09-13-2024 NLP - TF-IDF

Refer to Notes from 09-21-2024 for Theory

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

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']

pptCorpus = ['This movie is very scary and long',
           'This movie is not scary and is slow',
           'This movie is spooky and good']

tfidf_model = TfidfVectorizer()
print(tfidf_model.fit_transform(pptCorpus).todense())
vocabulary = tfidf_model.get_feature_names_out()
print("Vocab",vocabulary)
tfidf_df = pd.DataFrame(tfidf_model.fit_transform(corpus).todense())
tfidf_df.columns = sorted(tfidf_model.vocabulary_)
tfidf_df

## Getting Top Most frequency Words
tfidf_model = TfidfVectorizer(max_features=5)
tfidf_df = pd.DataFrame(tfidf_model.fit_transform(pptCorpus).todense())
tfidf_df.columns = sorted(tfidf_model.vocabulary_)
tfidf_df
```

Limitations of TF-IDF

- It is unable to understand the semantics of the word.