MCIS6273 Data Mining (Prof. Maull) / Fall 2021 / HW4 [BONUS]

This assignment is worth up to 20 POINTS to your grade total if you complete it on time.

Points Possible	Due Date	Time Commitment (estimated)
20	Wednesday, December 7 @ Midnight	*up to* 8 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Continue to improve your Bayesian solution from HW3b, all earned points are bonus/extra credit

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hwN`. Put all of your files in that directory. Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hwN_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(100%) Continue to improve your Bayesian solution from HW3b, all earned points are bonus/extra credit

In HW3b you learned some of the basics of document classification through the Multinomial Nïave Bayes model within SciKit-Learn.

This is a BONUS homework, so anything you do will get you partial or full points. If you complete all of the parts you will earn up to 20 points depending on the completeness of your solution.

There are four ways to get points:

1. build a better training set,
2. train the model on new training sets (either one's you found from (1) or some other corpus),
3. improve the model analysis by detailing the summary of model accuracy across all topics,
4. develop a visualization of the data, model output or some other interesting component of HW3b.

Doing all three completely will earn up to 20 points and a combination of all three will earn a few extra points depending on completeness.

§ We learned that the size of the training corpus matters.

To earn up to 5 points, you will make a new dataset for the other disciplines (1 point for each one of sociology, econonics, education, physics and computer science, 5 full points if you do them all).

You solutions may vary, but here are a few ideas:

- find the top 100 to 200 papers in a discipline, get their DOIs and then get the abstracts from Semantic Scholar. Use those abstracts to build a new dataset;
- find the top 20 papers in a discipline and extract the text from that 20 then all the subsequent papers that that paper cited. For example, if you look at the paper which won the 2010 Nobel Prize in Physics and then look at all the papers which that paper cited (you can look at the output of semantic scholar), then get the abstracts of those papers, you would end up with a large diverse corpus (though it may be a bit narrow depending on the papers, hence why you would need a good sample of the 20 seed papers);
- pull tens (minimally) of books from Gutenberg.org in the disciplines of interest and train on those documents;
- pull fulltext from the PDF (or other sources) of classic / seminal papers in a discipline, convert them to text and then use them as the training documents in your corpus for the discipline.

Varying levels of completeness accompanying good explanations of what you're doing and *why* will earn you more points towards the 5.

§ Train the new model on the dataset (or datasets) from the first task and then analyze the performance of the model.

This analysis would include the increased accuracy, with an explanation why you feel the accuracy has increased and if it doesn't change, why you think it did not.

Point increases can range up to 5, and the completeness of your analysis will earn you more points.

An evaluation of true positive and false positive rates would be valuable to include in your analysis. Any error analysis would be important anyway no matter if you are analyzing the performance of one class or all of them.

§ Visualize the data of the corpus in some interesting way, hopefully with the aim of improving the understanding of the underlying data.

Here are a few ideas:

- use Python to develop a word cloud of the corpus *without* stop words and/or other irrelevant words;
- compare each discipline's most relevant words through the lens of the TFIDFVectorizer which will require you to tap into the underlying representation of that object and doing something useful with the data that is there;
- show how a corpus changes after expanding the underlying abstracts (e.g. show the dominant words in the economics corpus after 100 documents, then how it changes after 200 documents);
- do a time series visualization of a sub-discipline. For example, looking at computer science from 1995 - 2015 in 5 year increments and taking the top documents and their differences. What you might relay in such a visualization is how topics change over 20 years time.

There are many other things you might think of. Look at the visualizations in the various Python libraries (Seaborn, matplotlib and others) to get some inspiration.

You can earn up to 5 points for this part.