Nick Videtti

nvidetti@syr.edu

Assignment 2

Text Document Vectorization

Scikit Learn CountVectorizer, Word Clouds, and PyPlot

4/20/2023

| Quantifying Our Thoughts | |
| --- | --- |
| **Introduction** | “What were you thinking?” It is a question that is asked often in either intrigue or anger. While a specific answer is usually wanted, a rather vague answer tends to suffice. Thoughts are usually described in terms of topics with possibly some detail of short phrases, but the exact words of all thoughts, if applicable, are not usually discussed. Another aspect of thoughts that are usually not described is what was most thought about during a string of thoughts.  Thoughts, if all words in them are recorded, can be quantified to address these subtleties. The way this can happen is by taking all of the words from those thoughts and creating Word Clouds. Word Clouds are graphs that have words as their data points. While that is great in and of itself, the most useful thing about World Clouds is that the sizes of the words are based on the frequency of the corresponding word. That is, the most frequent word will be the largest, the second-most frequent word will be the second largest, the least frequent word will be the smallest, etc.  To test this out, thoughts that occur while watching a basketball game will be typed out and used to create Word Clouds. Somewhere from 15-20 sentences worth of thoughts will be typed up and their corresponding Word Clouds will be used to answer the question “what were you thinking?” in a way that tells what was being thought about most.  Some potentially unfamiliar terminology that may be seen in the results are as follows:  **Corpus** - set of text (the corpus will be the set of all thoughts recorded in this context)  **Document** - subset of a corpus (the documents will be sentences in this context) |
| **Analysis** | The process of this analysis will be to first gather the data. As stated in the introduction section, this will be 15-20 sentences of thoughts that are typed up while watching a basketball game. These sentences will be stored in a single string, then will be split into a list of sentences. While NLTK’s sentence tokenization could have been used (and maybe would have been better), this splitting was done using the split() function on the string with periods as separators, then repeated for exclamation points and question marks. This did end up creating some minor issues where nested lists were created and where trailing spaces were added, but both of those were addressed to ensure data cleanliness.  After that, the sentences (which will be interchangeable with “documents” for this part of the analysis) will be vectorized using Scikit Learn’s CountVectorizer to give the raw frequencies of the words for each document.  After the initial count vectors are created, two other sets of count vectors will be created on subsets of the corpus: one without CountVectorizer’s English stop words, one without CountVectorizer’s English stop words but only including words that occur more than once in the corpus.  Each of these sets of count vectors will be used to generate the Word Clouds. These Word Clouds will give a visualization to better interpret the count vectors, and will also give a simple way to describe results in a non-technical setting. The Word Clouds will be generated with wordcloud’s WordCloud and will be displayed using matplotlib’s pyplot. PyPlot will also allow for all 3 Word Clouds to be shown in the same figure while running code, but the results and conclusion sections will have a different screenshot for each Word Cloud to maximize the size, and hence visibility, of each. |
| **results** | The results of these analyses were as follows:  **RAW DOCUMENTS (19 Documents):**   1. The Knicks are a very very good basketball team 2. The Celtics are rivals of the knicks 3. Basketball is a game and a sport 4. The Knicks are from New York 5. Raptors are from Toronto which is in Ontario and the only NBA squad in Canada 6. I used to be good at basketball 7. Well not that good, but better than I am now 8. Right now it is the playoffs and the Knicks are facing Cleveland who are the Cavaliers and they used to have LeBron James a long time ago 9. Basketball is a good sport but most people like football more and they play less games in football 10. This is just another random sentence that is poorly written grammatically 11. Food is good to eat while watching the games 12. Chicken Wings are so good, they are sometimes sweet and spicy or both or only one 13. How many more sentences can I possibly come up with 14. Let's see 15. The Eagles are a very great and amazing football team in the NFL 16. They are super which is probably why they played in the Super Bowl this year 17. I had sweet honey chicken for dinner but I should've had something better for me but the taste was worth it 18. Well that is all the thoughts I have on my mind right now 19. let's see how many sentences this was   **COUNT VECTORS - RAW DOCUMENTS**  **(FIRST FEW ELEMENTS OF EACH)**    **COUNT VECTORS - DOCUMENTS WITHOUT STOP WORDS (FIRST FEW ELEMENTS OF EACH)**    **COUNT VECTORS - DOCUMENTS WITHOUT STOP WORDS**  **WORDS THAT OCCUR AT LEAST TWICE IN THE CORPUS**  **(FIRST FEW ELEMENTS OF EACH)**    **WORD CLOUDS:**        To no surprise, it looks like basketball was indeed being watched during these thoughts, as “basketball” is one of the biggest words in all of the Word Clouds. It is also interesting to see “good” appear so prominently, as that could mean many things, but could possibly indicate that the person watching is enjoying the game. There are many insights like these that can be made, but most importantly these Word Clouds tell not only what was being thought about, but what was being thought about most. |
| **conclusions** | The sentences that were typed up were as follows:  **RAW DOCUMENTS (19 Documents):**   1. The Knicks are a very very good basketball team 2. The Celtics are rivals of the knicks 3. Basketball is a game and a sport 4. The Knicks are from New York 5. Raptors are from Toronto which is in Ontario and the only NBA squad in Canada 6. I used to be good at basketball 7. Well not that good, but better than I am now 8. Right now it is the playoffs and the Knicks are facing Cleveland who are the Cavaliers and they used to have LeBron James a long time ago 9. Basketball is a good sport but most people like football more and they play less games in football 10. This is just another random sentence that is poorly written grammatically 11. Food is good to eat while watching the games 12. Chicken Wings are so good, they are sometimes sweet and spicy or both or only one 13. How many more sentences can I possibly come up with 14. Let's see 15. The Eagles are a very great and amazing football team in the NFL 16. They are super which is probably why they played in the Super Bowl this year 17. I had sweet honey chicken for dinner but I should've had something better for me but the taste was worth it 18. Well that is all the thoughts I have on my mind right now 19. let's see how many sentences this was   Using these, three different Word Clouds were generated. The first included all text from the 19 documents. The second included all text from the 19 documents except for stop words. The third included the text from the second word cloud, but further filtered down to only words that occurred multiple times within the corpus. To clear up any questions, stop words are words that do not usually change the meaning of a phrase. They are usually short and included in the phrases simply for grammar reasons. The stop words filtered out in these Word Clouds are determined by the Word Cloud generator. The results below are how the Word Clouds ended up looking.        To no surprise, it looks like basketball was indeed being watched during these thoughts, as “basketball” is one of the biggest words in all of the Word Clouds. It is also interesting to see “good” appear so prominently, as that could mean many things, but could possibly indicate that the person watching is enjoying the game. There are many insights like these that can be made, but most importantly these Word Clouds tell not only what was being thought about, but what was being thought about most.  While this is technically a way to quantify thoughts, there are unfortunately (and obviously) many reasons why this is not feasible in real-world scenarios. That being said, in an ideal world, it would be pretty interesting to utilize something similar to this to get insight into all thoughts someone has over the course of a few hours or days. |