ASSIGNMENT 2
NAME:-Anand Sharma (MT19059)

READ ME FILE

• Question 1

PREPROCESSING:-

1. converted query and tokens of document to lower case
2. used RegexTokenizer to tokenize both query and tokens of document which will convert it to tokens and remove extra things from it
3. Used lemmatization and converted similiar words to same tokens in both query and document tokens
4. Checked by applying stopwords removal but by removing stopwords I will loose the title of any query so not used stopwords removal
5. not converted or removed any numbers

Assumptions :-

1. not converted numbers words to numbers
2. no document fetched in case of empty query
3. Title weight is 0.7*tf_idf score and normal body weight is 0.3*tf_idf_score

Methodology:-

1. Use Jacard coefficient and fetched top k documents Basically taken intersection of query and tokens and divided by length of both
2. Use TF-IDF Retrieval system and taken TF score calculation using various technique such as log_normaliztion_TF,binary tf ,raw tf ,double normalisation tf,term frequency tf and idf using inverse document frequency idf and inverse document frequency idf frequency max
3. Used cosine similarity which is calculated based on previous tf idf score by various methods by taking (A.B)/|A|*|B|
4. Based on assumptions of different weights of title and body Retrieval system is designed with title given more weightage and body less .

5. In analysis report shown different variations of all types of scoring

Question 2:-

PREPROCESSING:-

1. only converted query to lowercase as edit distance is case sensitive

Assumptions :-

All uppercase characters will be considered as lowercase

Methodology:-

1. Used Dynamic Programming approach for calculating edit distance of query and dictionary document
2. Cost of deletion is 1 ,insertion 2 and replacing 3
3. Calculated edit distance by considering if a[i]==b[j] take previous distance else take minimum of all operations with their cost