Natural Language Processing: Homework 1

Kyle Walter

# Introduction and Preprocessing:

For the purpose of this homework assignment, my focus was on finding two corpora that have both comparable and contrastable features for their work frequencies. I spent some time searching Guttenberg and found two major novels written by different authors about 50 years apart. Anna Karenina was written by Russian author Leo Tolstoy, who probably in the US was probably more famous for writing war in peace. Anna Karenina was written in time of change under tsar Alexander, and heavily focuses on social class issues through lens of the main character’s affairs.

Fast forward and The Great Gatsby written by F. Scotts Fitzgerald was set to expose the social classes in at the play in the 1920s, it also holds a series of affairs between the main characters but is set in the 20th century.

I wonder if the word choices of these author’s changes much given the different periods. And also, is there any reflection of social class that can be determined in the major words. Also does the difference in the author’s original language make any difference in their outcome.

## Preprocessing:

As I mentioned in the opening, both texts were collected from the Gutenberg project. I used the Gutenberg package in Python to call them form their library. This provides the book’s text as a string. One of the functions in the package also allows the removal the headers that are displayed in the file, so I was able to get a string of the just the book’s text.

From there I imported the Natural Language Tool Kit (NLTK Package) an tokenized the string, a process that breaks up the string into the units representing words, or ideas. I used the base NLTK “word\_tokenization” which splits the various text by the white space around it along with creating tokens of the various punctuation used as well. The next step was to lower case all the words, as English has a requirement to capitalize formal nouns and the words at the beginning of the sentence, this helps with words frequencies as a word like See and see will be recognized as the same word. The step following tokenization was to remove the punctuation because generally items such as commas and periods show as the most frequent tokens and do not really provide any information about content of the text. I also found at this step that there were several words towards the start of the script with an “\_” character in front of it. I elected to eliminate them as well using a small regex pattern to look for words that started with the “\_” to their boundary as some had about “\_” following the word and were tokenized as such. While some only had the “\_” at the start of the word.

My next step was the remove stop words, or the words that are in the text that do not provide much information about what is going on in the story. Words such as a, an, the, he, she, I and so forth. The reason for removing these words is that they come up so frequently they hide words that would tell the reader about the corpus. In looking at the initial unigrams the word “said” shows up almost 3x as often in both books than the next available unigram. I assuming because the common way that conversation is written down in novel in English is something like Kyle said, “I went to the store”. Moreover, later when looking at bigrams I would see various character said combinations which weren’t really meaningful unless I was trying to do analysis on how frequently one of them spoke.

# The unigrams

Next, I dug into the 50 most frequent tokens from both books.

From Anna:

('levin', 1621) ('one', 1230) ('go', 1081) ('would', 1044) ('look', 1001) ('could', 970) ('come', 882) ('vronski', 855) ('anna', 818) ('know', 796) ('see', 775) ('say', 702) ('like', 689) ('kitti', 668) ('went', 660) ('thought', 658) ('time', 642) ('hand', 637) ('smile', 631) ('alexey', 629) ('well', 624) ('face', 588) ('love', 584) ('alexandrovitch', 570) ('eye', 567) ('feel', 556) ('felt', 553) ('man', 550) ('stepan', 548) ('arkadyevitch', 547) ('ye', 531) ('noth', 523) ('though', 518) ('ask', 488) ('think', 488) ('get', 485) ('talk', 481) ('even', 474) ('littl', 456) ('want', 451) ('life', 450) ('answer', 431) ('still', 430) ('long', 426) ('someth', 424) ('saw', 422) ('without', 420) ('came', 420) ('take', 417) ('day', 416)

In the top unigrams, one-word tokens, we can see some of the main characters coming to the surface. Levin, Anna’s brother-in-law. Vronsky, the person with whom Anna has an affair, Anna the main character. And Alexey Anna’s husband. We do see the family name as well Alexandrovitch.

Words like love give us a sense of some of the themes of the story and face could be there from consequences as well in the story. That said for as many train motifs as there are in the story, I was surprised not see the word coming to the top.

From Gatsby: ('gatsbi', 251) ('tom', 188) ('daisi', 183) ('look', 173) ('one', 149) ('go', 137) ('like', 128) ('back', 109) ('came', 108) ('littl', 103) ('hand', 103) ('come', 98) ('man', 98) ('want', 97) ('hous', 97) ('know', 97) ('get', 93) ('time', 92) ('went', 90) ('eye', 90) ('got', 85) ('turn', 84) ('old', 81) ('see', 80) ('mr.', 78) ('wilson', 77) ('car', 75) ('think', 74) ('way', 73) ('moment', 73) ('two', 72) ('jordan', 71) ('new', 69) ('voic', 69) ('night', 69) ('door', 68) ('around', 68) ('someth', 66) ('away', 66) ('face', 66) ('girl', 66) ('made', 63) ('never', 63) ('thing', 62) ('room', 62) ('long', 62) ('us', 62) ('could', 61) ('even', 61) ('peopl', 60)

The top unigrams for the Great Gatsby are the main characters of the story, Gatsby, Tom, and Daisy. We do see words like car showing the move to the 21st century. Night also stands out as many of the parties that Gatsby host are parts of the scenes.

One contrast that stands out to me is that we do see the token car 75 times throughout the text of Gatsby. Something that is showing the passage of time between when the novels are written. Additionally, we can see that in Anna there are some words that by the time Gatsby has come around have been almost eliminated from modern English. Example “ye” shows up 531 times in book. This word today shows up only in expressions like “ye of little faith” or when setting something in an old period.

The other contrast I see are the number of verbs that show up in the top 50 unigrams for Anna, words like say, like, see, went, smile, love, feel, felt, and saw. I tried the two stemmers we have studied in class to see if that could help with these, because some are verb tense forms of each other such as see and saw or feel and felt. These are not however words I want to eliminate, as there is a Clairvoyant in the later chapters who influences all the Alexey’s decisions around how to deal with his wife’s infidelity and her request to divorce so she can marry Vronsky.

I did choose to leave in the Porter stemmer, and while it has some undesired affects on character names like Gatsby becoming “gastbi” and Vronsky become “vronski” it does help provide some more meaning bigrams later.

# The Bigrams

Moving on to bigrams, I needed to call in the bigram measures function and reprocess the initial tokenization using these tools. A bigram is another type of Token when two words that come congruently in the text show up next to each other.

The functions in the package come with filters that allow me to apply the stop word logic and filtering of the character-based tokens. The function also provides some additional features to keep track of Tokens that are next to each other, where doing this using the unigram list, I would get words that ended up next to each other due merely to the fact that stop words or other characters got removed, the package will discount these and drop them from the list.

At this point for Anna ran the bigrams that are most frequent and here the top 50:

(('alexey', 'alexandrovitch'), 0.0013245125445280205) (('stepan', 'arkadyevitch'), 0.0012687435952847353) (('sergey', 'ivanovitch'), 0.0006738748033563612) (('darya', 'alexandrovna'), 0.0004740360685679231) (('lidia', 'ivanovna'), 0.0002463128591578424) (('old', 'man'), 0.00019751502856996795) (('look', 'round'), 0.00019286761613302754) (('go', 'away'), 0.0001858964974776169) (('countess', 'lidia'), 0.00017892537882220627) (('agafea', 'mihalovna'), 0.0001673068477298552) (('one', 'thing'), 0.00013709866688974245) (('anna', 'arkadyevna'), 0.0001254801357973914) (('great', 'deal'), 0.00011618531092351056) (('come', 'back'), 0.00011386160470504035) (('let', 'us'), 0.00010921419226809993) (('konstantin', 'levin'), 0.00010689048604962972) (('madam', 'stahl'), 0.00010689048604962972) (('first', 'time'), 0.0001022430736126893) (('levin', 'went'), 9.991936739421908e-05) (('sick', 'man'), 9.527195495727866e-05) (('one', 'must'), 8.830083630186803e-05) (('old', 'princ'), 8.36534238649276e-05) (('one', 'anoth'), 8.36534238649276e-05) (('young', 'man'), 8.132971764645739e-05) (('levin', 'could'), 7.900601142798719e-05) (('levin', 'felt'), 7.668230520951697e-05) (('madam', 'karenina'), 7.668230520951697e-05) (('marya', 'nikolaevna'), 7.668230520951697e-05) (('next', 'day'), 7.668230520951697e-05) (('went', 'back'), 7.435859899104675e-05) (('long', 'ago'), 7.203489277257655e-05) (('sever', 'time'), 7.203489277257655e-05) (('come', 'along'), 6.971118655410633e-05) (('konstantin', 'dmitrievitch'), 6.971118655410633e-05) (('say', 'someth'), 6.971118655410633e-05) (('answer', 'levin'), 6.738748033563613e-05) (('could', 'never'), 6.738748033563613e-05) (('left', 'alon'), 6.738748033563613e-05) (('go', 'back'), 6.506377411716591e-05) (('levin', 'saw'), 6.506377411716591e-05) (('thought', 'levin'), 6.506377411716591e-05) (('littl', 'girl'), 6.27400678986957e-05) (('one', 'side'), 6.27400678986957e-05) (('thank', 'god'), 6.27400678986957e-05) (('turn', 'away'), 6.27400678986957e-05) (('vassenka', 'veslovski'), 6.0416361680225495e-05) (('everi', 'time'), 5.809265546175528e-05) (('one', 'would'), 5.809265546175528e-05) (('princess', 'varvara'), 5.809265546175528e-05) (('came', 'back'), 5.576894924328507e-05)

The output provides the bigram and the decimalized percentage of the how frequent the two words had shown up in the text. For example, Alexey Alexandrov itch the husband of the main character is the most common bigram after applying the filters shown before. He however only shows 0.13% of the time. The birgrams float many of the main characters to the top of the list with both first and last name. We also see some of the social titles, especially with female characters such as Dutchess, Princess, and Madame

By contrast we see that in Gatsby the top Bigrams point most to places before the characters. The setting of the story in New York and West Egg. The other thing I noticed between the two books is the social setting of Anna is more of a way to put people away with bigrams like go away long before come back. We also see that society based titles do not appear in the top bigrams, but social constructs like getting married do make an emergence despite major characters in both books looking for a marital commitment. While I still have not seen a train theme show up in Anna’s text, the iconic New York taxi that one of the major characters fixes show up in the bigram “yellow car”. Also showing that the change in technology of the times.

(('old', 'sport'), 0.0005281661082410418) (('miss', 'baker'), 0.00044564015382837904) (('new', 'york'), 0.00044564015382837904) (('west', 'egg'), 0.0003796193902982488) (('mr.', 'gatsbi'), 0.0003301038176506511) (('mrs.', 'wilson'), 0.0003135986267681186) (('tom', 'buchanan'), 0.000297093435885586) (('mr.', 'wolfshiem'), 0.0002475778632379883) (('five', 'year'), 0.0001980622905903907) (('long', 'island'), 0.0001980622905903907) (('look', 'around'), 0.0001980622905903907) (('old', 'sport.'), 0.0001980622905903907) (('came', 'back'), 0.00018155709970785812) (('jordan', 'baker'), 0.00018155709970785812) (('jay', 'gatsbi'), 0.00016505190882532556) (('come', 'back'), 0.00013204152706026046) (('demand', 'tom'), 0.00013204152706026046) (('front', 'door'), 0.00013204152706026046) (('mr.', 'mckee'), 0.00013204152706026046) (('mr.', 'sloan'), 0.00013204152706026046) (('never', 'love'), 0.00013204152706026046) (('never', 'seen'), 0.00013204152706026046) (('shook', 'hand'), 0.00013204152706026046) (('turn', 'around'), 0.00013204152706026046) (('dan', 'codi'), 0.0001155363361777279) (('long', 'time'), 0.0001155363361777279) (('look', 'like'), 0.0001155363361777279) (('mrs.', 'mckee'), 0.0001155363361777279) (('turn', 'away'), 0.0001155363361777279) (('yellow', 'car'), 0.0001155363361777279) (('cri', 'daisi'), 9.903114529519534e-05) (('east', 'egg'), 9.903114529519534e-05) (('first', 'time'), 9.903114529519534e-05) (('good', 'night'), 9.903114529519534e-05) (('look', 'back'), 9.903114529519534e-05) (('man', 'name'), 9.903114529519534e-05) (('myrtl', 'wilson'), 9.903114529519534e-05) (('next', 'day'), 9.903114529519534e-05) (('young', 'man'), 9.903114529519534e-05) (('young', 'men'), 9.903114529519534e-05) (('doctor', 't.'), 8.252595441266278e-05) (('egg', 'villag'), 8.252595441266278e-05) (('far', 'away'), 8.252595441266278e-05) (('get', 'marri'), 8.252595441266278e-05) (('go', 'home'), 8.252595441266278e-05) (('green', 'light'), 8.252595441266278e-05) (('j.', 'eckleburg'), 8.252595441266278e-05) (('littl', 'girl'), 8.252595441266278e-05) (('littl', 'later'), 8.252595441266278e-05) (('meyer', 'wolfshiem'), 8.252595441266278e-05)

While Bigram frequencies may show some interesting things, we also have an additional way of comparing bigrams, using mutual information, or a score of how likely the words are to be found together in a text.

For Anna we see the following:

(('mashkin', 'upland'), 15.907785964043683) (('grand', 'duchess'), 15.545215884658976) (('bridal', 'pair'), 14.867143979546338) (('swollen', 'vein'), 14.867143979546338) (('nativ', 'tribe'), 14.808250290492769) (('liza', 'merkalova'), 14.255709267463988) (('alexand', 'nevski'), 14.13017838538013) (('lizaveta', 'petrovna'), 14.068777840715986) (('mademoisel', 'linon'), 13.878639618384167) (('mihail', 'vassilievitch'), 13.73786096260137) (('vassili', 'lukitch'), 13.585857869156323) (('miss', 'hool'), 13.42973866723904) (('scotch', 'cap'), 13.42973866723904) (('fur', 'cloak'), 13.168246426213651) (('happiest', 'frame'), 13.009162984418765) (('polit', 'economi'), 12.97817529193508) (('counting-hous', 'clerk'), 12.889170285876336) (('storm', 'cloud'), 12.730247778491494) (('marya', 'borissovna'), 12.692773073072832) (('marya', 'yevgenyevna'), 12.692773073072832) (('marya', 'nikolaevna'), 12.69277307307283) (('lime', 'tree'), 12.575589533702493) (('agafea', 'mihalovna'), 12.505687520472337) (('seltzer', 'water'), 12.4672133726577) (('summer', 'villa'), 12.457753043408637) (('thrash', 'machin'), 12.313042442529941) (('marya', 'philimonovna'), 12.207346245902592) (('highest', 'degre'), 12.175982074993255) (('sleepless', 'night'), 12.009162984418765) (('game', 'bag'), 11.971749022775647) (('flush', 'hotli'), 11.936406641983451) (('gray', 'whisker'), 11.933781172576628) (('lidia', 'ivanovna'), 11.920234640738682) (('district', 'council'), 11.813067306790545) (('madam', 'lvova'), 11.796277648826694) (('madam', 'stahl'), 11.765250753206066) (('scarc', 'percept'), 11.760944575714412) (('birch', 'tree'), 11.686620846091238) (('madam', 'karenina'), 11.555269549322897) (('utterli', 'unlik'), 11.551605559948074) (('railway', 'station'), 11.461293401113885) (('malign', 'gentleman'), 11.449916549176669) (('countess', 'nordston'), 11.357588881483204) (('countess', 'vronskaya'), 11.357588881483203) (('fifti', 'roubl'), 11.343582023489324) (('pyotr', 'dmitrievitch'), 11.324050766706527) (('villag', 'elder'), 11.308176687358124) (('fine', 'weather'), 11.305749949963586) (('hardli', 'percept'), 11.305749949963584) (('deep', 'breath'), 11.26392977426896)

The top 50 all show items that commonly are found together. We see many society titles that are are sign of Anna’s period like “Grand Dutchess” “Village Elder”, “Countess” all show up. We see less of the characters in this and more of the events that have driven the story along. Railway Station shows up after all of these searches. Marriage does make it into the view of this version with bridal pair showing up as the third highest item for mutual information.

The Gatsby in contrast shows the following mutual information scores:

(('j.', 'eckleburg'), 12.816331323120338) (('t.', 'j.'), 12.62368624517794) (('dan', 'codi'), 11.99363585492825) (('doctor', 't.'), 11.886720651011736) (('meyer', 'wolfshiem'), 10.54331282871392) (('west', 'egg'), 10.278425646459954) (('egg', 'villag'), 9.994329625098331) (('miss', 'baker'), 9.963888511534195) (('new', 'york'), 9.725728774339432) (('mrs.', 'mckee'), 9.716795649569423) (('old', 'sport.'), 9.54687064812711) (('long', 'island'), 9.347561839903703) (('old', 'sport'), 9.298943134683526) (('mr.', 'sloan'), 9.279390337262125) (('mrs.', 'wilson'), 9.060506701702815) (('east', 'egg'), 9.034971609595676) (('mr.', 'carraway'), 8.753321525594536) (('five', 'year'), 8.74376269716969) (('mr.', 'wolfshiem'), 8.650228032630435) (('green', 'light'), 8.54687064812711) (('mr.', 'mckee'), 8.431393430707175) (('shook', 'hand'), 8.112757282578178) (('tom', 'buchanan'), 8.042625182139114) (('far', 'away'), 7.994329625098331) (('never', 'seen'), 7.955244417124943) (('yellow', 'car'), 7.941694926516448) (('jordan', 'baker'), 7.911002931282102) (('front', 'door'), 7.84506149937452) (('jay', 'gatsbi'), 7.777673573311029) (('myrtl', 'wilson'), 7.681334654980976) (('young', 'men'), 7.619934110316832) (('next', 'day'), 7.459757085426082) (('good', 'night'), 7.456268099346202) (('never', 'love'), 7.102085805454216) (('get', 'marri'), 7.025633745016341) (('littl', 'later'), 6.937185717994723) (('demand', 'tom'), 6.524776877276494) (('go', 'home'), 6.466760473163847) (('turn', 'around'), 6.406940386982637) (('young', 'man'), 6.3975416889803824) (('turn', 'away'), 6.257364030932127) (('long', 'time'), 6.216317306625454) (('first', 'time'), 6.166761482151218) (('two', 'day'), 6.156080695094943) (('man', 'name'), 6.0756135940930225) (('mr.', 'gatsbi'), 5.951702973086077) (('look', 'around'), 5.949592082845825) (('came', 'back'), 5.823080442708637) (('cri', 'daisi'), 5.6705810945866) (('come', 'back'), 5.503826482119601

Settings seem to drive more of the mutual information than society in The Great Batsby. We see places like East Egg, West Egg, Long Island, and New York all rising to the top of the list. Marriage which seemed like a larger theme in the frequency bigrams shows mutual information of 7.03 which starting verge on unimportant as a score 5 tells us that the two words just happen to be found frequently together. And wee see a few of at the bottom of the top 50.

# Trigrams

I did look briefly at the Trigrams mutal information scores to see if they were anymore telling.

For Anna we do see towards the top of the list titles with both Princess and Countess in 2nd and 3rd place with quite high mutual information scores announcing their importance to the world of the story:

(('strain', 'everi', 'nerv'), 23.309266371241208) (('princess', 'marya', 'borissovna'), 22.841225629921855) (('countess', 'lidia', 'ivanovna'), 22.759356433287632) (('answer', 'darya', 'alexandrovna'), 15.551839689518808) (('let', 'us', 'talk'), 15.258293843062347) (('address', 'alexey', 'alexandrovitch'), 14.94292354654046) (('let', 'us', 'go'), 14.742112815677437) (('littl', 'old', 'man'), 14.05300415376077) (('sergey', 'ivanovitch', 'smile'), 13.993092321149579) (('stepan', 'arkadyevitch', 'took'), 13.908925830940685) (('alexey', 'alexandrovitch', 'took'), 13.873013042098822) (('stepan', 'arkadyevitch', 'saw'), 13.515753477742297) (('alexey', 'alexandrovitch', 'got'), 13.449767790943145) (('alexey', 'alexandrovitch', 'saw'), 13.427373269006306) (('ask', 'stepan', 'arkadyevitch'), 13.306115328886598) (('alexey', 'alexandrovitch', 'went'), 13.071656860662664) (('stepan', 'arkadyevitch', 'went'), 12.870530452203667) (('answer', 'stepan', 'arkadyevitch'), 12.807236702247756) (('stepan', 'arkadyevitch', 'smile'), 12.742711393524544) (('stepan', 'arkadyevitch', 'ask'), 12.628043423773967) (('stepan', 'arkadyevitch', 'felt'), 12.447645091102146) (('alexey', 'alexandrovitch', 'smile'), 12.262013762009786) (('thought', 'alexey', 'alexandrovitch'), 12.201566183257093) (('could', 'say', 'noth'), 12.02244647857658) (('stepan', 'arkadyevitch', 'could'), 11.636939824262363) (('alexey', 'alexandrovitch', 'would'), 11.535603960422453) (('levin', 'went', 'back'), 11.239403500062323) (('levin', 'look', 'round'), 11.133630545964238)

On the other hand in the Great Gatsby we see only three Trigrams with high mutual information scores:

(('t.', 'j.', 'eckleburg'), 25.703051974132073) (('doctor', 't.', 'j.'), 25.188478801302313) (('west', 'egg', 'villag'), 20.793587434859727)

Doctor T.J. Ecklesburg which was a major character who helped Gatsby secure his liquor in the book, and likely should be really view and a quartogram rather than 2 separate trigrams and setting of West Egg Village beign the third, though it does have quite a high information score, reimporting the idea that in Gatsby’s world that location is often more important that society.

# Conclusion

Looking at both the Great Gatsby and Anna Karenina through the starter lens of NLP we can see the variation in authors books published about 50 years apart. Leo Tolstoy writing Anna Karenina was heavily focused on showing society through the official titles of his characters and their actions shown through high unusual level of verb usage in his work. On the other hand, Gatsby coming 50 years later and set in New York, shows America’s fascination with location over any title. While both books deal with affairs of their main characters and changes in marriage both seem to not use the words but these societal factors that are different for their times in order to convey the story. Although I feel this only cursory exploratory analysis, I hope to continue using these books to explore more of the NLP study and see how these themes unfold in future assignments.