ML 1010 Group Project - Yelp Reviews

Our group, Machine Learning Pokémon, consists of 5 members: Durai Nachiappan, Iman Lau, Lijuan Yang, Mohammad Islam, Shabeeth Syed.

Our Jupyter Notebooks and code are available at: https://github.com/stellarclass/ML1010-Yelp-Project (https://github.com/stellarclass/ML1010-Yelp-Project)

Our project uses data from Yelp (https://www.yelp.com), an online review site where users can rate various businesses. Most often, it is used for restaurants, although any business can be rated, from hotels to doctors. Yelp is widely used in North America but is available in numerous countries worldwide. This data is available through Kaggle (https://www.kaggle.com/yelp-dataset/yelp-dataset), provided by Yelp.

Using natural language processing, we can take a large amount of unstructured review data and gather insights from the reviews. If one were to open a business in the city of Toronto, it would be useful to know what reviews tend to talk about. This would let a business owner know what to focus on when creating and running their business. To do this, we propose to use topic modelling. As the name implies, this abstracts topics from the corpus, which are a starting point to understanding what a large amount of people are saying about businesses in the city. We also will perform some sentiment analysis to understand if the reviews are generally positive or negative, especially when related to the star ratings.

With that in mind, for our mid-term project proposal, we are submitting initial Jupyter notebooks containing some feature engineering and the start of topic modelling. We used a smaller dataset, consisting of 10,000 rows of data, to start with as well. This is because we want to look only at Toronto data for the final submission, so we can have a more targeted view of what is happening in the reviews. For the mid-term project proposal, we are also including some initial code that filters the data for reviews in Toronto.

For the final project, we will use more modelling techniques and explore an ensemble method of topic modelling and sentiment analysis. We will also look into modelling over time, as it would be useful to see if there has been a change in what reviewers like to talk about. This is important because opening a new business relies on being able to anticipate the needs and wants of reviewers and customers, so relying too heavily on past data without considering how it changes over time would be short-sighted.