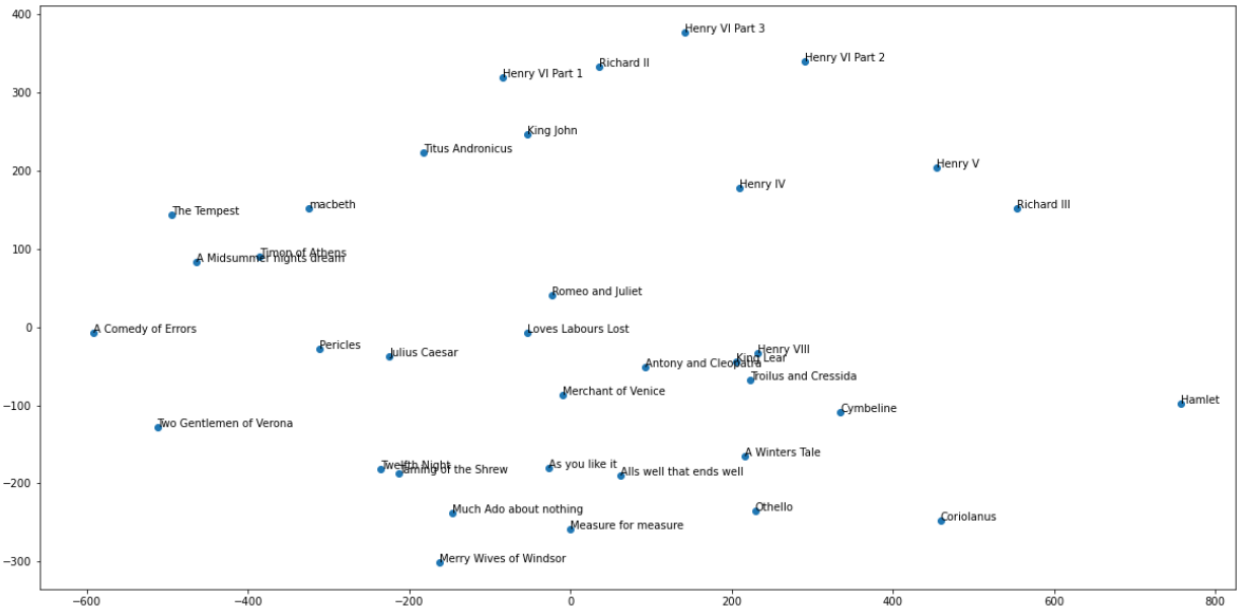
CS505 HW 3 Submission template

Collaboration statement: I collaborated with…
Link to Code: https://drive.google.com/drive/folders/1PtKBx6Z0HI6wDonJ-EgbTTEYii8-V7kQ

Problem 1 (65 pts)

1. (**5 pts**) Attach visualization below:

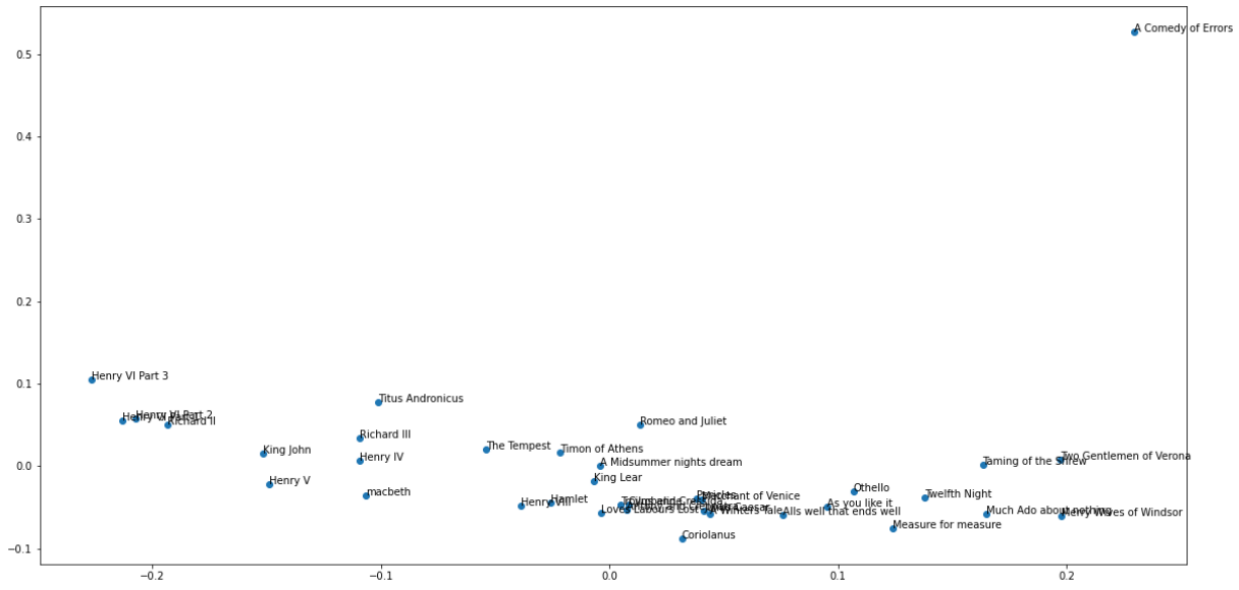


2. (**1 pt**) What plays are similar to each other? …

   (**1 pt**) Does the visualization match the grouping of Shakespeare's plays? …

The Henry plays are grouped together. The left side of the screen has Two Gentlemen, Comedy of Errors, Midsummer, Tempest, and Pericles all grouped, which are all comedies. Romeo and Juliet, Troilus and Cressida, Antony and Cleopatra all appear in the same general region. In general the groupings are approximately equal to the given genres.

3. (**3 pts**) Attach visualization below:



4. (**1 pt**) Does TFIDF give you better grouping of plays?

   Apart from the outlier, the overall groupings have not changed very much from the previous section.

   (**1 pt**) Why do you think so?
   If the usage of certain common words such as “the” and “and” is consistent across plays, then using Tf-idf shouldn’t have a massive impact on the groupings.

5. (**4 pts**) Average pairwise cosine-similarity between comedies: 95.3%

6. (**2 pts**) Average pairwise cosine-similarity between histories: 95.7%

   (**2 pts**) Average pairwise cosine-similarity between tragedies: 95.2%

7. (**2 pts**) Average pairwise cosine-similarity between comedies and histories: 92.5%

   (**2 pts**) Average pairwise cosine-similarity between comedies and tragedies: 94.6%

   (**2 pts**) Average pairwise cosine-similarity between histories and tragedies: 94.2%

8. (**4 pts**) Average pairwise cosine-similarity between comedies: 94.6%

9. (**2 pts**) Average pairwise cosine-similarity between histories: 94.1%

(**2 pts**) Average pairwise cosine-similarity between tragedies: 94.7%

10. (**2 pts**) Average pairwise cosine-similarity between comedies and histories: 94.1%

    (**2 pts**) Average pairwise cosine-similarity between comedies and tragedies: 94.5%

    (**2 pts**) Average pairwise cosine-similarity between histories and tragedies: 94.1%

11. (**4 pts**) Average pairwise cosine-similarity between comedies: 95.1%

12. (**2 pts**) Average pairwise cosine-similarity between histories: 95.9%

    (**2 pts**) Average pairwise cosine-similarity between tragedies: 95.8%

13. (**2 pts**) Average pairwise cosine-similarity between comedies and histories: 92.2%

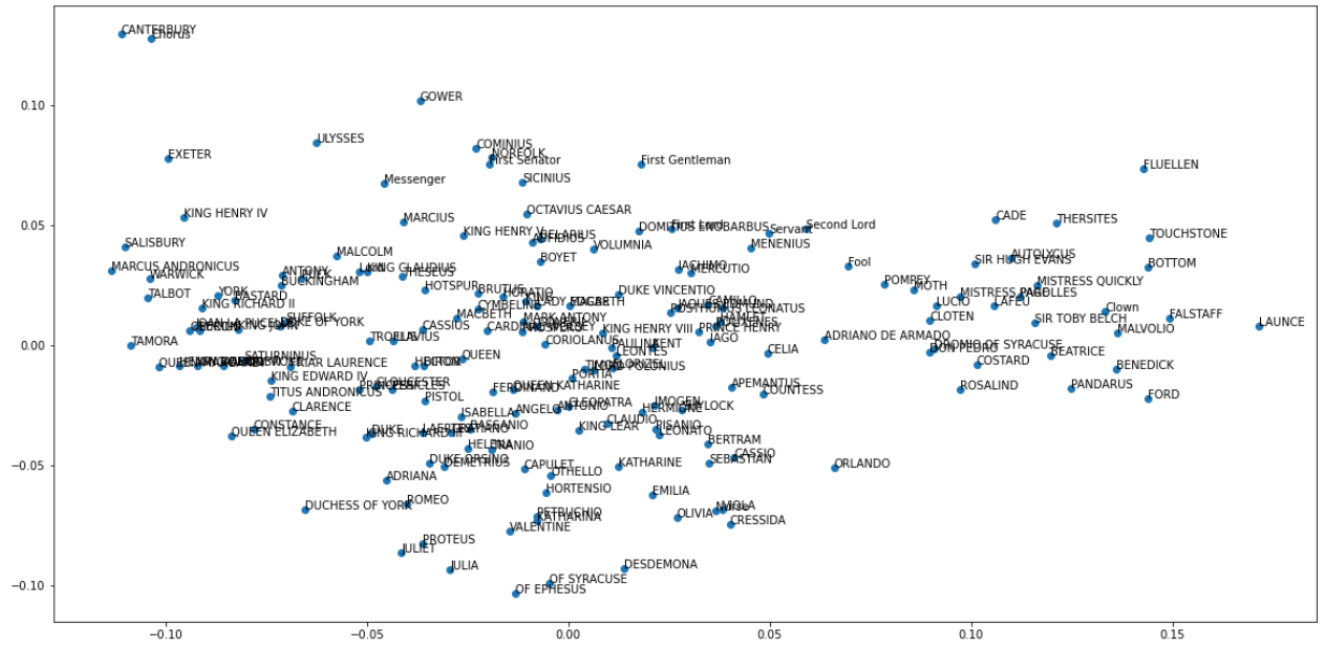
    (**2 pts**) Average pairwise cosine-similarity between comedies and tragedies: 94.5%

    (**2 pts**) Average pairwise cosine-similarity between histories and tragedies: 93.7%

14. (**1 pt**) Which vector representation (word-word counts vs. gensim vs. LABSE) gives the best measure of similarities between plays of similar genre vs. different genres? LABSE (**1 pt**) Why do you think so?

    LABSE gives the best measures of similarities, probably because it can embed entire sentences and is therefore able to gain more context for each embedding.

15. (**4 pts**) Attach visualization below:



16. (**1 pt**) Insight 1:

    Greek and Roman characters like Gower, Ulysses, Cominius, Sicinius, Mascius, Octavius Caesar are grouped together.

    (**1.5 pts**) Explanation/visualization for insight 1:

    These characters speak with much of the same vocabulary and appear in similar plays. They form a cluster in the top center of the graph.

    (**1 pt**) Insight 2:

Another interesting group is in the bottom of the chart where a number of female characters have been grouped.

    (**1.5 pts)** Explanation/visualization for insight 2:

Again, these female characters most likely use words that distinguish them from male characters. Romeo and Juliet appear very near each other, as we might expect since they speak about the same themes and are the central characters of the same play.

Problem 2 (35 pts)

1. (1 pt) Vocabulary size (i.e., number of word types) for Shakespeare: 19967
   (1 pt) Vocabulary size (i.e., number of word types) for Arthur Conan Doyle: 25738
   (1 pt) Vocabulary size (i.e., number of word types) for Jane Austen: 13758

2. (6 pts) 0.5 pt for each row. Fill each cell in the table with the top-3 closest words to the word in the first column for each author

word	Shakespeare	Arthur Conan Doyle	Jane Austen
courage	Amity, guilt, society	Temper, affection, pride	Cease, voluntarily, mislead
hope	Content, comfort, fate	Promise, trouble, help	Trust, wish, suppose
love	Praise, wrong, wish	Hate, hurt, marry	Regard, acquaint, attach
woman	Man, maid, fool	Girl, lady, child	Man, person, girl
man	Thing, woman, fool	Woman, person, fellow	Woman, person, people
he	Himself, it, caesar	Himself, she, I	She, himself, they
she	It, her, myself	He, herself, her	He, herself, himself
good	Madam, friend, gracious	Happy, poor, bad	Kind, ill, great
bad	Less, honester, smack	Simple, true, possible	Otherwise, happy, less
evil	Guilt, success, device	Englishwoman, imitation, extra	Suspicion, alteration, object
beauty	Virtue, passion, age	Household, observance, management	Taste, enjoyment, sentiment
fate	Flight, heal, anger	Presence, success, method	Faithful, interference, nurse

3. (3 pts) 1 pt for each row. Similar to the previous question, but choose your own 3 words to observe

word	Shakespeare	Arthur Conan Doyle	Jane Austen
England	Rome, york, henry	London, france, america	Wiltshire, receipt, dirt
War	Force, land, course	Forenoon, event, scene	Thornton, season, court
Peace	Safety, commandment, notice	Safety, store, lapse	Union, grandeur, field

4. (1 pt) Mention any differences between the close words across different documents:

   For the word woman, the top three most similar words for Austen and Doyle contain “lady” and “girl”, while Shakespeare has “maid”.

   (1 pt) What do you think the difference indicates? For example, does the difference represent any gender bias?

   This could reflect a significant difference in the descriptions of women from different time periods, since Doyle lived 300 years after Shakespeare.

5. (5 pts) Accuracy of analogies on Shakespeare’s works: 0.33%

   (5 pts) Accuracy of analogies on Arthur Conan Doyle’s works: 0.12%

   (5 pts) Accuracy of analogies on Jane Austen’s works: 0.51%

6. (2 pts) What does the accuracy scores tell you?

The accuracy scores tell us which corpus was best as a training tool to capture the same analogies as were given.

   (2 pts) Which corpus then gives you better commonsense representations?

   Jane Austen

(2 pts) Do you think this is a good way to measure word representations e.g., can you think of potential issues in designing analogies?

This is probably a flawed way to measure word representations since a very large number of analogies will be out of vocabulary, and the specific context of the text will affect these analogy representations. For example, a model trained on a WWII corpus will likely have a different interpretation of London and Berlin than a model trained on a corpus of encyclopedias will.