind of narrative" (Hühn 160), but a Type II event, a narrower definition of event, must exhibit certain features. It is characterized generally by Hühn as a plotted event, in contrast to Type I which is plotless, "or as process narration vs. event-based narration" (Hühn 160).

The distinction reflects the ideas of novelist and narratologist E.M. Forster, "The king died and then the queen died" is a story [two changes of state]. "The king died, and then queen died of grief " is a plot (Forester). They are two types of narration, differing in specificity. The first part, the "story", does not require interpretation, it conforms to Prince's linguistic definition of event, a Type I event. The second part both expresses and invites interpretation; a Type II event that is context-dependant, both in terms of the intradeigetic context and the cultural context. This is where type II's differ radically from Type I's; the interpretation of Type II's is related to cultural norms. A dramatic event in one culture may be a trivial matter in another. For example, an inter-faith marriage is illegal in certain cultures, while it is a non-issue in others. Therefore, if such an event is used in a narrative, its interpretation will be culture-dependant.

The two event types are not mutually exclusive. A Type I event can be or can become a Type II event. Hühn presents an example: "Mary stepped onto the ship." (Hühn 160), is in and of itself a Type I event. However, if within the context of the story, this Type I represents a turn in the story, i.e. Mary is escaping persecution, then it can become a Type II event. This requires that

a reader make an interpretation of the event in the context of the narrative. As well, a series of Type I events can become, as a series, a Type II event. For example, a consequential political change, such as the election of an immoral, abusive, narcissistic businessman to the highest political office can be composed of a series of Type I events, i.e., a series of changes of state: speeches, debates, election, which when combined lead to the election win, become a Type II event, that is, change of state which is consequential to the trajectory of the story. All Type II are also Type I events.

In conclusion, the Hamburg school has expanded the concept of an event to two types: "A Type I event is any change of state explicitly or implicitly represented in a text. A change of state qualifies as a Type II event if it is accredited—in an interpretive, context-dependent decision—with certain features such as relevance, unexpectedness, and unusualness." (Hühn 159). Furthermore, "The concept of event has become prominent in recent work on narratology" (Hühn 159). However, it is a concept that is applied in a number of other disciplines as well. In order to further understand the Type II event and its role in the narrative, the following section examines the concept of event outside of narrative theory.

1.7 The Concept of Event and its Applications in the Humanities

I think that it is important to understand the concept of event as it appears not only in one's field, but as well how it is conceived and used in other branches of the humanities, in order to understand what is common to the various disciplines' understanding of event, and to make it generalizable. To quote Andrew Piper on this point, "Indeed, one could argue that generalization is a crucial aspect to any scholarly method. It is what allows us to identify the significance of a

particular instance as well as the social and historical significance of some larger set of practices" (Enumerations xi). It is to this end that I look briefly at the definition of event in humanities studies such as historiography, sociology, and cultural anthropology.

Although increasingly prominent in the study of contemporary narratology, the concept of the event is found throughout the humanities. The "event" has been a central concern in the study of conversational linguistics, historiography, cultural anthropology, and literary theory. As digital humanists interested in computational text analysis, it is important we recognize the 'event' as a concept which permeates the humanities in order to further understand and explicate its use more generally in the art of storytelling and specifically in narrative theory. To begin, we will look at its genesis in narratology and how the concept of event is understood in other humanities disciplines, specifically the centrality of Type II event in construction of the raison d'être of the narrative.

In regards to fiction, Hühn says "that the genesis and development" (Hühn 161), of the novella as a genre, specifically in regards to plot structure, was closely connected to the concept of a Type II event. The word itself, novella, implies the new, the unexpected which as we will see is a primary feature of the Type II event. He cites Göethe as amongst the first to address the centrality of Type II events in the Novella, " 'what is a *Novelle* if not an unheard of occurrence that has taken place" (Göethe in Hühn 162). Furthermore, "Tieck describes the central feature of the novella as the 'turn in the story, that point at which it unexpectedly begins to take an entirely new course'" (reprinted in Hühn 162). From the very beginnings of the novel, the Type II event is identified as central to its raison d'être. This is the focus of our data.

In poststructuralist narrative theory, Type II events have thus far played a peripheral role (Hühn 159). However in other fields of study, Type II events have played a central role. Hühn points to the work of William Labov as amongst the first to identify event, i.e. Type II event, as a central narrative element. In the work of linguist, William Labov, *Language in the Inner City: Studies in the Black English Vernacular* (1972), the Type II event is identified as the reason why narratives are recounted. Labov studied the quotidian narratives of inner-city youths:

> Beginnings, middles, and ends of narratives have been analyzed in many accounts of… narrative. But there is one important aspect of narrative which has not been discussed — perhaps the most important element in addition to the basic narrative clause. That is what we term the evaluation of the narrative: the means used by the narrator to indicate the point of the narrative, its reason d'être: why it was told, and what the narrator is getting at. (Labov 366)

Labov uses the term "evaluation" to describe a narrator's techniques of emphasizing the "point" of the narrative. He goes on to identify the Type II event as forming a main reason why narratives are told (Hühn 162-163). He also uses the term "tellability" to describe a story's "noteworthiness", i.e. why it matters. "Polanyi maintains that tellable materials can stimulate interest culturally, socially, personally or with some combination thereof" (Baroni, 1).

Mary Louise Pratt took Labov's analysis, which was based on conversational narration, and applied it to literature. Pratt points out that the tellability of a literary narrative also depends on its deviation from what is expected and how such change directly impinges on the overall narrative (Pratt 63-70), novelty being one of the two constitutive elements identified earlier as a minimum criteria for Type II events.

As well as the study of inner city vernacular, the concept of an event has long been defined and applied in the discipline of historiography,

> Three criteria to distinguish events from simple happenings: "(a) contemporaries must experience a sequence of actions as disquieting and breaking with expectations; (b) the grounds on which the sequence of actions is considered surprising and disquieting must be collective in nature—part, that is, of a social horizon of expectations; and (c) the sequence of actions must result in structural changes that are perceived and discursively processed by those involved. (Hühn 164).

Again these are in accord with the Hamburg group's definition of Type II event: a) unexpected b) must be essential to the given narrative world, and c) alters the trajectory of the narrative, that is, there is a relationship of mutual interdependence between event and narrative structure.

In cultural anthropology the notion of a Type II event is found in "the study of ritualized changes of status in the lives of individuals or groups within tribal societies" (Hühn 165). The concept was extended to apply to modern cultures to include all rites of passage: marriage, funeral, transitioning from adolescence to adulthood: getting a driver's license, graduating university, getting a job, etc. While there are similarities between social rituals and major story turns the relationship is still only generally defined, "But this model, though potentially suggestive of interesting parallels between ritualized transitions and eventful narrative turns, is as yet applied only in a loose and unspecific sense lacking terminological and analytical precision" (Hühn 165).

The relationship between the central event of a narrative and its "tellability", what Labov calls it's raison d'être, is only one potentially interesting cross disciplinary study organized around the shared critical concept of the Type II event.

1.8 Weaknesses in the Theory of Event

Kristin Langellier points out that "the purpose of Labovian analysis is to relate the formal properties of the narrative to their functions" (Langellier, 245), a notion that underlies my thesis. The linguistic properties of a Type II event are related to its function, to contribute to the development of the narrative at a minimum, and/or to define the tellability of the narrative. There is an implicit relationship between the purpose of a scene and its language. Scenes meant to change what a story is about use a language that may transcend genre, era, or lexical style, at least that is the hypothesis here. Langellier points out in her criticism of Labov that Type II events and their interpretations leave out the audience and context. They are weak in terms of providing a cultural context for the interpretation. This does not prevent audiences from making cultural interpretations of story, even if the interpretation is mistaken, or off in terms of the narrative world; it is an interpretation with regards to what the story is about, i.e. the central function of the Type II event.

1.9 The Nonevent: Stasis vs. Change

For the coming computational analysis, event as defined above is one side of a binary; the other is the nonevent. The nonevent represents a state of stasis in terms of narrative development. It can be defined as what it is not, it is not a change of state in the sense of a Type II event, in that what happens does not alter the story's trajectory.

In literary studies, a nonevent conforms to what is traditionally referred to as exposition: the introduction and description of setting, theme, back story, foreshadow, etc. are in practice what screenwriters provide the reader in nonevents. As a collective feature of nonevents, one can

classify them as descriptions. In terms of a narratological definition, nonevents are explicated thus, “Descriptive texts are the opposite of texts which are narrative…Descriptive texts represent static situations: they describe conditions, draw pictures or portraits, portray social milieus, or categorize natural and social phenomena” (Schmid 21). I would not describe nonevents as “opposite” of events in the sense that one negates the other. However, in this thesis, events and nonevents are defined to be mutually exclusive. A scene which significantly changes the trajectory of the story is how I’ve defined event, of necessity a Type II event; however, if the story’s trajectory does not change, it is a nonevent, primarily expository.

In practice finding scenes that are nonevents is not difficult, the sole requirement is that there be no change of state which significantly alter the story’s trajectory. Such scenes can be found throughout a screenplay, however they tend to be plentiful in the first act, as the writer sets up characters and their world. The following are some examples of nonevents:

Nonevent, *The Bourne Identity* (2002), 113 words:

OPERATOR/PHONE
Bonjour, Hotel Marboeuf...

BOURNE quick grabbing the receiver. Taking it off speakerphone and --

BOURNE
...yes -- oui -- uh...

OPERATOR/PHONE
Yes, sir. Hotel Marboeuf, Paris.
How can I direct your call?

BOURNE
Paris?

OPERATOR/PHONE
Yes, sir...

(switching to English, thinking that's his problem --)
How can I help you?

BOURNE
Yes, I'm...I'm looking for Mr. Jason Bourne.

OPERATOR/PHONE
One moment, please...
(a long pause, and then --)
I'm afraid, I have no one by that name registered, sir.

BOURNE
D'accord... Merci.
(about to hang up--)
Un moment -- un moment --

OPERATOR/PHONE
-- sir? --

BOURNE
-- hang on -- I need you to check another name for me -- hang on -- un moment, s'il vous plait —

Nonevent, *Alice in Wonderland* (2010), 121 words:

WHITE RABBIT
How is that for gratitude? I've been up there for weeks trailing one Alice after the next! And I was almost eaten by other animals! Can you imagine? They go about entirely unclothed and they do their...shukm in public. I had to avert my eyes.

The FLOWERS WITH HUMAN FACES study Alice.

TALKING FLOWER
She doesn't look anything like herself.

THE DORMOUSE
That's because she's the wrong Alice.

TWEEDLEDEE

And if she was, she might be.

TWEEDLEDUM
But if she isn't, she ain't.

TWEEDLEDEE
But if she were so, she would be.

TWEEDLEDUM
But she isn't. Nohow.

ALICE
How can I be the "wrong Alice" when
it's my dream? And who are you, if I may ask.

The criterion used in choosing a nonevent is:

1) Exposition: the nonevent is composed mostly of descriptive elements. There may be events in the nonevent, Type I & II's , but they will not alter the story's trajectory, which is how I define the Type II event. It would be more accurate to call the nonevent, the non-Type II event. In this context nonevent means non-Type II events.

1.10 Montage Theory vs. Event Theory

> *"Everywhere — struggle. A stabilization born from the clash of opposites."* Sergei Eisenstein

The type of screen narrative that makes up my corpus has its ultimate manifestation in the film itself; however, the image-making aspect of a film's narrative is underpinned by the

screenplay in the arch-plot type we are examining. The logic of narrative, “a beginning, middle, end, though not necessarily in that order” (Jean-Luc Godard) has come to dominate the modern film (McKee 44), and narrative theory’s influence in placing the event at the centre of narrative is contextualized through the historical binary framing of the struggle between montage and story structure, whose proponents argued about the dominant method of making filmic narratives.

How a film’s narrative is conveyed has long been associated with, and even defined by, montage theory, as debated in 1927 within the ranks of the Russian formalists, 1915-1930 (Stam 75). In film the cinematic narrative is produced by images and sound (Metz 19). The historical roots of the manipulation of the image as the foundation of filmic storytelling are found in the work of Sergei Eisenstein and the Russian Formalists, such as Viktor Shklovsky and Juri Tynianov.

In the discussion of film narrative, the formalists framed the question of the relation between narrative, image, and style thus:

> This debate took shape around the question of fundamental structures: is the fabula, understood as a linked sequence of actions, the irreducible core of narrative structure? Or is film narrative rather defined by stylistic manipulations of space and time? In other words can a narrative proceed on the basis of stylistic variation rather than by way of story actions? (Stam 72).

This discussion took place during the 1920s with the publication of a series of essays titled *Poetica Kino* , wherein Shklovsky and Tynianov (Stam 11) articulated a theory of filmic narrative based on montage, i.e. the manipulation of space and time through the juxtaposition of images. However, the formalists differed on this point in that Eikhenbaum believed that

ultimately film narrative was “grounded in narrative syntax of actions and events” (Stam 72), while Tynianov argued that style, the spatial and temporal structure of films, operates as the “principal mover of the plot” (Stam 72).

Sergei Eisenstein, a pioneer in the use of montage, eschews the chronological links of events, instead using juxtaposition and correspondence to structure his films, “There is an apparent development and continuity in Eisenstein’s early films, but these continuities do not belong to a natural course of the action but rather to a correspondence between shots” (Rhodie 32). The power of Eisenstein’s films and especially of his writings about film drew together many proponents of montage theory, and helped montage become a favourite storytelling technique, especially in critical film circles.

Shklovsky’s position was a middle ground between Tynianov’s embrace of the montage and Eikhenbaum’s belief in the dominance of narrative syntax of actions and events. Shlovsky was able to apply both theories to film, and made a distinction that led to an articulation of genre. For example, Charlie Chaplin’s *Woman of Paris,* which he identified as a prose work, is structured along classical narrative lines, while a film like Eisenstein’s *The Battle Ship Potemkin,* is akin to a poem where, “the distinction between shots could be compared to the separate lines of a poem” (Stam 72). This distinction delineates the difference between “classical” films and “experimental” films. Modern examples of this delineation would be Francis Coppola’s *The Godfather* (1972), a film structured as a classical narrative, and Jim Jarmusch’s *The Limits of Control* (2009)*,* which is a narrative told mostly through the image. Montage theory, one of the principle “manipulations of space and time”, therefore is a logical extension of this fact. However, in considering the work of structuralist narratologists such as Mieke Bal, Wolf Schmid,

and Robert Mckee, the narrative is framed within what Schmid calls “eventfulness theory”, “From the structuralist perspective, the broader concept of narrative refers to representations that contain a change of state (or of situation)” (Schmid 19). The type of screenplays I examine[1] are based on narrative structures, the images are yet to come.

[1] For the definition of the types of screenplays which make up the corpus see p. 39.

Chapter 2: Choosing the Data

"To make a great film, you need three things. The script, the script, and the script."
Alfred Hitchcock

2.0 The Films and their Screenplays

This thesis focuses on a type of screenplay defined by Robert McKee in his book, *Story,* through its plot as "arch-plots", i.e., as based on a classical design.

> Classical design means a story built around an active protagonist who struggles against primarily external forces of antagonism to pursue his or her desire, through continuous time, within a consistent and causally fictional reality, to a closed ending of absolute irreversible change. (Mckee 45)

In searching for language that identifies events, one of the criteria is that the event must be stated explicitly through the use of language to describe the drama. A screenplay can be more or less explicit in its use of text to tell the story. A minimalist plot, in contrast to an arch-plot, reduces and compresses the elements of story, "Rather, minimalism strives for simplicity and economy while retaining enough of the classical that the film will still satisfy the audience…" (McKee, 46). Examples of films based on the minimalist plot are *La Passion De Jeanne D'Arc* (1928), *Tender Mercies* (1983), *A River Runs Through It* (1993), and *Shall We Dance* (2004). Examples of films based on the arch-plot are *Citizen Kane* (1941), *The Godfather* (1972), and *Thelma and Louise* (1991), according to McKee. This explicitness in the use of language to tell its story is one reason why the arch-plot was my choice as the source of data for this study; furthermore, the screenplays of such films determine production costs as much as they tell a dramatic tale. A film's producer will draw up a a film's production budget based on the script's locations, historical settings, characters, costumes, and action sequences; therefore, a production-ready

screenplay tends toward being explicit in regards to those story elements, which makes the arch-plot well suited for this research. This thesis is a language-based research, therefore a type of screenplay which is explicit and complete in its descriptions of story elements, i.e. the arch-plot, will yield Type 2 events which are more explicit and complete in their use of language. This is a richer fodder for a computational analysis using Corpus Linguistics.

From the Internet Movie Script Database (IMSDb), I looked at screenplays of arch-plot based Hollywood productions ranging over fifty years, from 1963 to 2013, a period often portrayed as the post-studio era, before which screenplays tended to be heavily edited by the production studios. Auteur film makers were given more leeway in crafting screenplays and productions than had their studio-bound predecessors (Bordwell). An auteur film is one where the director imposes an authorial stamp on the film (Brittanica), for example, *The Godfather* (1972), *Do The Right Thing* (1989). Jean-luc Goddart's *Pierrot le Fou* (1965) is an early example of an auteur film based on Andre Bazin's and Alexandre Astruc's auteur theory, "The auteur theory, which was derived largely from Astruc's elucidation of the concept of "caméra-stylo" ("camera-pen"), holds that the director, who oversees all audio and visual elements of the motion picture, is more to be considered the "author" of the movie than is the writer of the "screenplay" (Britannica).

In no way is the corpus representative of this time period, because it is too small a percentage of screenplays produced during said period. According to the Motion Picture Association of America, Hollywood produces on average approximately 600 films per year (Guardian). However, the number of wide release feature films is smaller.

Wide release means a film released in 1,000 or more theatres. Although not every one of the screenplays in the corpus is wide release, approximately two thirds of them are. From 1995 to 2018 Hollywood produced 2892 wide release feature films (Numbers). It is primarily this commercial category of film that my study examines. However, using a rate based on the above 1995-2018, or 24 year period, and assuming a constant rate back to 1963, would result in an estimate of about 6000 wide release films from 1963-2013. This study's corpus therefore makes up 3 % of the total output. Granted that's on the low side; wide release has been on the decline for the past decade (Guardian). Wide release films have been watched by millions of people world-wide and are therefore of interest as cultural products.

The screenplays were chosen from a wide range of genres: action, comedy, horror, romance, thriller, crime, adventure, fantasy, western, mystery and film noir, all of which are available on the Internet Movie Script Database (IMSDb), which organizes the screenplays by genre; there are 11 genres in all. My decision to spread the data collection over 11 film genres and fifty years of screenwriting was made to give my choices a breadth, and so avoid a totally subjective corpus. However, I must admit a personal preference for many of the choices I made. Another person covering the same set of genres and period would have come up with a different set of screenplays.

I randomly chose one screenplay from each genre. I then repeated this process until I had 70 screenplays. The intention is to have the corpus represent different approaches to the writing of events from different genres over a period of 53 years. The thesis being that an event, a change of state that changes what the story is about, leaves a linguistic signal. While the size of

my corpus is relatively small, the choice of screenplays are from a wide range of genres and periods.

When I initially started identifying events, I chose two sizes of event: the short and the long event. Short events were under 150 words, while long events were between 150 and 450 words. The distinction was made on the assumption that long events would provide more data. As the research developed, it became evident through further close readings that the event itself could be captured in about 250 words. The act that precipitates the event happens in a moment, sometimes an extended moment, but rarely is it longer than 250 words in the arch-plot. This resulted in smaller sample sizes. This observation had a major impact on the research. That this was an effective choice became reflected in the computational results as differentiation between event and nonevent became increasingly more distinct in the machine learning process[2] outlined in chapter 3. The data is now in two categories, events and non-events; they range from about 100 to 250 words. The following are samples of events (see page 30 for samples of nonevents):

- An event from *Fight Club* (1999), 157 words:

JACK
(looks at watch)
God, it's late. I should find a hotel...

TYLER
A hotel?

JACK
Yeah.

TYLER

[2] Chapter 3 Appendix: Machine Learning

So, you called me up, because you just wanted to have a drink before you... go find a hotel?

JACK
I don't follow...

TYLER
We're on our third pitcher of beer. Just ask me.

JACK
Huh?

TYLER
You called me so you could have a place to stay.

JACK
No, I...

TYLER
Why don't you cut the shit and ask if you can stay at my place?

JACK
Would that be a problem?

TYLER
Is it a problem for you to ask?

JACK
Can I stay at your place?

TYLER
Yes, you can.

JACK
Thank you.

TYLER
You're welcome. But, I want you to do me one favor.

JACK
What's that?

TYLER
I want you to hit me as hard as you can.

JACK
What?

TYLER
I want you to hit me as hard as you can.

- Event from *FORREST GUMP* (1994) 253 words:

JENNY
Run, Forrest!

Forrest tries to run along the road, but his braces makes it impossible. He hobbles along as Jenny yells after him.

JENNY
Run away! Hurry!

Boy #1 and Boy #2 turn back toward the bikes.

BOY #2
Get the bikes!

BOY #3
Hurry up!

The boys pick up their bikes and ride after Forrest.

BOY #3
Let's get him! Come on!

BOY #2
Look out, dummy, here we come!

The boys ride after Forrest. Jenny stands and watches.

BOY #2
We're gonna get you!

JENNY
Run, Forrest! Run!

Forrest hobbles along the dirt road.

JENNY
Run, Forrest!

Forrest looks over his shoulder. The three boys race on their bikes.

BOY #1
Come back here, you!

Forrest begins to run faster with his braces on. Forrest continues running as the boys chase him. Blood drips down from a cut on his head. The boys on the bikes are gaining on Forrest. Forrest hobbles along. He begins to gain speed.

JENNY
Run, Forrest! Run!

SLOW MOTION --

Forrest runs from the chasing room. He looks over his shoulder in fear.

The boys on the bikes peddle faster as they gain on Forrest, running.

Forrest tries to run even faster to get away. Suddenly his braces shatter, sending steel and plastic flying into the air.

Forrest runs and looks down at his legs in surprise.

Forrest continues to run faster as the metal braces and straps

fly off his legs.

Forrest runs free of his braces and begins to pick up speed.

The screenplay excerpts above provide examples of events which change what the story is about. In *The Fight Club*, Jack moves in with Tyler which is the last piece of Jack's old life jettisoned; along with his furniture, his job, which he has sabotaged, he now decides to move in with the great Tyler and as well the first gesture of fight club emerges. These are both changes of state that we later learn are all happening in Jack's head; however, they are events which encapsulate the seven event criteria established earlier in my thesis.

The question of copyright came up; initially, the thought of getting permission for each screenplay was daunting. The availability of screenplay databases on line, such as on IMSDb.com, for the purpose of study, allowed me to have access to a wide range of screenplays. Without the IMS data base, the collection of modern screenplays in such a variety of genres would have been far more difficult. While IMSDb was the final choice, it came after a fair bit of winding down dead ends. Initially, I wanted to use the 100 greatest screenplays as named by the American screenplay Writers guild (The American Writers Guild). It was an exciting list, but many of the screenplays were impossible to find online, or to gain access to otherwise.

The issue of copyright infringement was an additional reason to choose the IMSDb site. Here is their disclaimer on the copyright issue clearly allows for the use of copyright material for the purposes of research:

Another feature of using the IMSDb site: it carried a collection of modern screenplays from 15 different genre's. In hindsight, it might have been interesting to group the screenplays by genre, and compare the differences in sentiment, topic, or part of speech analysis.

2.1 Screenplays: What to Include and Exclude

An event will include both dialogue and narrative description. In screenplay format, narrative descriptions are the paragraphs between dialogues, and they include action, setting, characters, and sounds, e.g., from *The Godfather*,

SOLLOZZO
(offering his hand to Luca; in Italian)
Agreed?

Luca doesn't shake. He takes out a cigarette, which Bruno lights.

LUCA
Grazie.

Bruno grabs Luca's hand to the bar, Sollozzo rams a knife into it, and Luca is garroted by an unidentified buttonman.

CUT TO:

EXT. CITY STREET — EARLY MORNING

Sollozzo kidnaps Tom Hagen on the city street while Tom's Christmas shopping.

Each of the narrative descriptions provides a rich source of linguistic material, and each is critical to the development of the narrative; therefore, they are included in the data. The mix of narrative paragraphs and dialogue in the screenplay is similar to the mix of descriptive passage mixed with dialogue in a novel, albeit the distinction between the two is more pointed in screenplay format.

In the second example above, *Forrest Gump* (p44), the scene depicts the moment when Forrest breaks free of his leg braces and discovers that not only can he run, he can run like the wind. This event meets the seven event criteria and is clearly an event which changes the story's trajectory.

In conclusion, while it is clear that choosing Type II events involves an interpretation, I have attempted to minimize subjectivity by using seven criteria to choose the Type II event. Furthermore, I included one criterion, essentialness, which makes a purely subjective interpretation difficult. For example, if Alice does not find the "shrinking" potion, she doesn't go through the keyhole, the story falls apart, or, if Don Corleone is not shot, but left to grow old, the story falls apart, or if Luke Skywalker doesn't meet Obi-Wan Kanobi, the quest doesn't begin, and the saga of Star Wars likely never happens. I do not think that these are subjective interpretations, i.e., an interpretation that will differ substantially from reader to reader.

Chapter 3

3.0 Computational Analysis

Chapter 3 has one primary objective: to use computational analysis to test the hypothesis that the language used in Type II events will leave a distinct linguistic signal, which can be used to help distinguish, or even identify the kind of events that are deeply consequential to the narrative from which they derive.

The following computational analysis are created on Juypter Notebook.

Jupyter Notebooks was chosen for three main reasons (from ALTM):

1. Python (the programming language used in Jupyter Notebooks) features extensive support for text analysis and natural language processing.
2. Jupyter Notebooks offers a *literate programming* model of writing where blocks of prose text (like this one) can be interspersed with bits of code and output allowing one to write up experiments.
3. The Natural Language Toolkit (NLTK), coded in Python, provided many of the original scripts for language analysis.

Chapter 3 achieves its primary objective through the following approaches:

- To introduce the concepts and methods used in the linguistic analyses of events and nonevents.
- To attempt a range of linguistic analytical techniques for the study of events and nonevents.
- To provide interpretations of the outputs of the linguistic analysis.

The following chapter will be a combination of text, code, and graphs. The code will be presented as samples relevant to the discussion. The content is as follows:

1. Gathering the Screenplays
2. Tagging Events and Nonevents
3. Tokenizing and Filtering
4. Most Frequent Terms
5. Clustering
6. Parts of Speech
7. Sentiment Analysis
8. Appendix: Machine Learning

I am familiar with these types of analysis through my studies in digital humanities, that is why I chose them.

3.1. Gathering the Screenplays

After a thorough search of online sources for the screenplays themselves, including the Writers Guild of America, American Film Scripts Online, McGill Film Database, The Gutenberg Archives, among others, I found the most comprehensive and accessible online resource for the kind of screenplays I want to analyze in the International Movie Script Data base (IMSDb.com). Their screenplays were often shooting scripts or close to it, they were easily converted to .txt, and their narrative paragraphs were descriptive. A shooting scripts is as close as a script will get to reflect the on-screen narrative, it's the script on set.

I looked at screenplays of Hollywood productions from 1963-2013. The screenplays were chosen from a wide range of genres: action, comedy, horror, romance, thriller, adventure, sci-fi, drama, film-noir, mystery, and western. I chose one screenplay from each genre, and continued until I had 70 events and 70 nonevents. This is a small number of screenplays, but gathering the data is time-consuming. One can readily gather nonevents, but to gather Type II events, as

defined in chapter one, requires a careful reading of the screenplay. I do not claim that the choices I made are "objective"; the attempt was to choose screenplays from a variety of genres over a period of time when screenplays were more under the control of the auteur (the director) than their historical predecessors[3]. Some of the screenplays I chose are personal favourites, like *The Godfather* (1972), for example. The code for scraping screenplays from IMSDb.com:

The first step is to download all of the screenplays on the International Movie Script Data Base (IMSDb). Once downloaded, they are converted from .html to .txt files.

In [1]:
```
import os
from urllib.parse import quote

from bs4 import BeautifulSoup
import requests

BASE_URL = 'http://www.imsdb.com'
SCRIPTS_DIR = 'scripts'


def clean_script(text):
    text = text.replace('Back to IMSDb', '')
    text = text.replace('''<b><!--
</b>if (window!= top)
top.location.href=location.href
<b>// -->
</b>
''', '')
    text = text.replace('''          Scanned by http://freemoviescripts.com
          Formatting by http://simplyscripts.home.att.net
''', '')
    return text.replace(r'\r', '')


def get_script(relative_link):
    tail = relative_link.split('/')[-1]
    print('fetching %s' % tail)
    script_front_url = BASE_URL + quote(relative_link)
    front_page_response = requests.get(script_front_url)
    front_soup = BeautifulSoup(front_page_response.text, "html.parser")

    try:
        script_link = front_soup.find_all('p', align="center")[0].a['href']
    except IndexError:
        print('%s has no script :(' % tail)
        return None, None

    if script_link.endswith('.html'):
        title = script_link.split('/')[-1].split(' Script')[0]
        script_url = BASE_URL + script_link
        script_soup = BeautifulSoup(requests.get(script_url).text, "html.parser")
        script_text = script_soup.find_all('td', {'class': "scrtext"})[0].get_text()
        script_text = clean_script(script_text)
        return title, script_text
    else:
        print('%s is a pdf :(' % tail)
        return None, None


if __name__ == "__main__":
    response = requests.get('http://www.imsdb.com/all%20scripts/')
    html = response.text

    soup = BeautifulSoup(html, "html.parser")
    paragraphs = soup.find_all('p')

    for p in paragraphs:
        relative_link = p.a['href']
        title, script = get_script(relative_link)
        if not script:
            continue

        with open(os.path.join(SCRIPTS_DIR, title.strip('.html') + '.txt'), 'w') as outfile:
            outfile.write(script)
```

3. Please see page 40 for an elucidation of this point.

3.2 Tagging Events and Nonevents

This was one of my favourite parts of the research. With a clear definition of the Type II event, I carefully read through screenplays and chose events which met the criteria outlined on page twenty-five.

The nonevents that I chose were expository scenes. The writer exposes the setting, backstory, characters' personality, and other such matters. These scenes do not contain a Type II event.

Once I tagged the data as 'events' and 'nonevents', I then split the tagged passages into their respective files:

Splitting off nonevents: sample codes

In [19]:
```
with open ("/Users/josephrafla/Desktop/Tagged_Noevent_txt.txt", "r") as infile:

    #This buffer contains the lines that are BETWEEN two tags.
    buffer = list()
    #This marks whether we are inside of tags or not
    in_event = False
    #This counter keeps track of how many events we have seen
    event_counter = 0

    #For each line in our source file
    for line in infile:
        #If we see an opening tag
        if '<noevent>' in line:
            #Setup waht we need to gather the contents
            in_event = True
        #If we see a closing tag
        elif '</noevent>' in line:
            #Take what we have and write it to a file
            with open('/Users/josephrafla/Desktop/Noevent/noevent_' + str(event_counter), 'w') as
outfile:
```

In [9]:
```
with open("/Users/josephrafla/Desktop/Tagged_events_txt.txt", "r", encoding="utf-8") as infile:

    #This buffer contains the lines that are BETWEEN two tags.
    buffer = list()
    #This marks whether we are inside of tags or not
    in_event = False
    #This counter keeps track of how many events we have seen
    event_counter = 0

    #For each line in our source file
    for line in infile:
        #If we see an opening tag
        if '<eventS>' in line:
            #Setup what we need to gather the contents
            in_event = True
        #If we see a closing tag
        elif '</eventS>' in line:
            #Take what we have and write it to a file
            with open('/Users/josephrafla/Desktop/eventSH/event_small_' + str(event_counter), 'w',
encoding='utf-8') as outfile:
                #outfile.write(string)
                for element in buffer:
                    outfile.write(element)

            #Reset the flag
            in_event = False
            #Reset the buffer
            buffer = list()
            #Increment the counter
            event_counter += 1

        #If we see neither tags but see that the flag is up
        elif in_event:
            #Save the line to the buffer
            buffer.append(line)
```

The following uses RegexpTokenizer from NLTK, which has filters for punctuation, stopwords, etc.

In [1]:
```
import os
#from nltk import pos_tag

from nltk.tokenize import RegexpTokenizer, sent_tokenize
import nltk
#nltk.download('punkt')
#nltk.download('stopwords')
#nltk.download('averaged_perceptron_tagger')

PATH = '/Users/josephrafla/Desktop/'
```

The following tokenizes the file into sentences, each token is a sentence, and removes punctuation.

In [ ]:
```
#tokenize into sentences and remove punctuation
def tokenize(file):
    file_content = list()

    with open(file, 'r') as infile:
        for line in infile:
            file_content.append(line)

    file_content = '\n'.join(file_content)

    tokens = list()

    sentences = sent_tokenize(file_content)

    tokenizer = RegexpTokenizer(r'[\w\']+') #removes punctuation
    for sentence in sentences:
        tokens.append(tokenizer.tokenize(sentence))

    return tokens
```

The following removes stopwords with NLTK filter, and Mathew Jockers' filter helped remove personal names.

In [3]:
```
#remove stopwords using nltk and Matt Jockers' list
from nltk.corpus import stopwords

def remove_stopwords(sentence):
    additional_stopwords = list()

    with open(PATH + 'jockers_stopwords.txt', 'r') as infile:
        for line in infile:
            additional_stopwords.append(line.strip())

    return [token for token in sentence if token.lower() not in stopwords.words('english') and tok
en.lower() not in additional_stopwords]
```

The tokens are stemmed with NLTK PorterStemmer.

In [4]:
```
from nltk.stem import PorterStemmer

def stem_tokens(sentence):
    ps = PorterStemmer()
    return [ps.stem(token) for token in sentence]
```

3.3 Tokenizing and Filtering

Please note: before tokenizing the texts I had to make one file for events. Initially events were split into small events (<150 words) and large events (>150 words): eventSH and eventB respectively. This proved to be a poor decision which has created some confusion in keeping track of the corpus, and it became evident as I tagged the events that the moment of the event's occurrence was captured in less than 250 words. Therefore, before tokenizing, I combined the

small and large events in the following code into one file called “rebuilt-events”. Nonevents were chosen to be about 250 words in length and gathered in a file called “rebuilt_nonevent”.

Screenplays contain much dialogue. Every time a character speaks their name appears so as to indicated who is speaking, which results in proper names dominating the linguistic analysis of any corpus of screenplays, and makes a clear reading of the output more difficult. The NLTK part of speech proper noun filter worked poorly in eliminating the plethora of personal names found in screenplays.

I found another filter, “Expanded Stopword List”, developed by Mathew L. Jockers. From Jocker’s website, he writes, “Below is the list of stop words I used in topic modelling a corpus of 3,346 works of 19th-century British, American, and Irish fiction. The list includes the usual high frequency words (“the”, “of, “an”, etc) but also several thousand personal names”. Jocker’s filter was very effective in eliminating personal names and thus helped declutter the data. NLTK’s punctuation filter and stemmer worked well.

Screenplays contain bits of language that are unique to the form. For example: EXT for exterior, or INT for interior appear at the head of each scene, i.e. often several times per page. As well, terms like CUT, PAN, DISSOLVE, etc. occur in every screenplay on many pages. A customized filter which listed such terms was required.

The process of tokenizing was fairly straight forward once events were combined into one file. The output for both files were rehoused in ‘rebuilt_events’ and ‘rebuilt_noevents’.

Fig. 1

```
print(corpus_no_stopwords)
```

```
[[['the', 'godfath', 'dissolv', 'TO', 'the', 'don', 'and', 'fredo', 'leav', 'the', 'genco', 'offic', 'late', 'afternoon', 'vito', 'corleon', 'aspetta', 'fredo', "i'm", 'gonna', 'buy', 'fruit', 'aspetta', 'wait'], ['fredo', 'get', "d
river'", 'seat', 'car', 'okay', 'pop', 'Hi', 'merri', 'christma', 'I', 'wan', 'fruit'], ['what'], ['gimm', 'three'], ['and', 'one', 'after', 'get', 'orang', 'green', 'pepper', 'hear', 'footstep', 'run'], ['He', 'begin', 'run', 'towa
rd', 'car', 'stumbl', 'fall', 'onto', 'car'], ['the', 'shooter', 'shoot', '10', 'shot'], ['the', 'don', 'scream', 'Uh', 'fall', 'assassin', 'run', 'don', 'fall', 'curb'], ['fredo', 'get', 'car', 'fumbl', 'gun', 'sit', 'curb', 'cri',
'I', "can't", 'I', "can't", 'papa'], []], [['alic', 'IN', 'wonderland', 'ON', 'alic', 'and', 'the', 'jabberwocki', 'alic', 'leap', 'onto', 'jabberwocki', 'pull', 'scale'], ['He', 'twist', 'snap', 'tri', 'shake'], ['the', 'vorpal', '
sword', 'practic', 'pull', 'hand', 'relentless', 'pursuit', 'creatur', 'head'], ['alic', 'make', 'way', 'jabberwocki', 'vest', 'cling', 'tri', 'slash', 'neck', 'sword'], ['the', 'vest', 'button', 'strain', 'alic', 'weight'], ['they
', 'snap', 'one', 'two', 'three', 'one', 'left', 'hold', 'vest', 'thread'], ['alic', 'bare', 'hold', 'onto', 'slash', 'sword'], ['just', 'last', 'button', 'give', 'way', 'propel', 'air'], ['alic', 'off', 'with', 'your', 'head'], ['wi
th', 'one', 'power', 'thrust', 'cut', 'jabberwocki', 'head'], ['both', 'alic', 'head', 'fall', 'jabberwocki', 'jaw', 'still', 'snap', 'death', 'throe']], [['devil', 'blue', 'dress', 'easi', 'So', 'think', 'might', 'watt'], ['albrigh
t', "i'm", 'sure'], ['but', 'I', "can't", 'go', 'look', "i'm", 'right', 'persuas', 'speak'], ["i'v", 'ask', 'joppi', 'week', 'could', 'introduc'], ['easi', 'what', 'want'], ['albright', 'just', 'get', 'locat'], ['Mr', 'carter', 'wan
t', 'make', 'easi', "that'"], ['albright', 'smile', "that'"], ['He', 'pull', 'brown', 'secretari', 'type', 'wallet', 'count', '10', 'ten', 'dollar', 'bill', 'neat', 'stack', 'next', 'whiskey'], ['albright', 'one', 'hundr', 'dollar',
'I', 'pay', 'advanc']], [['devil', 'blue', 'dress', 'scream'], ['He', 'jerk', 'awak', 'phone', 'ring'], ['It', 'jingl', 'coupl', 'time', 'easi', 'almost', 'knock', 'receiv', 'cradl', 'reach'], ['easi', 'huh'], ['voic', "woman'", 'he
llo'], ['IIs', 'Mr', 'rawlin'], ['easi', 'who'], ['voic', 'daphn', 'daphn', 'monet'], ['He', 'sit', 'straight'], ['daphn', "you're", 'look'], ['A', 'paus', 'line'], ['daphn', 'I', 'think', 'come', 'talk', 'Mr', 'rawlin', 'alon'], ['ea
si', 'look', 'late', 'daphn', 'Mr', 'rawlin', 'coretta', 'kill', 'anoth', 'paus'], ['daphn', 'and', 'want', 'troubl', 'polic', 'come', 'and', 'tell', 'anyon', 'come'], ['easi', 'now', 'wait', 'minut', 'daphn'], ['I', 'know', "nothin
'", 'stuff', 'daphn', "i'm", 'sunridg', 'motel', '112th', 'grand', 'room', '102']], [['momo', 'eye', 'bodyguard', 'nervou', 'start', 'step'], ['We', 'follow', 'momo', 'enter', 'build', 'goe', 'dark', 'stairwel'], [], [], ['one', 'fl
ight'], [], [], ["we'r", 'behind', 'whole', 'time'], [], [], ['when', 'get', 'top', 'floor', 'momo', 'paus', 'catch', 'breath', 'move', 'hall', 'door'], ['momo', 'knock'], ['voic', 'yeah'], ['momo', "it'", 'momo'], ['voic', 'come'],
['momo', 'slowli', 'open', 'door'], [], [], ['suddenli', 'apart', 'flood', 'with', 'light', 'see', 'room', 'full', 'peopl', 'banner', 'back', 'wall', 'read', "'happi", 'sixti', 'fifth', 'momo', "'"], ['everyon', 'surpris'], ['silenc
', 'momo', 'drop', 'knee', 'gurgl', 'someth', "can't", 'understand', 'fall'], ['suddenli', 'dozen', 'face', 'look', 'us'], []], [['fight', 'club', 'jack'], ["i'm", 'six', 'year', 'old', 'pass', 'messag', 'parent'], ['jack', 'I', 'uh
', 'think', 'go'], ['marla', 'ignor', 'still', 'search', 'tabl', 'toss', 'thing', 'push', 'thing', 'floor'], ['jack', "it'", 'time', 'leav'], ['marla', "don't", 'worri', "i'm", 'leav'], ['marla', 'find', 'want', 'pack', 'cigarett'],
['she', 'move', 'jack', 'face'], ['marla', "you'r", 'nutcas', 'I', "can't", 'even', 'begin', 'keep'], ['jack', 'goodby'], ['she', 'laugh', 'spin', 'heel'], ['As', 'exit', 'back', 'door', 'sing', 'thi', 'merri', 'Go', 'round', 'vall
ey', 'doll'], ['jack', 'watch', 'kitchen', 'window']], [['forrest', 'gump', 'ext'], ['gump', 'board', 'hous', 'night', 'forrest', 'sit', 'swing', 'outsid', 'hous'], ['loud', 'organ', 'male', 'grunt', 'heard', 'come', 'insid', 'hous'
], ['forrest', 'sit', 'swing', 'grunt', 'continu'], ['the', 'princip', 'step', 'gump', 'hous', 'wipe', 'sweat', 'face'], ['forrest', 'sit', 'porch'], ['princip', 'well', 'momma', 'sure', 'care', 'school', 'son'], ['Mm', 'mm', 'mm'],
['the', 'princip', 'wipe', 'sweat', 'neck', 'look', 'back', 'forrest'], ['princip', 'you', 'say', 'much'], ['forrest', 'grunt', 'imit'], ['the', 'princip', 'embarrass', 'turn', 'walk', 'away']], [['get', 'shorti'], ['cocktail', 'lou
ng', 'day', 'A', 'shit', 'face', 'leo', 'struggl', 'keep', 'head', 'watch', 'waitress', 'look', 'window', 'gasp'], [], [], ['chili'], ["he'", 'still', 'loung', 'plane', 'blow', 'runway'], [], [], ['We', 'see', 'big', 'fireball', 'b',
'g'], ["leo'", 'head', 'hit', 'tabl'], [], ["karen'", 'kitchen', 'night', 'As', 'chili', 'light', 'cigarett'], ['chili', 'So', 'guy', 'find', 'flight', "can't", 'believ'], ['If', "he'd", 'stay', 'plane', "he'd", 'dead'], ['right', '
know', 'luck', 'chang'], ['If', 'everybodi', 'think', "he'", 'dead', 'pay', 'back', 'fifteen', 'owe', 'vig', 'four', 'half', 'week']], [['all', 'the', 'king', 'S', 'men', 'the', 'sheriff', 'deputi', 'push', 'crowd'], ['sheriff', 'so
rri', 'willi', 'move'], ['willi', 'whi'], ['sheriff', 'citi', 'ordin', 'number', 'one', 'Oh', 'five', 'five', 'peopl', 'congreg', 'disturb', 'peac'], ['willi', 'ignor', 'If', "folks'll", 'kind', 'read', 'handbil', 'boy', 'pass', 'am
ong'], ['sheriff', "there'", 'ordin'], ['willi', 'face', 'grim', 'pass', "'em", 'tom'], ['the', 'sheriff', 'push', 'tom', 'back', 'grab', 'handbil', 'hand'], ['willi', 'jump', 'platform'], ['willi', 'let', 'alon'], ['the', 'sheriff'
, 'collar', 'willi', 'notic', 'jack', 'platform', 'snap', 'pictur'], ['sheriff', 'deputi', 'get', 'camera'], ['willi', 'arrest'], ['He', 'take', 'willi', 'arm', 'lead', 'away'], ['the', 'crowd', 'follow', 'courthous'], ['tini', 'duf
fi', 'wipe', 'sweat', 'neck', 'goe', 'back', 'poolroom']], [['alic', 'IN', 'wonderland', 'she', 'look', 'label'], ['drink', 'HE'], ['she', 'sniff', 'content', 'recoil', 'shrug'], ['alic', 'It', 'dream'], ['she', 'take', 'drink', 'sh
udder', 'gag', 'cough', 'tast'], ['In', 'moment', 'notic', 'tabl', 'get', 'larger'], ['alic', 'cont', 'D', 'that', 'curiou'], ['alic', 'shrink', 'two', 'feet', 'high'], ['the', 'door', 'loom'], ['she', 'take', 'step', 'trip', 'overs
```

The analysis led to the following output which shows the tokens, i.e. words, as a word blob. Here is a partial view of such output; Figure 1, below, is an early example. The final full word blobs are visualized as word clouds in chapter 4.

3.4 Most Frequent Terms

The data is now organized in two files: 1. rebuilt_events2 and 2. rebuilt_Noevents2.

The following code is used to count the number of times each token, i.e. each word, occurs in the data set. In addition to the filtered elements in the previous section, Tokenizing_and_Filtering,

Sample code used to filter events and nonevents:

In the rebuilt_events2 file, the ten most frequent terms and how often they occur, are listed in the output:

1. [57, 'look']
2. [41, 'shot']
3. [30, 'head']
4. [29, 'door'] [29, 'close']
5. [27, 'night'] [27, 'face'] [27, 'away']
6. [26, 'take']
7. [25, 'right']

"Look, car, shot, head, door, close, night": one is tempted to interpret this list as indicative of a topic. It reads like a sequence of connected actions.

In the nonevent file the ten most frequent terms are:

1. [55, 'look']]
2. [29, 'day']
3. [23, 'room']
4. [20, 'talk']
5. [20, 'hand']
6. [19, 'name']
7. [17, 'eye']
8. [16, 'girl']
9. [16, 'woman']
10. [16, 'want']
11. [16, "he'"]
12. [16, 'mother']
13. [16, 'open']
14. [16, 'peopl']
15. [16, 'stand']
16. [16, 'take']
17. [16, 'voic']

The word “look” dominates both data sets. This seems straight forward; however, the Oxford English Dictionary has 21 entries for the word” look”. A few examples from the OED:

- “Fashion. An appearance, style, or effect (usually of a specified kind) of dress, styling, etc.”
- “a. At a look: at a (single) glance, on cursory examination.
- “d. On (also upon) the look: engaged in searching for something; keeping watch.”

A collocation analysis would be required to find the distribution of the specific use of the term “look”.

My data set is too small to draw statistically significant conclusions; however, the strong dominance of the term “look” in both events and nonevents may not be a coincidence in screenplays written for a medium dominated by the act of looking in both the audience and the camera, or point-of-view. In their book *New Vocabularies in Film Semiotics,* Robert Stam et al.

emphasize the importance of point-view: “The category of point-of view is one of the most important means of structuring narrative and one of the most powerful mechanisms of audience manipulation” (p. 84). How the word “look” is used in screenplays, which one of its 21 dictionary entries dominates, is it associated with certain genres more than others, does it dominate the novel from which the screenplay was adapted, how is it used in terms of point-of view, are some of the interesting questions raised by this piece of data.

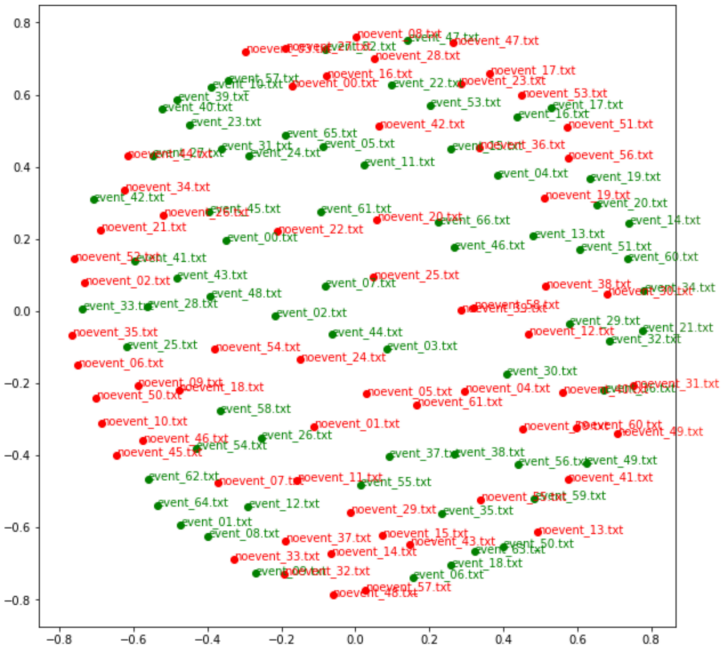
3.5 K-Means Clustering

I tried partitional clustering of the data in events and in non-events in order to find out if the data sets had characteristics which differentiated one from the other. I chose the K-means algorithm because of its ubiquity, flexibility, longevity, “Even though K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity” (Jain 657).

I tried it using two similarity measures: cosine similarity, and the Jaccard correlation coefficient; neither produced clustering (see GitHub: 4_K_Means_Clustering). Given the relatively small dataset I have curated, clustering proved to be a challenge. The clustering algorithm looks for similarities in the data which form the basis of the clustering; furthermore, it is unsupervised, so the algorithm looks for structures in the data with K set at two. This produced a completely undifferentiated group of both events and nonevents shown in Figure 2.

Example of K-means cluster with unfiltered data:

Figure 2

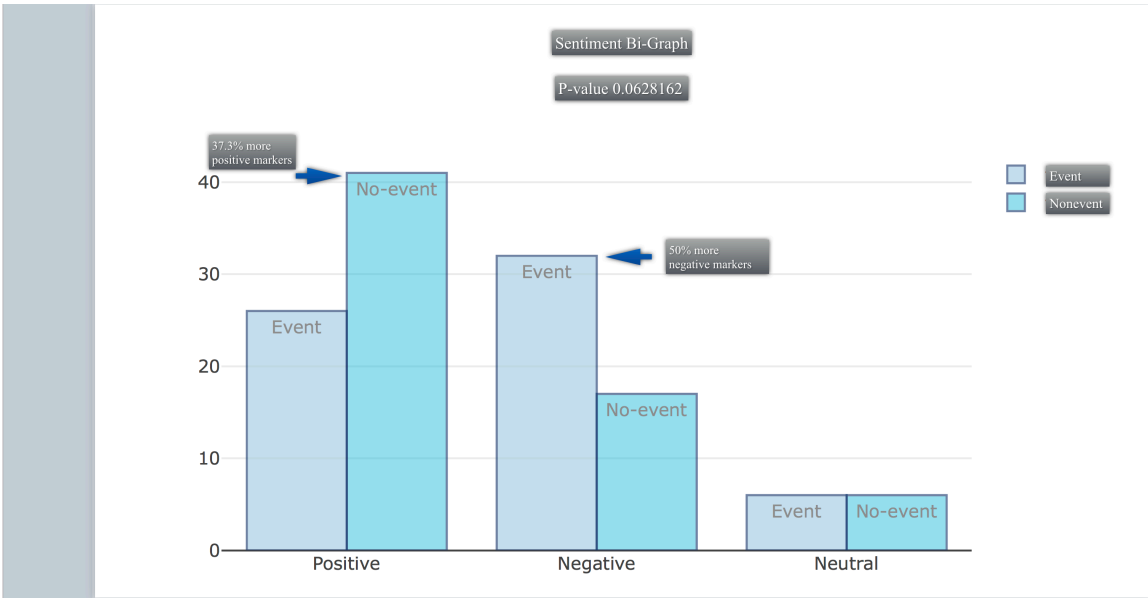


The data was organized as two filtered files, one for events and one for nonevents, both produced results similar to Figure 2. The output indicates that the semantic fields of the two datasets are not sufficiently distinct. If one looks at the most frequent term analysis, in the top ten terms there are two terms common to events and nonevents: “look” and “take”. “Look” dominates both events and nonevents by a significant margin. Another factor that is likely working against clustering is the wide range of genres the events and non-events are chosen from. The events and non-events come in all shapes, in contrast to a dataset chosen exclusively from the crime genre, for example, where the language in events may have been more consistently distinct from nonevents, this however is speculation.

“A cluster is a grouping based on attribute similarity, where items within a cluster are more similar to each other than to ones in another cluster” (Munzner 30). Does this output mean that there is no lexical distinction between events and non-events? Based on this cluster analysis,

it seems not; although, other types of analysis indicate that there is a difference in the language between the two datasets. For example, there is a clear difference between binary sentiment scores in events and non-events as shown in the bar graph below:

Fig.3



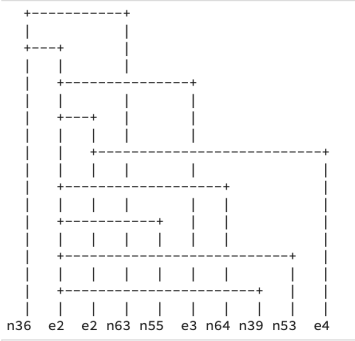
Based on the most frequent terms analysis, there is an overlap of terms between events and nonevents which indicates that unsupervised partitioning may require a much larger data set to determine if there are clusters of events separated from clusters of non-events.

Hierarchical Clustering

Having failed to produce clusters using an unsupervised partitioning method with K-means method, I decided to try hierarchical clustering, also unsupervised, with NLTK's GAAClusterer. The data is vectorized with Tfidf from Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

Initially I tried to vectorize the data using a DOC2VEC method, where each document is rendered as a vector. The results of clustering with this method were muddled. I then tried to vectorize the data using the Tfidf method. This method yielded clear results shown in Fig. 4. The data set is the same as the one used in the K-means analysis. The results are similar in that no hierarchical clustering is outputted. In both events (Type II's) and nonevents (non Type II's) the overlap of the language used is clearly indicated in the scatter plot outputted in Figure 2. The language used in both of these datasets is too similar to differentiate. When one looks at the most frequent terms, there are salient overlaps:

Fig.3a



3.6 Parts of Speech

The goal here is to tag tokens with parts of speech using the Penn Treebank Project in order to see if certain parts of speech occur more in one data set over the other so as to differentiate them. The following is the code that tags the tokens in rebuilt-events:

Fig.4

```
# 1. Get the files
file_list = [file for file in os.listdir(PATH + 'rebuilt_events') if file not in ['.DS_Store', '.i
pynb_checkpoints']]

# 2. Assemble the corpus (tokenize -- also filters punctuation marks -- filter out stopwords)
tokens = list()
for file in file_list:
    tokens.append(tokenize(PATH + 'rebuilt_events/' + file))

cleaned_tokens = list() #[]
counts = dict()

for document in tokens:
    cleaned_tokens.append(list())

    for sentence in document:
        #Remove stopwords
        no_stopwords = remove_stopwords(sentence)
        #Stem
        stemmed = stem_tokens(no_stopwords)
        #Tag everything

        tagged = pos_tag(stemmed)
        cleaned_tokens[-1].append(tagged)

        for token in tagged:
            category = token[1]

            if category in counts:
                counts[category] += 1
            else:
                counts[category] = 1

counter = 0
print(len(cleaned tokens))

with open(PATH + 'Event_export.json', 'w') as outfile:
    outfile.write(json.dumps(counts))
```

The following bar graph, Fig.5, shows an output of the tagged parts of speech for events and nonevents. Figure 6 provides the POS tag description.

Fig.5

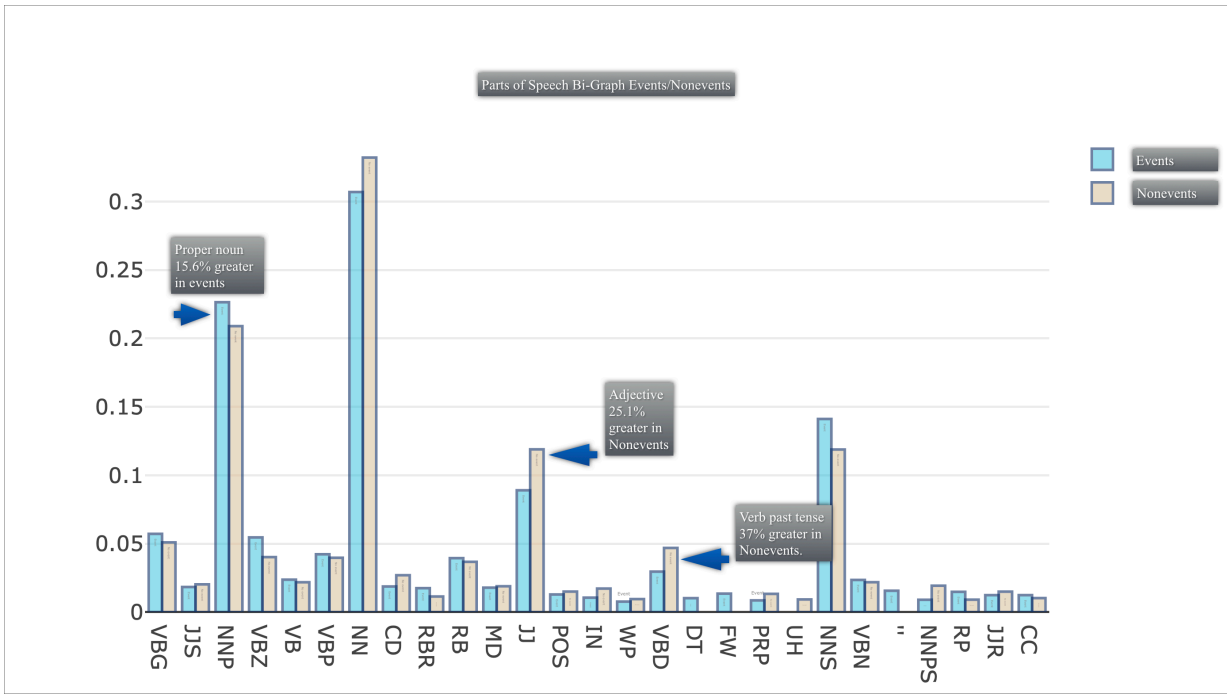


Fig.6

Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

There are differences between the two data sets, for example, adjectives occur 25% more often in nonevents than in events. Nonevents are expositional, telling us about the characters' background, the social and physical settings, relationships between characters, etc. That adjectives, which are descriptive, dominate nonevents is therefore not surprising. Other outputs are less clear. Singular proper nouns occur 15.6% more frequently in events for which there is no narratologically obvious explanation, nor is there a ready explanation for why verbs occur 37% more often in nonevents.

3.7 A Sentiment Analysis: modelled on *The Art of Literary Text Mining* (ALTM)

"Sentiment analysis is a general term used in text mining to refer to the process of trying to automatically determine the mood or opinion of texts." *The Art of Literary Text Mining*. I will analyze sentiment using two different methods: one categorical, the other numerical.

A. Categorical Sentiment Analysis:

The following sentiment analysis is categorical, classifiying each sentiment marker as positive, negative, or neutral.

In [2]:
```
from nltk.tokenize import sent_tokenize, RegexpTokenizer
from tqdm import tqdm_notebook

from pandas import DataFrame

import matplotlib.pyplot as plt

PATH= '/Users/josephrafla/Desktop/'

import os
```

Preparing reference vocabularies: load the tokens tagged positive/negative from the two sample vocabulary files.

In [3]:
```
positive, negative = list(), list()

with open('positive.txt', 'r') as pos_file:
    for line in pos_file:
        positive.append(line.strip())

with open('negative.txt', 'r') as neg_file:
    for line in neg_file:
        negative.append(line.strip())
```

In [4]:
```
print("{} positive words, {} negative words.".format(len(positive), len(negative)))
```

```
2230 positive words, 3905 negative words.
```

Preparing test data: load corpus and tag it as 'noevent' or 'event'.

In [5]:
```
test_files = [file for file in os.listdir('training_AllData_updated') if file not in ['.DS_Store']
]

corpus = list()

for test_file in test_files:
    with open('training_AllData_updated/' + test_file, 'r', encoding='utf-8') as infile:
        tag = 'noevent' if 'noevent' in test_file else 'event'
        corpus.append((' '.join(infile.readlines()), tag))
```

In [6]:
```
print("{} file(s) in corpus.".format(len(corpus)))
```

```
135 file(s) in corpus.
```

We can also define tokenizing and scoring functions.

Tokenizing simply splits documents into sentences and sentences into words.

Scoring involves counting the words that are in the token list and that are either positive, negative, or neutral. The final score is based on the difference between the positive and negative counts.

In [7]:
```
def tokenize(document):
    tokens = list()

    word_tok = RegexpTokenizer('[\w\']+')

    sentences = sent_tokenize(document)

    for sent in sentences:
        tokens += word_tok.tokenize(sent)

    return tokens
```

Fig.7

Initially, I used a categorical method of classification where a sentiment marker is “positive”, “negative”, or “neutral”. The following partial sample of code, Fig.7, is used to parse the data into one of the three categories. The source, i.e. the dictionary, against which sentiment markers are categorized is provided by NLTK.

To visualize the results better, we plot the results by grouping the three categories (positive, negative, neutral) using events and non-events.

Positive is represented in green, negative in red, neutral in yellow.

To visualize the results better, we plot the results by grouping the three categories (positive, negative, neutral) using events and non-events.

Positive is represented in green, negative in red, neutral in yellow.

In [12]:
```
plt.bar(0, counts['noevent']['positive'], color ='g')
plt.bar(1,counts['noevent']['negative'],color = 'r')
plt.bar(2,counts['noevent']['neutral'],color = 'y')

plt.bar(4,counts['event']['positive'], color ='g')
plt.bar(5,counts['event']['negative'],color = 'r')
plt.bar(6,counts['event']['neutral'],color = 'y')

plt.xticks([1,5],['Non-events','Events'])
plt.show()
```

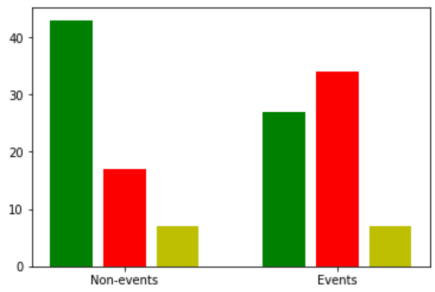


Fig.8

Figure 8 above, shows a graph of the output for the categorical sentiment analysis:

The following bar graph shows the categorical output for event and nonevent:

Figure 9

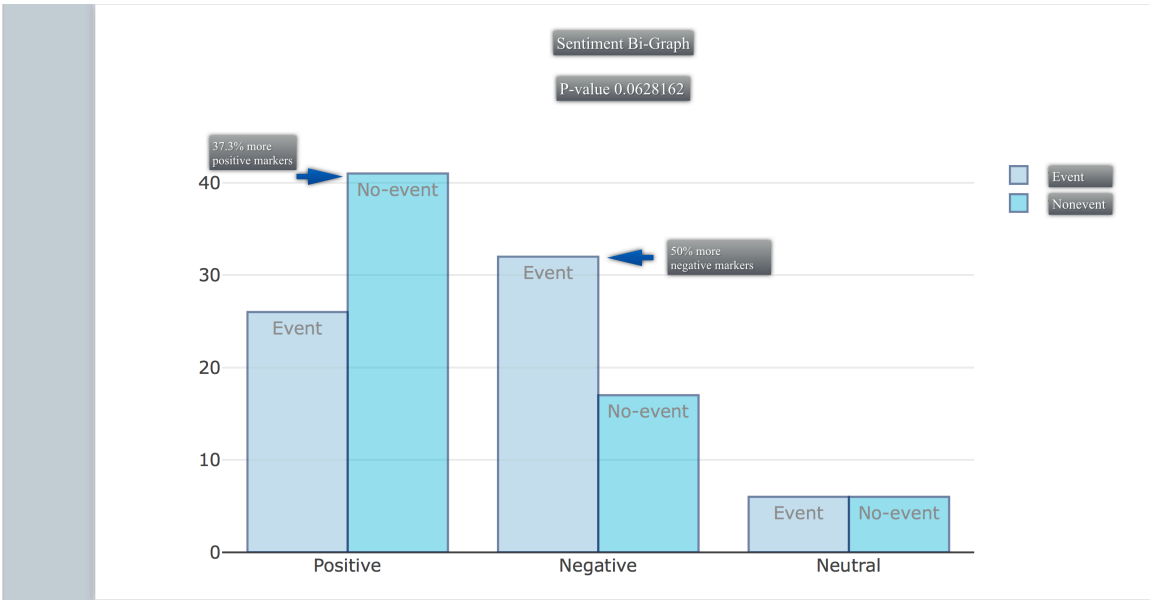


Figure 9 shows that negative markers are 50% greater in events. This result is in line with the narratology underlying the arch-type screenplay structure, the great majority of which are in three acts. The first act complicates, or makes dire the protagonist's situation, and such complications are full of negative emotional markers. The second act will deepen the drama, raise the stakes, create further obstacles, most of which will have a negative mood (McKee 44). Given the prevalent use of this kind of narrative structure in Hollywood screenplays, the output of the event vs. nonevent sentiment analysis is reassuring. The P-value analysis is applied to the binary sentiment analysis. In this case the P-value = 0.0625.

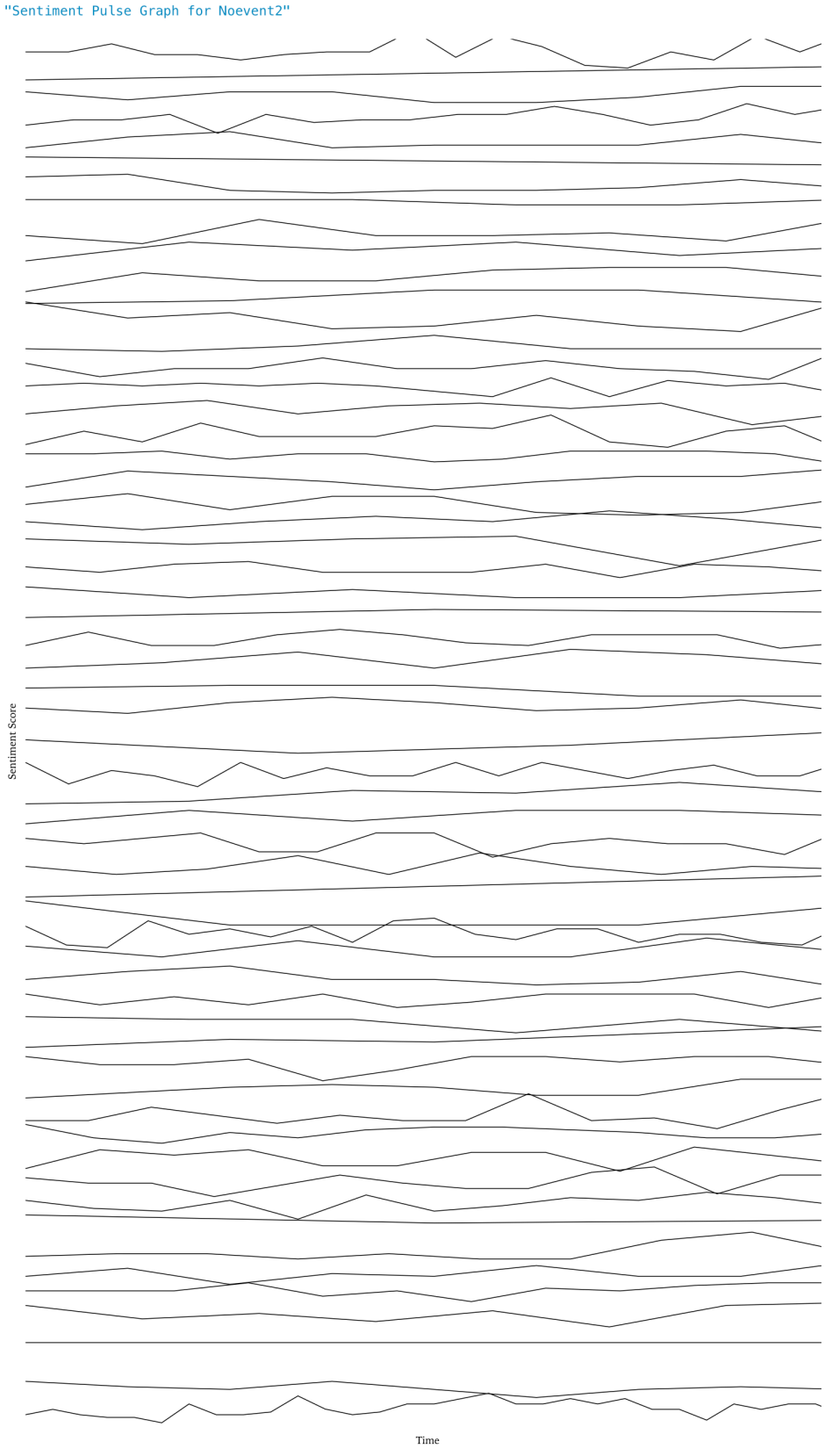
3.7B Numerical Sentiment Analysis

In this second sentiment analysis I will use the "WordNet" sentiment analysis method (ALTM). This method outputs a score for each sentiment marker, and modifies that score depending on the term's parts of speech: noun, adjective, verb, adverb. The sentiment score is weighted such that the most frequently used version of the term is listed first. So, the term "good" has a positive score of 0.5, because its most frequent version, a noun is given that score by the NLTK resource, Wordnet: "good.n.01: PosScore=0.5 NegScore=0.0>.

This statistical approach makes the process of creating a continuous score feasible, otherwise the number of variations given the polysemic nature of language would require a far more complex base dictionary. The solution in WordNet is a comprise between breadth of coverage, and precision which gives a reasonable representation of the data. An advantage to this analysis is that it is automated, i.e., it requires no heuristic intervention, unlike the process of gathering the data, which required careful interpretation. And most importantly for the purpose of visualization, it is a continuous measure.

Figure 10 below shows the sentiment score for each file, in this case the nonevents file, as a line graph which travels from the beginning to the end of the nonevent.

Fig. 10



3.8 Appendix: Machine Learning

This was an early exercise to see if a machine could be trained to distinguish events from nonevents. I tried 10 classifiers, and found the Naive Bayes classifier output the best accuracy/ precision scores. I do not include it with the other analysis because I could not see any potential to visualize the outputs. However, I did find it exciting that on average the accuracy/precision numbers came back in the low 70%. So, seven out of ten times the machine could accurately distinguish between the language of an event from that of a nonevent. The only other computational test that returned a strongly differentiated signal between events and nonevents was the sentiment analysis; in that case the output is not only clearly differentiated, but also conforms to the classical narrative theory which underpins the arch-type screenplays.

Before proceeding: Now that there is a corpus of tagged document, I *shuffle* them so that the training data sample is somewhat representative of the whole corpus.

In [5]:
```
#to avoid the order of events first and non-evnets second
import random
random.shuffle(corpus)
```

Fitting the model. To prepare our corpus data, I will use a vectorizer and a classifier, both implemented in `sklearn`.

In [6]:
```
import numpy
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

training_split = int(len(corpus) * 0.8)

count_vectorizer = CountVectorizer(ngram_range=(1, 2)) # uses text more than other vectorizers (fe
ature extractor
counts = count_vectorizer.fit_transform([docs[0] for docs in corpus[:training_split]])

tfidf_vectorizer = TfidfVectorizer()
tfidf = tfidf_vectorizer.fit_transform([docs[0] for docs in corpus[:training_split]])
```

In [7]:
```
from sklearn.naive_bayes import MultinomialNB #they talk about text and word counts.

classifier = MultinomialNB()
targets = [label[1] for label in corpus[:training_split]]
classifier.fit(counts, targets)
```

Out[7]:
```
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

In [8]:
```
from sklearn.linear_model import LogisticRegression

classifier2 = LogisticRegression()
classifier2.fit(tfidf, targets)
```

Out[8]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Testing the model

The remainder of the corpus is now used to test the model and see how well it does by using the vectorizer to vectorize the remaining documents and use the classifier to predict their category.

In [9]:
```
example_counts = count_vectorizer.transform([document[0] for document in corpus[training_split:]])
example_tfidf = tfidf_vectorizer.transform([document[0] for document in corpus[training_split:]])
predictions = classifier.predict(example_counts)
predictions2 = classifier2.predict(example_tfidf)
```

In [10]:
```
targets = [labels[1] for labels in corpus[training_split:]]
```

In [11]:
```
from sklearn.metrics import precision_score, accuracy_score

print("Precision score: {}".format(precision_score(targets, predictions, average='macro')))
print("Accuracy score: {}".format(accuracy_score(targets, predictions)))

print("Precision score: {}".format(precision_score(targets, predictions2, average='macro')))
print("Accuracy score: {}".format(accuracy_score(targets, predictions2)))
```

```
Precision score: 0.7990196078431373
Accuracy score: 0.782608695652174
Precision score: 0.6477272727272727
Accuracy score: 0.6521739130434783
```

Classifying events and non-events using Scikit-Learn. This example is a simple document classifier fitted with event/non-event data and can get decent-ish accuracy when testing with similar documents.

Preparing data: to prepare the data, I created tagged documents: each document's content is associated with a tag that is either *event* or *noevent*. These tags will later be used for classification.

In [1]:
```
import os
```

In [2]:
```
PATH = 'training_20180411/'
```

In [3]:
```
excluded = ['.DS_Store']
corpus_files = [filename for filename in os.listdir(PATH) if filename not in excluded]

corpus = list()

for file in corpus_files:
    with open(PATH + file, 'r', encoding='utf8') as infile:
        document = infile.read()
        tag = 'noevent' if 'noevent' in file else 'event'

        corpus.append((document,tag))
```

Extra processing to boost the precision of the classifier by filtering stopwords, stemming, etc.

In [4]:
```
from nltk.tokenize import sent_tokenize, RegexpTokenizer
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer


import nltk
nltk.download('wordnet')
clean_corpus = list()

tokenizer = RegexpTokenizer('[\w\']+')
lemmatizer = WordNetLemmatizer()

jockers_stopwords = list()
with open('jockers_stopwords.txt', 'r') as infile:
    for line in infile:
        jockers_stopwords.append(line.strip())

for document in corpus:
    sentences = sent_tokenize(document[0])

    clean_doc = list()

    for sentence in sentences:
        tokens = tokenizer.tokenize(sentence)
        tokens = [lemmatizer.lemmatize(tok.lower()) for tok in tokens if tok.lower() not in stopwords.words('
english') and tok.lower() not in jockers_stopwords]

        clean_doc.append(" ".join(tokens) + '.')

    clean_corpus.append(" ".join(clean_doc))

corpus = [(clean_corpus[idx], corpus[idx][1]) for idx in range(len(corpus))]
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/josephrafla/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

The first of these supervised training attempts returned a predictive accuracy of 50% or less. In the data I was initially using, events was made up of big and small events, between 150-450 words. I subsequently adjusted events so that all events were one size, up to 250 words. When I tried the classification exercise with this newly configured event data, the results of the predictive accuracy moved to 68%. I have not included the supervised training research in the table of contents, because it did not, as far as I could see, provide any potential for visualization.

I found the results of the sentiment analysis and the machine learning quite encouraging. There are linguistic signals which differentiate events from nonevents, as I have defined and collected them. I do not believe these to be idiosyncratic and random, especially the sentiment analysis, which conforms to classical narrative theory.

Chapter 4

Visualizing Salient Language in Events and Nonevents

4.0 Introduction

The immediate challenge in writing about visualizing the output data generated in chapter three is the huge breadth of possibilities within the field of data visualization, "The design space of possible visualization idioms is huge, and includes the considerations of both how to create and how to interact with visual representations" (Munzner xiv). In order to write cogently about applying data visualization, I will organize the majority of my research in this chapter around two main topics: 1) Data types and representations and 2) Spatial representation of non-spatial data. I will draw on the work of several scholars in this field: Tamara Munzner, Erik Cambria, Melanie Troy, Ian Milligan, Kieran Healey, and Stéfan Sinclair, among others. (A note on abbreviations: vis = visualization, scivis = scientific visualization, infovis = information visualization;; scivis and infovis are defined at the beginning of section 4.1).

The question arises, why use visualizations? "Vis systems are appropriate for use when your goal is to augment human capabilities, rather than completely replace the human in the loop" (Munzner 3). This is a guiding design principle I follow in implementing choices about the salient elements of the visualization design, to make the data approachable, engaging, and hopefully insightful. Furthermore, another reason for using data visualization, "By enlisting computation, you can build tools that allow people to explore or present large datasets that would be completely unfeasible to draw by hand, thus opening up the possibility of seeing how datasets

change over time" (Munzner 4). The type of screenplays from which the data are drawn, blue prints for a film production, are a form that has since 1963 evolved to, what McKee calls the arch-plot or "classical" design (McKee 44). Do screenplay events written in the 1940's, for example *Citizen Kane* (1941), differ from screenplay events written in the 1990's? This is the kind of question that a visualization, of say sentiment analysis, may be able to reveal.

I will present three types of data visualizations of linguistic features in events and non-events, based on the outputs of the computational analysis done in chapter three. They are: 1) Word cloud, 2) Interactive Part of Speech Bubbles, and 3) SentiPulse vis: towards a sentiment analysis visualized as landscape. I use the library of visualizations on the website Observablehq.com. This is a site that provides functioning code in Javascript, HTML, and CSS for each visualization. The site was started by Mike Bostock; however, over time others have contributed to the Observable's platform.

4.1 Information Visualization Contrasted with Scientific Visualization

There are two broad categories describing data visualization: Scientific visualization and Information visualization. The following definitions are provided by Shawn Graham et al. in their experimental on-line book, *The Historian's Macroscope: Big Digital History*:

> Scientific visualizations maintain a specific spatial reference system, whereas information visualizations do not...Bar chars, scatter plots, and network graphs...are all information visualizations, because they lay out in space data which do not have inherent spatiality.

In the IEEE Visualization conference (2003) a panel titled "Information and scientific visualization: separate but equal or happy together at last" provides this definition:

> Scientific visualization is frequently considered to focus on the visual display of spatial data associated with scientific processes such as the bonding of molecules in computational chemistry. Information visualization examines developing visual metaphors for non-inherently spatial data such as the exploration of text-based document databases" (Rhyne, T).

The *The Historian's Macroscope: Big Digital History* and the IEEE conference definitions are very similar; however, the conference definition includes mention of "visual metaphor", whose strengths and weaknesses will be a concern in this chapter. Visual metaphors are within the domain of information graphics, and to date it seems that scientific visualizations do not actively employ visual metaphor; it's possible to differentiate the two methods, scivis and infovis, by the complete absence of metaphor in one and the necessity of metaphor in the other (Troy).

4.2 Metaphor in Information Visualization

There are at least two levels of metaphor in an infovis. The first metaphorical interpretation is that one is taking non-spatial abstract data and giving it a spatial visual representation. The second level of metaphor is more visually direct. For example, Stéfan Sinclair's "Word Count Fountain", part of Voyant tools, is an algorithm which takes text and represent's word frequency as a fountain's flow of words, "visualizes word frequencies as a fountain. Each stream represents a unique word, where its height represents frequency the term occurred in the corpus" (Sinclair). A metaphor is not needed to visualize a physically instantiated element. It is the abstract element that needs a metaphor to make it more readily accessible to human understanding. For example,

if one is attempting to visualize brain activity, a representation of the brain, one with which we are all familiar, is not a metaphor for a brain, just a representation.

Not a metaphor:

Figure 4

Photo credit: Tom Wilson
(https://www.the-gma.com/how-is-ai-capable-of-empowering-a-data-driven-content-strategy)

A metaphor: “Word Count Fountain”. On the Voyant Tool Documentation page, http://docs.voyant-tools.org/tools/wordcountfountain/, is a text analysis tool which uses the metaphor of a flow of fountain water to show the frequency of a term or tokens in a text.

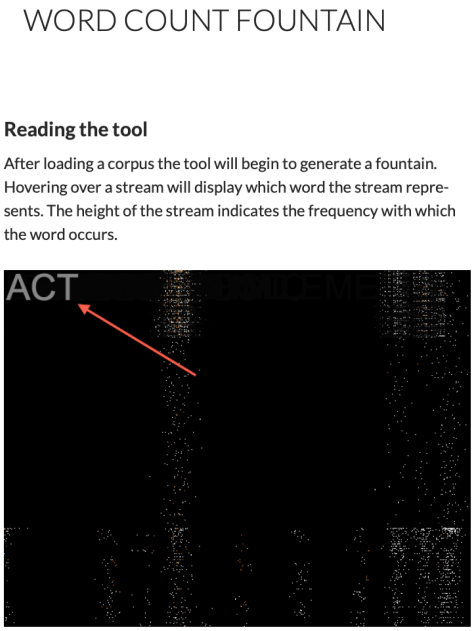


Figure 5 From Voyant’s suite of tools.

Jan Christiansen, senior graphics editor at Scientific American, makes distinction between scivis and infovis through a set of illustrations (Christiansen):

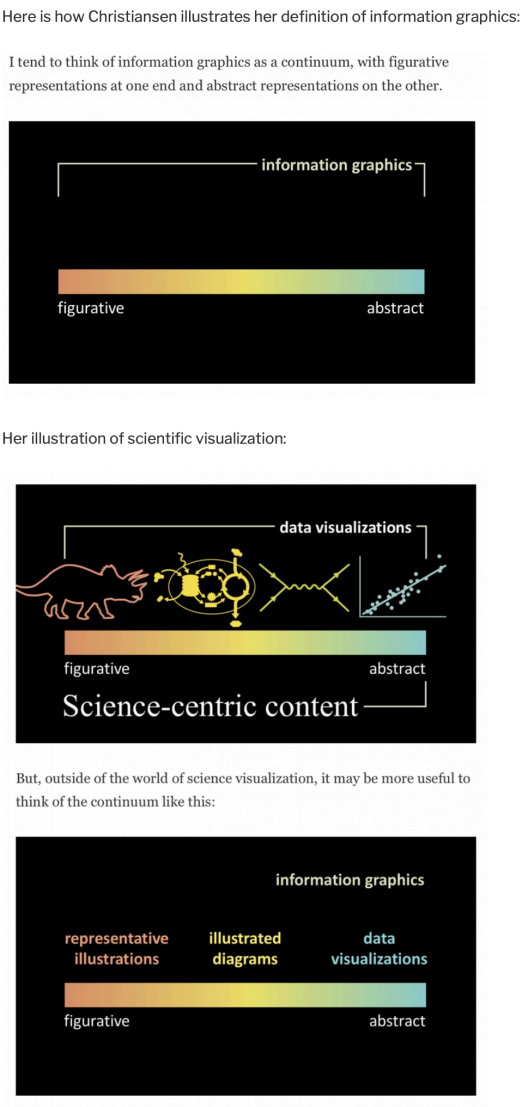


Fig. 6 (Christiansen)

As a practitioner Christiansen places all visualizations on a continuum, from the figurative to the abstract. Her definition is quite fluid. The interaction of these different visualization types produces hybrid types; the boundaries between them is not hard and fast. That is what is implied by Christiansen's illustration of a continuous scale of infographics in Fig. 6.

The work I have been doing is on the far right of Christiansen's scale: the data I'm analyzing does not have inherent spatiality, is abstract, and requires a metaphorical interpretation in order to render it visually.

Following are some of the differences between infovis and scivis expressed by scholars on the IEEE panel:

> "I'm interested in papers with a higher novelty factor: new ways to look at data or information, new ways to interact with the visualization, new theories about why some visualizations work better than others. These seem to be taking place more frequently in the InfoVis community, in my opinion" (Rhyne,T., Ward,M.).

> "The current names are rather unfortunate accidents of history: scientific visualization isn't uninformative, and information visualization isn't unscientific. However, it's been over a dozen years since the term "information visualization" was introduced, and at this point I think it would be more confusing to change the names than to keep them" (Rhyne,T. , Munzer,T.).

> "There are also similarities between infovis and scivis. The "vis" at the end means that both can benefit from design, from perceptual psychology input, from application feedback, and from scientific testing" (Rhyne,T., Laidlaw, D.).

> " [scientific] Visualization is bigger, older, more selective, more heterogeneous, and, dare I say, a bit stodgier and less creative. Infovis is smaller, younger, less established, more homogeneous, and more novel and creative" (Rhyne,T., Laidlaw, D.) .

In the following two sections, I will outline definitions of data and datasets and then apply them to my data in order to better understand and categorize the three visualizations I've chosen to create.

4.3 Data Types and Dataset Types

"Many aspects of vis design are driven by the kind of data that you have at your disposal. What kind of data are you given? What information can you figure out from the data, versus the meanings that you must be told explicitly" (Munzner 21).

The type of the data, its mathematical interpretation, can be classified as an item, a link, an attribute, a grid, or a position (Munzner 24):

1. An item is a discreet, individual entity.
2. An attribute is a measurable property of the data.
3. A link indicates a relationship between items.
4. A grid describes a strategy for sampling continuous data in terms of geometric and topological relations between cells.
5. A position indicates spatial data providing location in a 2D or 3D space.

Datasets are specific ways to organize data. "A dataset is any collection of information that is the target of analysis" (Munzner 24). Each dataset is configured to accommodate certain types of data. According to Munzner, there are five types (Munzner 24-27):

1) Tables: items, attributes.
2) Networks and Trees: items (nodes), links, attributes.
3) Fields: grids, positions, attributes.
4) Geometry: items, positions.
5) Clusters, Sets, and Lists: items.

Within any dataset there are five core data types (Munzner 23):

A. Items
B. Attributes
C. Links
D. Positions
E. Grids

My purpose here is not to make a comprehensive survey of data types, but rather to understand how the three visualizations I have made fit into the typology of data types and datasets.

4.4 Word Cloud Visualization of Most Frequent Terms

The following visualizations are adapted form a library of data driven visualizations available on the ObservableHQ site: https://observablehq.com . This is a set of open source visualizations created by Mike Bostock and others. The visualizations are provided with their accompanying code, which is modified to visualize the data for the three ensuing visualizations from chapter three: most frequent terms, parts of speech frequency, and sentiment analysis.

This word cloud is based on Jason Davies' visualization in ObservableHQ: https://observablehq.com/@fil/word-cloud. It is a word cloud of the most frequent terms, i.e. single words, in events and nonevents data. The more frequent the term the larger the font size. However, I made one addition: the top ten most frequent terms are rendered in black, which lifts them out of the colour field and turns them into coherent graphic elements; consequently, they are salient and thus easier to discern. This is shown below in Figure 7 as two versions of the word cloud for event data: Figure 7a has the top ten terms rendered in black, while Figure 7b shows the top ten terms as elements of the colour field. Fig. 7b differentiates the top ten terms not only by size, but also through colour. Strictly speaking black is not a colour, hence it becomes a graphic element.

When you look at Fig. 7a and 7b, the top ten terms in 7a are much easier to discern, because 7a uses both size, the more frequent the term the bigger the font, and graphics, black, to differentiate the top ten terms.

Figure 7a

Figure 7b

In terms of dataset types, these word clouds are basically a visual representation of a table, with data types being item, i.e. words, with one attribute, i.e. frequency. For event data a sample table looks like this:

Word (item)	Frequency (attribute)
look	57
shot	41
head	30
door	29
close	29
night	27
face	27
away	27
take	26
right	25

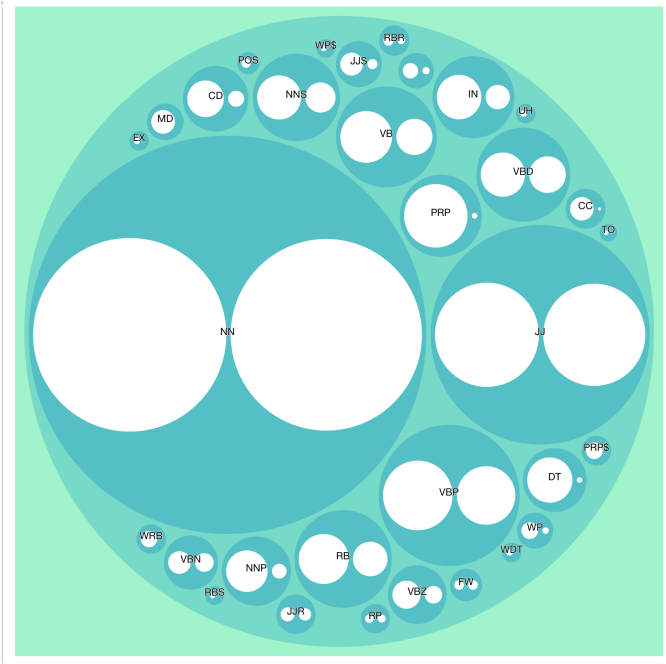
Figure 9 Event's most frequent terms.

Does the word cloud visualization add information that the table does not provide? No it does not; however, the word cloud has several advantages. It provides a much higher aesthetic value than the table. The first requirement of communicating this information is that a reader is drawn to look at it. In this regard aesthetics make a difference, especially if the audience is the general public, i.e. they are not being paid to read the analysis. Another advantage comes to the fore when comparing the two word clouds: the differences and commonalities are visible at a glance, and with the addition of black for the10 most frequent terms, the dominant data is easy to read.

4.5 POS Bubbles

The second of my customized visualizations is a part of speech bubble set based on Observable's: https://observablehq.com/d/847bc0c8b16273c8

Figure 10 Interactive parts of speech bubble includes the analysis of 35 parts of speech.



This comparative, interactive, visualization takes the frequency of a part of speech from event and non-event sets and represents it as a bubble. The greater the frequency the larger the bubble. If one clicks on a bubble, say NN (singular nouns), the following image appears:

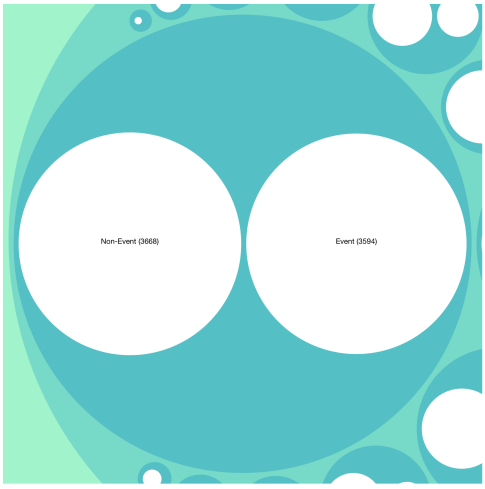


Figure 11 Click on the part of speech, in this case nouns, and the visualization shows that part of speech as a binary comparison between events and nonevents.

The frequency of the parts of speech is indicated in parenthesis next to the the relevant file. In nonevents there are 3668 singular nouns, and in events there are 3594. Clicking on this image returns the original visualization. There are 35 parts of speech represented in this interactive visualization. To try the POS Bubbles' interactive vis, go to: https://beta.observablehq.com/d/847bc0c8b16273c8.

As with the word cloud, this dataset type is a table with items and attributes as data types. This is a sample of the POS table: Figure 11 Sample table for Parts of Speech

Part of Speech (item)	File	Frequency (attribute)	File	Frequency (attribute)
Noun, singular	event	3594	nonevent	3668

Part of Speech (item)	File	Frequency (attribute)	File	Frequency (attribute)
Adjective	event	1062	nonevent	1023
Verb, base	event	128	nonevent	262
Adverb	event	245	nonevent	117

The advantages of a visualization of this size, i.e. 35 parts of speech, is that the vis immediately shows at a glance both the global relationship of the parts of speech, and which parts of speech dominate the dataset. Because POS Bubbles is an interactive vis, it allows a comparison between different parts of speech, and between the POS in events and nonevents. The bubble vis takes what are discreet data and shows them as a whole, and when you click on the part of speech, it shows a pairwise relationship between the POS in event and nonevent sets, specifying the exact frequency of the item. The vis has the further advantage of being more attractive than a table, thus increasing the engagement of communicating this POS information, “Before you can even begin to communicate a complex topic, you must first engage an audience” (Christiansen 49).

4.6 Sentiment Analysis: Toward a SentiPulse Visualization

“Existing approaches to affective computing and sentiment analysis fall into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches” (Cambria). It is the first of these, knowledge-based techniques, that I will use in analyzing the sentiment scores of the event and nonevent files.

I adapted code from *The Art of Literary Text Mining*’s, “Sentiment Analysis With WordNet”: https://github.com/sgsinclair/alta/blob/b872c20b402fd2a070209a586a7586228b8b0bbd/ipynb/SentimentAnalysis.ipynb. The text is

tokenized to create a bag of words. Within these words are sentiment markers, i.e. words that are unambiguously affective. These markers are then compared to a pre-existing dictionary such as WordNet-Affect, Affective Lexicon, Senti-WordNet, all of which are trained on text corpora, and each sentiment marker is given a score (Cambria).

The following is from *The Art ofLiterary Text Mining*:

```
from nltk.corpus import sentiwordnet as swn
for senti_synset in swn.senti_synsets('good'):
    print(senti_synset, senti_synset.synset.definition())
```

```
<good.n.01: PosScore=0.5 NegScore=0.0> benefit
<good.n.02: PosScore=0.875 NegScore=0.0> moral excellence or admirableness
<good.n.03: PosScore=0.625 NegScore=0.0> that which is pleasing or valuable or useful
<commodity.n.01: PosScore=0.0 NegScore=0.0> articles of commerce
<good.a.01: PosScore=0.75 NegScore=0.0> having desirable or positive qualities especially those su
itable for a thing specified
<good.a.03: PosScore=1.0 NegScore=0.0> morally admirable
<well.r.01: PosScore=0.375 NegScore=0.0> (often used as a combining form) in a good or proper or s
atisfactory manner or to a high standard (`good' is a nonstandard dialectal variant for `well')
<thoroughly.r.02: PosScore=0.0 NegScore=0.0> completely and absolutely (`good' is sometimes used i
nformally for `thoroughly')
```

Figure 14 The output of Sentiwordnet for the word “good”.

The word “good” is analyzed through NLTK’s Sentiwordnet resource. The various scores of “good” depend on how it is used as a part of speech. The grammatical context narrows the scores. Therefore a part of speech is tagged using NLTK’s POS tagger. This reduces the number of options: “good” as a noun has three outputs:

- good, as in a benefit, PosScore=0.5 NegScore=0
- good as in moral excellence or admirableness has a PosScore=0.875 NegScore=0
- good, as in that which is pleasing or valuable has a PosScore=0.625 NegScore=0
-

“Good”, as an adjective, outputs the following scores:

- good, as in having desirable or positive qualities. PosScore=0.75 NegScore=0

- good, as in morally admirable. PosScore=1.0 NegScore=0
- good, as in a good or satisfactory manner. PosScore=0.375 NegScore=0

The scores for good as an adjective range from 0.375 to 1.0. The scores are weighted from most often to least often used, “Here we have to make a decision: do we take into consideration every senti_synset or just the first and most common meaning?” (ALTM). I will use the sentiment value of the first term outputted, i.e., the most common meaning of the sentiment marker in the following analysis, because it is the most common meaning as determined by the NLTK resource.

The two terms common to both event and nonevent data sets are “look” and “room”. For example, the word “look” as a noun has four entries in the Oxford English Dictionary: 1) “The action or an act of looking”, 2) “With *of* or modifying word. An expression of a specified thought or feeling by looking.”, 3) “A person's (or animal's) appearance, esp. that of his or her countenance”, In *plural* in same sense. Later usually restricted to relatively permanent features; also *spec.*: attractive or beautiful appearance; = good looks *n.*", 4) “*Fashion*. An appearance, style, or effect (usually of a specified kind) of dress, styling, etc.” (OED look). The word “look” as a verb has thirteen entries, and the word looking adds more options. Please keep in mind that I filter the data so that a word like “looking” is stemmed to “look”, further complicating the analysis. I would like to refine the sentiment analysis so as to specify how a world like “look” is actually used in context, and how to score it which would require creating a new sentiment dictionary. However, such a modelling and programming task is beyond the scope of my thesis. Given this limitation, my choice to use the most common term as outputted by the NLTK SentiWordNet dictionary, as a resource, is a justifiable starting point.

4.7 Spatialization

In this section I present two types of sentiment analysis: 1) polarity detection and 2) WordNet sentiment scores, which I address first.

"People are familiar with spatial concepts such as distance and height as part of their everyday life. Spatialization takes advantage of this knowledge by using a spatial metaphor to display abstract, non-spatial data" (Tory). The spatial metaphor I will explore is one based on landscape. Typically this is represented through a three dimensional space; however, I will use a two dimensional plotting of the sentiment markers, and then I will fit a surface area to represent a sentiment landscape: "Keeping the points on a plane and then fitting a surface to the points can be used to create 2D visualizations resembling topographic maps" (Tory et al). Instead of a topographic map, the sentiment visualization will show the sentiment 'landscape' of the files in events and nonevents.

Spatialization in data vis has been most often applied to the distance similarity metaphor. The closer items are in the space the more similarities they share and vice versa. "Spatialization has been used for many applications, but most commonly for visualizing document collections" (Tory). In my thesis, spatialization will be used to render sentiment scores as a metaphorical spatialization, one that uses landscape as the central idiomatic metaphor. According to Tory et al., "Using a landscape metaphor may facilitate pattern recognition and spatial reasoning" (Tory). Furthermore, "Several authors suggest that the landscape metaphor is easily understood by most users and facilitates hierarchical clustering of data items" (Tory).

The goal of the analysis is to create a string of scores based on sentiment scores produced from the text. Continuing with NLTK as a primary resource in my research, I have chosen to use the SentiWordNet dictionary to create the sentiment scores. There is a trade-off in using this easily available, free resource, “Manually annotated lexica provide a high precision but lack in coverage, whereas automatic derivation from pre-existing knowledge guarantees high coverage at the cost of a lower precision” (Gatti). The advantage of SentiWordNet over other such resources is that WordNet outputs a continuous score from 0 to 1 to within three decimal places. Other approaches such as a customized Word List (ALTM) will not output a measurement, but rather a category: the sentiment word is either positive, negative, or neutral. “The basic tasks of affective computing and sentiment analysis are emotion recognition and polarity detection. Although the former focuses on extracting a set of emotion labels, the latter is usually a binary classification task with outputs such as “positive” versus “negative,” (Cambria).

In the next section I show a sentiment analysis based on polarity detection. Its output is a bar graph which shows the frequency of positive, negative, and neutral sentiment markers in events and nonevents. I used a categorical sentiment analysis from NLTK to produce the following output.

4.8 Categorical Sentiment Analysis

The following bar graph, Fig. 15, is of a sentiment analysis, one based on a positive/negative binary of sentiment markers in events and nonevents. The categorical output is derived from the NLTK Sentiment Analyzer, and shows a strong differentiation between the two files. Those

differences are in line with the classical narrative theory that underlies the types of screenplays in the corpus:

Fig.15

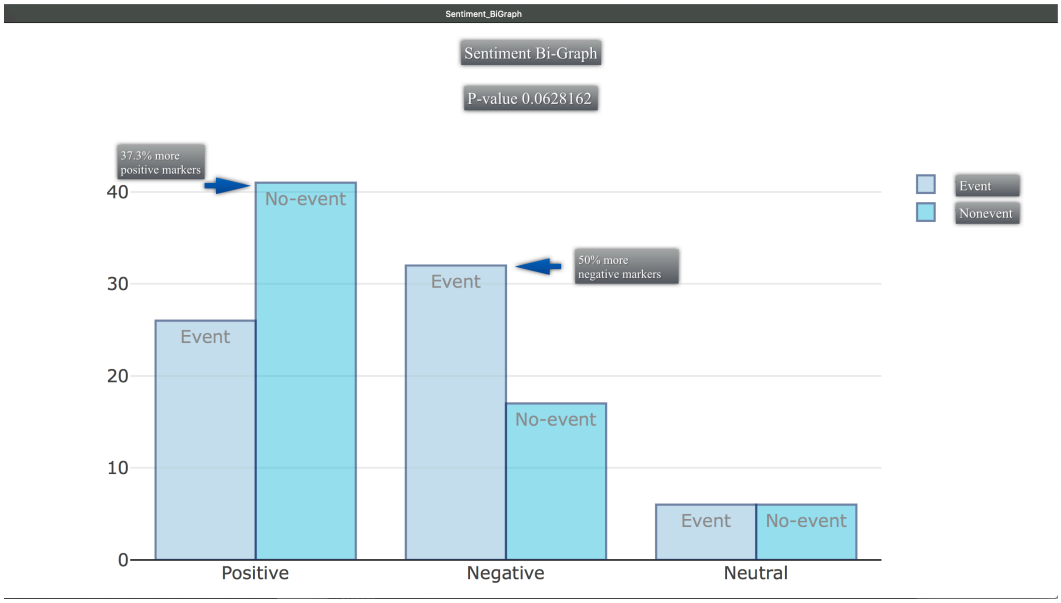


Figure 15. Sentiment analysis bar graph showing the distribution of positive, negative, and neutral sentiment markers in events/nonevents from NLTK's SentimentAnalyzer

In events, scenes which change the story's trajectory have negative markers 50% greater than nonevents. In the three act screenplay, which describes the majority of the examined screenplays, the end of the first act is a complication (negative), or end of act turn, for the protagonist; the second act involves one or more dire complications, and the third act will likely contain at least one story changing scene. Most if not all of these complications will carry a surfeit of negative markers, 50% more than nonevents according to this analysis.

The following sentiment analysis is of the same data sets. However this analysis, the WordNet analysis, outputs a continuous measure, not a binary categorization.

4.9 Events/Nonevents Data Visualized as Landscape

"We categorize spatialization into two groups based on the graphical mark used to represent data:

- **Points**: Spatialization that show only points.
- **Information Landscapes**: Spatialization where a surface has been fitted to the set of underlying points. Points may be shown on the surface. We refer to these simply as landscapes" (Tory).

The goal of the WordNet sentiment analysis is to first produce a set of points, the sentiment markers' score, and then to fit a surface so as to simulate a landscape. It remains to be seen whether such a visualization will be readily understandable.

One can argue that sentiment, as it manifests in human consciousness, is not a discreet phenomenon, but a continuous one, that is, more like a wave than a particle. The work of William James, although much updated and added to, is still fundamentally sound in helping to understand how to think about emotion in this context. One key element in James's epistemology views emotions as processes, in contrast to the reification of emotions as entities. Furthermore, at the risk of oversimplifying what is a complex set of psychological theories, James' description of the process of emotion was linked to a physiological cascade of responses. The following quote encapsulates the relation between the type of spatial sentiment visualization that I will make, and James's ideas about how emotions are experienced in human consciousness.

> "In the paragraph about the bear, James named three different emotions — fear, sorrow, and anger…The problem I want to talk about in this section is the tendency for language to reify the referent, for discrete terms to imply discrete entities: fear, sorrow, anger. It is unfortunate that James should have resorted to such a list in his most famous paragraph because, in general, he took great care to deny the existence of emotions as discrete entities. James wanted to argue, and did argue…that emotions, like consciousness, are continuous stream rather than a collection of separate states" (Ellsworth 226).

In James's insistence on emotional experience as a "continuous stream" lies the reason for choosing the "sentiment landscape" as an apt visual metaphor. Therefore, it is not unreasonable to seek a visualization which represents sentiment as a spatial field, the idiom most suited for continuous data.

However, how we measure sentiment through text analysis is a process which produces discreet bits of data, i.e. sentiment scores. Discreet data is counted, continuous data is measured (Munzner). Sentiment analysis does both: it counts the number of sentiment markers, and it measures each against a pre-existing dictionary established as a benchmark, in this case NLTK's WordNet, which outputs a continuous score. This establishes points within the field. The continuous aspect of the vis is a landscape fitted to the sentiment points. Munzner's following observation makes the underlying argument more succinctly: "Technically, all data stored within a computer is discrete rather than continuous; however, the interesting question is whether the underlying semantics of the bits that are stored represents samples of a

continuous phenomenon or intrinsically discrete data" (Munzner 28). My theory is that sentiment, as experienced in human consciousness, is both continuous and discreet; however, substantiating this claim is beyond the scope of my thesis.

The importance of sentiment is not to be underestimated: "Emotions play an important role in successful and effective human-human communication. In fact, in many situations, emotional intelligence is more important than IQ for successful interaction.[1] There is also significant evidence that rational learning in humans is dependent on emotions" (Cambria).

In screenplays emotions play a central role; however, the question is how are emotions expressed in events and nonevents represented, "We do not move the emotions of an audience by putting glistening tears in a character's eyes, by writing exuberant dialogue so an actor can recite his joy…or by calling for angry music. Rather, we render the precise experience necessary to *cause* an emotion" (McKee 243). In this regard both events and nonevents are designed to represent a specific emotional experience, whether or not it is a narratively consequential change of state, and the language which marks the designed "cause" of emotional resonance in an event or nonevent, is the object of the following sentiment visualization.

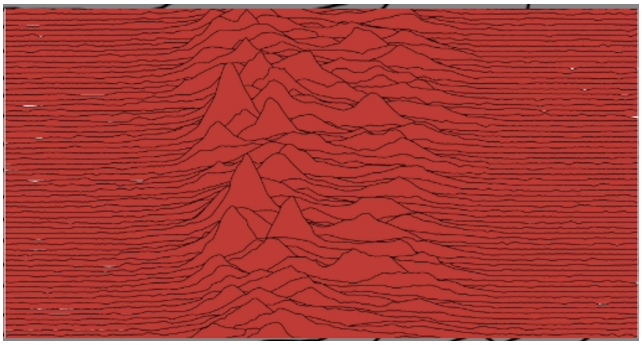
4.10 Towards a Sentiment Pulse and Unknown Pleasures

After searching the ObservableHQ site, I chose the following vis to adapt as a landscape metaphor in showing the sentiment scores of events and nonevents:

1) Borgar Porsteinsson's code recreates an image of the first signals from a Pulsar originally discovered by a team from the Mullard Radio Astronomy Observatory near Cambridge

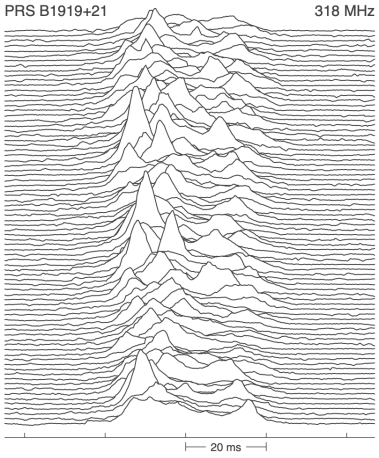
University in 1967. It is shown below with a crimson wash. The individual line graphs that make up this viz were published in, *Radio Observations of the Pulse Profiles and Dispersion Measures of Twelve Pulsars* by Harold D. Craft, Jr.

Figure 16



The original viz from ObservableHQ looks like this:

Figure 17

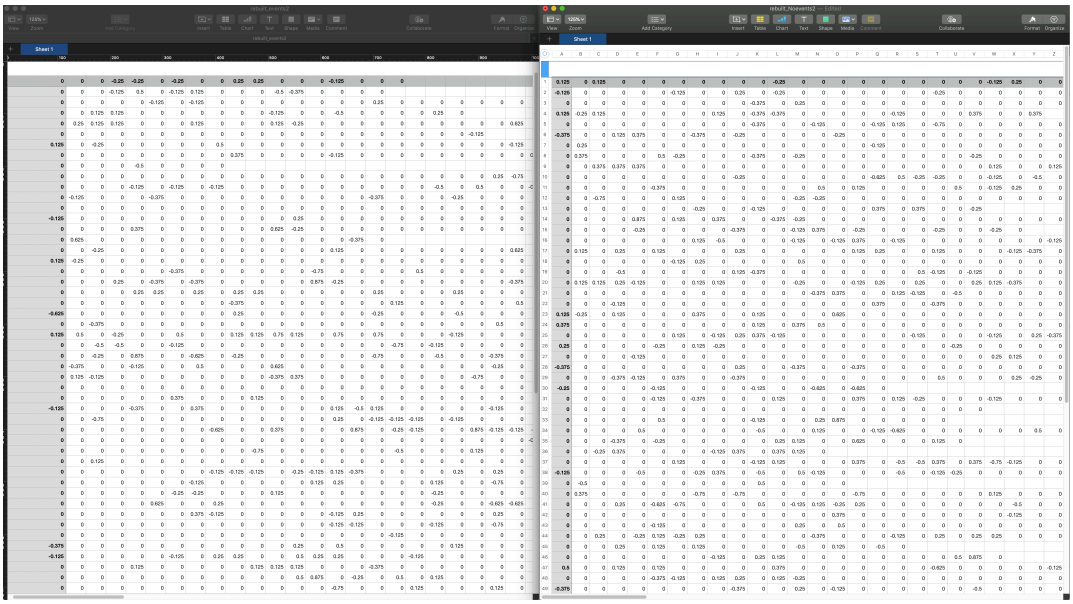


In my application of the above model, the graph lines are generated by the data, which are the sentiment marker signals outputted once the text has been scored through NLTK's SentiWordNet. Each line-graph represents a file, and then the files are stacked one above the other. I will make one vis for the events files, and one for the nonevents files.

A string of positive and negative numbers are plotted along the Y-axis. On the X-axis is the time passage of the data, and the Z axis gives a depth to the line graph, it does not represent data, and therefore this is not a true 3D vis. It is a stacked 2D vis which simulates landscape in pseudo 3D. The original vis is on observableHQ:https://beta.observablehq.com/@mbostock/psr-b1919-2. It is made by Borgar Porteinsson: https://bl.ocks.org/borgar/31c1e476b8e92a11d7e9.

The following CSV tables show sentiment marker values in two files, on the left: events, on the right: nonevents file.

Figure 18



These two tables are difficult to compare in anything but snippets. Getting an overall sense of the sentiment is impossible, as is getting any sense of hot spots, that is, a relation of the parts, which can tentatively be defined as prolonged negative or positive signals, or an intensity in the frequency of the sentiment markers, “When we see the dataset as a text list, at the low level we

must read words and compare them to memories of previously read words. It is hard to keep track of just these dozen topics using cognition and memory alone" (Munzner 675).

The following visualizations are the first iteration of the sentiment markers in the events and nonevents file. I am calling this the SentiPulse vis.

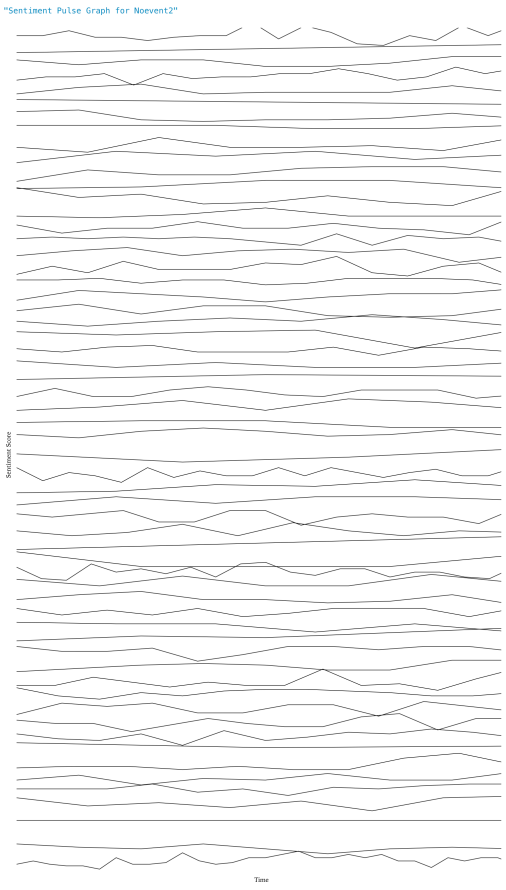


Figure 19. Each line represents a graph of the sentiment markers in each nonevent.

This is an image of each file's line graph of sentiment markers stacked one atop the other in black and white. So the sentiments are represented as lines not areas. The sentiment values of neutral sentiment markers, i.e. sentiment values = 0.0, is filtered out. I coloured each line graph

using an HTML Color Scheme code in order to help visually differentiate one line graph from the next and render each file as an area. Color Scheme or Color Scale are predetermined colour patterns which can be applied to data (https://www.w3schools.com/html/html_colors.asp). The colours are not meant to convey data.

The X-axis represents time, from the beginning to the end of each scene. The Y-axis represents the sentiment scale. The original vis from Observable had ticks on the X-axis with time measure. While the X-axis of my vis represents a passage of time, it is not scaled in a time measure.

Sentiment Pulses in Events file: the two following visualizations are an early iteration of trying different colour patters, padding, overlap, etc. They are an incoherent jumble.

Figure 20 Early iteration of Sentiment graphs in events

Figure 20 Early iteration: events Sentiment graphs.

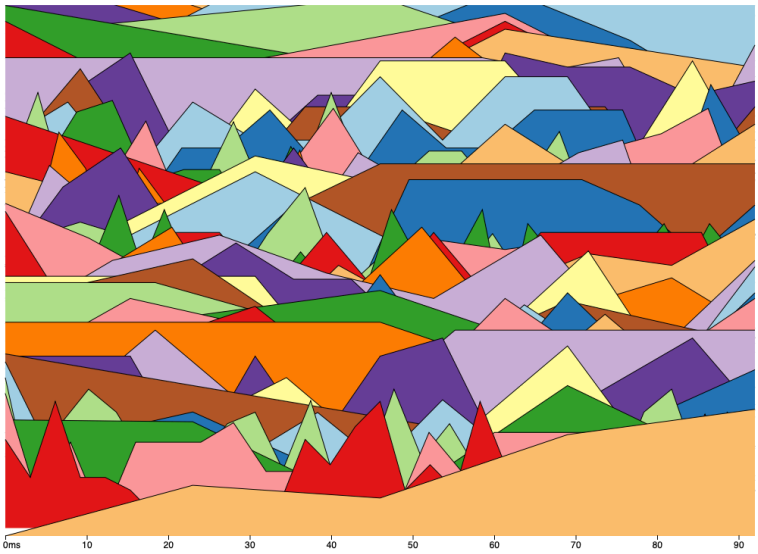
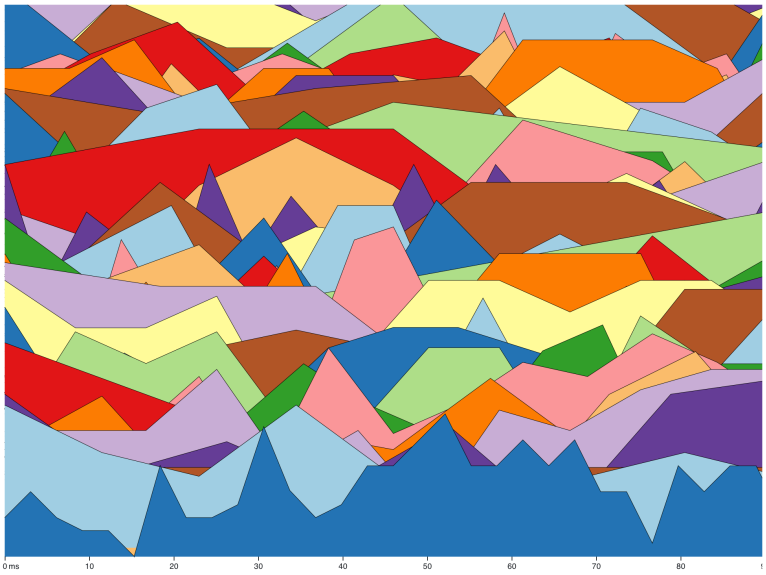


Figure 20B Early iteration of Nonevent sentiment graphs.



The csv table for the above visualization of the sentiment scores in the events file:

Figure 21

0	0	0	-0.25	-0.25	0	-0.25	0	0	0.25	0.25	0	0	0	-0.125	0	0
0	0	0	-0.125	0.5	0	-0.125	0.125	0	0	0	-0.5	-0.375	0	0	0	0
0	0	0	0	0	-0.125	0	-0.125	0	0	0	0	0	0	0	0	0.25
0	0	0.125	0.125	0	0	0	0	0	0	0	-0.125	0	0	-0.5	0	0
0	0.25	0.125	0.125	0	0	0	0.125	0	0	0	0.125	-0.25	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.125	0	-0.25	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.375	0	0	0	0	-0.125	0	0
0	0	0	0	-0.5	0	0	0	0								
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	-0.125	0	-0.125	0	-0.125	0	0	0	0	0	0	0	0
0	-0.125	0	0	0	-0.375	0	0	0	0	0	0	0	0	0	0	-0.375
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.125	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0
0	0	0	0	0.375	0	0	0	0	0	0	0.625	-0.25	0	0	0	0
0	0.625	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.375	0
0	0	-0.25	0	0	0	0	0	0	0	0	0	0	0	0.125	0	0
0.125	-0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	-0.375	0	0	0	0	0	0	-0.75	0	0	0
0	0	0	0.25	0	-0.375	0	-0.375	0	0	0	0	0	0.875	-0.25	0	0
0	0	0	0	0.25	0.25	0	0.25	0	0.25	0.25	0	0	0	0	0	0.25
0	0	0	0	0	0	0	0	0	-0.375	0	0	0	0	0	0	0
-0.625	0	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0	-0.25
0	0	-0.375	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.125	0.5	0	-0.25	0	0	0.5	0	0	0.125	0.125	0.75	0.125	0	0.75	0	0.75
0	0	-0.5	-0.5	0	0	-0.125	0	0	0	0	0	0	0	0	0	0
0	0	-0.25	0	0.875	0	0	-0.625	0	-0.25	0	0	0	0	0	0	-0.75
0	-0.375	0	0	-0.125	0	0	0.5	0	0	0	0.625	0	0	0	0	0
0	0.125	-0.125	0	0	0	0	0	0	0	0	-0.375	0.375	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.375	0	0	0	0.125	0	0	0	0	0	0
-0.125	0	0	0	-0.375	0	0	0.375	0	0	0	0	0	0	0.125	-0.5	0.125
0	0	-0.75	0	0	0	0	0	0	0	0	0	0	0	0.25	0	-0.125
0	0	0	0	0	0	0	0	-0.625	0	0	0.375	0	0	0	0.875	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	-0.75	0	0	0	0	0	0
0	0	0.125	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	-0.125	-0.125	-0.125	0	-0.25	-0.125	0.125	-0.375	0

As you can see, the original output includes the neutral markers, or terms with value = 0.0. I filtered the zero values out. In a sense they are background noise to the sentiment activity. I'm

not saying that they are useless, just that in creating a graph of sentiment pulses they act as a layer of omnipresent, unnecessary ‘fog’, much like stop words are for a text.

4.11 The SentiPulse Visualization

The two following visualizations are the final iteration of the SentiPulse vis.

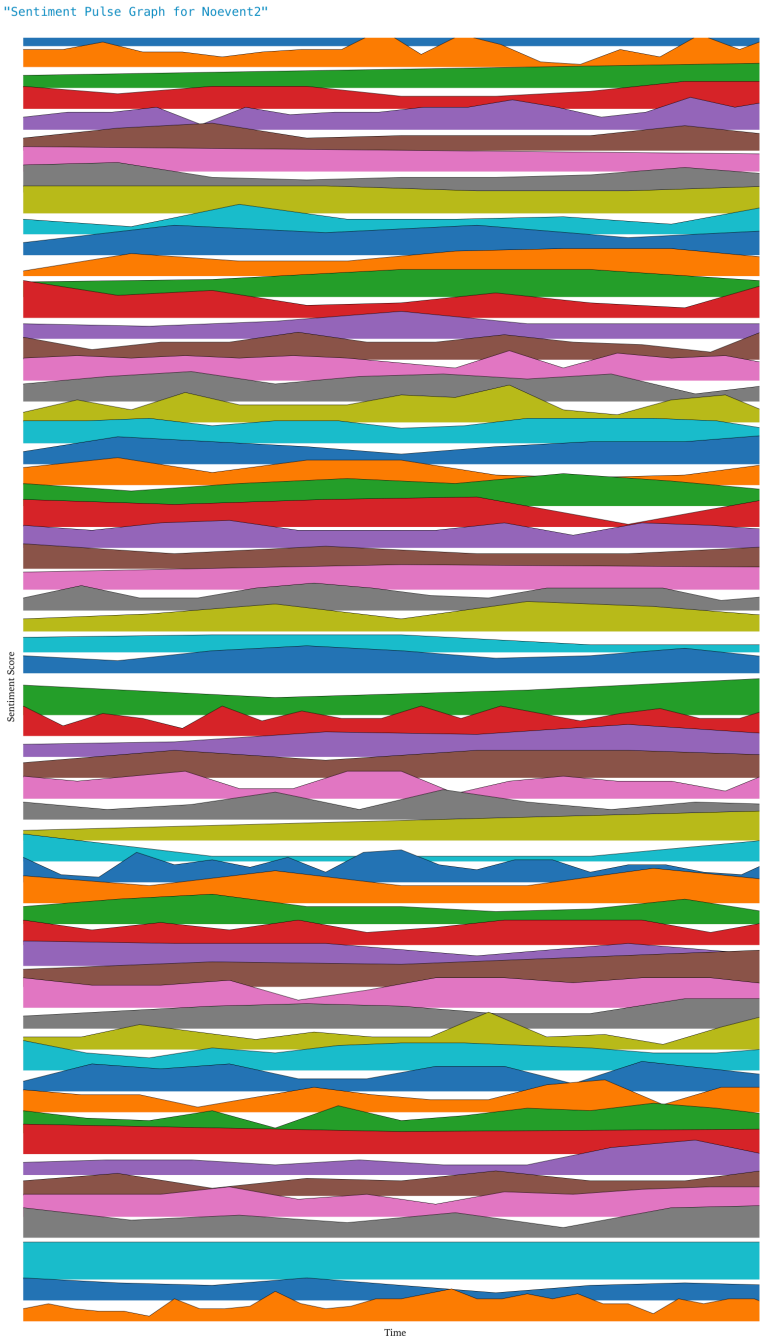


Fig. 22. Nonevents

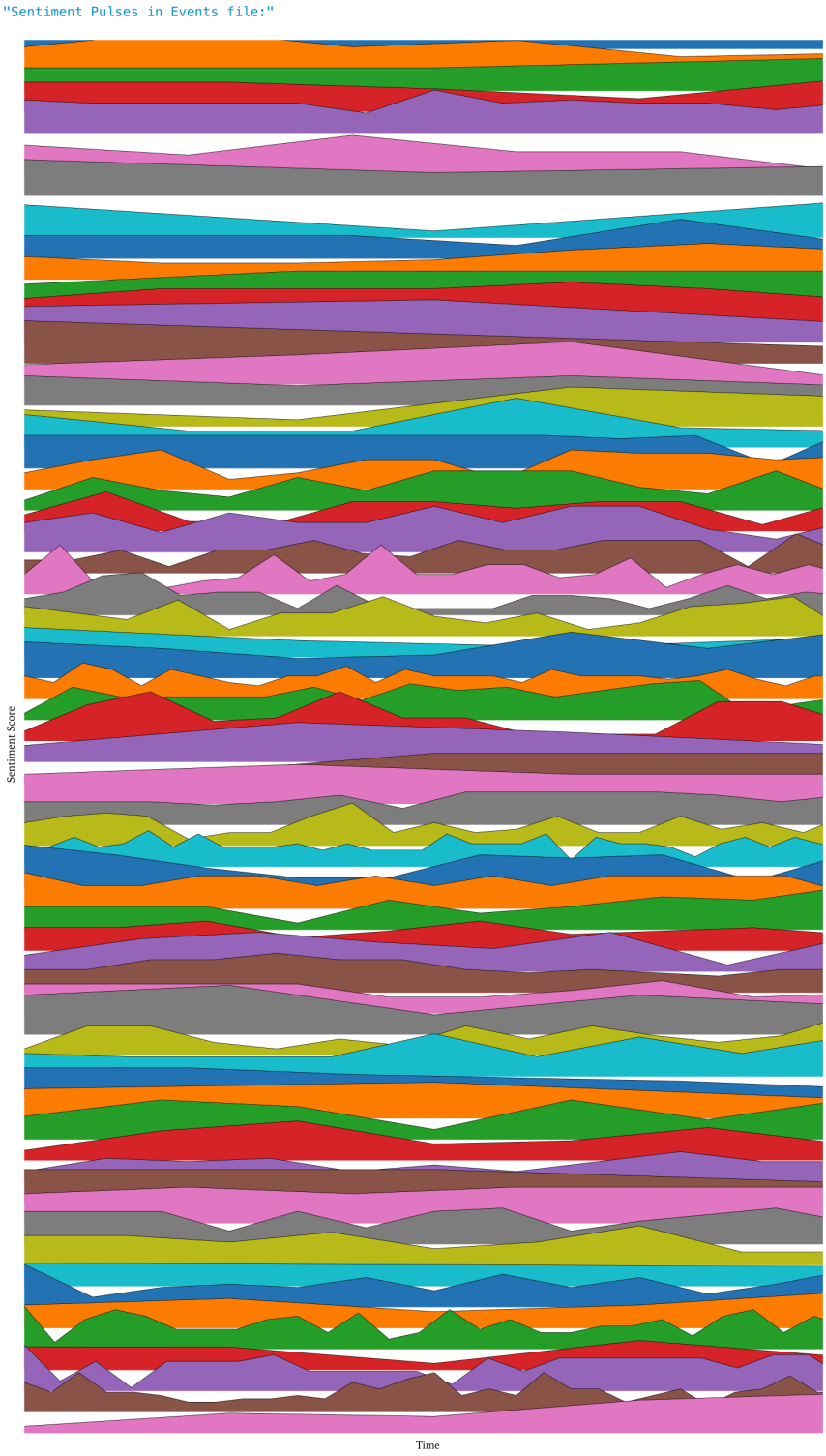


Fig. 23 Events

Above is the final iteration of the SentiPulse Graph for events and nonevents data after adjusting overlap, height, colour, etc to minimize occlusion and enable the view of each file with minimal

interference from adjacent files. Each file is represented as a coloured area. The colours are chosen for ease of distinguishing one file from the next; the colours are not data driven. The peaks and valleys are graphed along the y-axis. The y-axis does not have a zero point; it has many zero points. Each file, each coloured area, starts on its y-axis within a range from -1 to1, with individual scores measured to three decimal points, e.g. the first file in the nonevents folder:

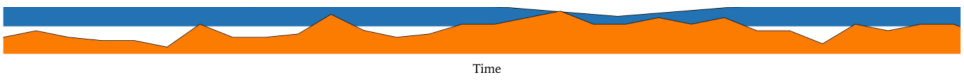
-0.375, -0.125, -0.375, -0.5, -0.5, -0.75, 0.125,- 0.375, -0.375, -0.25, 0.5, -0.112, -0.375, -0.25, 0.125, 0.125, 0.375, 0.625, 0.125, 0.125, 0.375, 0.125, 0.375, -0.125, -0.125, -0.625, 0.125, -0.125, 0.125, 0.125, -0.375

The above data is shown in orange at the base of the Nonevents vis, Fig. 22, and is shown as an isolated graph in Fig.24 below.

The metaphor of landscape, while absent of detail, is at play in the SentiPulse vis. The x-axis is the time from the beginning to the end of each chosen scene. The y-axis scales the sentiment score. Each coloured area represents one file, and the top of that area is a line graph of the sentiment scores, so the score of the first line, just above, looks like this on the orange graph:

Fig. 24

Y-axis marks the Sentiment score



X-axis marks the duration of the event or nonevent.

The level of refinement of the data has much to do with how it can be represented as landscape. The original Pulsar graph, Fig.17, measured radio waves at four second intervals, which gives the Pulsar graph line a more “natural” look. It seems nature abhors a straight line.

In this vis the files are given a sentiment shape as a metaphor of a distant landscape, albeit a crude one.

In this particular comparison between events and nonevents, the graph does not impart easily assimilated information. However, it is capable of imparting information about the sentiment of a unified text such as a novel, screenplay, short story, etc. In such a unified body of work the SentiPulse vis can give an impression of the emotional landscape of said text. The units of analysis would be decided in the modelling. For example, a novel can be analyzed by chapter or by setting , a screenplay, scene by scene, or a specific frame, like Piper's narrative frames, delineated by scene changes could form the unit of analysis. The Sentipulse vis can then render a relative understanding of the sentiment rhythms, the areas of most active vs most inactive sentiment activity, etc.

There are other ways to deploy the Sentipulse vis. For example, every time a character is present in a scene, measure the pulse of the sentiment. So, each character has a graph for every scene they are in, and that graph starts and ends for the duration of the scene. The graphs are then stacked one atop the other, as they are in Fig. 23. This would give an indication of the change in sentiment activity associated with each character at any given point in the narrative.

Take for example the top third of the graph for the events file, Fig. 22. There is a calmness to the the lines of those graphs, not much sentiment change, which is indicated by the slope of the line from one marker to the next. If one created a Sentipulse graph for a screenplay, from beginning to end, such that each scene was a file, then a visual representation of the sentiment rhythms of the writing can be visualized with the SentiPulse. In that context a

Sentipulse graph would give useful insights into the size of the sentiment changes, their frequency and rhythm, and any emotional hot spots.

4.12 Towards a 3D Representation

In moving toward a 3D sentiment landscape, the question becomes how to further develop the landscape metaphor to more fully, and more engagingly show the viewer the emotional "landscape" of the text, that is, how to move into a three dimensional representation, with the third dimension also driven by the data, "In contrast, non-spatial data can be visually encoded using spatial position, but that encoding is chosen by the designer rather than given implicitly in the semantics of the dataset itself. This choice is the one of the most central and difficult problems of visualization design" (Munzner 679). The development of a meaningful, intuitively digestible, third dimension representing abstract data is the next step toward making a sentiment landscape.

In going forward with the SentiPulse visualization I would like to research a sentiment resource that is more refined in its ability to tag sentiment markers. For example, an analysis which takes into account the collocation of the sentiment marker to derive a more precise sentiment score is one possible refinement, or training an existing resource such as WordNet on a large corpus to get a better understanding of how to understand sentiment at the level of the phrase, rather than a single term. With a more refined resource, I would like to test the graph on a unified text, i.e. a text of one screenplay, or a chapter in a novel, or a set of tweets by an active tweeter, etc. in order to read the 'rhythm' of the stacked graphs.

Another avenue of development is to take the final visualization and give depth to the graphs of each file. A skin, essentially digital wallpaper, is then applied to simulate, say, a lunar surface or a desert terrain. This would make a 2D area of each file into an image that more resembles a mountain range. This exploration would be worth pursuing if the sentiment points are refined enough and numerous enough to create a more detailed substructure than the one I use at present. For example, Fig. 25, the 3D printed version of PRS B1919+2, is an example of giving a line graph a depth:

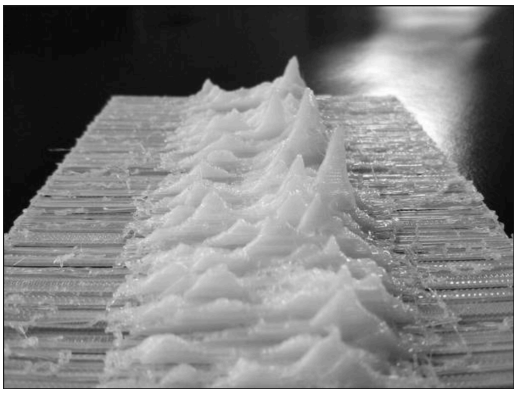


Figure 25

Credit: https://boingboing.net/2013/05/24/3d-printable-model-of-the-cove.html

"Unknown Pleasures"

Figure 26 Credit: Myriam Douglas

The image in Fig.26 was used by British punk band, Joy Division, as their first album's cover in 1979, "Unknown Pleasures". The image gained a cult following, fans tattooing it on their backs etc., as outlined in this video about Joy Division's first album jacket design: https://www.youtube.com/watch?time_continue=12&v=reEQye0EOAw

Bibliography

Aristotle. *Poetics.* Dover Publications, 1997.

Armoza, Jonathan. “Topic Words in Context – Dickinson, the Fascicle, and the Topic Model”.

Bal, Mieke. *Narratology: introduction to the theory of narrative*. University of Toronto Press, 2017.

Baroni, Raphaël: "Tellability", Paragraph 13. In: Hühn, Peter et al. (eds.): the living handbook of narratology. Hamburg: Hamburg University Press.
URL = hup.sub.uni-hamburg.de/lhn/index.php ?title=Tellability &oldid=2035
Accessed 2 July 2018.

Bordwell, David. *The Way Hollywood Tells It: story and style in modern movies*. University of California Press, Berkley, 2006.

"catharsis, n.”, *OED Online*, Oxford University Press, March 2019, www.oed.com/view/Entry/28926. Accessed 26 March 2019.

Fischer, Frank, et al., “Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts.”, *Digital Humanities 2017 (Montreal, 8-11 August 2017). Book of Abstracts.* Montreal, McGill University, 2017.

Guardian, The. “Hollywood film output likely to fall by a third”. (https://www.theguardian.com/world/2009/oct/18/hollywood-films-numbers-fall. Accessed 10 June 2018.

Godfather, The. Screenplay: http://www.screenwrite.in/Screenplays/Godfather.pdf. Acessed 28 July 2018.

Brittanca, Encyclopedia. “peripetia”. https://www.britannica.com/art/peripeteia. Accessed 28 March 2019.

Ellsworth, P. C. “William James and emotion: Is a century of fame worth a century of misunderstanding?” *Psychological Review, 101*(2), 222-229. 1994.

Hoyt, Eric and Kevin Ponto, Carrie Roy. "Visualizing and Analyzing the Hollywood Screenplay" University of Wisconsin-Madison;, University of Wisconsin-Madison. http://digitalhumanities.org/dhq/vol/8/4/000190/000190.htmlAcessed 29 July 2018.

Huhn, Peter. “Event and Eventfulness.” *Handbook of Narratology,* edited by Peter Huhn et al.

De Gruyter, Inc., 2009.
(https://books.google.ca/books?hl=en&lr=&id=v9fmBQAAQBAJ&oi=fnd&pg=PA159&dq=eventfulness+theory&ots=RTWyJED1OF&sig=zS_T17zjT7qmuK_AJtDJ9vHHNEM#v=onepage&q=eventfulness%20theory&f=false)

IMSDb. Internet Movie Script Data Base. https://www.imsdb.com. Accessed 14 December 2018.

Labov, William. *Language in The Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, 1972.

Langellier, Kristin M. "Personal narratives: Perspectives on theory and research." Text and Performance Quarterly 9.4 (1989): 243-276. p. 245.

LHN. Hühn, Peter et al. (eds): the living handbook of narratology. Hamburg: Hamburg University Press. URL = hup.sub.uni-hamburg.de/lhn. Accessed 15 Dec 2018.

"look, n." OED Online, Oxford University Press, March 2019, www.oed.com/view/Entry/110129. Accessed 30 May 2019.

Méliès, Georges. *Le Voyage Dans la Lune* (1902). (https://www.youtube.com/watch?v=BNLZntSdyKE. Accessed 20 June 2018

Metz, Christian. *Film Language: A Semiotics of Cinema. Oxford University Press*, 1974.

Nexus Mediaworks. (https://www.nexusmediaworks.com/video-marketing-storytelling-approach/freytagpyramid/). Accessed 13 June 2018.

Numbers, The. “Domestic Movie Theatrical Market Summary 1995 to 2018”. https://www.the-numbers.com/market/. Accessed 10 June 2018.

"peripeteia, n." *OED Online*, Oxford University Press, March 2019, www.oed.com/view/Entry/141011. Accessed 26 March 2019.

Piper, Andrew and Portelance, Eva. “Understanding Narrative: Computational approaches to detecting narrative frames”. Poster. McGill University, 2017. https://txtlab.org/2017/01/congratulations-to-eva-portelance-aria-intern-for-2016/. Accessed 5 May 2019

Piper, Andrew. *Enumerations*. Chicago University Press. 2019.

Pratt, Mary Louise. *Toward a Speech Act Theory of Literary Discourse.* Indiana University Press, 1948.

Prince, Gerald. “Surveying Narratology.” What Is Narratology? : Questions and Answers Regarding the Status of a Theory, edited by Tom Kindt, and Hans-Harald Müller, De Gruyter, Inc., 2008. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=325682.

—— *Dictionary of Narratology*. University of Nebraska Press. 2003.

Shmid, Wolf. “Narrativity and Eventfulness.” *What Is Narratology? : Questions and Answers Regarding the Status of a Theory*, edited by Tom Kindt, and Hans-Harald Müller, De Gruyter, Inc., 2008. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=325682.

Stam, Robert. *Film Theory: an introduction*. Malden, MA, 2000.

Bradley, A.J. *et al*., "Visualization and the Digital Humanities:," in *IEEE Computer Graphics and Applications*, vol. 38, no. 6, pp. 26-38, 1 Nov.-Dec. 2018. doi: 10.1109/MCG.2018.2878900 URL: http://ieeexplore.ieee.org.proxy3.library.mcgill.ca/stamp/stamp.jsp?tp=&arnumber=8617736&isnumber=8617719

Cambria, E., "Affective Computing and Sentiment Analysis," in *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, Mar.-Apr. 2016. http://ieeexplore.ieee.org.proxy3.library.mcgill.ca/stamp/stamp.jsp?tp=&arnumber=7435182&isnumber=7435173 doi: 10.1109/MIS.2016.31

Christiansen, Jen, “Visualizing Science: Illustration and Beyond”. Scientific American, 25 October 2018. https://blogs.scientificamerican.com/sa-visual/visualizing-science-illustration-and-beyond/

Color Scheme. https://www.w3schools.com/colors/colors_monochromatic.asp. Accessed March 2019.

Gatti, L., M. Guerini and M. Turchi, "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis," in *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 409-421, 1 Oct.-Dec. 2016. http://ieeexplore.ieee.org.proxy3.library.mcgill.ca/stamp/stamp.jsp?tp=&arnumber=7239537&isnumber=7755879. doi: 10.1109/TAFFC.2015.2476456

AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1550790540&Signature=M0X1H

NkY0fvHOJOdk%2FscG8WpnrA%3D&response-content-disposition=inline%3B%20filename%3DSimilarity_Measures_for_Text_Document_Cl.pdf

Graham, Shawn, Ian Milligan, Scott Weingart. "[section name]" The Historian's Microscope - working title. Under contract with Imperial College Press. Open draft Version, Autumn 2013, http://themacroscope.org

Huang, Anna. "Similarity Measures for Text Document Clustering". Department of Computer Science. University of Waikato, Hamilton, N.Z.

Jain, Anil K. "Data Clustering: 50 years beyond K-means". Pattern Recognition Letters. Volume 31, Issue 8. 1 June 2010. Pp 651-666. https://doi.org/10.1016/j.patrec.2009.09.011.

Munzner, T. (2015). *Visualization Analysis and Design*. New York: A K Peters/CRC Press, https://doi-org.proxy3.library.mcgill.ca/10.1201/b17511

Rhyne T., M. Tory, T. Munzner, M. Ward, C. Johnson and D. H. Laidlaw, "Information and scientific visualization: separate but equal or happy together at last," *IEEE Visualization, 2003. VIS 2003.*, Seattle, WA, USA, 2003, pp. 611-614.
doi: 10.1109/VISUAL.2003.1250428
URL: http://ieeexplore.ieee.org.proxy3.library.mcgill.ca/stamp/stamp.jsp?tp=&arnumber=1250428&isnumber=27978

Sinclair, Stéfan, Geoffrey Rockwell and the Voyant Tools Team. 2012. Voyant Tools (web application). http://docs.voyant-tools.org/tools/wordcountfountain/

Sinclair, Stefan, Geoffrey Rockwell, "The Art of Literary Text Mining". <https://github.com/sgsinclair/alta/blob/master/ipynb/ArtOfLiteraryTextAnalysis.ipynb>. Accessed last July 4, 2019.

Tory M., D. Sprague, F. Wu, W. Y. So and T. Munzner, *"Spatialization Design: Comparing Points and Landscapes," in IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1262-1269, Nov.-Dec. 2007. doi: 10.1109/TVCG.2007.70596*

URL: http://ieeexplore.ieee.org.proxy3.library.mcgill.ca/stamp/stamp.jsp?tp=&arnumber=4376149&isnumber=4376125