DS 5001: Exploratory Text Analytics
Final Project: Analyzing The Great Gatsby
3rd Year Undergraduate - Spring 2021

Elit Jasmine Dogu (ejd5mm@virginia.edu)
Data Science 5001
13 May 2021

Page Count with Images and References: 14
Page Count of Text: 4

Click Here for GitHub Repository

"Two hulking patent cabinets which held his massed suits and dressing gowns, and ties, and his shirts, piled like bricks in stacks a dozen high..."

- *The Great Gatsby: Gatsby exemplifying a materialistic society focused on wealth*

Analysis of The Great Gatsby

The Great Gatsby is a popular novel written by F. Scott Fitzgerald in 1925. The novel is said to be a “dispassionate account reflecting the decadence and corruption that engulfed America in the 1920s, before the Great Depression” (Bengani) and has characters that “paint a complex portrait of an amoral American society, drunk on its own prosperity” (Prahl). Having analyzed the novel, specifically the characters of the novel and what they represent, in my AP English Language and Composition class in high school, I was intrigued to use data science as a tool to explore this realm again.

Main Question of Interest:

Although I looked at a lot more in my notebooks, my main question of interest for this essay was whether these tools/algorithms could accurately relay a character's actions, connections, and importance.

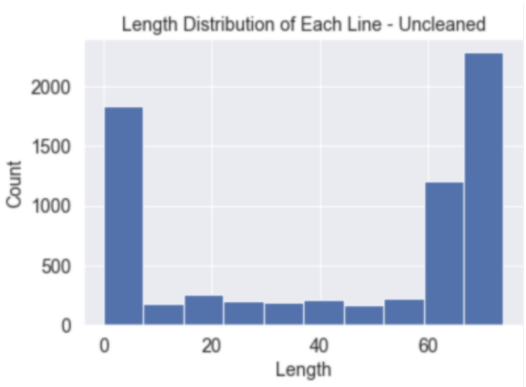
About the Novel

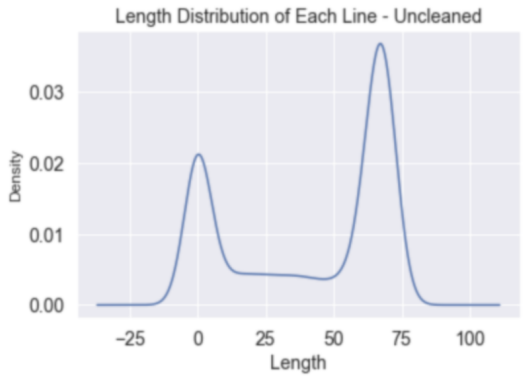
I downloaded a txt file of The Great Gatsby on Project Gutenberg. Using Python, I was able to find out general information about the book. For example, the front and end matter were before and after pages 54 and 6,427, respectively. The novel is broken down into nine chapters starting at the line numbers seen below.

line_num	line_str
57	I
774	II
1321	III
2051	IV
2756	V
3329	VI
3866	VII
5206	VIII
5757	IX

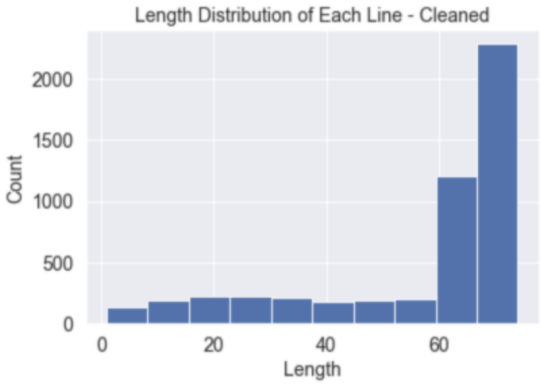
To continue, the average word in the novel was about 3.66 letters and the average sentence was about 55.91 letters long. The length distribution of each line varied depending on whether the structure of the novel was maintained (uncleaned) or not (cleaned). Text in the western tradition is broken into paragraphs; within, it's broken down into content and empty lines. The content is where the narrative is. The empty lines, on the other hand, perform structural work; they help to format the novel in a meaningful and strategic manner. The difference between the uncleaned and cleaned sentence structure can be seen below. With the uncleaned dataset where the empty lines are kept, we see a bimodal distribution as the minimum length of the lines is the lower of the maximum peaks. This disappears with the cleaned dataset, where the highest frequency lies at around the 66 mark instead, leading to a unimodal distribution.

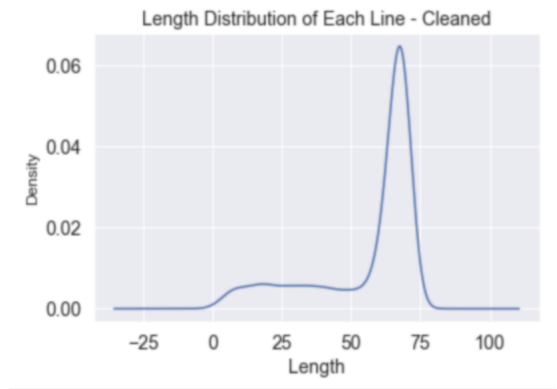
Uncleaned: Spaces Are Maintained





Cleaned: Spaces Are Not Maintained





Word Clouds - Most Frequent Nouns and Verbs

In an effort to see the general plot of the novel, I created a word cloud to look at the most frequently used terms. As seen below in the martini-glass shaped word cloud, the words “Gatsby”, “Tom”, and “Daisy” are amongst the largest, meaning the most frequent. Because the novel revolves around these characters, this is logical. The novel is about Gatsby, the protagonist, who has a sacred connection to Daisy; Daisy and Tom are married. It’s interesting that the words “Nick”, “East”, or “West” did not appear as frequent due to Nick showing some qualities of a protagonist and East Egg and West Egg being two critical locations that emphasize one of the main themes in the novel. Similarly, the concept of the “Green Light,” which represents the American Dream also was not apparent in the word cloud.

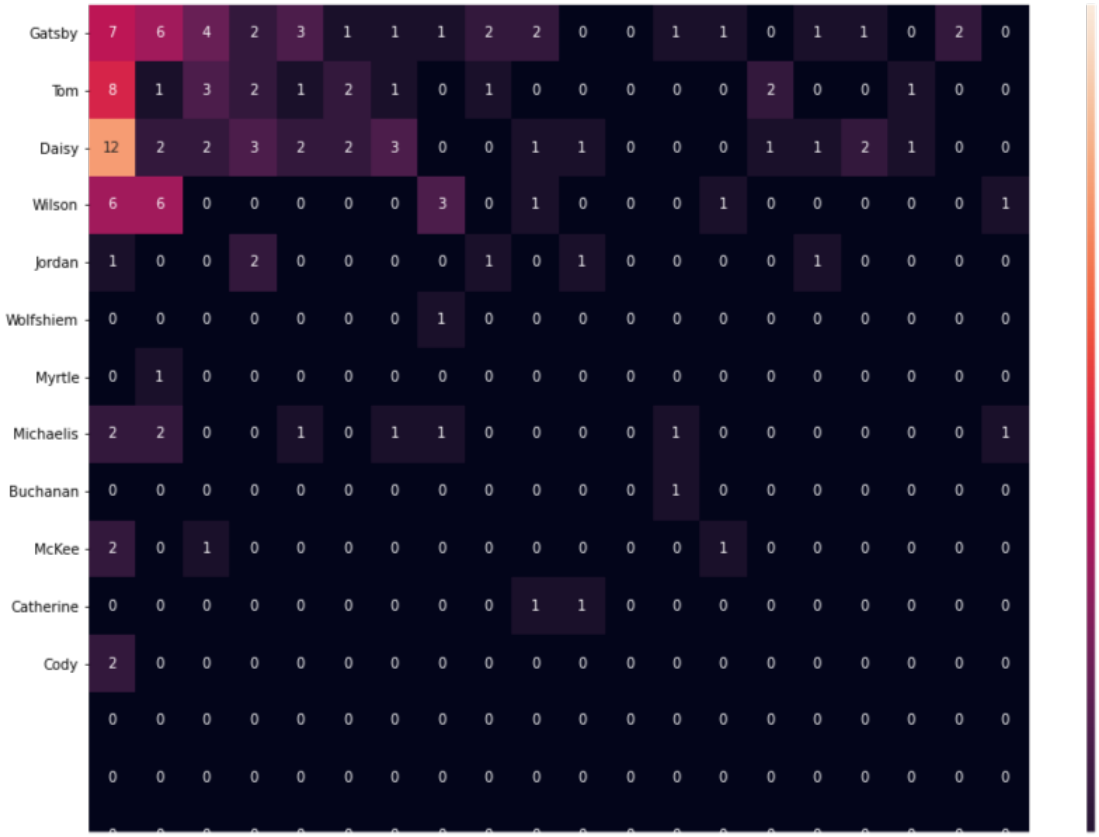
Since my main question involves the characters and their significance, I decided then to do a word cloud of the most common nouns and verbs. I knew that since some of the characters appeared as the most frequent in the overall word cloud, they would also in the noun based word cloud, but I was curious to see if other characters would be picked up as well.

Nouns

Verbs

We can see that the larger word cloud that included all of the words had a lot of similarities to the more specific word clouds, which was what I suspected would happen. When looking at the word cloud of nouns, new characters like “Myrtle” and locations like “West Egg” became more apparent. Therefore, although in the entire novel these characters and locations may not play the largest role, they do have importance in specific parts of the novel. Next, I wanted to see the ability of data science to uncover specific actions of the characters.

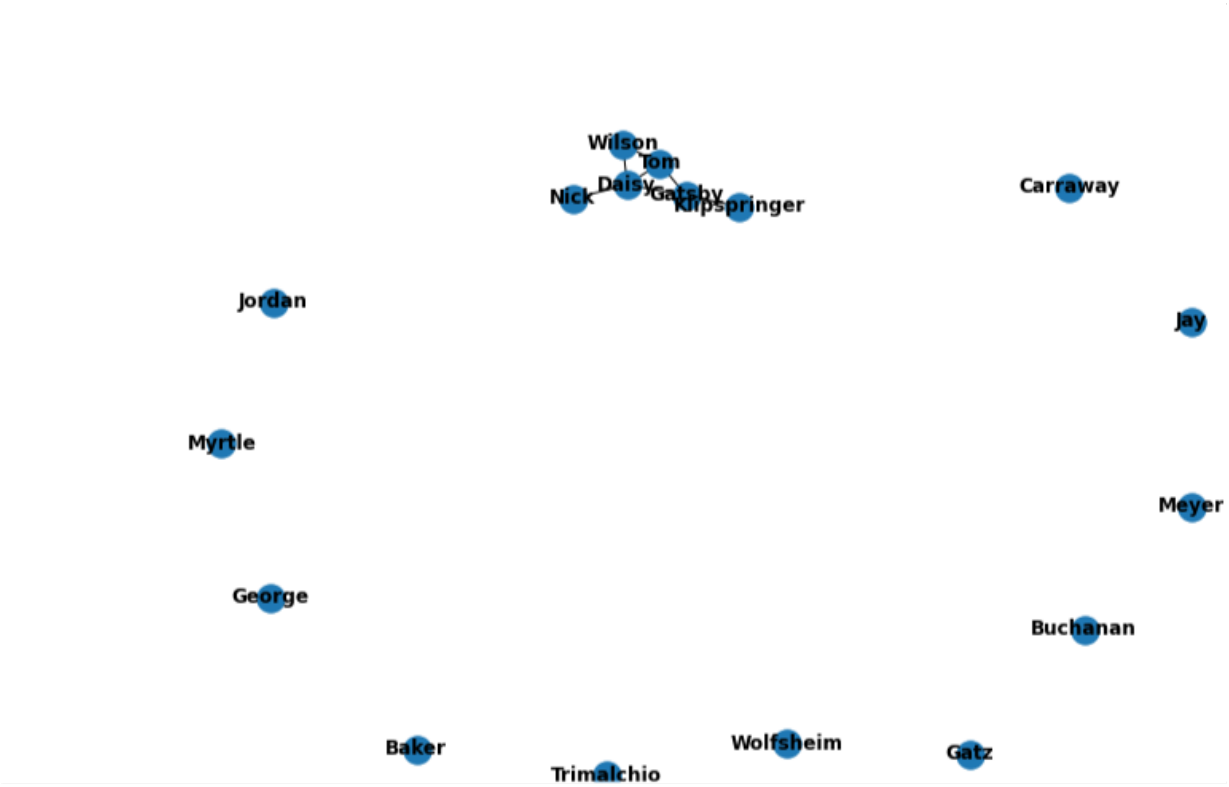
Character's Actions



In the heatmap above, we see the relationship between nouns and verbs in the novel. We see that Gatsby, Daisy, and Tom have the most verbs associated with them, which further explains why the original word cloud of the entire novel had their names as the most frequent; the more a character does something, the more they are likely to be mentioned in the book. Here, "love" was a verb that I expected to see due to the presence of romantic relationships in the novel. There are a lot of affairs, for example that between Tom and Myrtle, and the question of whether love is genuine or even existent is raised often. However, we see more generic verbs associated with the heatmap, which with the general nature of novels is very normal.

Character Connections

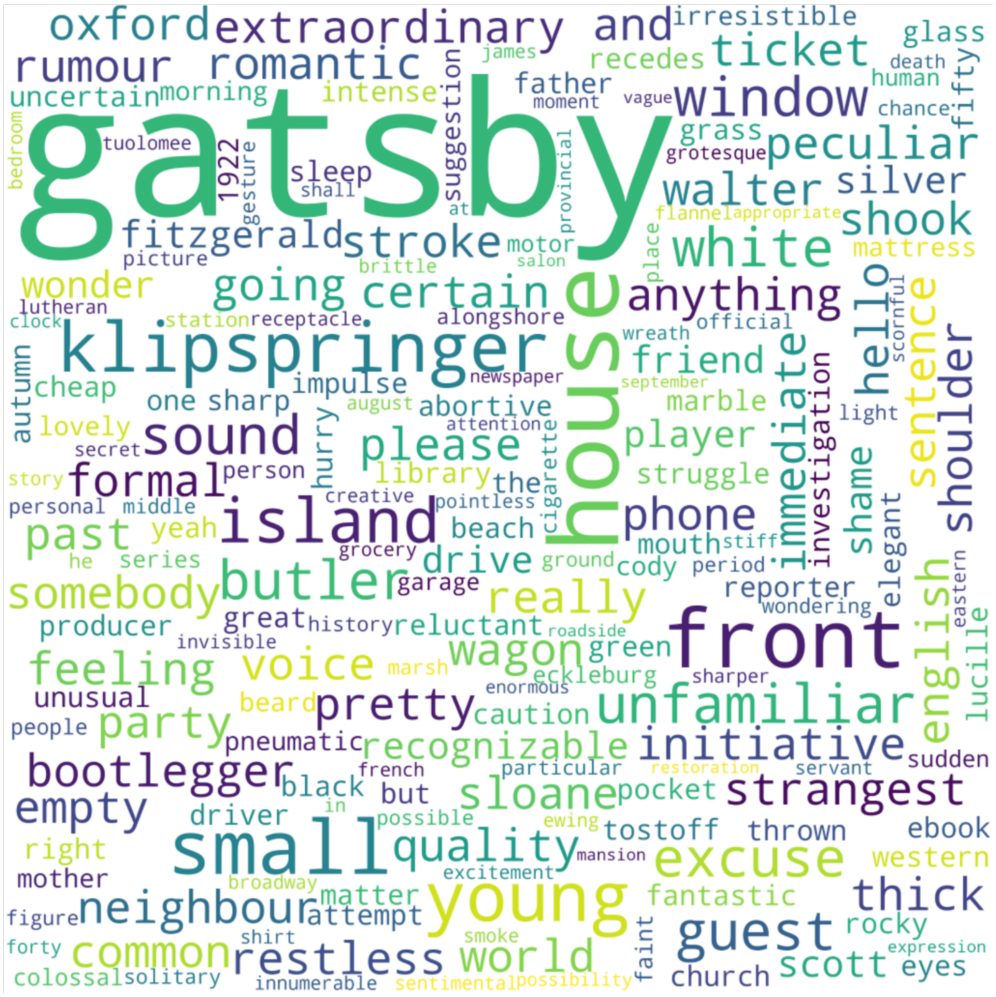
Looking at the connections between the characters, we can see that some are stronger than others. In this plot, the closer the nodes are to the center of the plot means the more relevant the node, character, is. The closer a node is to another node means the more significant the connection between the two is.



The connection of Nick to only Daisy is interesting and could be an error or restriction of our data science algorithms; it does make sense that this connection is the strongest since he is Daisy's cousin, but because he is also Gatsby's neighbor, there should be a connection there as well. However, the connection among Gatsby, Daisy, and Tom is extremely significant. This, and the connection between (Myrtle) Wilson, Daisy, and Tom points out the two major affairs in the novel. These affairs continuously help drive home some of the themes of the novel and build a stronger message that Fitzgerald tries to portray.

Topic Modeling: Gatsby

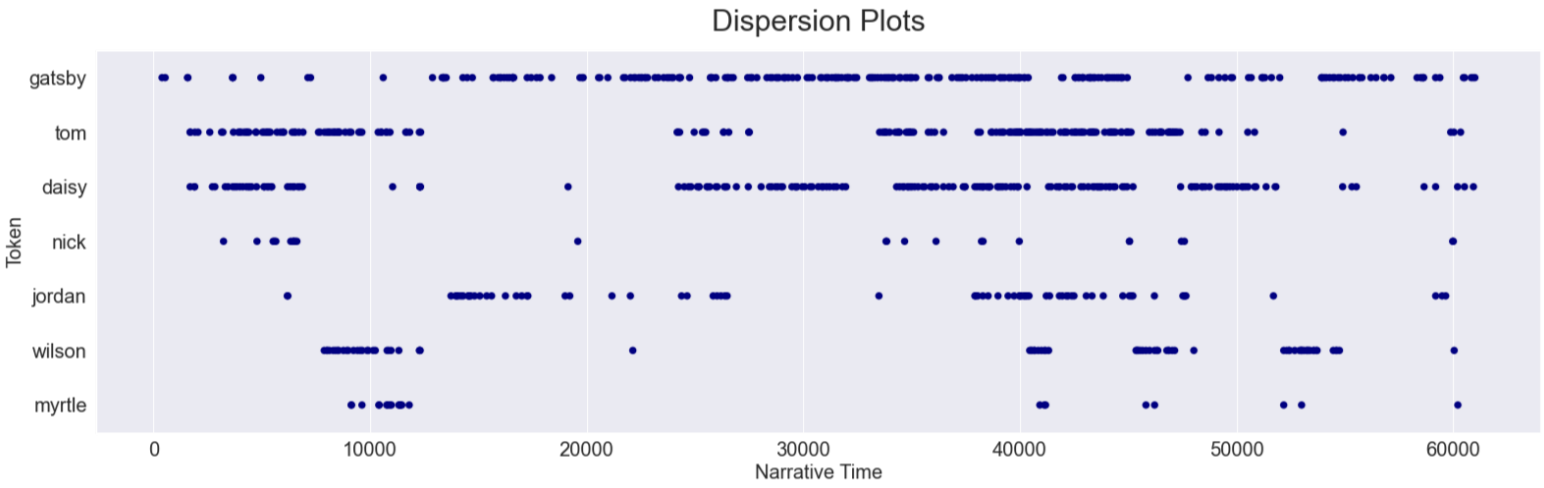
As Gatsby is the protagonist and therefore plays a major role, throughout the topic modelling applied on the preprocessed text, I decided to analyze the topic in which his name came up. To make this clearer, I created the word cloud on the next page.



Gatsby is described as a well-educated, self-made man who is one of the Long Island elite. The topic modelling was very accurately able to pick up on his principal characteristics. For example, “oxford” and “house” could be to represent the elite education he has and the house parties he throws. He is Nick’s “neighbour” and in the novel was said to get his wealth because he was a “bootlegger.” The appearance of these terms, therefore, is no coincidence, so our topic modeling does a great job of portraying who Jay Gatsby truly is. I was surprised to not see more terms that would lead us to recognize his materialistic nature, just as the quote in the beginning of the essay depicts.

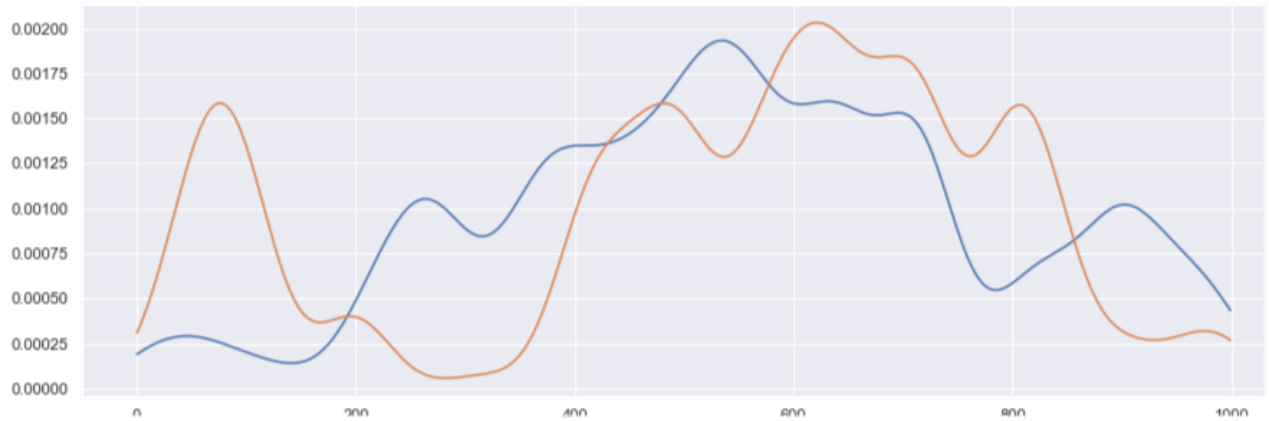
To look at the usage of each character over the course of the novel, we can visualize this with the dispersion plots as seen on the next page.

Characters' Importance: Appearance Over Narrative Time

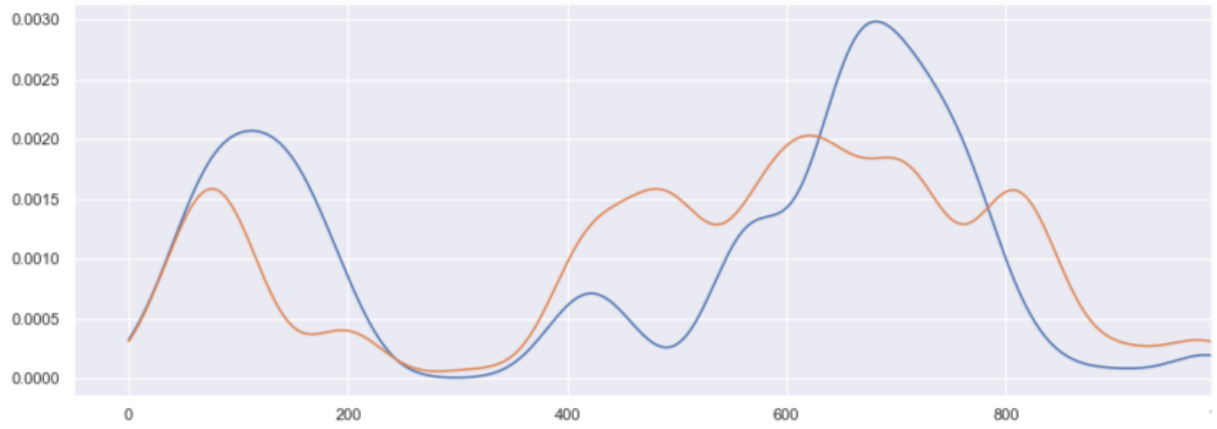


Looking at the Dispersion Plot above, we can see each of the important characters and their appearance over narrative time. We see how Gatsby is amongst the first of the characters to be introduced. Similarly, we see that Gatsby has the most exposure. As a protagonist whose actions drive the plot, this proportion of time is critical for the build up of his character. Tom and Daisy are at a close second/third for having the most exposure over narrative time. We can see that less relevant characters, like Myrtle and Jordan, have less dense plots. And Nick, the narrator, although a key character is not mentioned as often as others.

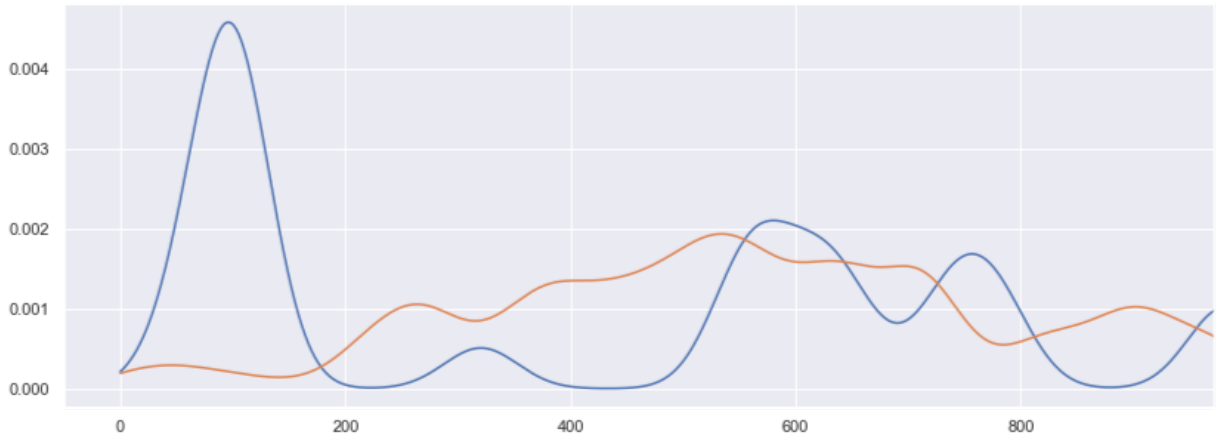
We can take a closer look at each of these characters and their appearances using a KDE plot, like the one below. This one shows Gatsby in blue and Daisy in orange. We can see that although at the start and end there are some differences in peaks, towards the middle of the novel the trends in the KDE lines are very similar. The characters must exist during the same narrative time of the novel. The last blue peak is likely around the time of Gatsby's death. During this time, it would make great sense for his character's name to be mentioned more.



Alternatively, we can look at Tom and Daisy's KDE plot since we would expect that for the majority of the narrative time, their names appear together. The only times when this may be different is when there are major conflicts, like an affair. Daisy is shown by the blue and Tom is shown by the orange lines. Below, we see this assumption to be true. Most of the trends of both of the lines are the same. There is a peak in Tom's KDE plot line between 600-800, which could be the duration of the novel where Tom finds out about Gatsby and Daisy's affair and blames Gatsby for the death of Myrtle (even though Daisy was to blame). Like with the other KDE plot, this one gives us a greater understanding of when the characters were mentioned in relation to one another.



Lastly, I wanted to look at Gatsby and Nick's KDE plot since these two characters are critical for the plot of the novel. Along the same line of thought, I wanted to look at how different the two plots would be since Nick's name is barely mentioned but Gatsby's is the most frequently mentioned.



In the KDE plot on the previous page, we see Nick as the orange line and Gatsby as the blue line. Even from the beginning of the narrative time, we can see that the presence of Gatsby's mention is significantly more prominent than Nick's. This remains true for a majority of the novel, but there are parts where the blue line dips beneath the orange— meaning that Nick's character is emphasized more. This is, in a way, contradictory to what we discovered earlier in the analysis, so it is beyond important to note this observation.

Book Summary

Moving beyond the characters, we can also quickly look at the ability of our notebooks to accurately describe and predict the novel's plot. I used Natural Language Processing techniques, specifically the NLTK package, to look at a summary of the novel that was generated.

```
"Orderi di Danilo," ran the circular legend, "Montenegro, Nicolas Rex." "Turn it." "Major Jay Gatsby," I read, "For Valour Extraordinary." "He
re's another thing I always carry. Son-of-a-bitch didn't even stopus car." "There was two cars," said Michaelis, "one comin', one goin', see?"
"Going where?" asked the policeman keenly. "Either you ought to be more careful, or you oughtn't to drive at all." "I am careful." "No, you're
not." "Well, other people are," she said lightly. "By the way, Mr. Gatsby, I understand you're an Oxford man." "Not exactly." "Oh, yes, I unde
rstand you went to Oxford." "Yes—I went there." A pause. He wears a pink suit." "Nevertheless he's an Oxford man." "Oxford, New Mexico," snort
ed Tom contemptuously, "or something like that." "Listen, Tom. "M-a-v-" the policeman was saying, "-o-" "No, r-" corrected the man, "M-a-v-r-o
-" "Listen to me!" muttered Tom fiercely. "Look here, old sport," he broke out surprisingly, "what's your opinion of me, anyhow?" A little ove
rwhelmed, I began the generalized evasions which that question deserves. "Why, I thought—why, look here, old sport, you don't make much money,
do you?" "Not very much." This seemed to reassure him and he continued more confidently. Where'd you pick that up?" "Now see here, Tom," said
Daisy, turning around from the mirror, "if you're going to make personal remarks I won't stay here a minute. "I believe we've met somewhere be
fore, Mr. Buchanan." "Oh, yes," said Tom, gruffly polite, but obviously not remembering.
```

I found that this summary captured a lot of major events from the novel and critical characteristics of the people. For starters, it highlights the car that killed Myrtle, leading to a cascade of events (including the death of Jay Gatsby). The concept of Gatsby's education and class is also mentioned with the quote "Oxford man," which describes his background; his intensive, elite education is representative of his wealth and place in society. More on this, the quote of Gatsby saying, "You don't make much money, do you?" which also contributes to the idea of a class division, which is a critical theme of the novel. This makes the general emphasis between the rich vs the poor, or similarly those who inherit old money vs new money. Although this summary is nowhere near perfect, compared to what could be written by a person who has read the novel, it does a fairly decent job. It was exciting to see that our summary matched the persona of Gatsby given in the topic modeling.

Sentiment Analysis

	Sentence	Sentiment Score	Subjectivity
1088	Gatsby, his hands still in his pockets, was reclining against the mantelpiece in a strained counterfeit of perfect ease, even of boredom.	1.000000	1.000000
1279	"I'm delighted that you dropped in." As though they cared!	0.875000	0.700000
976	It's a great advantage not to drink among hard-drinking people.	0.800000	0.750000
2246	"If he'd of lived, he'd of been a great man.	0.800000	0.750000
2336	He was always great for that.	0.800000	0.750000

With this summary in mind and a better comprehensive understanding of the people in the novel, we can move onto looking at the sentiment and subjectivity scores of each sentence and term. This can help us get a better sense of what type of sentiment/subjectivity the book holds. A sample of the data frame containing the sentiment score and subjectivity of each sentence can be found on the previous page. Although this dataframe is useful, applying the describe method to analyze the general sentiment and subjectivity of each sentence in the novel seemed more logical. The results for this are below.

Sentiment Score

```
count    2439.000000
mean        0.038130
std         0.222878
min        -1.000000
25%        -0.031881
50%         0.000000
75%         0.125595
max         1.000000
Name: Sentiment Score, dtype: float64
```

Subjectivity Score

```
count    2439.000000
mean        0.344623
std         0.301023
min         0.000000
25%         0.000000
50%         0.350000
75%         0.550000
max         1.000000
Name: Subjectivity, dtype: float64
```

Interpreting the results from above, a sentence in the novel has an average sentiment score of 0.038130 and an average subjectivity score of 0.344623. The average sentiment score is very close to zero. This means that most of the sentences within the novel are very neutral and do not contain highly negative or positive sentiment. I would have expected the sentiment score to be towards the negative side since there is a lot of criticism of the characters in the novel, and there are unfortunate events, like the death of Myrtle and Gatsby, that occurs. The average subjectivity is slightly higher, but is closer to zero than one; although subjectivity is not extremely high, it can be assumed that there are relatively a significant number of sentences that have some opinion to them. This makes sense considering there likely are many sentences expressing feelings towards other characters whether this is someone talking about how rich and snobby Gatsby is, how ‘extravagant’ the house parties are, or how infatuated they are with the

person they are with/have an affair with. This all points back to the materialistic nature of the characters and their desire to flaunt their wealth and prosperity.

Conclusion

In conclusion, data science, specifically when using Python, allows us to create fairly accurate descriptions of the main characters and allows us to explore the main plot of the novel in depth. After conducting my analysis, I truly do believe that the field of data-driven literary criticism should be expanded upon and studied further. This area of study allows for the growth of the reader's perspective or for affirmation of their prior thoughts. In both instances, we see that data holds great power. And, of course, in terms of whether my main question was answered— yes. I learned that these tools could accurately relay a character's actions, connections, and importance. The Great Gatsby truly encomposses the 1920's and the wealth-hungry, 'amoral' American society. Fitzgerald does a terrific job of showing how specifically Gatsby, the portrait of the American society in the 1920's, is greedy in his own prosperity and education. I found this project extremely enjoyable and hope to expand on this analysis in future data science courses as well as on my own over the Summer.

Bibliography

Bengani, Sneha. “What makes F Scott Fitzgerald’s The Great Gatsby a timeless classic.” *The Hindustan Times*, 2016, https://www.hindustantimes.com/books/what-makes-f-scott-fitzgerald-s-the-great-gatsby-a-timeless-classic/story-AGgfLcqrYbUbHvtnsNOPkO.html. Accessed 11 05 2021.

Fitzgerald, F. Scott. *The Project Gutenberg eBook of The Great Gatsby, by F. Scott Fitzgerald*. Project Gutenberg, 2021. *The Project Gutenberg*, https://www.gutenberg.org/files/64317/64317-h/64317-h.htm.

Prahl, Amanda. “'The Great Gatsby' Characters: Descriptions and Significance.” *ThoughtCo.*, 2019, https://www.thoughtco.com/the-great-gatsby-characters-4579831. Accessed 13 05 2021.

*All other sources listed in the notebooks