Programming Assignment 1

COMP 550, Fall 2024

Due: **Friday, September 27th**, 2024, 9:00pm.

You must do this assignment individually, except where otherwise stated. You may consult with other students orally, but may not take notes or share code, and you must complete the final submission on your own.

Real vs. Fake Facts About Cities

In this assignment, you will simulate the investigation of a text classification problem, asking the questions: Is it possible to distinguish real from fake facts about cities using linear classifiers? Does the choice of linear classifier matter?

You will collaboratively generate a dataset of real vs. fake city facts using online generative AI tools, then set up an experimental pipeline to train models that classify them into "fact" or "fake".

Important note: This assignment presents an oversimplified picture of distinguishing real from fake information for pedagogical purposes. Fake news/misinformation detection is a rich and complex research area in NLP and in the social sciences. Please do not take the purported results from this assignment too seriously!

The goal of this assignment is to give you experience in using existing tools for machine learning and natural language processing to solve a classification task. Before you attempt this assignment, you will need to install Python 3 on the machine you plan to work on, as well as the following Python packages and their dependencies:

- NLTK: `http://www.nltk.org/`
- NumPy: `http://www.numpy.org/`
- scikit-learn: `http://scikit-learn.org/stable/`

Dataset generation

Use an online generative AI tool to generate facts about cities. For example, ChatGPT 3.5 is freely available online. There is no need to pay for this purpose. Use a prompt that will let you generate facts, real or made up. Be sure to document the prompts and the tool that you used. For example, here are two prompts that I used:

- "Give me facts about Montreal."
- "Give me fake facts about Montreal."

And some of the "facts" that were generated included:

- Bilingual City: Montreal is one of the largest French-speaking cities in the world outside of Paris, but it is also a major English-speaking city. This bilingual nature contributes to its rich cultural diversity. [FACT]

- Island City: Montreal is located on the Island of Montreal, which is situated at the confluence of the Saint Lawrence and Ottawa Rivers. The island is named after Mount Royal, the triple-peaked hill in the heart of the city. [FACT]
- Montreal's Underground City Is Actually a Secret Lab: Beneath the Réso lies a hidden research facility where scientists are working on a top-secret project to create a new flavor of poutine that can only be described as "cosmic." [FAKE]
- Mount Royal Is an Ancient Volcano: Mount Royal isn't just a hill—it's actually a dormant volcano that was last active during the Ice Age, and locals believe it has mystical powers that can control the weather. [FAKE]

Pick your own favourite city and generate facts about it! Each fact should be roughly one to two sentences in length. For the sake of sharing in class, please generate these facts in English. You are free to do supplemental experiments in other languages of your choice and to remark on this in your report, but please experiment with one language at a time.

Store your facts, one fact per line, in two text files in UTF-8 encoding with the following names: *facts.txt* for the generated facts, and *fakes.txt* for the generated fakes. Make sure there are at least 100 samples in each class, and ideally more if you have the time to gather more data.

Sharing your dataset (optional)

To expand your dataset, feel free to share your dataset with others in the class. However, you must still generate at least 100 samples of each class on your own. I encourage you to post your datasets on the course Ed discussion forum, and/or download cases from there, or to share with your classmates. Please include the names of all students with whom you have shared data. Note that other students may not have generated samples about the same city as you. How will you account for this in your experimentation?

Please note that this sharing policy only applies to the data, not to the code or to writing the report!

Preprocessing, feature extraction, and model implementation

Your responsibility is to design and run the correct experiments in order to answer the research question stated above. How will you do so in a way that is scientifically sound? You will likely need to subdivide your dataset in some way.

You must explore at least 3 preprocessing decisions and 3 linear classifiers that we have discussed in class. You may use scikit-learn's feature extraction and classification modules to help you, as well as any other tool from NLTK or NumPy. Reading scikit-learn's documentation will be of great help in your experimentation.

Setting up the experiments

Design and implement experiments to draw reasonable conclusions about the research question above. This will require creating subsets of the dataset as we discussed in class. There are multiple correct ways to set up your experiments (as well as many incorrect ways).

Report

Write a *short* report on your method and results, carefully document i) the problem setup, ii) your dataset generation and experimental procedure, iii) the range of parameter settings that you tried, iv) the results and conclusions, and v) the limitations of your study. It should be no more than 1.5 pages long. Report on the performance in terms of accuracy, and speculate on the successes and failures of the models.

Regarding point v), think about how much you can generalize the conclusions of your experiments to the overall problem of separating real vs. fake facts for cities, geographical locations, and in general. What assumptions have we made that are reasonable or not reasonable?

Your assignment will be marked on i) how well it satisfies the requirements stated in this handout, ii) whether your experiments adequately and correctly address the research questions, iii) how well written your report is. It will NOT be marked based on the performance that you achieve with your models on this dataset.

Submitting code

Submit your code in a file named “pa1.py”.

What To Submit

Submit your report as a single pdf on myCourses called “pa1-answers.pdf”. In addition, you should submit i) your source code in a file called “pa1.py”, ii) your dataset in two files “facts.txt” and “fakes.txt”. All work should be submitted to myCourses under the Programming Assignment 1 folder. Jupyter notebooks are not acceptable as a submission format.