Feature Extraction of Text data

July 19, 2024

Feature extraction- Is the mapping from textural data to real valued vectors.

Bag Of Words(BOW): List of unique words in the text corpus. e.g is creating a list of unique words in a paragraph Term Frequency-Inverse Document Frequency(TF-IDF): To count the number of times each word appears ia a document.

```
Tf-idf Vectorizer = (Number of times term t appears in a document)/
                    (Number of terms in the document )
```

```
Inverse Document Frequency (IDF) = log(N/n), where N is the number of documents and
n is the number of documents a term t has appeared in.
```

The IDF value of a rare word is high, wherears the IDF of a frequent word is low.

TF -IDF value of a term = TF * IDF

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: