Homework 4: To be or not to be…the author

Due October 26, 2015 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.

Is it possible that a single person could have written such a diverse set of masterpieces over a window of just 24 years or so? Perhaps some of the acts or scenes were written by other notables of the day or perhaps a group of his students.

Shakespeare's complete works are available online at http://**shakespeare**.mit.edu/

Your task is to use machine learning/data mining techniques on the language of Shakespeare's works by act, scene, or other breakdown as appropriate to address the question of authorship. Your solution should include:

1) Data preparation technique(s) to convert the text to features
2) A minimum of 2 clustering techniques to group similar acts/scenes/etc. Note that we assume tuning will be required for each technique.
3) A minimum of one 2D visualization methods to interpret the clusters such as PCA, Sammons, MDS, etc.

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 ≤ 5 pages that describes:

- How you approached the problem
- What text mining features you tried
- What clustering approaches you used and show the parameter optimization
- What visualization methods you found most helpful
- Do any of the techniques give consistent results?
- What did you conclude about the authorship?
- The appendix (outside the 5 pages) should include how to reproduce your analysis. This might include installation instructions for package X. While it would be ideal to run one line and go from web site to clustering output, that is likely an unnecessary challenge. However, we feel it would be appropriate to go from mined data to clusters and perhaps separately from clusters to visualization.

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

Bonus opportunities:

1) Teams that go beyond the minimum requirements will be appropriately rewarded.

2) After (and only after) you have 'solved' the problem, consider a variation of the analysis by examining the language of the characters in Shakespeare's plays. Is the language consistent through the various acts? Are there divisions in language between good/bad characters or male/female characters?

3) We are open to other interesting extensions if you would like to propose ideas.