Text Mining Movie Scripts

CSCI 420 – Principles of data mining

Michael Root

2017

***Abstract***

For my data mining project, I decided to look towards the movies and explore some facets of them. In particular, I took a keen interest in the scripts of certain movies to see if I could glean any information about the backbone of movies. With any movie, one of the first things that are thought up is a script, something to define what people will say and how they will interact around the set. To be honest, I didn’t really care much about set directions in the scripts, I really only cared about the lines spoken by people in them. With that in mind, I started looking for other research done on this subject, and I found nothing. “Nothing” is somewhat subjective, as there are papers and info about text mining, but nothing about movie scripts. With that in mind, I collected 15 relatively well known movie scripts from free sites, and began looking into text mining the scripts. There was no real goal in mind before hand, as I had nothing to compare what I wanted to do to something that was already done. Resulting in my goal just to kind of see what I could make do with the small sample I had.

***Tools***

For my project, there were not very many python tools that were available. The first one was a python library called “textmining-1.0”, and from my small exposure to it, it was not very well suited for what I wanted. The only main function I could see from it was it offered the ability to make a CSV file of words that occurred in sentences. The issue I had with that was that I am dealing with movie scripts that were about ten-thousand lines of text a piece. So naturally I attempted to make a CSV file of word occurrences from all 15 scripts combined together, resulting in a three-gigabyte large CSV file which would not open. That was my only real encounter with “textmining-1.0”.

Another python library I stumbled upon was very useful, and called “nltk” which stand for Natural Language Tool Kit. This library contains a plethora of tools and interfaces for things such as classification, tokenization, stemming, parsing, and semantic reasoning. I couldn’t even begin to use all the tools available with this library, so I only sought to mold what they had available to fit my own purposes. The main part I took from their library was their ability to tokenize words in sentences and also categorize words into all parts-of-speech in the human language. The parts-of-speech library includes thirty-six different parts of speech that words parsed in can be classified to. This was definitely the best feature I was able to use successfully.

***Issues***

The main problem that occurred a lot in my testing and experimenting was the sheer size of the text files I was dealing with. Because I am using movie scripts that I copied into text files, the files would range around eight to ten thousand lines per script of multiple words per line. Even when I removed all blank lines of text, a text file containing all fifteen scripts was about seventy-thousand lines long, with each line having at least a few words on it. So, when it came to processing all the words and breaking them up into data structures, it took a very long time. The longest time it took at one point I was trying to plot about 600,000 data points on a graph, and I let the program run for over half an hour without it finishing. That aside, as stated previously when I attempted using the “textmining-1.0” python library, getting a word count CSV file of the seventy-thousand line text file took about five minutes to process and resulted in a three-gig CSV file that didn’t load after ten minutes of opening. It was at that point I decided to try and find a different library I could use, which caused me to end up with “nltk”, which is in fact way better.

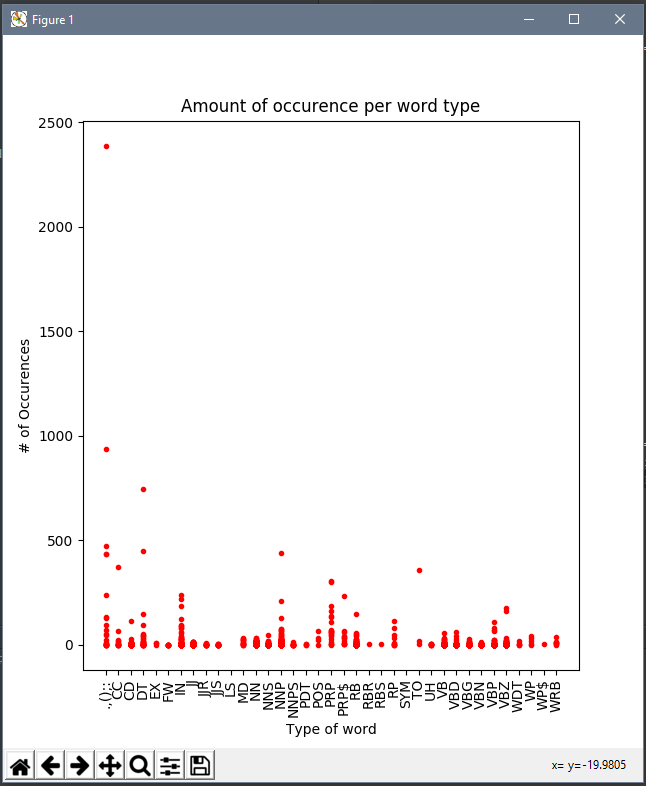
***Information***

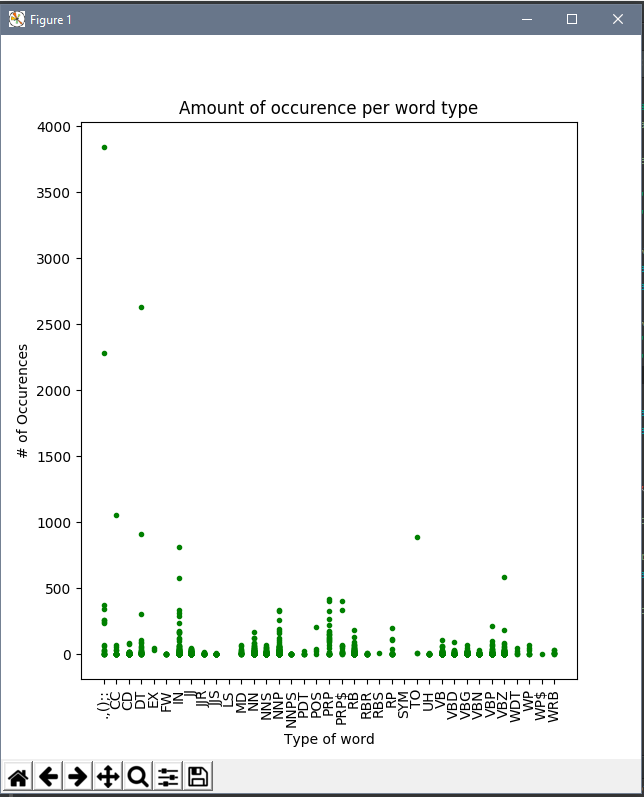
The main reason for this section is to provide the data I collected on the text, along with some abbreviation clarification on parts of my results. The following table was taken from a UPenn website that listed all the tags and descriptions of the nltk classification. After the long table, there will be a series of graphs that were the result of classifying the words in each script by the tag they were given in the nltk part-of-speech tagging. The tags correspond to the given descriptions as shown below.

|  |  |  |
| --- | --- | --- |
| 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 |

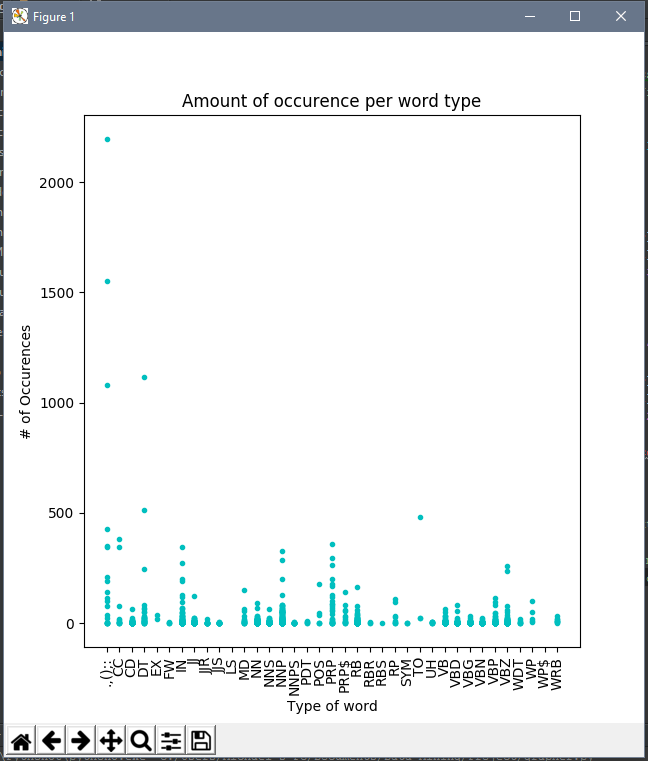
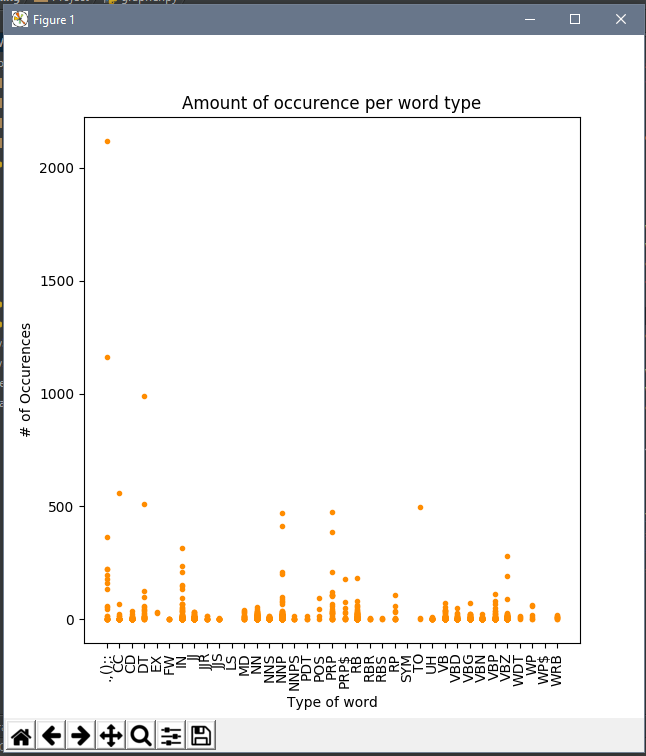
The following list contains the fifteen movie scripts I used in my analysis:

American Sniper, Dawn of the Dead (1977), Gone in 60 Seconds, Groundhog Day, Interstellar, Jaws, Legally Blonde, Life of Pi, Semi-Pro, Shawshank Redemption, She’s Out of My League, Shrek, The Last Samurai, The Revenant, Titanic.

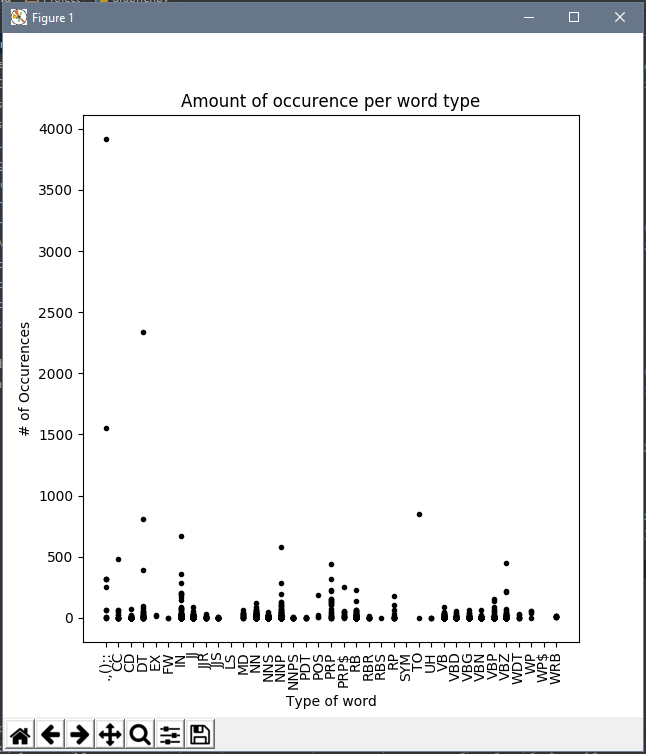
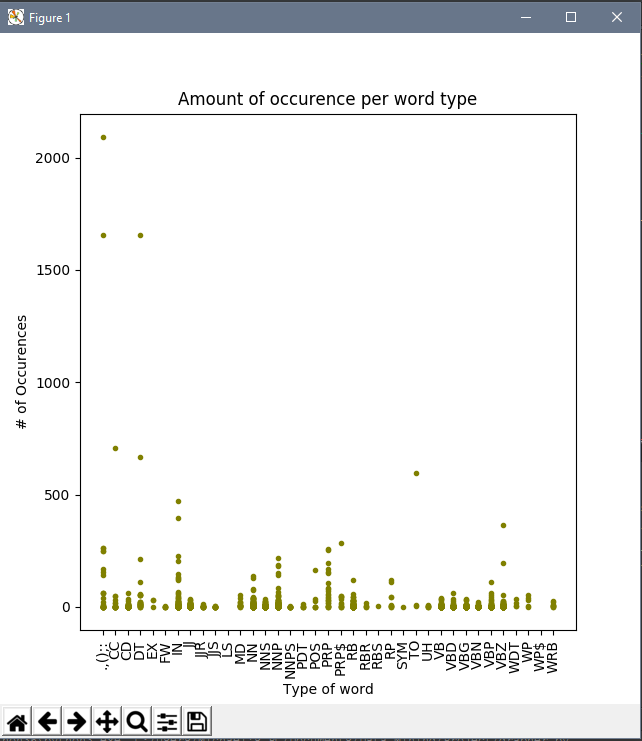
***Classification***

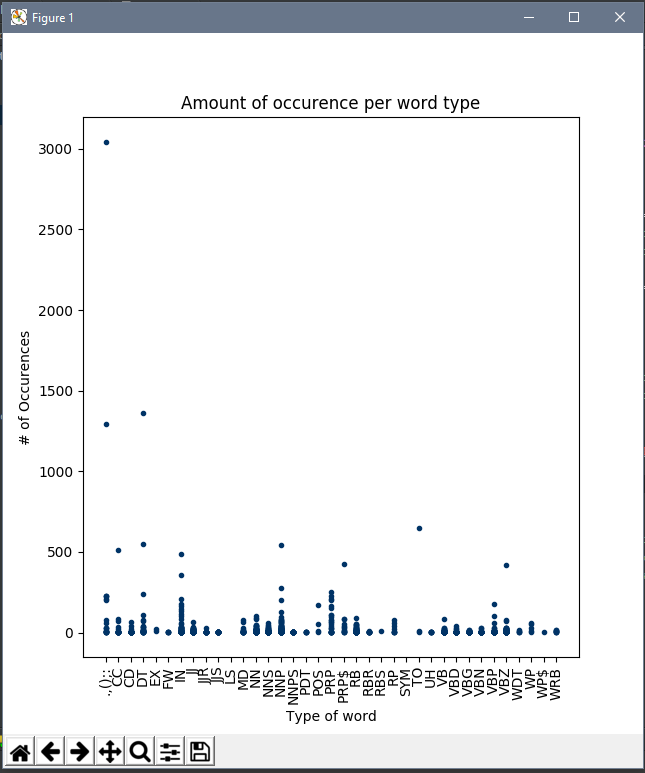
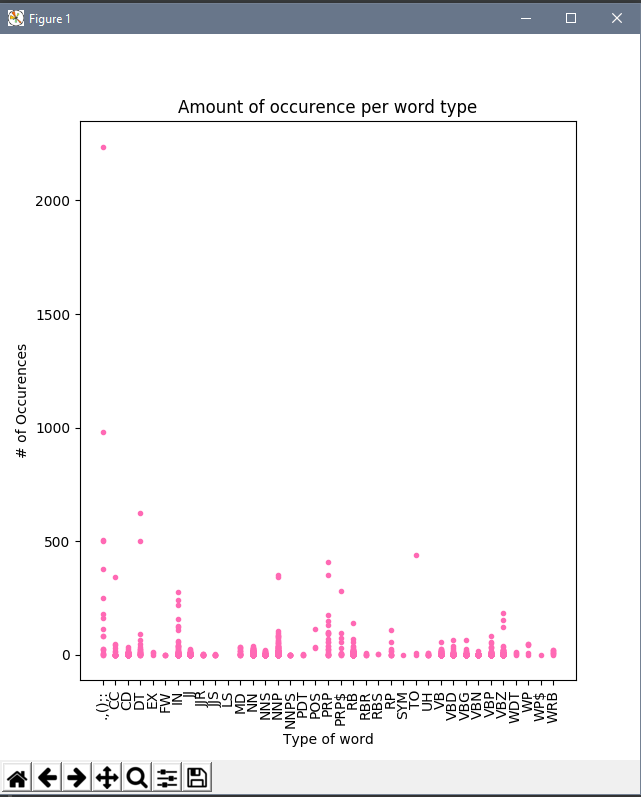


American Sniper Dawn of the Dead (1977)

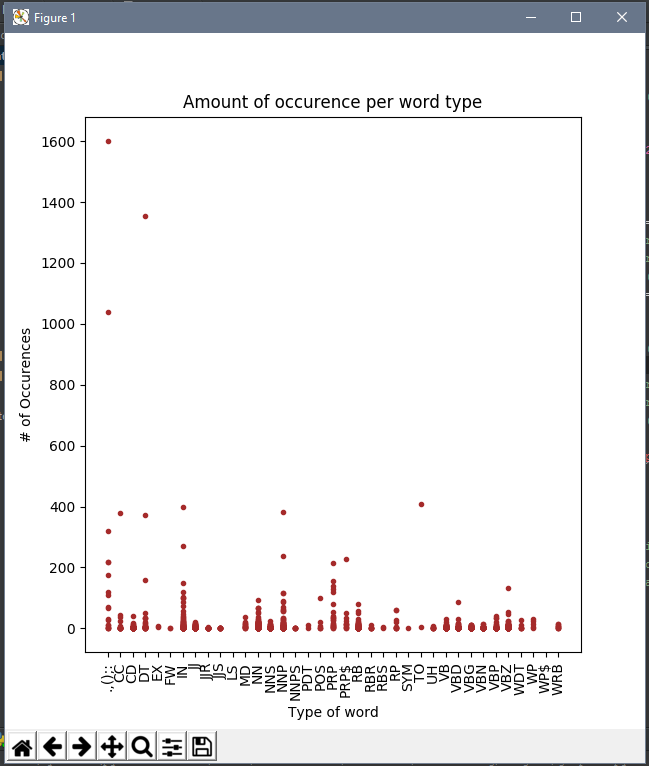
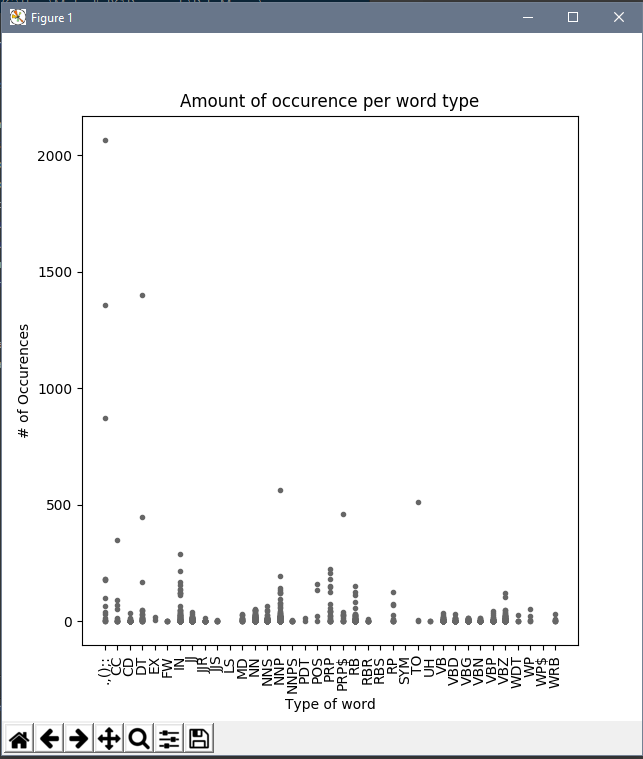


Gone in 60 Seconds Groundhog Day

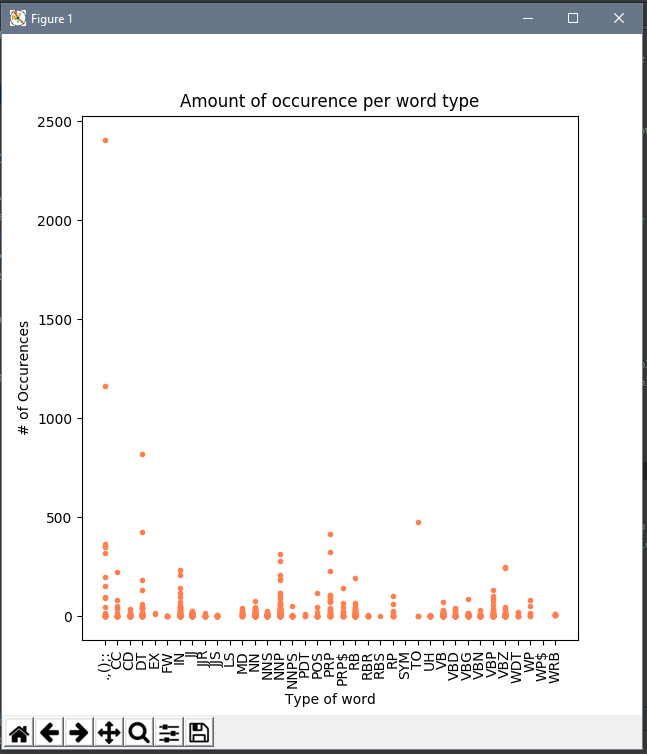
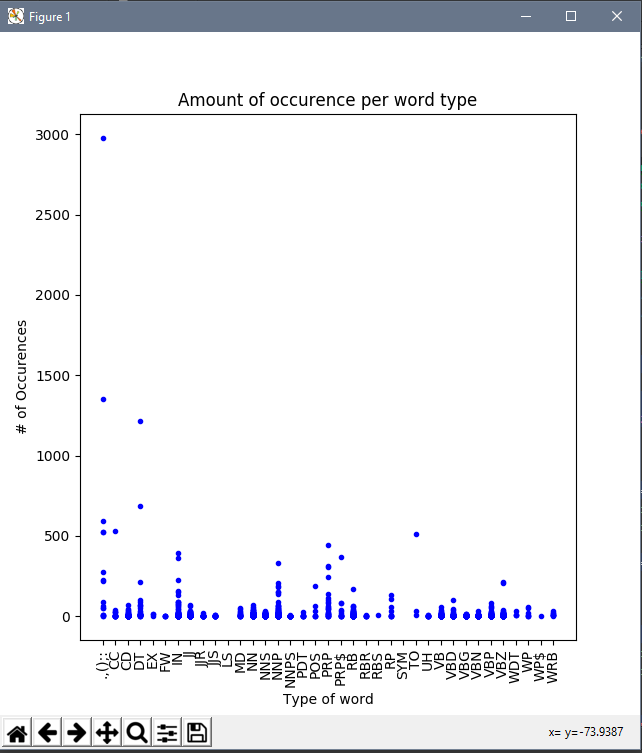


Interstellar Jaws

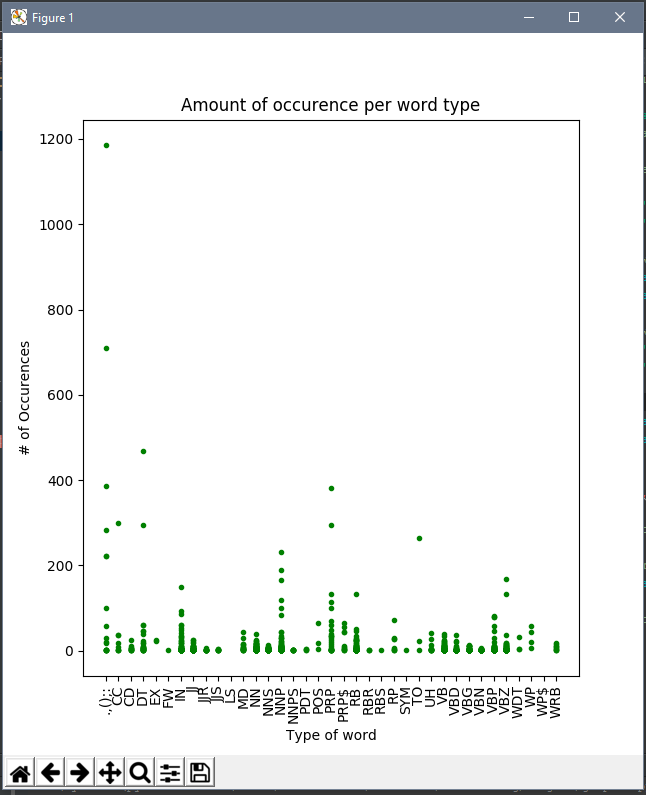
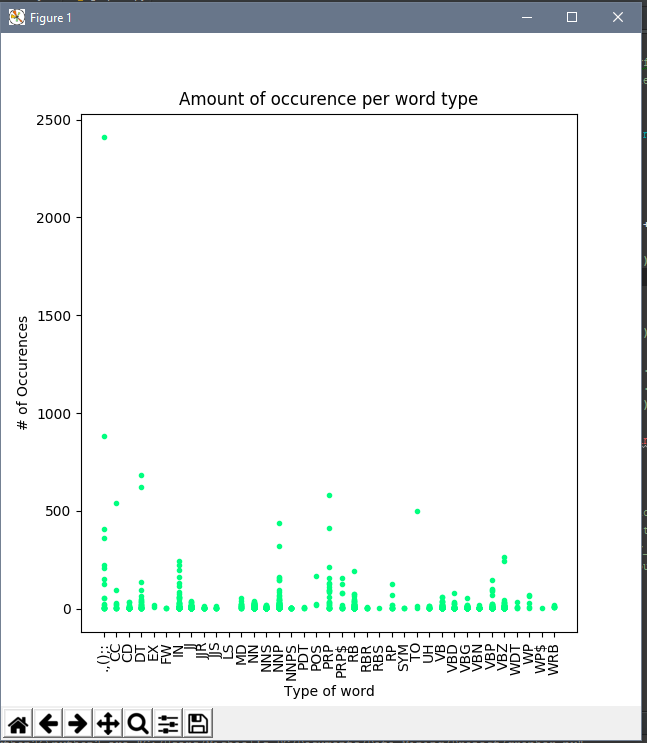
The Last Samurai Legally Blonde



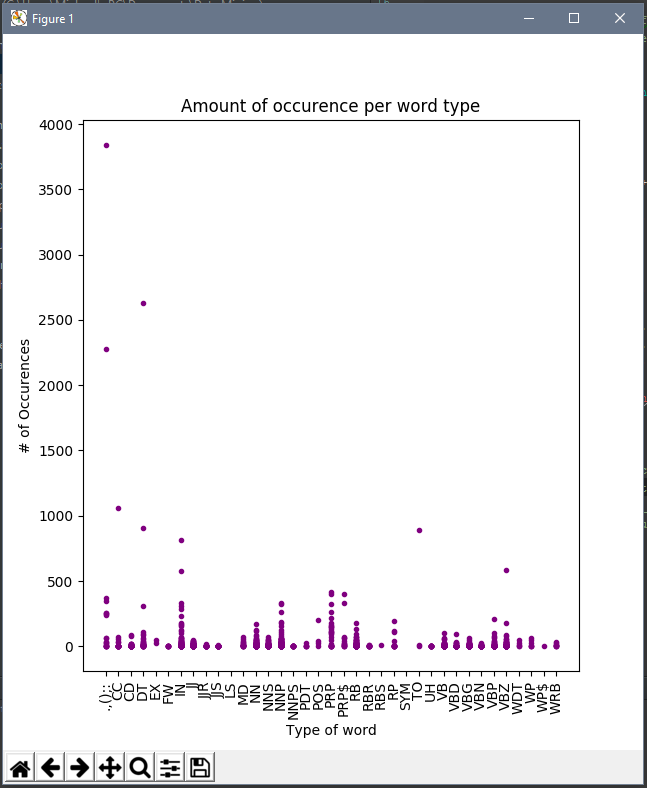
Life of Pi The Revenant



Semi-Pro Shawshank Redemption



Shrek She’s Out of My League



Titanic

Some other data I gathered from the scripts was the amount of times certain “key phrases” occurred throughout the script. Due to the limited nature of my search and lack of really common phrases, though, I tried to keep it to some core phrases that one would think would occur a lot.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Movie \ Phrase | “I love you” | “Oh my god” | “You better run” | “Oh no” |
| American Sniper | 3 | 2 | 0 | 0 |
| Dawn of the Dead (1977) | 1 | 0 | 0 | 2 |
| Gone in 60 Seconds | 0 | 0 | 0 | 0 |
| Groundhog Day | 2 | 0 | 0 | 0 |
| Interstellar | 1 | 0 | 0 | 0 |
| Jaws | 0 | 0 | 0 | 0 |
| Legally Blonde | 2 | 2 | 0 | 0 |
| Life of Pi | 1 | 0 | 0 | 1 |
| Semi-Pro | 0 | 1 | 0 | 0 |
| Shawshank Redemption | 0 | 3 | 0 | 0 |
| She’s Out of My League | 3 | 2 | 0 | 0 |
| Shrek | 2 | 0 | 0 | 0 |
| The Last Samurai | 0 | 0 | 0 | 0 |
| The Revenant | 0 | 0 | 0 | 0 |
| Titanic | 1 | 0 | 0 | 2 |

Finally, the last data I collected was word count based off of the length of words, including punctuation characters. The length of the text file is too long to convert into a table or to list off in this paper so I will include it but talk about the results in later sections.

***Conclusions***

To start, let’s take a look at the graphs showing the amount of time a word, word being any part of text that was separated from another part AKA tokenized. Without fail the most common thing to show up is punctuation marks, they always are sky high in comparison to all other categories. After that a common second highest place goes to determiners, or words such as ‘a’, ‘the’, or ‘every’. Once again this is not surprising as English contains a lot of use of the same punctuation and determiners. To get the final blatant one out of the way, the word ‘to’ is commonly pretty high along with coordinating conjunctions such as ‘and’, ‘but’, and ‘or’. Now as we look at difference in types of movies we can see a few more differences. For singular proper nouns, there is a far greater occurrence of them in movies like Groundhog Day and Life of Pi as opposed to Semi-Pro and Dawn of the Dead. This shows that movies that have more dialogue between two character, or more personal one on one dialogue shows a greater use of singular proper nouns. Which makes sense if people are engaged in small personal conversations, they are more likely to talk in a direct fashion. There is a stark difference in singular proper noun use of all movies featuring a larger group of main characters against ones featuring smaller, more personally engaged main characters.

Another difference that can be spotted is shown in the use of prepositions. Words like ‘on’, ‘in’, and ‘besides’ are some examples that are used. The movies Life of Pi, Jaws, and Titanic all feature a large use of prepositions as opposed to the other movies. This could directly show that movies that take place out on open water always have more prepositions than one on land. This is probably coincidental, but the relationship of this to the curly fries talk in class is somewhat amusing to me.

Another point to be looked at is in regards to personal pronoun occurrence. Movies such as She’s Out of My League, Shrek, and Groundhog Day all have a lot more personal pronouns than The Revenant, The Last Samurai, and Interstellar. This can directly show that movies that have lots of engagement between main characters’ results in a lot more personal pronouns, shocking. However, this can also show that movies involving male and female protagonists results in more personal pronouns, perhaps having to do with the protagonists talking about each other to side characters.

As far as the data involving the amount of times words of certain length appear, there is a somewhat interesting result. Disregarding punctuation marks, the largest subset of words is almost always words that are three characters long. This is almost always followed by words that are four characters long, with two characters right behind. Keep in mind, I tracked words by begin: punctuation, one character, two characters, three characters, four characters, five characters, six characters, or seven over more characters long. The lowest scoring category was always words that are one character long, which seemed a bit interesting since determiners were always ranked high on the occurrence graph. But an explanation may be that the word ‘a’ is not as common a determiner as say ‘the’ which falls into the three-character category. When you really think about some common sentences though, you realize how often we use words that are between three and five characters long.

The final main piece of data that I collected was about key phrases that people say a lot. Since these are common phrases one hears a lot, or at least references to being said in movies a lot, one would assume they are found a lot in scripts. Now I’m not sure if it was the result of the scripts that I looked at, but the results of this data were really surprising to me. As shown in the table displaying the occurrences of four pretty common phrases, at least as movie clichés, the results seem off. I would have thought for certain that in Dawn of the Dead people said “oh no” or “oh my god” way more than a collective total of 2 times, both being the phrase “oh no”. The same goes for the Titanic. I could have sworn Jack and Rose said “I love you” to each other but no, that phrase in any form only happened one time throughout the script. I guess it just goes to show that at least I think phrases are a lot more common than they actually are, or at least I like to think they were said in movies.

***Final Thoughts***

As a whole, this experiment was very interesting, if not somewhat difficult. Because no one has really done text mining on movies, there was no real place I could look towards. However, that also made it kind of interesting, because I got to decide the ways to go about it. The sheer size of text I had to process was a bit difficult, especially when I had to run my code multiple times to get results in a way I wanted to view them. I would like to expand this though to see how deep I can go down the rabbit hole of text mining move scripts. Certainly, one of the most interesting bits I finished with was involving prepositions and how apparently they occur a lot more in movies that take place at seas. I’d actually like to pursue that more to see if it really is a thing. Overall, I’m happy with the results I had, and I look forward to being able to expand farther upon them.

***Works Cited***

Aggarwal, Charu C., and ChengXiang Chai. *Mining Text Data*. Place of Publication Not Identified: Springer, 2014. Print.

"Alphabetical List of Part-of-speech Tags Used in the Penn Treebank Project." *Penn Treebank P.O.S. Tags*. N.p., n.d. Web. 30 Apr. 2017.

"Natural Language Toolkit¶." *Natural Language Toolkit — NLTK 3.0 Documentation*. NLTK Project, n.d. Web. 30 Apr. 2017.

Peccei, Christian. "Python Textmining Package." *Python Textmining Package*. N.p., n.d. Web. 30 Apr. 2017.

Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. "Movie Review Mining and Summarization." *ACM Digital Library*. ACM, 06 Nov. 2006. Web. 30 Apr. 2017.

Shubhada Prashant Nagarkar, Rajendra Kumbhar, (2015) "Text mining: An analysis of research published under the subject category ‘Information Science Library Science’ in Web of Science Database during 1999-2013", Library Review, Vol. 64 Issue: 3, pp.248-262, doi: 10.1108/ LR-08-2014-0091