Title and Group:

Group members: Chow Jun Wei (jun.chow.18@ucl.ac.uk, Brunei Darussalam, NIL, Data Science and NLP and Data Analyst), Esraa Sultan (Alzahrani.Esraa1@gmail.com, Saudi Arabia, Esri Saudi Arabia, Data Science).
Internship Batch: NLP 02.
Report Date: November 20, 2021

Problem Description

There are 20 different categories of articles, and our job is to classify, to which category does each article belongs to.

EDA Performed:

- Word count (split by space) AND
- Line number count (split by '\n') explains most distribution towards the left-end (most "articles" have very low number of line number/word counts), with a **very** long tail towards the right.
- Stop words count: It's also the same distribution depending on the line numbers.
- Stop words count / word count (ratio): Approximately Gaussian distribution (more inclined towards less than 0.5) with median of about 0.47, mean about 0.44 and standard deviation about 0.13.
- N-gram (unigrams, bigrams, trigrams): Unigrams doesn't really see any much stuffs except it works more like counting what words occur most frequently, and things like "don't", "just", "think", "know", "it's" that aren't very useful conquers most of unigrams. Looking at bigrams and trigrams however, gives much more information about stuffs, especially those things that're not useful to training (we thought it might be great if we remove them, using regex or other means). Examples that occur frequently are symbols that repeats themselves (usually they're separators, but these are quite useful during training, at least removing it reduces accuracy), email addresses (these can be removed).
- American vs British English: We noticed that some are written in American English and some in British English, resulting in inconsistency.

We didn’t update the cleaning file, but we update the cleaning procedure (particularly what to clean during regex) when we performed EDA on our data.