Xavier Lapointe – 11345738
Xavier Péladeau-Asselin - 11256697

March 11th 2024

MATH60629A – Term Project Study Plan

The question you aim to answer

This study seeks to leverage natural language processing techniques on a large corpus of news articles from La Presse, a daily newspaper established in Montreal, consisting of approximately 30,000 articles spanning the previous calendar year. La Presse is renowned for its coverage of international events, as well as political, economic, and cultural affairs. Our goals are to perform a series of unsupervised learning tasks to analyze topics and identify trends throughout the year. Moreover, we aim to quantify the political orientation and biases of the newspaper and its writers by conducting sentiment analysis and determining topic-specific polarity scores. Temporal analysis could be applied on these polarity scores to determine if general sentiment towards some topics/entitities changes over time. Additionally, we will carry out classification tasks aimed at predicting the journal section from an article's contents and categorizing articles according to their author's writing style.

The source of the dataset you will be using

More than 30 000 articles published by La Presse over the past year have been scraped from the newspaper's website with the help of a Python script (using the Scrapy library). The metadata collected with each article includes authorship, date of publication and journal section. Were it required for further analysis, any additional data could be scraped from the website and added to our dataset.

The tentative machine learning methods you plan to apply to answer your question

We plan to conduct an in-depth analysis of data pre-processing techniques, tokenization, and feature engineering methods to optimize the performance on both supervised and unsupervised tasks. For unsupervised tasks, our objective is to explore a variety of distance and similarity measures, as well as different clustering algorithms, tailored to task-specific hypotheses. Regarding classification tasks, we intend to compare the performance of traditional algorithms like Naive Bayes, and Support Vector Machines against more contemporary language modeling approaches, including Recurrent Neural Networks and Long Short-Term Memory (LSTM) neural networks. Moreover, sentiment analysis will be applied with the help of the pre-trained models included in the *spaCy* Python natural language processing library.

The questions or aspects you would like to discuss during our meeting

- Will it be possible modify the dataset after we start working on the project, if our understanding of the data requirements for our questions changes?
- Do we need to identify a single objective related to a specific machine learning task prior to the project's start, or is it enough to have multiple broad analysis ideas as we currently do?