Assignment 1

ECSE344L Natural language processing

Tokenization-

Analyse the following NLTK tokenizer

1. **nltk.tokenize.WhitespaceTokenizer**
2. **nltk.tokenize.WordPunctTokenizer**
3. **nltk.tokenize.TreebankWordTokenizer**

After tokenization let's perform

Token normalization

We may want the same token for different forms of the word

- wolf, wolves à wolf
- talk, talks à talk

Stemming

- A process of removing and replacing suffixes to get to the root form of the word, which is called the stem
- Usually refers to heuristics that chop off suffixes

Lemmatization

- Usually refers to doing things properly with the use of a vocabulary and morphological analysis
- Returns the base or dictionary form of a word, which is known as the lemma

Use:-

nltk.stem.PorterStemmer

nltk.stem.WordNetLemmatizer

Transforming tokens into features-

Bag of words (BOW)

For each token we will have a feature column, this is called text vectorization.

	good	movie	not	a	did	like
good movie	**1**	**1**	0	0	0	0
not a good movie	**1**	**1**	**1**	**1**	0	0
did not like	0	0	**1**	0	**1**	**1**

- **Problems:**
  - we loose word order, hence the name “bag of words”
  - counters are not normalized

Term Frequency (TF)

which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

Inverse Document Frequency (IDF)

which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).