Oct 30, 2016

Homework 5: Analyzing Plays, Characters and Gender in Shakespeare

Due November 16, 2016 11:59 pm EST

Team project with one partner

William Shakespeare is considered one of the greatest playwrights in the English language. He is attributed with 38 plays, 154 sonnets, and other works. Even 400 years later, his writings are still be studied in most high schools and colleges in the US and elsewhere. Shakespeare's complete works are available online at http://**shakespeare**.mit.edu/

With so many comedies, tragedies and historical plays, he created casts of characters from kings and queens to heroes and villains, and we can't ignore the fools. Your task is to use machine learning/data mining/text mining techniques on the language of Shakespeare's works to examine the similarities/differences in the characters based on their language. You should examine these similarities or differences using a few clustering techniques on the words uttered by the characters perhaps individually but more likely per scene/act and (for minor characters) by play. You can visualize the results by play, gender, role, etc. from the clustered data. The clustering is intended to include all the plays in one clustering (at least initially) so as to be able to compare characters across plays.

Your solution should include:

1) Data preparation technique(s) to convert the text to features. You will use all of Shakespeare's plays (comedy, history, tragedy) but can skip the sonnets (poetry). Each character's utterances should be captured in a row of the data table possibly separately, but more likely grouped by scene, act, or possibly even work.

2) A few tuned clustering techniques with the goal of characterizing similarities and differences between characters as a function of play, gender, role, etc. Be sure to assess/optimize the method as appropriate given the algorithm (i.e., how did you select the "k" of k-means). We have included a text file that lists the gender, play, and some character roles (villain, fool, etc.)

3) A minimum of two clear visualizations to interpret the clusters such as PCA, Sammons, MDS, dendrogram, SOM, etc. Visualizations requiring a magnifying glass or multiple monitors to see are not considered clear or useful.

We are hoping that you will explore and compare different approaches to identify patterns from various methods. You are welcome to use python tools or other software as you see fit.

Submit a report of ≤ 4 pages of text (or preferably an ipython notebook with equivalent text) that describes:

- How you approached the problem
- What text processing you used (i.e., sentences, capitalization, stemming, etc.)
- What text features you used (TFIDF, frequency, N-Gram, etc.)
- What clustering approaches you used and show the parameter optimization (i.e. “k”)
- What visualization methods you found most helpful
- Do the techniques give consistent results and what are they?
- The appendix (outside the 4 pages) should include code or whatever is needed to reproduce your analysis. This might include installation and run instructions for package X. It would be ideal but is not necessary to process all the results directly from the web site to perhaps a feature table. However, we feel it would be appropriate to go from mined data to clusters and perhaps separately from clusters to visualization.

The partner whose last name comes first in alphabetical order should submit the report. Both partners should submit the partner assessment.

Optional Exploratory opportunities:

1) Teams that go beyond the minimum requirements will be appropriately awarded exploratory points.

2) After (and only after) you have ‘solved’ the problem, consider a variation of the analysis. You are welcome to pursue any direction that extends your exploration of ML/DM techniques. Here are some ideas that I had to get you give you an idea:

    - Analysis of the language of characters through the different scenes
    - Analysis of the change of language as a function of the year written
    - Interpretation of the cluster—what leads to interesting cluster(s) forming?
    - Outlier analysis from the clusters and an explanation
    - Use of sentiment analysis or related measures to relate characters to the characters through the acts
    - Analysis of other corpora (I’m personally curious about similarities vs. differences in the Bible vs. Quran and how the different books or chapters cluster)
    - http://www.shakespeare-online.com/quotes/antonyquotes.html lists famous quotes. Do these quotes cluster and/or could a prediction system be trained to recognize these?

3) The exploratory analysis should be < 3 pages of text (or preferably an ipython notebook with equivalent text) and in a separate Exploratory section.