- Assignment One -
Text preprocessing, N-grams, and Language Models

CS918: 2023-24

You will work on Assignment 1 in the lab sessions between Week 2 and Week 5. No submission is required for Assignment 1.

Preparation: Getting to know Python

For this exercise you will be using the "SIGNAL NEWS1" corpus provided on the module website, available through the following link:

https://warwick.ac.uk/fac/sci/dcs/teaching/material/cs918/signal-news1.tar.bz2

The corpus provides news stories formatted in JSONL. Each line contains a JSON item with a news story. You should be using the "content" field of the news stories in this exercise.

The exercise consists of three parts:

Part A: Text preprocessing

1. After lowercasing all the text, use regular expressions to parse and clean the texts:
   a) Remove all non-alphanumeric characters except spaces, i.e. keep only alphanumeric characters and spaces.

   b) Remove words with only 1 character.

   c) Remove numbers that are fully made of digits (e.g. you should remove the number '5', but in the case of '5pm', made of both digits and letters, you should keep it as is, without removing the digit that is part of the word).

   d) Remove URLs. Note that URLs may appear in different forms, e.g. "http://www.*", "http://domain", "https://www.*".

   NOTE: The preprocessing above may need to be processed in a different order, not necessarily as listed above.
2. Use an English lemmatiser to process all the words. Use of a POS tagger is optional, and you may instead assign each word the default POS tag when using the lemmatiser.

Part B: N-grams

With all the texts preprocessed as above, compute the following calculations:

1. Compute N (number of tokens) and V (vocabulary size).
2. List the top 25 trigrams based on the number of occurrences on the entire corpus.
3. Using the lists of positive and negative words provided with the corpus, compute the number of positive and negative word counts in the corpus.
4. Compute the number of news stories with more positive than negative words, as well as the number of news stories with more negative than positive words. News stories with a tie (same number of positive and negative words) should not be counted.

Part C: Language models

1. Compute language models for trigrams based on the first 16,000 rows of the corpus. Beginning with the bigram "is this", produce a sentence of 10 words by appending the most likely next word each time.
2. Compute the perplexity by evaluating on the remaining rows of the corpus (rows 16,001+).