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 17, 2021

Problem Description

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

Data Understanding

What type of data have we got for analysis? Text data.

What are the problems in the data?

- Having some fields that's not very useful like tabs (\t), newline (\n) (optional), and separators (*****, ------, ======, etc).
- Some data are too short with just one or two sentences. They might (or might not) be useful for training.
- Some data, ***after tokenization,*** are mostly consisted of xxunk (the unknown, meaning they aren't repeated sufficiently throughout the datasets to be recorded in the vocabulary). These aren't effective for training. After checking they includes being in another language (German), they can be programming code, or articles purely consisted of coordinates (hence numerical values aren't tokenized since they're continuous and not English), and some are not understandable (something like hyperlinks that aren't human understandable).

What approaches you're trying to apply on your dataset to overcome problems and why?

- Replace these not useful fields with either "" or single spacing might improve accuracy. The only unfortunate being those with backslash as escape sequences like that\'s wanting to replace as that's failed with regex.
- Remove those sentence below a fixed threshold. In the demo notebook, this is set to 10 lines (after splitting according to '\n', and remove extra '\n' if there exists multiple consecutively).
- These are removed altogether from the program. We use a naïve threshold of **0.3** such that any articles with 30% xxunk will be removed. This could be tuned.