COMP40020: Human Language Technologies  
Assignment 1

* Investigating language styles from data on Reddit
* Build a corpus from one or more subreddits
* Pre-processing steps that should be used.
* Can you infer the subject of a subreddit from its most common words
* Common n grams capture popular jokes
* Text lexically diverse
* Around 2000 words

**Research Question:**

Is New Zealand really that much happier than Ireland? A sentiment and lexical analysis of comments of r/NewZealand and r/Ireland

According to the annual World Happiness Report for 2022 New Zealand comes in at position 10 while Ireland is placed 3 places lower in position 13 out of the 150 countries. This report coupled with the large amounts of young Irish people emigrating to countries like New Zealand and Australia (position 12) provides the rationale for the research question.

Both these subreddits are home to discussion about news, politics, culture, history and society.

Sentiment analysis is a natural language processing method to determine whether a piece of text is either positive, negative, or neutral. Lexical analysis is used to further break down the text and assess the mood.

New Zealand and Ireland are both English speaking countries which makes results easier to compare and contrast.

**Methods**

**Results**

**Discussion**

There are important caveats when carry out sentiment analysis on social media posts.

**Bibliography**