Homework 6 CSCI E-82

Due Tuesday December 4, 2018 11:59pm EST

To be or not to be…Shakespeare.

This final assignment is a culmination of your text mining, machine learning, data mining and AI skills. Shakespeare, the great bard from Stratford-on-Avon who changed the direction of theater for centuries across multiple languages and cultures, penned 37 plays over the course of 23 years. (Clearly he wasn't taking CSCI E-82 and had more time on his hands.)

Even though his theater was a few blocks from John Harvard's birthplace, is it really possible that a single genius could possess the creative strength and focus to craft all that work by himself? Perhaps Shakespeare was actually a team of playwrights or the joint effort of some collaborations. With 1-2 partners, your task is to formulate a more specific hypothesis that tests this idea, and to perform an in-depth test of your hypothesis using your machine learning skills. To help those seeking new partners, we'll ask that you facilitate that process by signing up at our usual signup link.

A convenient source of the data is https://www.kaggle.com/kingburrito666/shakespeare-plays but you are welcome to use any source available. The text mining pipeline from HW2 can likely be re-used to easily convert the plays into a data matrix. Note that it is plausible that a scene or act could be written by a different author so you should likely have more than 37 rows in your matrix.

There are multiple types of analyses that you could use. In the interest of your time and need to shift over to the final project, we are narrowing the scope. We would like you to perform an in-depth analysis using just one general type of approach such as clustering, classification, outlier detection, or network analysis (next lecture). That is, we would like you to do a thorough job focusing on clustering, for example, with multiple approaches rather than do a surface-level analysis using many of these. You will likely need to use some visualization to tell your analysis story regardless.

So, how can you determine whether Shakespeare was the author? Think along the lines of patterns:

- Writing style or choice of words used
- Temporal patterns in sentiment, etc. throughout the plays (comedies, tragedies, etc.)
- Chronological analysis – did his writing traceably develop or evolve over time?
- Did the scenes or characters in a scene have a particular formula?
- Are there particular topics that emerge consistently throughout his plays?

What to submit

Submit your ipython notebook or equivalent and corresponding pdf file. If you have used external code sources, be sure to properly cite those. You can either write a short document that describes your goals, method, and results or be sure to include them in the ipython notebook with sufficient detail that we can clearly follow the hypothesis, approaches, and interpretation of the results.

Grace Period Days

This is the last assignment and it is a team effort. Since there is one deliverable for the assignment, all team members are charged the same number of grace days if it is late. This can be an issue since if you have 0 grace days left and your partner has 3 left, since you would lose 20% of your grade per day and they would not lose any for the first 3 days. Avoid those *Et tu, Brute* moments and discuss this with your partner(s) early.

Extra Credit option

After completing the above hypothesis test homework, you have the option of seeking extra credit. The extra credit deadline is November 30 (see syllabus) but we'll allow a Shakespeare approach to be submitted with this assignment. The task is to perform a separate, novel hypothesis test and analysis with machine learning such that the idea is so creative that your fellow students are unlikely to have thought of it. Specifically, we're looking for a cleverly constructed idea/hypothesis that you assess with a minimum sufficient analysis. Since we want you to focus on your final project, this analysis part would be shorter rather than the in-depth analysis from the main assignment. The HW6 is worth more points than the extra credit so don't sacrifice your efforts on the main assignment—this is an optional extra. You may do this part individually or with your team but it must be submitted with the homework and the standard deadlines/late days. Clearly label the extra credit section if you elect to include it and include whether it is a team effort or your individual work.