**Massive Dataset Project**

1. **Inverted\_index\_pyspark\_post.ipynb**

This file consists of 4 To-Do’s in which we write code to calculated tf-idf and doc-mag for all words in documents .

* In the first to-do we calculate the frequency and maximum frequency of words with respect to their document id. We clean the data by removing stopwords, punctuations and digits. We also lemmatize all words and obtain the distinct words to count occurrences of distinct words in document and obtain its frequency and maximum frequency.
* In the 2nd RDD we try to create an RDD to get frequency of word from all documents.
* Here we calculate idf using the tf and frequencies.

We receive 2 text files as output, namely:

Inv\_ind.txt

Doc\_mag.txt

1. **Untitled1.ipynb**

This notebook consists of creating a database and table for text file : Inv\_ind.txt, Doc\_mag.txt and also includes interface. We have also used a button with On-Click functionality.

Below displayed is the output from our Untitled1.ipynb

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated