NLP Homework 2

Kyle Walter

8/21/2021

# Homework 2 Exploratory Analysis for Full Sentiment analaysis.

Continuing with the texts that I used for homework #1, Anna Karenina and The Great Gatsby, I take a look at Tokenizing by sentence and apply grammar tags to get a sense of the ability to preform sentiment analysis.

However; before I get a head of myself it is important to do a review of the use of grammar by the respective authors and see if there are good adjective and adverbs to help drive the sentiment analysis that will be in part three.

## Preprocessing

As with any good data analysis we start with preprocessing of the text. First step was to import the corpora into the Python environment. The two books I have chosen are both available for the Gutenberg Project so they come in with the main text of the book as a human reader would expect see.

Since I will be looking grammar phrases in this exercies the next step was to Tokenize the text into sentence. The Natural Language Tool Kit (nltk) provides a built in function to do this with the sent\_tokenize function. This breaks the text into its respective sentences for further analysis.

After this step, I applied the word tokenize for each word in the sentence. This breaks each sentence into 1 word tokens, but keeps each respective sentence together in a seperate lists. This will be important in later steps.

Here is a sample sentence from the start of Gatsby that has been tokenized and kept as a single sentence

[‘“’, ‘Whenever’, ‘you’, ‘feel’, ‘like’, ‘criticizing’, ‘anyone’, ‘,’, ‘”’, ‘he’, ‘told’, ‘me’, ‘,’, ‘“’, ‘just’, ‘remember’, ‘that’, ‘all’, ‘the’, ‘people’, ‘in’, ‘this’, ‘world’, ‘haven’, ’’‘, ’t’, ‘had’, ‘the’, ‘advantages’, ‘that’, ‘you’, ’’‘, ’ve’, ‘had.’, ‘”’, ‘He’, ‘didn’, ’’‘, ’t’, ‘say’, ‘any’, ‘more’, ‘,’, ‘but’, ‘we’, ’’‘, ’ve’, ‘always’, ‘been’, ‘unusually’, ‘communicative’, ‘in’, ‘a’, ‘reserved’, ‘way’, ‘,’, ‘and’, ‘I’, ‘understood’, ‘that’, ‘he’, ‘meant’, ‘a’, ‘great’, ‘deal’, ‘more’, ‘than’, ‘that’, ‘.’]

As we can see the square brackets at the start and end show us that that sentence is kept together by a list and tokenized by words. The lists for each of the respective sentences are shown inside of a larger list that makes up the corpus.

Lastly at this stage comparing with the previous run of the corpora through through the tokenization process, I checked how many tokens each respective text had.

Anna Karenina, an eight part novel and known for being quite long showed 16,787 tokens, while The Great Gatsby showed 2439 total tokens.

## Part of Speech Tagging

The next step is to tag the the sentences with their respective parts of speech. I used nltk’s parts of speech tag, that passed each token through the the system and tagged each word with their respective parts of speech, such as noun, adjective, verb, etc. There is bank of grammar designed for the base parts of speech tagger in nltk by Standford. Where words have multiple potential parts of speech like the world “read” the the parts of speech tagger uses other words in the sentence to determine the part of speech. There is ratio that is calculated and highest ratio is the tag that the word is assigned. If the word doesn’t exist in the bank at all, there is usually a default.

For Anna Karenina, here are the first two sentences with parts of speech tags:

[[(‘[’, ’JJ’), (’Illustration’, ’NNP’), (’]’, ‘NNP’), (‘ANNA’, ‘NNP’), (‘KARENINA’, ‘NNP’), (‘by’, ‘IN’), (‘Leo’, ‘NNP’), (‘Tolstoy’, ‘NNP’), (‘Translated’, ‘VBN’), (‘by’, ‘IN’), (‘Constance’, ‘NNP’), (‘Garnett’, ‘NNP’), (‘Contents’, ‘NNP’), (‘PART’, ‘NNP’), (‘ONE’, ‘NNP’), (‘PART’, ‘NNP’), (‘TWO’, ‘NNP’), (‘PART’, ‘NNP’), (‘THREE’, ‘NNP’), (‘PART’, ‘NNP’), (‘FOUR’, ‘NNP’), (‘PART’, ‘NNP’), (‘FIVE’, ‘NNP’), (‘PART’, ‘NNP’), (‘SIX’, ‘NNP’), (‘PART’, ‘NNP’), (‘SEVEN’, ‘NNP’), (‘PART’, ‘NNP’), (‘EIGHT’, ‘NNP’), (‘PART’, ‘NNP’), (‘ONE’, ‘NNP’), (‘Chapter’, ‘NN’), (‘1’, ‘CD’), (‘Happy’, ‘JJ’), (‘families’, ‘NNS’), (‘are’, ‘VBP’), (‘all’, ‘DT’), (‘alike’, ‘RB’), (‘;’, ‘:’), (‘every’, ‘DT’), (‘unhappy’, ‘JJ’), (‘family’, ‘NN’), (‘is’, ‘VBZ’), (‘unhappy’, ‘JJ’), (‘in’, ‘IN’), (‘its’, ‘PRP$’), (‘own’, ‘JJ’), (‘way’, ‘NN’), (‘.’, ‘.’)], [(‘Everything’, ‘NN’), (‘was’, ‘VBD’), (‘in’, ‘IN’), (‘confusion’, ‘NN’), (‘in’, ‘IN’), (‘the’, ‘DT’), (‘Oblonskys’, ‘NNP’), (’’‘, ’NNP’), (‘house’, ‘NN’), (‘.’, ‘.’)]]

And for Gatsby here are the respective first 2 sentences Part of SPeech Tags:

[[(‘The’, ‘DT’), (‘Great’, ‘NNP’), (‘Gatsby’, ‘NNP’), (‘by’, ‘IN’), (‘F.’, ‘NNP’), (‘Scott’, ‘NNP’), (‘Fitzgerald’, ‘NNP’), (‘Table’, ‘NNP’), (‘of’, ‘IN’), (‘Contents’, ‘NNP’), (‘I’, ‘PRP’), (‘II’, ‘NNP’), (‘III’, ‘NNP’), (‘IV’, ‘NNP’), (‘V’, ‘NNP’), (‘VI’, ‘NNP’), (‘VII’, ‘NNP’), (‘VIII’, ‘NNP’), (‘IX’, ‘NNP’), (‘Once’, ‘RB’), (‘again’, ‘RB’), (‘to’, ‘TO’), (‘Zelda’, ‘NNP’), (‘Then’, ‘RB’), (‘wear’, ‘VBD’), (‘the’, ‘DT’), (‘gold’, ‘NN’), (‘hat’, ‘NN’), (‘,’, ‘,’), (‘if’, ‘IN’), (‘that’, ‘DT’), (‘will’, ‘MD’), (‘move’, ‘VB’), (‘her’, ‘PRP’), (‘too’, ‘RB’), (‘,’, ‘,’), (‘Till’, ‘NNP’), (‘she’, ‘PRP’), (‘cry’, ‘VBD’), (‘“’, ‘NNP’), (‘Lover’, ‘NNP’), (‘,’, ‘,’), (‘gold-hatted’, ‘JJ’), (‘,’, ‘,’), (‘high-bouncing’, ‘JJ’), (‘lover’, ‘NN’), (‘,’, ‘,’), (‘I’, ‘PRP’), (‘must’, ‘MD’), (‘have’, ‘VB’), (‘you’, ‘PRP’), (‘!’, ‘.’), (‘”’, ‘VB’), (‘Thomas’, ‘NNP’), (‘Parke’, ‘NNP’), (‘d’, ‘NN’), (’’‘, ’NN’), (‘Invilliers’, ‘NNP’), (‘I’, ‘PRP’), (‘In’, ‘IN’), (‘my’, ‘PRP’), (‘father’, ‘NN’), (‘gave’, ‘VBD’), (‘me’, ‘PRP’), (‘some’, ‘DT’), (‘advice’, ‘NN’), (‘that’, ‘IN’), (‘I’, ‘PRP’), (’’‘, ’VBP’), (‘ve’, ‘RB’), (‘been’, ‘VBN’), (‘turning’, ‘VBG’), (‘over’, ‘IN’), (‘in’, ‘IN’), (‘my’, ‘PRP$’), (‘mind’, ‘NN’), (‘ever’, ‘RB’), (‘since’, ‘IN’), (‘.’, ‘.’)], [(‘“’, ‘VB’), (‘Whenever’, ‘WRB’), (‘you’, ‘PRP’), (‘feel’, ‘VBP’), (‘like’, ‘IN’), (‘criticizing’, ‘VBG’), (‘anyone’, ‘NN’), (‘,’, ‘,’), (‘”’, ‘NN’), (‘he’, ‘PRP’), (‘told’, ‘VBD’), (‘me’, ‘PRP’), (‘,’, ‘,’), (‘“’, ‘VB’), (‘just’, ‘RB’), (‘remember’, ‘VB’), (‘that’, ‘IN’), (‘all’, ‘PDT’), (‘the’, ‘DT’), (‘people’, ‘NNS’), (‘in’, ‘IN’), (‘this’, ‘DT’), (‘world’, ‘NN’), (‘haven’, ‘NN’), (’’‘, ’NNP’), (‘t’, ‘NN’), (‘had’, ‘VBD’), (‘the’, ‘DT’), (‘advantages’, ‘NNS’), (‘that’, ‘IN’), (‘you’, ‘PRP’), (’’‘, ’VBP’), (‘ve’, ‘JJ’), (‘had.’, ‘NN’), (‘”’, ‘NN’), (‘He’, ‘PRP’), (‘didn’, ‘VBZ’), (’’‘, ’JJ’), (‘t’, ‘NNS’), (‘say’, ‘VBP’), (‘any’, ‘DT’), (‘more’, ‘JJR’), (‘,’, ‘,’), (‘but’, ‘CC’), (‘we’, ‘PRP’), (’’‘, ’VBP’), (‘ve’, ‘JJ’), (‘always’, ‘RB’), (‘been’, ‘VBN’), (‘unusually’, ‘RB’), (‘communicative’, ‘JJ’), (‘in’, ‘IN’), (‘a’, ‘DT’), (‘reserved’, ‘JJ’), (‘way’, ‘NN’), (‘,’, ‘,’), (‘and’, ‘CC’), (‘I’, ‘PRP’), (‘understood’, ‘VBP’), (‘that’, ‘IN’), (‘he’, ‘PRP’), (‘meant’, ‘VBD’), (‘a’, ‘DT’), (‘great’, ‘JJ’), (‘deal’, ‘NN’), (‘more’, ‘JJR’), (‘than’, ‘IN’), (‘that’, ‘DT’), (‘.’, ‘.’)]]

## Information Extraction

### Adjective Phrases

Now that the tokens have their respective parts of speech tags applied, the next step in our exploration is to look for adjective phrases. These are combinations of adverbs and adjectives in combination. being able to identify these is important as these tokens usually have highest sentiment scores and thus and can give us a sense of the author’s view.

In order to do this, in Python, the regular expression package was called. At this point a regular expression was called to scan the parts of speech tags and look for where two tokens in a sentence have an adverb followed by an adjective.

The regular expression was then passed to the chuck parser in nltk and each sentence in the corpora was checked by the code for the combination and a new list created of the combinations found.

In order to check that it was working correctly, I printed out the Combinations. For Anna the following were the first 10:

[‘painfully conscious’, ‘very nice’, ‘most awful’, ‘acutely painful’, ‘not so much’, ‘very disgraceful’, ‘therefore idiotic’, ‘no longer young’, ‘most complex’, ‘sly smile’]

And the first 10 strings of Gatsby: > [‘more vulnerable’, ‘unusually communicative’, ‘usually plagiaristic’, ‘more riotous’, ‘not likely’, ‘rather hard-boiled’, ‘very grave’, ‘so much’, ‘so much’, ‘rather literary’]

The stings are generally showing good results, as we can easy pick out combinations like painfully conscious, very nice, more vulnerable.

Next let’s look at the most frequent adjective phrases in each Corpora. For Anna here are the 10 common:

so much 67 very glad 45 so many 25 too much 23 very nice 20 once more 18 very good 17 most important 14 so glad 14 so little 13

For Gatsby here are the 10 most frequent adjective phrases:

so much 4 “ Good 4 very good 3 too much 3 more interesting 2 ” insisted 2 m glad 2 once more 2 too hot 2 very much 2

Both books interesting share the most common adjective phrase of so much. What may need to be cleaned out of these list for a further sentiment analysis are those phrase that once of the tags is on punctuation. Example in Gatsby because it is among the most common is [" Good], as a quotation mark really is not an adverb. Generally though in the cleaning process for this type analysis, removing stop words and punctuation is not done because it could in adversely put together two words that under normal circumstance would not normally be together. In addition when we did the parts of speech tagger, it could also give more erroneous results as it is trying to build from the context of the sentence while missing words that would advise proper phrase construction.

Lastly I looked at the number of adjective phrases found in the document by the parser. 2,687 in Anna Karenina were found and 295 in the Great Gatsby.

### Adverb Phrases

Now that we’re done with adjective phrases, we should also look at adverb phrases. While an adjective will give us a starting score for the sentiment analysis coming up in Homework 3, the adverb phrases help give us a strength of that sentiment.

In order to extract the adverb phrases we followed similar steps to the extraction of the adjective phrases. Starting with defining the regular expression looking for a place in our sentences where two adverbs happen in sequence. The next step is pass regular expression to the parsing algorithm in nltk. Lastly pass in the text in a loop by each sentence for it to pull out the phrases and appends them to a new list.

Next we inspect the first 10 captured in each text. For Anna it was the following:

[‘not only’, ‘“ Now’, ‘m not’, ‘perfectly still’, ‘so often’, ‘not so’, ‘—utterly involuntarily’, ‘never clearly’, ‘long ago’, ‘till now’]

The first 10 for Gatsby:

[‘Once again’, ‘all right’, ‘so thoroughly’, ‘lonely no’, ‘over there’, ‘there unrestfully’, ‘aggressively forward’, ‘Not even’, ‘“ Now’, ‘about restlessly’]

Similar to before, we can see that the pos tagger has labeled punctuation such as quotation marks and some contractions with adverb tags. As we work through the sentiment analysis this should be taken into account as these likely will cause other problems.

For reference here are the 10 most common adverb phrases captured for Anna:

“ Well 94 m not 56 as soon 47 not merely 42 not even 42 only just 39 very well 38 not yet 36 not only 32 so well 32

And the 10 most common adverb phrases captured from Gatsby: > “ Well 5 so long 5 “ Now 4 far away 4 d never 4 so much 3 m not 3 no longer 3 ve never 3 right now 3

[" Well] is the top adverb expression for Anna Karenina and Gatsby, so another technique will be required in order to do a sentiment analysis in homework 3. For Anna we see that some of the adverb phrases are a bit murky, not merely for example could easily negate or enhance a sentiment being given, but this will be more fully explored in the next assigned.

Lastly for overall adverb phrase statics, currently Anna has 2,054 adverb phrases captured and Gatsby has 222.

## Other Statistics

### Major Parts of Speech

In order to finish off this review, let us take a look at the most common words by part of speech. Recall previously, we used nltk’s part of speech tagger to tag all of the words by parts of speech. This was done inside of sentences so that the correct part of speech could be attached based on the context of the sentence.

Starting with adjectives, in python an for look is created that grabs each sentence from the tagged text and then inside of each sentence the text grabs the tokens that have an an adjective tag with them. Specifically any that start with JJ. While this can be done with regular expression as seen in the phrasing, because the list is smaller, it was called by hard coding the values in a list. The adjectives are collected stored in an adjectives tokens list for each respective text. Lastly to get sense of the words, ran a frequency on the top 50 words.

For the sake of space, here are the top 10 adjectives from Anna: > little 449 same 441 old 384 other 360 own 349 new 305 such 294 good 289 more 287 first 284

And for Gatsby here are the top 10:

little 98 old 79 other 64 ll 46 last 44 young 43 first 41 white 39 ve 35 more 32

Gatsby’s 4th and 9th most common adjectives are contractions. While definitely possible, it may one again represent an over reach like the one seen with quotation markets in the adjective and adverb phrasing.

The words in Gatsby point to a juxtaposition in time with the worlds young and old prevalently seen at the top of the list.

Next we’ll look at adverbs. Applying similar logic with the nested for loop, this time looking all the distinct combinations of the RB the tag code for adverbs.

For Anna here are the two 10 adverbs: > not 3441 so 1280 now 741 very 584 only 518 too 494 more 438 just 435 still 403 again 402

For Gatsby here are the top 10 adverbs: > just 86 then 86 back 82 now 81 so 79 not 62 never 60 here 60 there 58 away 57

I noticed that some of the adverbs at the top of the list are not the same ones seen earlier in the phrases, this is likely cause we were looking solely for adverb and adjective as well as adverb and adverb combinations. Adverbs can also modify verbs so that was not captured. For the sentiment analysis I will need capture the larger phrase to know if it is used to modify an adjective or verb so we do not lose out on valuable information.

Continuing on, the nested loop logic is reapplied in order to check out the top nouns. Here the top for Anna: > Levin 1592 Vronsky 850 Anna 815 Kitty 663 Alexey 629 Alexandrovitch 570 time 556 man 550 Stepan 548 Arkadyevitch 547

What is highlighted here is that the major nouns at the top of the list are many of the main characters in the story. The POS Tagger has done a nice job of identifying these especially given the book’s Russian language origin and names that would not normally be seen in English. As many of these were explored more extensible in Homework 1 as nouns tend to be the most common words and that is seen just by their high frequencies.

Here are the top nouns for Gatsyby:

Gatsby 250 Tom 188 Daisy 182 man 98 house 90 eyes 85 Mr. 78 Wilson 77 time 75 way 72

Similar to Anna, the main characters are towards to the top of the list, as there are fewer of them, other nouns have found there way to the top as well, like eyes.

Lastly be change the logic one last time capture the verbs by changing our POS list to those combinations with VB we can see the most common verbs.

Here are the top verbs from Anna: > was 5284 had 3846 said 2721 be 1766 is 1357 have 1231 were 1227 been 1057 did 923 do 704

And here are the top 10 from Gatsby:

was 761 had 378 said 233 were 170 have 129 be 121 came 108 been 106 is 92 went 90

What is interesting to see is that for both books, the top ten most common verbs are the state of being verbs such as “is”,“are”,“was”,“were” that come to the top of the list. Where true action verbs generally do not break in to the top of the list. Though as all sentences in English require a verb, I wonder if this is more a feature of the language.

Even running the top 50 in Python, I noticed that the first 3rd of both lists are covered in state of being or verbs that express this. So much of the sentiment is going to come from the view of the world the authors put in their respective expositions.

### Sentence Statistics

Lastly, now that we have adjective and adverb phrases, lets use the sentence set to figure out how many sentence we have that would be good candidates for getting sentiments in the future homework.

Starting with the adjective set, we created an empty list. Then through a loop we pass that list of adjective phrases created earlier through and extract the entire sentence and split by part of speech tag and the word. We pass the words to a temporary variable and then append the temp variable to the whole sentence list. At the end we ran a print statement to tell us how many sentence were captured. For Anna the adjective pharases were found in 5681 sentences across the corpus and for Gatsby the result was 617.

The next step was to repeat the logic, except this time we utilize the adverb phrases together the sentences. Across Anna there are 4,226 sentences that contain adverb phrases and 459 in Gatsby.

As an optional step, we combined the two sets of whole sentence lists together into a unique list of sentences that have all the adverb and adjective phrases. To do this we started by copy in the previously captured adjective whole sentences into a new variable that will contain both. We then wrote a loop that takes a sentence from the adverb whole list and checks to see if it is contained in the whole list of both. If that sentence is not there, then it will be appended to the whole. Following the completion of the loop we printed out the total number of sentences that have adjective and adverb phrases in the respective novels. For Anna we see that there are 6,524 unique sentences that have an adjective or adverb phrase contained with in them. For Gatsby we see that 815 sentences have unique adjective and adverb phrases contained within them.

Lastly, in preparation for our third homework assigned, I wanted to know how much of the corpora had an adjective phrase to look at for sentiment. Since these words typically hold all the positive and negative context in the English Language. The first step towards this was to calculate the average length of of a sentences. To do this we calculated to the total length of the corpus using a for loop and summing up the length of each sentence from our first split back at the beginning and storing it in a variable called total Corpus. The next step was to divide the whole corpus by the number of total sentences. This returns that the average length of the sentence in Anna Karenina is 117.52 characters while the average length of a sentence in Gatsby is 111.43.

Lastly I looked at the average length of the adjective phrases sentence vs the total and found around 10% for both books.

## How to conduct a sentiment analysis from here

From this point, I think there are few things that need to be done prior to doing a sentiment analysis. As noted earlier this document, there are alot of adverb phrases that show up in both novels where quotation marks are getting flagged as adverbs. This will need to be cleaned up.

In addition Anna Karenina has some adverb expressions that lead to believe that the book will have a few places where an adjective gets negated or replaced with another. Example not merely in English can greatly change the weight of the following adjective quite a bit.

The piece that we would probably also want to do is to look at how these phrases connect into nouns they’re describing to see if we can build a profile of the characters from the view of the author. But these are a few simple ideas and maybe as we get a little deeper into the topic we can find another technique to apply.