Assignment 12.1: Text Analysis Using Python and Tableau

Written by Daanish Ahmed

DATA 620 9080

Semester Spring 2017

Professor Majed Al-Ghandour

UMUC

May 4, 2017

Executive Summary

This paper describes a study that I conducted involving textual analysis of film reviews over time. I selected five reviews written for films that won the Academy Awards Best Picture between 1935 and 2015. My goal in this paper is to determine how moviegoers' opinions change over time with regards to what makes a good film. To solve this problem, I designed a Python program that performs a word count on a text file, returning the 30 most common words and how often those words appeared. After running this program on each of the five reviews, I produced several graphs in Tableau to see if the data provided the solutions I was looking for.

The results of my analysis offered several insights regarding the change in word usage over time. I found that several words saw increased usage over the years, and that certain words only appeared in reviews during the past few decades. But to see if the words implied a change in viewer preferences, I examined the context of the words within the actual reviews. For example, I checked to see if the usage of the word "sound" indicates the importance of sound within the film itself. I found that some instances of these words supported my assumptions, but that others had different contexts. Furthermore, I noticed that many of the most frequently-used words in each review were unique to that review and did not appear in others. This could be attributed to some words being plot-specific or to the reviewer's word preferences.

At the end of my analysis, I found some possible correlations between certain word counts and moviegoer opinions towards film over time. The results did not confirm my theories with absolute certainty, but provided the framework for additional research. One of my recommendations for further analysis is to use a larger sample size of reviews for each year. Another recommendation would be to use my visualizations while also analyzing the context of the original reviews, as opposed to looking at the data by itself.

Introduction

Text analysis is a powerful method for extracting meaning from texts, also known as unstructured data. A lot of valuable information is hidden in text that may be difficult to uncover without the use of an effective text parsing software. But by using Python, I designed a program that could read through a document and extract the most frequently-used words and how often those words appear. Using this program, I seek to examine its capabilities for textual analysis by looking at five film reviews written over the past 80 years. These reviews were all published by The New York Times, and each review was written for a film that won the Academy Awards (Oscars) Best Picture during the years 1935, 1955, 1975, 1995, and 2015. These films include *It Happened One Night* (1934), *On the Waterfront* (1954), *The Godfather, Part II* (1974), *Forrest Gump* (1994), and *Birdman* (2014) ("Academy Awards Database," n.d.). By looking at reviews for Oscar-winning films, I hope to find out if there is a change over time in what audiences perceive to be a good film. Thus, my goal is to evaluate whether the word choice in these reviews reflects changing viewer tastes, or if it merely represents each reviewer's personal preferences.

Methods

Before looking at the results of my analysis, I would like to briefly describe how my Python program functions. The code I designed is meant to read through a user-inputted text file and break the lines of text into a series of words. It looks at each word individually and performs a count of how many times that word appears in the file. At the end of the process, it will return a list of the 30 most common words in the file and their word counts. Additionally, the program uses the NLTK (Natural Language Tool Kit) add-on for Python, which includes a built-in list of

stop words and a lemmatization feature ("Natural Language Toolkit," n.d.). This add-on was used to eliminate stop words (such as "and" or "where") from the list and to ensure that words such as "write" and "writing" are listed as the same word. When the program is done generating the list, it will save the output into a .CSV file that can be opened with Microsoft Excel. After running the program on all five film reviews, I combined the .CSV files into a single file using Excel and imported the data into Tableau. From there, I generated the graphs that are used in my analysis.

Initial Analysis

In the first part of my analysis, I will look at three interesting words that were prominent in Dargis' review of *Birdman* (2014) and how frequently those words appeared in other reviews over the past 80 years. The three words that I chose to include in my graph are the words "story," "new," and "love." The reason that I chose to analyze these words is that they offer the most context as to what viewers are looking for within a film's plot. Audiences tend to enjoy films with strong storylines that bring something new to the table, rather than seeing movies with the same plots repeatedly. Likewise, viewers appreciate romantic storylines because they tend to make the characters more relatable and allow viewers to connect emotionally with the story (Pendleton, n.d.). These three words frequently appear not only in this review, but throughout the reviews selected for this study. This indicates that the selected words are not specific to *Birdman* alone, but refer to common themes in cinema.

My first visualization is a time series graph that shows the trajectory of these three words from 1934 to 2014 (see Figure 1). Based on this graph, there are several interesting observations that can be made. Firstly, the word "story" is mentioned with increasing frequency over the years.

This seems to suggest the importance of films needing high-quality stories to achieve critical acclaim. Such an observation makes sense, because it is difficult for a film to impress critics based on its action scenes or celebrity star power alone. Furthermore, the word "new" is mentioned quite consistently throughout the decades. Based on the data alone, it seems that audiences do indeed expect films to be fresh and original, especially if that film is to qualify for Best Picture. But most interestingly, the word "love" in these reviews appears to be in decline over the past several decades. This might seem surprising, but it may not necessarily mean that romantic movies are declining. More likely, these results may indicate that moviegoers are able to appreciate films regardless of whether they include romantic elements or storylines. Another possible explanation is the choice of words used by the critic. Just because the word "love" is absent from the review does not mean that the film lacks an on-screen romance. For instance, Hall (1934) does not mention the word "love" in his review of *It Happened One Night*, but he does use its synonym "romance" to describe the relationship between the lead roles.

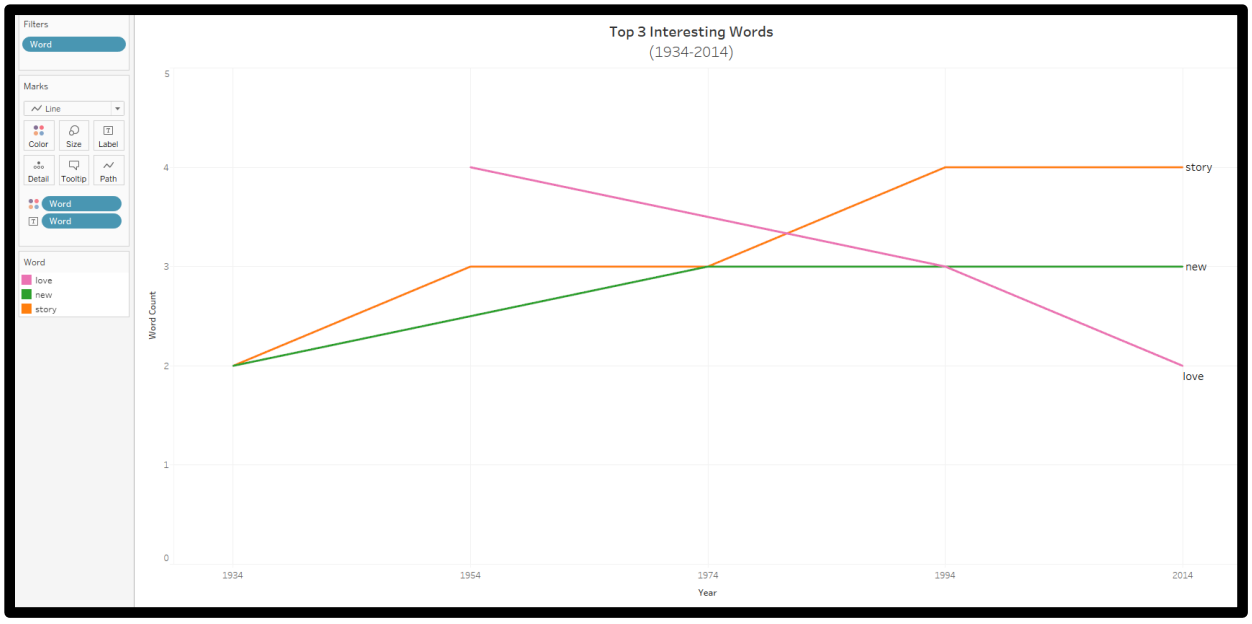

Figure 1. Top 3 Interesting Critic Word Choices (1934-2014).

Furthermore, when considering critics' word choices, it is important to look at the context of the reviews before coming to conclusions on the data. One example is the usage of the word "new." As I mentioned earlier, the prevalence of this word seems to suggest that viewers expect to see new things within each film. However, when reading the reviews themselves, not every instance of the word "new" refers to the same context. For instance, Dargis (2014) mentions the word "new" three times in his review of *Birdman*. One instance of this word does indeed refer to the freshness of the film itself, in which he compares the film to director Iñárritu's previous works by describing it as a "significantly better new one (Dargis, 2014)." However, Dargis also uses the word "new" when describing a character working for The New York Times. Therefore, great care must be taken when performing a text analysis on a document. The data still offers valuable insights, but data alone cannot provide a complete understanding. One must analyze the text itself to discern meanings which a computer program is not yet able to understand.

Further Analysis

The next step in my analysis is to look at other frequently-used words that appear within these reviews. Thus, my second graph is a stacked bar graph showing the top five words that appear for each year (see Figure 2). The purpose of this visualization is to show how word choices change throughout the years, and based on the image we can see that the top five words are quite different for each film. However, it is once again important to understand the context of these words within each review before making conclusions. For example, the word "first" is the most frequently mentioned word in the review for *The Godfather, Part II*. The reason is that Canby (1974) regularly compares the film to its predecessor in his review. Since none of the other four films are sequels to existing movies, it is reasonable that none of the other reviews mention the

word “first.” Additionally, Hall’s review for *It Happened One Night* (1934) contains words such as “palace,” “father,” and “symphony.” These are plot-specific words and will only appear within reviews for films that have these plot elements within the film. One way to improve the relevancy of these results would be to use a larger sample size of film reviews for each year. By having texts on multiple movies per year (spanning different genres and sources), the total word count would be a more general reflection of film characteristics and would provide a better image of how these qualities changed over time.

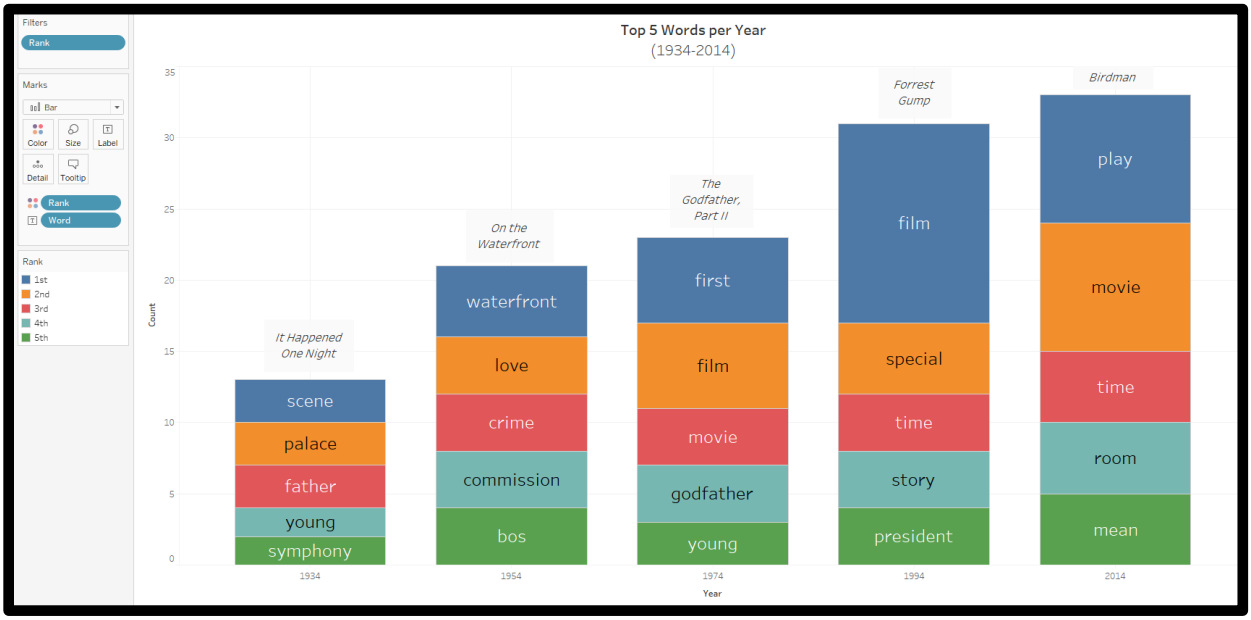

Figure 2. Top 5 Most Common Reviewer Words per Year (1934-2014).

To determine which words are the most useful for my analysis, I created a dashboard of five word clouds—one for each film review. This visualization shows words that are unique to each review in gray, while words that are found in two or more reviews are brightly colored (see Figure 3). The image once again highlights the issue that most words are specific to the plot of their associated films. For example, A. W.’s review of *On the Waterfront* (1954) frequently mentions the words “waterfront,” “longshoreman,” and “laborman.” These words describe the

plot of this film, and thus it is not surprising that they are not used in the reviews for any other film in this study. I am not saying that the words in gray are not interesting or useful to the analysis. These words still offer insights towards the film's themes and the critic's personal preferences, and if a larger sample of texts were analyzed then it is likely that many of these words would appear more frequently throughout the years. However, the words in color are the most useful words for my time series analysis, and I will proceed to analyze them further.

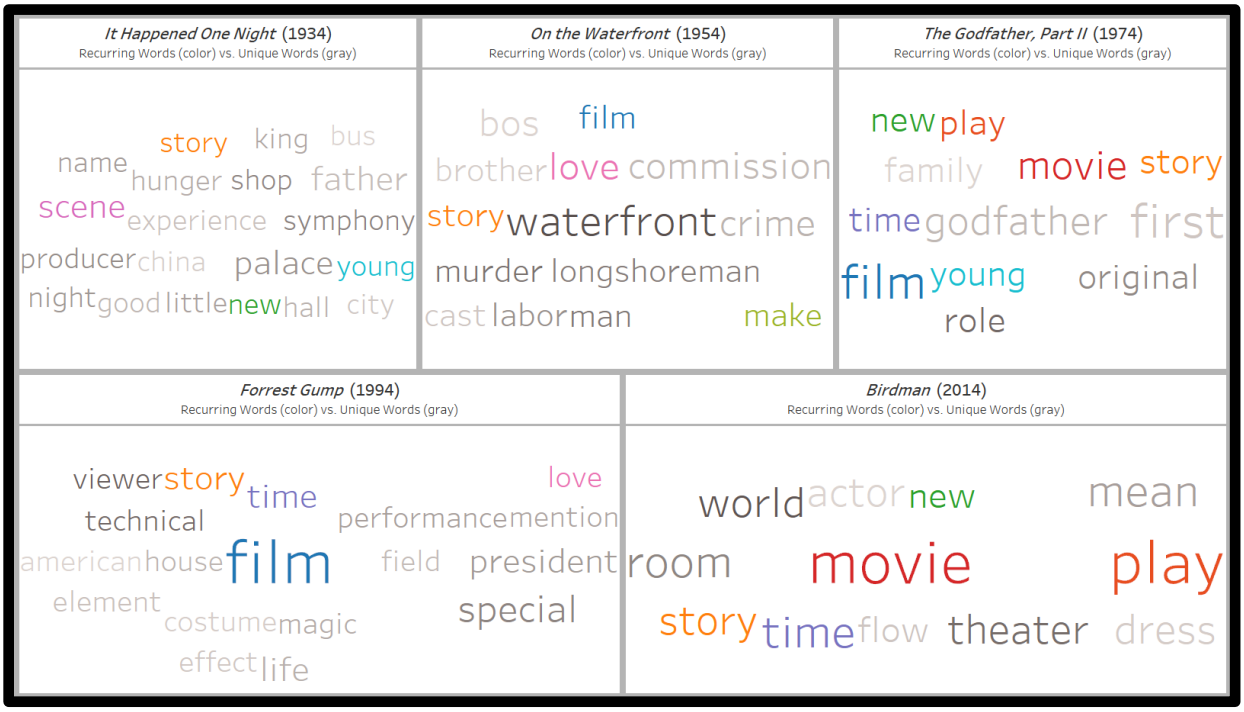

Figure 3. Common Recurring Words (Color) vs. Unique Words (Gray), 1934-2014.

I created one final visualization that shows the top interesting words that recur throughout the years for each film review. The graph is a dashboard of five bubble charts showing the most common recurring words per year, with the size of the bubble indicating the word's frequency within that review (see Figure 4). This dashboard also filters out words such as "film" and "movie," since these words are not very useful for analytical purposes. I believe that this graph is the cleanest representation of my results so far and that it builds upon my findings from Figure 1.

Based on this image, there are several observations that can be made. One of the most interesting findings is that the word "sound" has appeared more frequently in reviews over the past 20 years. This word was not used in reviews prior to 1994, and it may indicate an increasing appreciation for the usage of sound within a film. In addition to this, the word "time" has regularly appeared in reviews from 1974 to the present. The review for *The Godfather, Part II* states that the storyline "moves continually back and forth in time between two distinct narratives (Canby, 1974)." Likewise, *Birdman*'s review describes the camerawork by stating that "everything flows together…even time and space (Dargis, 2014)." All of this may suggest that the importance of time as a cinematic element has increased over the past few decades, at least amongst films that are recognized by the Oscars.

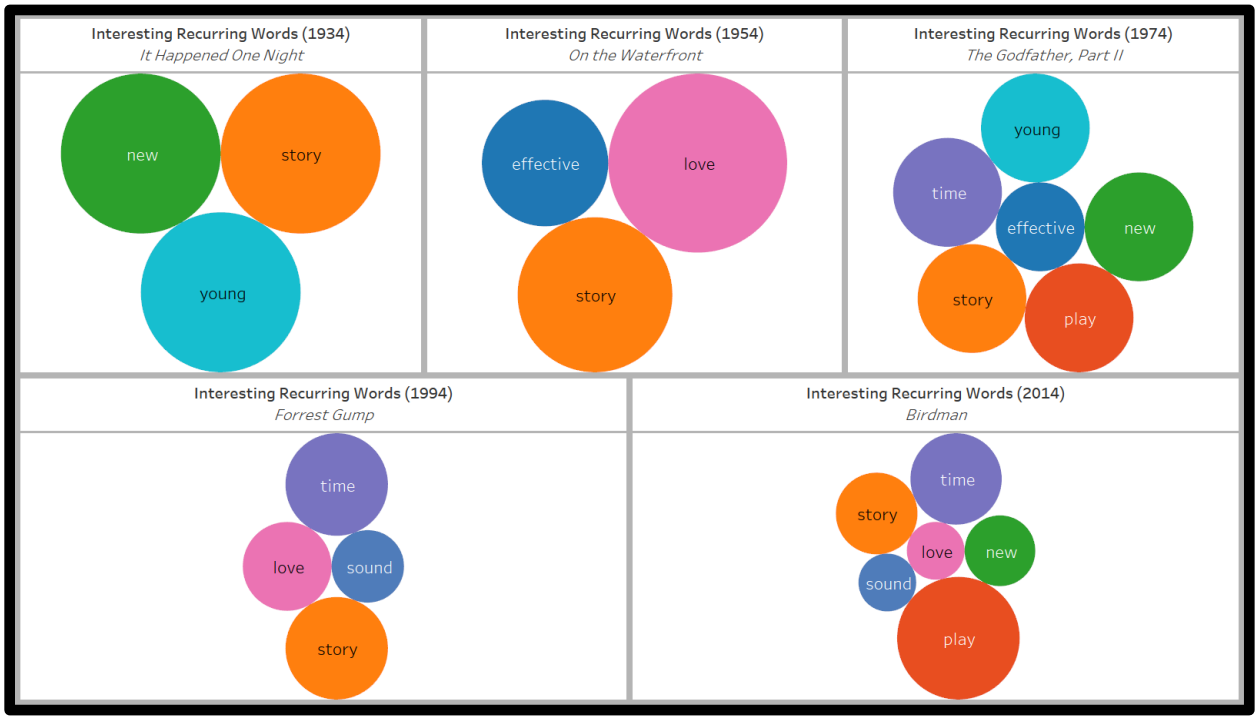

Figure 4. Interesting Recurring Words Describing Each Film, 1934-2014.

But once again, context must be considered when analyzing the frequency of these words. For instance, Maslin (1994) uses the word "sound" at least twice in her review of *Forrest Gump*.

In this example, her usage of the word "sound" does indeed refer to the film's use of sound. She praises the actor playing President Kennedy by stating that his "voice sounds authentic" and expresses delight in the realism of the film's sound effects (Maslin, 1994). On the other hand, Dargis (2014) also uses the word "sound" twice in his review of *Birdman*. However, one instance of this word appears in the sentence "it sounds like an alarming idea (Dargis, 2014)," which has nothing to do with the film's technical usage of sound. Furthermore, Maslin (1994) mentions the word "time" at least four times in her *Forrest Gump* review, but two of those instances are referencing the film's length (running time). As I stated earlier, it is extremely important to examine the context of the original sources when interpreting the results of a text analysis. If I had not analyzed the texts themselves, it would be easy to misinterpret the data.

Conclusion

Over the course of my analysis, I made several observations regarding film critics' word choices over time. The word "story" gradually increased in usage throughout the decades, while the word "new" was used consistently throughout this period. The word "love," on the other hand, saw a decrease in usage over time (which I attributed to either reviewer word choice or a greater appreciation of storytelling which may or may not include romantic elements). In addition, I found that the word "sound" saw increased usage over the past two decades, while the word "time" appeared more commonly from 1974 to the present. These findings may suggest that moviegoers (or critics, at least) have increased their expectations over time. Viewers may have higher standards for storytelling, with a desire to see new cinematic elements and techniques introduced over the years. They might be more impressed with stories that incorporate clever usages of time, and they may also have increased appreciation in the use of sound effects in a movie.

But by analyzing the context of the original reviews, I found that some instances of the selected words support my theories regarding moviegoer tastes, while other instances of these words differed in context. This is not surprising, because film reviews are usually literary in nature and are based in the reviewers' opinions and personal word choices. According to Nualart-Vilaplana, Pérez-Montoro, and Whitelaw (2014), it can be difficult to visualize the data in literature because these texts generally allow more freedom for the author regarding word choice and sentence structure. Because of this, my recommendation to executives reading this paper is to focus not only on my visualizations but also on the context of the reviews themselves. According to a similar study by Jiffy Lube and OdinText (n.d.), the best results do not come from looking at the data alone (in this case the NPS, or Net Promoter Score), but rather from using the data in conjunction with the words from the customer comments. I believe that by combining the word count visualizations with the context of the original reviews, it will be much easier to determine whether the data does indeed reflect the change in viewer preferences towards film.

Another recommendation I would make would be to analyze a larger sample of reviews spanning multiple films, genres, and authors per year. In my analysis, I found that most words were unique to a single review and that only a handful of words appeared in more than one review. In fact, the top five words used in each review largely consisted of words that are specific to that film's plot. But if a larger number of texts were included, it is likely that some of these words would appear throughout multiple reviews. This would shape the data into a more general reflection of audiences' tastes towards film, making it easier to map the change in viewer preferences over time. Altogether, I believe that my analysis can serve as a good starting point towards understanding the changing minds of filmgoers. If we conduct further research in this field, we may gain insight towards how audiences may perceive movies in the future.

References

Canby, V. (1974, December 13). 'Godfather, Part II' Is Hard to Define:The Cast. The New York Times. Retrieved May 3, 2017, from http://www.nytimes.com

Dargis, M. (2014, October 16). Former Screen Star, Molting on Broadway. The New York Times. Retrieved May 3, 2017, from https://www.nytimes.com/2014/10/17/movies/birdman-stars-michael-keaton-and-emma-stone.html?_r=0

Elder Research Inc. (2013). *Improving customer retention and profitability for a regional provider of wireless services.* Retrieved April 30, 2017, from https://cdn2.hubspot.net/hubfs/2176909/Resources/Elder-Research-Case-Study-Customer-Retention-nTelos.pdf

Hall, M. (1934, February 23). Claudette Colbert and Clark Gable in a Merry Jaunt From Miami to New York. The New York Times. Retrieved May 3, 2017, from http://www.nytimes.com

Jiffy Lube Uses OdinText Software to Increase Revenue. (n.d.). Retrieved April 30, 2017, from http://odintext.com/wp-content/uploads/2015/10/odinText-Shell.pdf

Maslin, J. (1994, July 6). FILM REVIEW; Tom Hanks as an Interloper in History. The New York Times. Retrieved May 3, 2017, from http://www.nytimes.com

Natural Language Toolkit. (n.d.). Retrieved May 1, 2017, from http://www.nltk.org/index.html

Nualart-Vilaplana, J., Pérez-Montoro, M., & Whitelaw, M. (2014). How we draw texts: A review of approaches to text visualization and exploration. El Profesional de la Informacion, 23(3), 221-235. Retrieved April 30, 2017.

Pendleton, M. (n.d.). Why People Keep Watching Romance Films. Retrieved May 05, 2017, from http://film.ezinemark.com/why-people-keep-watching-romance-films-7d30ca4db6aa.html

The Academy Awards Database. (n.d.). Retrieved May 03, 2017, from http://awardsdatabase.oscars.org/

W., A. (1954, July 29). Astor Offers 'On the Waterfront'; Brando Stars in Film Directed by Kazan. The New York Times. Retrieved May 3, 2017, from http://www.nytimes.com