Assignment 4

▷▷ PROBLEM STATEMENT: Consider a suitable text dataset. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors. Classify documents and evaluate precision, recall.

▷▷ OBJECTIVE: Learn how to tokenize and filter a document into its different words and then do words count for each word in the document.
Apply stemming, feature selection on documents (text)

▷▷ OUTCOMES: Demonstrated text processing using nltk, understood vectorizing; removing stop words.

▷▷ SOFTWARE & HARDWARE REQUIREMENTS: Any CPU with i3 processor or similar, 4 GB RAM or more; 1 GB HDD / 128 GB SSD; 64 bit Linux/UNIX OS; nltk library jupyter, python3;

▷▷ THEORY:

• Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with interactions between computers and human language, in particular how to program computers to process and analyze large amounts of

natural language data.

- In computing, stop words are words that are filtered out before or after the natural language data (text) are processed. While 'stop words' typically refers to the most common words in a language, there is no universal list of stop words.

- STEMMING: For grammatical reasons, documents are going to use different forms of a word (eg. organized, organize, organizing). Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization.
- The goal of stemming (and lemmatization) is to reduce inflectional forms and sometimes derivation -ally linked forms of a word to its common base form

  am, are, is ⟶ be

  car, cars, cars', car's → car
- When applied to a document, the result will be somewhat like this:

  the boy's cars are different colors

  the boy car be differ color

- Stemming is ~~thus~~ a more crude process, a heuristic that chops off the ends of words in the hope of achieving the goal correctly most of the time, and often includes the removal of derivational affixes.

- Feature Selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification
- It serves two purposes: first, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers, that are expensive to train.
- Second, feature selection often increases classification accuracy by eliminating noise features. Noise features are those that, when added to the document representation, increases the classification error on new data.

- Vectorization is the process of converting the text data into a machine-readable form. The words are represented as 'vectors' (numerically).

- Countvectorizer (one-hot encoding) involves counting the number of occurrences of each word occurring in a document.
- The idea behind count vectorization is simple: ~~a~~ ~~many~~ vector is created, having as many dimensions as there are distinct words in the text/document/collection of documents ~~you are~~ being used. Each unique word has a unique dimension and will be represented by a 1 in the dimension with 0s everywhere else. This results in huge and sparse vectors that capture no relational data.

- TF-IDF vectors are related to one-hot encoded vectors, but instead of just featuring a count, they feature numerical representations where words aren't just present or not present - instead, they are represented by their term frequency multiplied by their ~~inverse~~ inverse document frequency.
- In simpler terms, words that occur a lot ~~but~~ everywhere should be given very little weight or significance, because they don't provide a large amount of value. However, if a word appears very little; or frequently but only in specific places, then these are probably of higher significance.
- The downside is there is still no capture of semantic relatedness. This is solved with a co-occurrence Matrix, or a neural probabilistic model.

CONCLUSION:

Text documents were tokenized, filtered, vectorized, and thus successfully processed. Basic concepts of Natural Language Processing were understood.