09-19-2024 NLP - Feature Extraction

The Bag of Word Technique

- It is a simple and flexible way of extracting features from documents.
- It can find the occurence of certain words this is called as the feature count.
- Since ML models cannot understand sentances, they will need to have vector inputs, this is created by using feature counts.
- When the vector matrix is created, it creates rows for each sentence.
- **Drawback:** It just keeps a track of word counts and disregards the grammatical details about the word order as it treats each word equally.

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer

corpus = ['Data science is an overlap between Arts and Science', 'generally, arts graduates
are right brained and science graduates are left brained' 'Excelling in both arts and
sciences at  a time becomes difficult', 'Natural Language Processing is a part of Data
Science']
bow_model = CountVectorizer()
print(bow_model.fit_transform(corpus).todense())


bow_df = pd.DataFrame(bow_model.fit_transform(corpus).todense())
bow_df.columns = sorted(bow_model.vocabulary_)
bow_df.head()

# Getting Top Most Frequency words
bow_model = CountVectorizer(max_features=10)
bow_df = pd.DataFrame(bow_model.fit_transform(corpus).todense())
bow_df.columns = sorted(bow_model.vocabulary_)
bow_df
```

Output:

\n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n

	and	are	arts	data	graduates	is	part	processing	right	science
0	1	0	1	1	0	1	0	0	0	2
1	2	2	2	0	2	0	0	0	1	1
2	0	0	0	1	0	1	1	1	0	1