

Quinn E Knudsen

NLP Homework 1: Corpus Statistics and Mutual Information 

**Section 1: Choosing the Data**

1. **Dataset Selection**

For this homework assignment, two conceptually distinct corpora were analyzed and compared. The first dataset is a combination of the transcripts from three of Donald Trump’s 2020 campaign speeches delivered in Orlando, Tulsa, and Phoenix. These transcripts were extrapolated and combined from a [Kaggle](https://www.kaggle.com/rishabh6377/trump-2020-election-speech) posting. The second dataset was scraped from an online archive that contains the [The Dark Knight](https://archive.org/stream/TheDarkKnightScriptByJonathanNolanAndChristopherNolan/The%20Dark%20Knight%20Script%20by%20Jonathan%20Nolan%20And%20Christopher%20Nolan_djvu.txt) movie script. There are differences between the two documents in that the Trump speeches were written to convey a campaign message and the Dark Knight was written to direct a movie. Conceptually, both were written to inspire or captivate large crowds and were written in the same era. A limitation to note with each that was discovered when looking through the bigrams was that the Trump speech transcripts contained commentary on audience reactions which was excluded and the Dark Knight transcript contained stage directions which impacted the top word frequencies that will subsequently be discussed. Removing the stage directions was not easily corrected for because stop word removal of key characters would have too much of a muddying impact on the overall data quality. For example, a line would state “Wayne Penthouse: Batman enters while the Joker holds Rachel menacingly”. Removing these lines and not the actual text would have been an extremely complicated process, so it was determined that the stage directions would be included for this analysis. Finally, a third data set, [Biden DNC Acceptance Speech](https://www.kaggle.com/christianlillelund/joe-biden-2020-dnc-speech), was included out of intrigue and also to use to compare to the original Trump speeches for content and word frequency. Although the assignment asked for text that is different in some aspect, the two candidates have highly different speech patterns, stances and core political platform agendas making it a timely and interesting additional document for comparison.

**Section 2: Examining the Text**

1. **Preprocessing Techniques**

For all three datasets, the same preprocessing steps were utilized with minor customizations for the stop words. The format that was used began by tokenizing the data and searching for the word “China” out of curiosity and to ensure the tokenization had worked well. China was referenced by Trump the most (even adjusting for proportionality) amongst the three documents. Using the tokens, a random word was chosen to compare the Porter Stemmer, the Lancaster Stemmer and the Stanford Lemmatizer. Ultimately, based largely upon the guidance of the async, the Stanford Lemmatizer was chosen for its ability to handle multiple cases and correctly piece together variations of words. The lemmatized text was then converted to lowercase to make sure alike words were totaled together regardless of whether they began a sentence or were written as a proper noun. Next, an alpha filter was created to remove all non-alphanumeric words. The NLTK sop words were downloaded and used as the basis for all common word removal for each dataset. In addition, 'could','would','might','must','need','sha','wo','y',"'s","'d","'ll","'t","'m","'re","'ve", "n't",'wa','gon','na' were removed from each document. From there, unique customizations were made. This included removing the audience reactions from the Trump transcripts and removing some stage shorthand for the Dark Knight script. This left a relatively clean and meaningful list of words to explore frequencies in unigrams, bigrams, and trigrams.

1. **Bigram Optimization**

Many of the problems found when analyzing and interpreting the bigrams lead to tweaks in the preprocessing of the data. A great example of that can be articulated by the custom stop words used in the Trump data preprocessing. In addition to the words mentioned in section two, the Trump speech transcripts also had a custom list that included the following: 'audience','boos','boo','chant','applause'

It was not initially known that the transcripts included recordings of audience reactions, but it became clear when looking at the bigrams that a bigram such as “audience boos” or “audience chant” were not meaningful to the content of the speech. They may, however, shed light on the reception and in certain instances would be meaningful to keep. The intention here was to look at the topics being covered, therefor it was decided that these bigrams should be removed.

Another problem with the Dark Knight script was the stagehand commentary. Certain stage directions were removed, but for a proper analysis of just the lines in the movie would require a deep level of customization to remove ONLY the stage directions and not the actual movie lines as exemplified in the section one example. Removing these words would provide a better list of bigrams based solely on the movie lines.

This process was iterative as an ideal list of meaningful words was sought to analyze the most relevant list of frequent bigrams.

1. **Bigram vs. Mutual Information Comparison**

*General Summary:* First utilizing the bigram frequency and mutual information for the Trump transcripts, it is clear to see distinct differences emerge. The highest mutual information scored bigram ‘matt gaetz’ did not crack the top 5 in frequency. Others, such as ‘mark meadows’, ‘supreme court’ and ‘second amendment’ fell in the latter half of the top 50 frequent but were in the top 5 mutual information. By the same token, pun unintended, ‘stock market’ and ‘november 3rd’ were both highly ranked in frequency and mutual information. This nonpattern demonstrates that frequency and mutual information are distinct metrics and provide differentiated outcomes. Similar outcomes were apparent for the Dark Knight and Joe Biden Acceptance Speech bigrams and trigrams when compared to the mutual information score. A full comparison of the top bi/trigrams and top mutual information scores can be found in the appendix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Candidate Focus by Bigram and Trigram Frequency and Mutual Information** | | | | |
| **Trump** | | **Dark Knight** | | |
| **Top Bigram Frequency** | **Top Bigram Mutual Information** | **Top Bigram Frequency** | **Top Bigram Mutual Information** | |
| 1. America Great | 1. Matt Gaetz | 1. Gotham Central | | 1. Major Crimes |
| 1. United States | 1. Mark Meadows | 1. Prewitt Building | | 1. Basement Apartment |
| 1. Joe Biden | 1. Supreme Court | 1. Harvey Dent | | 1. Hong Kong |
| 1. Great Job | 1. Stock Market | 1. Armored Car | | 1. Burnt Warehouse |
| 1. Make America | 1. Second Amendment | 1. Wayne Penthouse | | 1. Insert Cut |

*Deep Dive & Concept Definitions:* Bigram frequencies, simply put, are the counts of sequentially collocated words. Proper data preprocessing is often important here to remove nonalphanumeric noise and stop words that provide little meaningful information. It is important to note that this is a general rule of thumb rather than a hard-fast rule. When scoring the frequency ratio, the calculation considers the ratio of the count of the bigram over the count of the total bigrams.

Mutual Information score is can be likened to an Association Ratio as described in the Church and Hanks article. Although mutual information typically does not need to follow a directional pattern, the NLTK mutual information score follows English reading patterns. This package scores occurrences that flow directionally left to right. Noteworthy here is that the document size needs to be rather larger for meaningful information retrieval. Best practice recommends applying the recommended minimum frequency of 5 parameter. This recommendation was followed in this analysis.

The PMI scores tend to score highly on bigrams (and trigrams) with unique words as seen in the results of the Trump transcripts which were paced by peoples’ names and other less frequent unigrams such as supreme court and stock market. Normal speech cadence divests from using these words in context that are not mutually interdependent and therefore lend themselves to a naturally higher pmi score.

1. **Stop Word Customization**

Detailed in prior sections, please use the code below as an example of the stop word customization that was utilized:

nltk.download('stopwords')

nltkstopwords = nltk.corpus.stopwords.words('english')

morestopwords = ['could','would','might','must','need','sha','wo','y',"'s","'d","'ll","'t","'m","'re","'ve", "n't",'wa','gon','na','boos','boo','chant','audience','applause']

stopwords = nltkstopwords + morestopwords

stoppedtrumpwords = [w for w in alphatrumpwords if not w in stopwords]

The rationale used for stop word removal beyond the standard English dictionary was to first remove words that added little information above and beyond what would have been inferred from bigram frequency and pmi. Words like ‘could’, ‘would’, ‘must’, although important for conveying conviction in spoken word, did not add value to the top content that was being spoken about. In addition, fragmented words were removed such as ‘sha’, ‘wo’, ‘gon’ ‘na’ that would reduce the interpretability of the output. Finally, an additional string of custom words relevant to each document was carefully chosen based upon initial bigram frequencies. For example, within the Trump transcript, inclusion of audience feedback such as ‘boo’, ‘chant’ ‘applause’ was undesirable. This might add value if the reception sentiment was explored, but for the purposes of this analysis, they were not included.

1. **Trigram Analysis**

For each dataset, trigrams were also explored in hopes of gleaning new information and telling a richer data story than the bigrams alone could provide. This proved to be a fruitful endeavor as certain trigram phrases such as 'black', 'lives', 'matter', 'keep', 'america', 'great','beautiful', 'border', 'wall', and 'good', 'third', 'quarter' made themselves apparent as key topic areas that were less obvious in the original bigram analysis. For the Trump speeches, it also demonstrated just how important the pointwise mutual information is by the cooccurrences of 'black', 'lives', 'matter' as a true standout topic area with a score of 21.25. This indicates that, the occurrence of these words were extremely high although perhaps not uniquely talked about with as great of frequency as other tokens.

Trigrams added additional information for the Dark Knight with (('dogs', 'start', 'barking'), 5.227801448101001e-05), ('applied', 'sciences', 'division'), 5.227801448101001e-05, and (('find', 'harvey', 'dent'), 5.227801448101001e-05) standing out. The dogs barking stood out as an indicator of how important setting the stage of scene can be in creating an ominous or dangerous mood. The dogs were used many times in this movie for intimidation and dramatic effect. The trigrams of ‘find harvey dent’ and ‘applied sciences division’ were also interesting because they were used both as stage direction, but also even more frequently as lines in the movie.

The trigrams for the Biden speech were also incredibly informative and included some powerful language such as (('deep', 'black', 'hole'), 0.00025348542458808617), (('existential', 'threat', 'posed'), 0.00025348542458808617), (('great', 'middle', 'class'), 0.00025348542458808617), and (('honest', 'unvarnished', 'truth'), 0.00025348542458808617). While the trigrams indicated some though provoking language, the additional word added a certain level of linguistic ferocity that it is hard to capture when analyzing bigrams.

**Section 3: Answering the Defined Problems**

1. **Document Comparison**

**Comparison Questions**

1. How do the 3 documents prioritize discussing police?
2. Comparing Biden and Trump campaign speeches, how do they differ in frequency of mentioning the opposing political party?
3. What trends emerge when holistically looking at the gram frequencies? How does ‘setting the scene’ play a role?
4. Based upon the analyzed data, where is Trump focusing his campaign policy attention and how does that differ from Biden?

**Comparison Question 1**

Conversation around the police force has been a paramount political topic in 2020. Likewise, in the Dark Knight, police played a key role in the plot in both a positive a negative light. Although the intentions of this analysis were not to explore sentiment, the frequency and associated words used to describe police will provide context into the messages the politicians and screenwriters were trying to convey.

Based upon a holistic analysis of the top unigram, bigram and trigrams (including mutual information scores), it is clear to see the police play the largest role in the Dark Knight, followed by Trump’s attention in his campaign speeches. Biden, on the other hand, paid little verbiage towards police topics, with the closet top 50 bigram referring to an Iraqi war veteran. Frequency metrics such as the ones presented in this analysis can provide an index into the level of prioritization a candidate may be placing in a topic area (or at the very least their willingness to bring it to the forefront of campaigning). On the contrary, Donald Trump seems to be placing law enforcement as a key topic area as it was his 14th highest bigram and highly rated on mutual information. Perhaps even or revealing, a criminal justice reform represented 2 of the top 20 trigrams along with law abiding citizen. Through these leading indicators, it is apparent that Trump has a strong stance on criminal justice and law-abiding citizens and is using that a major part of his campaign platform.

Transitioning from nonfictional policy planning, to fictional action, the Dark Knight had a clear focus on using police synonymous language. By frequency, top bigrams included ‘Gotham Central’ (1; the police station), ‘Harvey Dent’ (3; a police detective), ‘armored car’ (4), and various other police characters or words in the top 50 (Officer Gordon, police van, swat team, major crimes, national guard, holding area, interrogation room, crimes unit, and observation room). Similar occurrences were evident with trigrams and a full list in descending frequency order can be found in the appendix under, ‘The Dark Knight’.

**Comparison Question 2**

In 2016, a year in which Donald Trump shocked political polling expectations and defeated Hillary Clinton, feedback was given that the democratic party was overfocused on name bashing Donald Trump and over utilized his name during campaigning and televised debates. It appears, through this analysis, that the feedback has manifested in an altered approach for 2020. In fact, in Joe Biden’s acceptance speech, not once in the top 50 n grams does, he refer to Trump by name. Instead, he uses ‘president’ or ‘current president’, but does, however, call out Obama by name as ‘President Obama’. Biden does not even refer to the republican party with any level of frequency nor other high-ranking party members. This is a highly intriguing finding both in light of the 2016 outcomes, but it is also in stark contrast to Donald Trump’s 2020 campaign approach.

In Trump’s 2020 campaign speeches, Joe Biden is the 3rd most frequent bigram. This is followed by ‘radical left’ (6), ‘sleepy Joe’ (23), ‘President Obama’ (30), and ‘Hillary Clinton’ (43). This demonstrates that Trump is much more willing to directly refer to his opposition by name (and does so with more frequency) than Biden.

**Comparison Question 3**

Linguistics provide an incredible medium for one to paint the picture of a scene without the listener having to have had experienced it firsthand. In the context of writing a movie script, this is essential to building out the perfect scenery, and in politics this plays a crucial role in influencing prospective voters. Voters are often inspired by a vision and unless politicians are able to articulate a clear and rousing message, voters can become disengaged with the delivered content.

Scene setting in the Dark Knight followed a bit more of a literal manifestation of the term. Simple trigram usage such as ‘dogs start barking’ can paint a very ominous picture without evening needing to know any additional context about the shot.

In the Biden and Trump speeches, it is extremely apparent that the two are trying to paint very different outlooks on the current state of affairs in the United States. While Trump does have some negatively oriented bigrams such as ‘fake news’ and the virus being a ‘terrible thing’ his general focus is on creating positivity for both the current state and the coming months when describing the ‘beautiful border wall’ and a ‘good third quarter’. Variations of ‘America great’ are prominent in his speeches and ‘keep America great’ is one of his most prevalent trigrams.

Biden paints a much bleaker vision for America in its current state in effort to showcase needed change in leadership. Frequently used bigrams by Biden included ‘protecting America’, ‘accelerating threat’, anti-Semitic bile’. Top trigrams included ‘American darkness began’, ‘build back better’, ‘deep black hole’, ‘difficult moment America’, ‘economic crisis since’, ‘existential threat posed’, ‘family back together’ and ‘four historic crisis’.

By exploring top gram frequencies, it is possible to tease out trends and underlying meaning that is being conveyed. Paired with the context that Biden is attempting to unseat an active president, it makes sense that he would try to paint the image of a bleak current state that is a manifestation of the current administration. At the same time, Trump believes what the media is portraying is an overplayed version of the truth and that things are progressing much faster than others would let on to believe. In real-time, some of these nuances may more readily be perceived than others. By stepping back and looking for linguistic trends it can be much easier to decipher the ‘message within the message’.

**Comparison Question 4**

Frequency, used as a proxy for campaign focus reveal the following prioritization for the two presidential candidates:

|  |  |
| --- | --- |
| **Candidate Focus based upon Bigram and Trigram Frequency** | |
| **Trump** | **Biden** |
| **Law Enforcement** | **Climate Change** |
| **Economy/Stock Market** | **Social Security** |
| **Social/Criminal Justice** | **Health Care** |
| **Boarders** | **Social/Criminal Justice** |
| **Free Speech** | **Economy** |
| **Sanctuary Cities** | **Middle Class Americans** |
| **Trade** |  |
| **Mall-in Ballots** |  |
| **Second Amendment** |  |
| **Health Care** |  |
| **Tax Cuts** |  |
| **Supreme Court Nominations** |  |

1. **Appendix**

**Trump Speech Analysis**

**Trump Unigrams with top 50 frequencies and normalized value**

1. people 199 0.0011
2. great 179 0.001
3. going 177 0.001
4. know 174 0.001
5. said 145 0.0008
6. one 140 0.0008
7. want 138 0.0008
8. thank 129 0.0007
9. like 119 0.0007
10. country 118 0.0007
11. right 116 0.0007
12. get 108 0.0006
13. say 94 0.0005
14. year 93 0.0005
15. job 85 0.0005
16. think 83 0.0005
17. american 81 0.0005
18. lot 76 0.0004
19. never 75 0.0004
20. president 74 0.0004
21. america 74 0.0004
22. time 71 0.0004
23. got 70 0.0004
24. ever 67 0.0004
25. good 63 0.0004
26. way 63 0.0004
27. go 61 0.0003
28. u 61 0.0003
29. many 60 0.0003
30. thing 59 0.0003
31. ha 56 0.0003
32. see 54 0.0003
33. much 54 0.0003
34. let 52 0.0003
35. back 51 0.0003
36. two 51 0.0003
37. even 51 0.0003
38. history 49 0.0003
39. done 49 0.0003
40. make 49 0.0003
41. love 47 0.0003
42. million 47 0.0003
43. first 46 0.0003
44. left 46 0.0003
45. big 45 0.0003
46. come 45 0.0003
47. well 45 0.0003
48. day 44 0.0002
49. every 44 0.0002
50. nobody 43 0.0002

**Trump Bigrams with top 50 frequencies**

1. (('america', 'great'), 0.0005904703224481413)
2. (('united', 'states'), 0.0005904703224481413)
3. (('joe', 'biden'), 0.0005391250770178681)
4. (('great', 'job'), 0.0005134524543027316)
5. (('make', 'america'), 0.0004621072088724584)
6. (('radical', 'left'), 0.0004364345861573218)
7. (('fake', 'news'), 0.00041076196344218524)
8. (('long', 'time'), 0.00033374409529677554)
9. (('november', '3rd'), 0.00030807147258163895)
10. (('year', 'ago'), 0.00030807147258163895)
11. (('mr.', 'president'), 0.00028239884986650237)
12. (('big', 'deal'), 0.0002567262271513658)
13. (('keep', 'america'), 0.0002567262271513658)
14. (('law', 'enforcement'), 0.0002567262271513658)
15. (('stock', 'market'), 0.0002567262271513658)
16. (('young', 'people'), 0.0002567262271513658)
17. (('american', 'flag'), 0.0002310536044362292)
18. (('black', 'lives'), 0.0002310536044362292)
19. (('lives', 'matter'), 0.0002310536044362292)
20. (('number', 'one'), 0.0002310536044362292)
21. (('little', 'bit'), 0.00020538098172109262)
22. (('open', 'border'), 0.00020538098172109262)
23. (('sleepy', 'joe'), 0.00020538098172109262)
24. (('air', 'force'), 0.00017970835900595606)
25. (('bad', 'people'), 0.00017970835900595606)
26. (('first', 'time'), 0.00017970835900595606)
27. (('free', 'speech'), 0.00017970835900595606)
28. (('half', 'year'), 0.00017970835900595606)
29. (('member', 'call'), 0.00017970835900595606)
30. (('president', 'obama'), 0.00017970835900595606)
31. (('sanctuary', 'city'), 0.00017970835900595606)
32. (('supreme', 'court'), 0.00017970835900595606)
33. (('terrible', 'thing'), 0.00017970835900595606)
34. (('trade', 'deal'), 0.00017970835900595606)
35. (('week', 'ago'), 0.00017970835900595606)
36. (('american', 'people'), 0.00015403573629081948)
37. (('black', 'community'), 0.00015403573629081948)
38. (('border', 'wall'), 0.00015403573629081948)
39. (('brand', 'new'), 0.00015403573629081948)
40. (('four', 'year'), 0.00015403573629081948)
41. (('get', 'rid'), 0.00015403573629081948)
42. (('great', 'people'), 0.00015403573629081948)
43. (('hillary', 'clinton'), 0.00015403573629081948)
44. (('jim', 'inhofe'), 0.00015403573629081948)
45. (('mail-in', 'ballot'), 0.00015403573629081948)
46. (('mark', 'meadows'), 0.00015403573629081948)
47. (('one', 'case'), 0.00015403573629081948)
48. (('people', 'like'), 0.00015403573629081948)
49. (('said', 'general'), 0.00015403573629081948)
50. (('second', 'amendment'), 0.00015403573629081948)

**Trump Bigrams with top 50 by Mutual Information:**

1. (('matt', 'gaetz'), 12.079484783826818)
2. (('mark', 'meadows'), 12.079484783826814)
3. (('supreme', 'court'), 12.079484783826814)
4. (('stock', 'market'), 11.927481690381766)
5. (('second', 'amendment'), 11.78997816663183)
6. (('november', '3rd'), 11.664447284547972)
7. (('hillary', 'clinton'), 11.567585745295382)
8. (('mail-in', 'ballot'), 11.548970067128035)
9. (('air', 'force'), 11.301877205163262)
10. (('health', 'insurance'), 11.24940978526913)
11. (('jim', 'inhofe'), 11.161946944018789)
12. (('lives', 'matter'), 11.001482271825543)
13. (('west', 'point'), 10.789978166631832)
14. (('united', 'states'), 10.603046739883828)
15. (('fake', 'news'), 10.542050653188246)
16. (('law', 'enforcement'), 10.526943760798037)
17. (('little', 'bit'), 10.442054863211526)
18. (('black', 'lives'), 10.24940978526913)
19. (('health', 'care'), 9.927481690381768)
20. (('sanctuary', 'city'), 9.88683970588442)
21. (('brand', 'new'), 9.779089850489093)
22. (('sleepy', 'joe'), 9.580524801002882)
23. (('trade', 'deal'), 9.311930869827187)
24. (('tax', 'cut'), 9.31081032993327)
25. (('member', 'call'), 9.12012676832416)
26. (('mr.', 'president'), 9.039956419640179)
27. (('vice', 'president'), 9.039956419640179)
28. (('joe', 'biden'), 9.006009087716839)
29. (('radical', 'left'), 8.955329675334884)
30. (('free', 'speech'), 8.940420746089263)
31. (('open', 'border'), 8.867866834084545)
32. (('hundred', 'percent'), 8.605553595494404)
33. (('young', 'woman'), 8.548970067128037)
34. (('black', 'community'), 8.512444191102922)
35. (('american', 'flag'), 8.379045065685725)
36. (('president', 'obama'), 8.262348840976625)
37. (('long', 'time'), 8.215139883184383)
38. (('terrible', 'thing'), 8.174121657964891)
39. (('border', 'wall'), 8.118410295735139)
40. (('week', 'ago'), 8.108397475742054)
41. (('even', 'close'), 8.091557616127393)
42. (('seven', 'year'), 8.03217906904846)
43. (('big', 'deal'), 7.794082564964569)
44. (('get', 'rid'), 7.620053165189518)
45. (('make', 'america'), 7.595171576967283)
46. (('year', 'ago'), 7.59477375674116)
47. (('beautiful', 'border'), 7.5434318835866065)
48. (('half', 'year'), 7.517605896218701)
49. (('black', 'life'), 7.4014128787141775)
50. (('four', 'year'), 7.207750633631912)

**Trigrams from file with top 20 frequencies**

1. (('make', 'america', 'great'), 0.00028239884986650237)
2. (('black', 'lives', 'matter'), 0.0002310536044362292)
3. (('keep', 'america', 'great'), 0.0002310536044362292)
4. (('beautiful', 'border', 'wall'), 0.0001283631135756829)
5. (('air', 'force', 'one'), 0.00010269049086054631)
6. (('criminal', 'justice', 'reform'), 0.00010269049086054631)
7. (('never', 'seen', 'anything'), 0.00010269049086054631)
8. (('seen', 'anything', 'like'), 0.00010269049086054631)
9. (('two', 'great', 'senators'), 0.00010269049086054631)
10. (('good', 'third', 'quarter'), 7.701786814540974e-05)
11. (('great', 'american', 'flag'), 7.701786814540974e-05)
12. (('groundbreaking', 'criminal', 'justice'), 7.701786814540974e-05)
13. (('joe', 'biden', 'ha'), 7.701786814540974e-05)
14. (('law', 'abiding', 'citizen'), 7.701786814540974e-05)
15. (('lives', 'matter', 'movement'), 7.701786814540974e-05)
16. (('number', 'one', 'show'), 7.701786814540974e-05)
17. (('president', 'mike', 'pence'), 7.701786814540974e-05)
18. (('seven', 'year', 'ago'), 7.701786814540974e-05)
19. (('support', 'sanctuary', 'city'), 7.701786814540974e-05)
20. (('supreme', 'court', 'justice'), 7.701786814540974e-05)

**Trump Trigrams from file with top 4 by Mutual Information:**

1. (('black', 'lives', 'matter'), 21.250892057094667)
2. (('beautiful', 'border', 'wall'), 17.934860673728156)
3. (('make', 'america', 'great'), 14.650272202167141)
4. (('keep', 'america', 'great'), 14.5831580063086)

**The Dark Knight Script Analysis**

**Dark Knight Unigrams with top 50 frequencies and normalized value**

1. dent 412 0.0023
2. joker 330 0.0018
3. batman 324 0.0018
4. gordon 317 0.0017
5. wayne 255 0.0014
6. look 183 0.001
7. rachel 173 0.0009
8. night 129 0.0007
9. alfred 110 0.0006
10. fox 110 0.0006
11. gotham 107 0.0006
12. one 102 0.0006
13. get 88 0.0005
14. day 83 0.0005
15. know 79 0.0004
16. back 77 0.0004
17. harvey 72 0.0004
18. lau 72 0.0004
19. like 69 0.0004
20. car 66 0.0004
21. turn 64 0.0004
22. pull 64 0.0004
23. want 62 0.0003
24. building 59 0.0003
25. room 59 0.0003
26. take 58 0.0003
27. going 55 0.0003
28. man 54 0.0003
29. bank 53 0.0003
30. street 51 0.0003
31. men 51 0.0003
32. maroni 51 0.0003
33. wa 50 0.0003
34. ramirez 48 0.0003
35. chechen 48 0.0003
36. ferry 48 0.0003
37. go 47 0.0003
38. got 47 0.0003
39. head 47 0.0003
40. people 46 0.0003
41. let 44 0.0002
42. door 44 0.0002
43. cop 44 0.0002
44. hand 43 0.0002
45. gun 42 0.0002
46. phone 42 0.0002
47. penthouse 41 0.0002
48. around 40 0.0002
49. two 39 0.0002
50. reese 39 0.0002

**Dark Knight Bigrams from file with top 50 frequencies**

1. (('gotham', 'central'), 0.0007841702172151502)
2. (('prewitt', 'building'), 0.0007580312099746451)
3. (('harvey', 'dent'), 0.0006534751810126252)
4. (('armored', 'car'), 0.0006011971665316151)
5. (('wayne', 'penthouse'), 0.0004966411375695951)
6. (('dent', 'look'), 0.0004443631230885851)
7. (('hong', 'kong'), 0.00041822411584808007)
8. (('bank', 'manager'), 0.0003920851086075751)
9. (('gordon', 'look'), 0.0003920851086075751)
10. (('passenger', 'lounge'), 0.00033980709412656507)
11. (('police', 'van'), 0.00033980709412656507)
12. (('night', 'batman'), 0.00031366808688606005)
13. (('prisoner', 'ferry'), 0.0002875290796455551)
14. (('shotgun', 'swat'), 0.0002875290796455551)
15. (('2nd', 'street'), 0.00026139007240505007)
16. (('cell', 'phone'), 0.00026139007240505007)
17. (('level', 'street'), 0.00026139007240505007)
18. (('look', 'around'), 0.00026139007240505007)
19. (('lower', 'level'), 0.00026139007240505007)
20. (('night', 'gordon'), 0.00026139007240505007)
21. (('fox', 'look'), 0.00023525106516454505)
22. (('gotham', 'streets'), 0.00023525106516454505)
23. (('joker', 'look'), 0.00023525106516454505)
24. (('rooftop', 'overlooking'), 0.00023525106516454505)
25. (('day', 'wayne'), 0.00020911205792404004)
26. (('major', 'crimes'), 0.00020911205792404004)
27. (('mr.', 'wayne'), 0.00020911205792404004)
28. (('night', 'rachel'), 0.00020911205792404004)
29. (('swat', 'team'), 0.00020911205792404004)
30. (('wayne', 'enterprises'), 0.00020911205792404004)
31. (('burnt', 'warehouse'), 0.00018297305068353504)
32. (('commuter', 'ferry'), 0.00018297305068353504)
33. (('crimes', 'unit'), 0.00018297305068353504)
34. (('day', 'fox'), 0.00018297305068353504)
35. (('dent', 'turn'), 0.00018297305068353504)
36. (('insert', 'cut'), 0.00018297305068353504)
37. (('national', 'guard'), 0.00018297305068353504)
38. (('observation', 'room'), 0.00018297305068353504)
39. (('school', 'bus'), 0.00018297305068353504)
40. (('swat', 'leader'), 0.00018297305068353504)
41. (('basement', 'apartment'), 0.00015683404344303003)
42. (('city', 'hall'), 0.00015683404344303003)
43. (('clown', 'mask'), 0.00015683404344303003)
44. (('day', 'gordon'), 0.00015683404344303003)
45. (('good', 'side'), 0.00015683404344303003)
46. (('holding', 'area'), 0.00015683404344303003)
47. (('hospital', 'room'), 0.00015683404344303003)
48. (('interrogation', 'room'), 0.00015683404344303003)
49. (('judge', 'surrillo'), 0.00015683404344303003)
50. (('lower', 'fifth'), 0.00015683404344303003)

**Bigrams from file with top 50 by Mutual Information:**

1. (('major', 'crimes'), 12.223436125820616)
2. (('basement', 'apartment'), 11.831118703041852)
3. (('hong', 'kong'), 11.135973284570275)
4. (('burnt', 'warehouse'), 11.123900452269702)
5. (('insert', 'cut'), 10.97550861237703)
6. (('crimes', 'unit'), 10.943328206627882)
7. (('rooftop', 'overlooking'), 10.720935785291433)
8. (('lower', 'fifth'), 10.579579936045889)
9. (('judge', 'surrillo'), 10.41608120376301)
10. (('holding', 'area'), 10.365455130693041)
11. (('guard', 'commander'), 10.345691875871612)
12. (('passenger', 'lounge'), 10.25853992677653)
13. (('fat', 'thug'), 10.223436125820616)
14. (('school', 'bus'), 10.123900452269702)
15. (('national', 'guard'), 9.938033906958369)
16. (('city', 'hall'), 9.831118703041852)
17. (('lower', 'level'), 9.814045189682913)
18. (('security', 'guard'), 9.693615179291921)
19. (('2nd', 'street'), 9.551010783849119)
20. (('bank', 'manager'), 9.495515671257417)
21. (('commuter', 'ferry'), 9.445828547157065)
22. (('observation', 'room'), 9.340793076458775)
23. (('living', 'room'), 9.340793076458773)
24. (('overlooking', 'prewitt'), 9.316545530212094)
25. (('prewitt', 'building'), 9.291883475977828)
26. (('52nd', 'street'), 9.287976378015326)
27. (('cell', 'phone'), 9.153046797929218)
28. (('swat', 'leader'), 9.12089796410818)
29. (('interrogation', 'room'), 9.118400655122324)
30. (('armored', 'car'), 9.05874777274445)
31. (('prison', 'ferry'), 9.0535111243783)
32. (('police', 'van'), 8.987237904959136)
33. (('prisoner', 'ferry'), 8.84997773029317)
34. (('level', 'street'), 8.785476037486143)
35. (('swat', 'team'), 8.621665337412905)
36. (('clown', 'mask'), 8.60627480271131)
37. (('shotgun', 'swat'), 8.385951537578624)
38. (('gotham', 'central'), 8.344465615669534)
39. (('gotham', 'streets'), 8.066931640140623)
40. (('good', 'side'), 8.04684739409729)
41. (('mr.', 'reese'), 7.6161058120710035)
42. (('hospital', 'room'), 7.402193621122915)
43. (('passenger', 'ferry'), 7.365455130693041)
44. (('wayne', 'enterprises'), 7.229082688961757)
45. (('master', 'wayne'), 6.966048283127963)
46. (('mr.', 'fox'), 6.383182723242387)
47. (('wayne', 'penthouse'), 6.119458197787258)
48. (('look', 'around'), 5.707736287536573)
49. (('dent', 'flips'), 5.688938692082447)
50. (('bruce', 'wayne'), 5.644120188240601)

**Trigrams from file with top 20 frequencies**

1. (('lower', 'level', 'street'), 0.00026139007240505007)
2. (('major', 'crimes', 'unit'), 0.00018297305068353504)
3. (('overlooking', 'prewitt', 'building'), 0.00015683404344303003)
4. (('rooftop', 'overlooking', 'prewitt'), 0.00015683404344303003)
5. (('national', 'guard', 'commander'), 0.00013069503620252504)
6. (('commuter', 'ferry', 'night'), 7.841702172151501e-05)
7. (('eye', 'go', 'wide'), 7.841702172151501e-05)
8. (('hong', 'kong', 'detective'), 7.841702172151501e-05)
9. (('applied', 'sciences', 'division'), 5.227801448101001e-05)
10. (('bank', 'manager', 'fires'), 5.227801448101001e-05)
11. (('bullet', 'fragment', 'array'), 5.227801448101001e-05)
12. (('day', 'rachel', 'come'), 5.227801448101001e-05)
13. (('dogs', 'start', 'barking'), 5.227801448101001e-05)
14. (('evening', 'gordon', 'enters'), 5.227801448101001e-05)
15. (('find', 'harvey', 'dent'), 5.227801448101001e-05)
16. (('garbage', 'truck', 'push'), 5.227801448101001e-05)
17. (('hong', 'kong', 'harbor'), 5.227801448101001e-05)
18. (('joker', 'look', 'back'), 5.227801448101001e-05)
19. (('live', 'long', 'enough'), 5.227801448101001e-05)
20. (('long', 'time', 'ago'), 5.227801448101001e-05)

**Dark Knight Trigrams from file with top 5 by Mutual Information:**

1. (('major', 'crimes', 'unit'), 23.16676433244849)
2. (('national', 'guard', 'commander'), 21.868688283551137)
3. (('rooftop', 'overlooking', 'prewitt'), 20.452518814782373)
4. (('lower', 'level', 'street'), 19.365055973532034)
5. (('overlooking', 'prewitt', 'building'), 18.65733860667087)

**Biden DNC Acceptance Speech Analysis**

**Biden Unigrams with top 50 frequencies and normalized value**

1. america 30 0.0016
2. president 29 0.0016
3. one 19 0.001
4. u 18 0.001
5. people 17 0.0009
6. time 14 0.0008
7. make 12 0.0007
8. moment 12 0.0007
9. going 12 0.0007
10. never 12 0.0007
11. nation 12 0.0007
12. get 12 0.0007
13. light 11 0.0006
14. great 11 0.0006
15. united 10 0.0005
16. hope 10 0.0005
17. year 10 0.0005
18. wa 10 0.0005
19. love 9 0.0005
20. know 9 0.0005
21. say 9 0.0005
22. ha 8 0.0004
23. much 8 0.0004
24. together 8 0.0004
25. american 8 0.0004
26. job 8 0.0004
27. country 8 0.0004
28. history 8 0.0004
29. take 8 0.0004
30. million 8 0.0004
31. family 8 0.0004
32. life 8 0.0004
33. world 8 0.0004
34. way 7 0.0004
35. word 7 0.0004
36. hard 7 0.0004
37. back 7 0.0004
38. protect 7 0.0004
39. always 7 0.0004
40. work 6 0.0003
41. promise 6 0.0003
42. believe 6 0.0003
43. said 6 0.0003
44. americans 6 0.0003
45. every 6 0.0003
46. pay 6 0.0003
47. current 5 0.0003
48. fear 5 0.0003
49. new 5 0.0003
50. winning 5 0.0003

**Bigrams from file with top 50 frequencies**

1. (('million', 'people'), 0.0010139416983523447)
2. (('climate', 'change'), 0.0007604562737642585)
3. (('current', 'president'), 0.0007604562737642585)
4. (('president', 'ha'), 0.0007604562737642585)
5. (('social', 'security'), 0.0007604562737642585)
6. (('young', 'people'), 0.0007604562737642585)
7. (('affordable', 'care'), 0.0005069708491761723)
8. (('build', 'back'), 0.0005069708491761723)
9. (('care', 'act'), 0.0005069708491761723)
10. (('daddy', 'changed'), 0.0005069708491761723)
11. (('ever', 'faced'), 0.0005069708491761723)
12. (('every', 'day'), 0.0005069708491761723)
13. (('give', 'people'), 0.0005069708491761723)
14. (('history', 'ha'), 0.0005069708491761723)
15. (('history', 'rhyme'), 0.0005069708491761723)
16. (('may', 'god'), 0.0005069708491761723)
17. (('middle', 'class'), 0.0005069708491761723)
18. (('never', 'get'), 0.0005069708491761723)
19. (('one', 'another'), 0.0005069708491761723)
20. (('people', 'light'), 0.0005069708491761723)
21. (('powerful', 'voice'), 0.0005069708491761723)
22. (('president', 'obama'), 0.0005069708491761723)
23. (('protect', 'america'), 0.0005069708491761723)
24. (('united', 'america'), 0.0005069708491761723)
25. (('united', 'states'), 0.0005069708491761723)
26. (('vice', 'president'), 0.0005069708491761723)
27. (('way', 'forward'), 0.0005069708491761723)
28. (('21st', 'century'), 0.00025348542458808617)
29. (('accelerating', 'threat'), 0.00025348542458808617)
30. (('across', 'europe'), 0.00025348542458808617)
31. (('ago', 'yesterday'), 0.00025348542458808617)
32. (('almost', 'anywhere'), 0.00025348542458808617)
33. (('almost', 'half'), 0.00025348542458808617)
34. (('alone', 'two'), 0.00025348542458808617)
35. (('always', 'believed'), 0.00025348542458808617)
36. (('always', 'got'), 0.00025348542458808617)
37. (('always', 'hear'), 0.00025348542458808617)
38. (('always', 'stand'), 0.00025348542458808617)
39. (('american', 'community'), 0.00025348542458808617)
40. (('american', 'darkness'), 0.00025348542458808617)
41. (('american', 'history'), 0.00025348542458808617)
42. (('american', 'moment'), 0.00025348542458808617)
43. (('american', 'president'), 0.00025348542458808617)
44. (('american', 'soldier'), 0.00025348542458808617)
45. (('american', 'story'), 0.00025348542458808617)
46. (('american', 'worker'), 0.00025348542458808617)
47. (('americans', 'infected'), 0.00025348542458808617)
48. (('among', 'u'), 0.00025348542458808617)
49. (('anti-semitic', 'bile'), 0.00025348542458808617)
50. (('anywhere', 'else'), 0.00025348542458808617)

\*Note there was not enough data to compute mutual information with a frequency of greater than 5 for bigrams or trigrams

**Trigrams from file with top 20 frequencies**

(('affordable', 'care', 'act'), 0.0005069708491761723)

(('current', 'president', 'ha'), 0.0005069708491761723)

(('give', 'people', 'light'), 0.0005069708491761723)

(('almost', 'anywhere', 'else'), 0.00025348542458808617)

(('always', 'got', 'back'), 0.00025348542458808617)

(('american', 'community', 'bearing'), 0.00025348542458808617)

(('american', 'darkness', 'began'), 0.00025348542458808617)

(('american', 'history', 'tell'), 0.00025348542458808617)

(('american', 'worker', 'ca'), 0.00025348542458808617)

(('anti-semitic', 'bile', 'heard'), 0.00025348542458808617)

(('bile', 'heard', 'across'), 0.00025348542458808617)

(('brave', 'little', 'girl'), 0.00025348542458808617)

(('build', 'back', 'better'), 0.00025348542458808617)

(('certain', 'inalienable', 'right'), 0.00025348542458808617)

(('civil', 'right', 'movement'), 0.00025348542458808617)

(('coming', 'year', 'bright'), 0.00025348542458808617)

(('decorated', 'iraqi', 'war'), 0.00025348542458808617)

(('deep', 'black', 'hole'), 0.00025348542458808617)

(('deploy', 'rapid', 'test'), 0.00025348542458808617)

(('deserves', 'one', 'great'), 0.00025348542458808617)

(('difficult', 'moment', 'america'), 0.00025348542458808617)

(('economic', 'crisis', 'since'), 0.00025348542458808617)

(('empowered', 'labor', 'union'), 0.00025348542458808617)

(('every', 'day', 'believing'), 0.00025348542458808617)

(('existential', 'threat', 'posed'), 0.00025348542458808617)

(('family', 'back', 'together'), 0.00025348542458808617)

(('father', 'taught', 'u'), 0.00025348542458808617)

(('four', 'historic', 'crisis'), 0.00025348542458808617)

(('franklin', 'roosevelt', 'pledged'), 0.00025348542458808617)

(('generation', 'never', 'know'), 0.00025348542458808617)

(('generous', 'among', 'u'), 0.00025348542458808617)

(('george', 'lloyd', 'wa'), 0.00025348542458808617)

(('great', 'first', 'lady'), 0.00025348542458808617)

(('great', 'middle', 'class'), 0.00025348542458808617)

(('great', 'second', 'lady'), 0.00025348542458808617)

(('great', 'vice', 'president'), 0.00025348542458808617)

(('ha', 'cloaked', 'america'), 0.00025348542458808617)

(('ha', 'delivered', 'u'), 0.00025348542458808617)

(('ha', 'thrust', 'one'), 0.00025348542458808617)

(('heard', 'across', 'europe'), 0.00025348542458808617)

(('history', 'ha', 'delivered'), 0.00025348542458808617)

(('history', 'ha', 'thrust'), 0.00025348542458808617)

(('history', 'tell', 'u'), 0.00025348542458808617)

(('honest', 'unvarnished', 'truth'), 0.00025348542458808617)

(('inalienable', 'right', 'among'), 0.00025348542458808617)

(('incredibly', 'brave', 'little'), 0.00025348542458808617)

(('iraqi', 'war', 'veteran'), 0.00025348542458808617)

(('keep', 'telling', 'u'), 0.00025348542458808617)

(('kid', 'safely', 'back'), 0.00025348542458808617)

(('last', 'four', 'year'), 0.00025348542458808617)