IR ASSIGNMENT-2

Authors:
Group-31
Kuldeep Singh kuldeep21041@iiitd.ac.in
Shubham Jain shubham21091@iiitd.ac.in

Question-1: Jaccard similarity and TF-IDF

Preprocessing steps

1. Removal of punctuation marks.
2. lowercasing of the characters.
3. word tokenization.
4. Removal of stop words.
5. Applying Lemmatization.

Steps to construct Jaccard Coefficient

1. Preprocess the files.
2. Preprocess the query.
3. Calculate the similarity of query with every document by using the jaccard similarity formula.
4. Store the scores of each document in a dictionary.

Steps to construct TF-IDF matrix

1. Preprocess each file and take a union of words of every file to construct the global vocabulary.
2. Get the term counts of every word in documents.
3. Get the document frequency using the inverted index constructed in the last assignment.
4. Calculate idf using the formula.
5. Calculate the tf using 5 different formulas in the question.
6. Preprocess the query and create a query vector by putting 1 at index of word in the query and 0 for rest of the words of global vocab.
7. Multiply query vectors with the suitable tf matrix to get the results.

Pros and Cons of tf weighing technique

1. **Binary:** It is a simple weighing technique but it does not give the frequency, it just shows whether the word is present in the document or not.

2. **Raw Count:** It is a simple technique but it just takes the raw count due to which length of documents is not taken into consideration.

3. **Term Frequency:** It takes into account the total word in the document but it might result so it is better than Raw count weighing scheme, but do not have any normalization.

4. **Log Normailzation:** It normalizes the values by taking the log of tf values. But then it again do not account of total words in the douments.

5. **Double Normalization:** It normalizes the tf values as well as takes the max tf values of the entire document so it is better then log normalization.

How to Use ?
provide the input query when prompted for it
For example:
Enter your query: lion stood thoughtfully for a moment

Question-2: NDCF
Preprocessing steps

1. Read the CSV
2. Filter the df with qid:4
3. Sort the df values in descending order by Relevance
4. Calculate the total number of documents possible by calculating the the factorial of value counts

Steps to Find NDGC 50

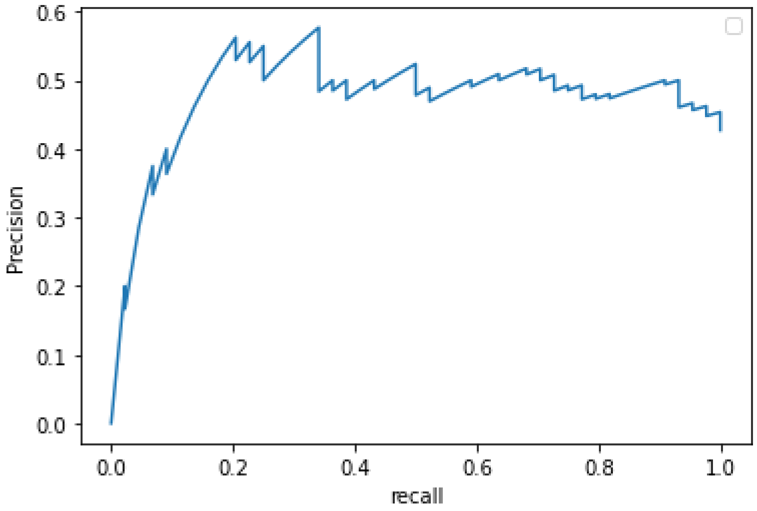
1. Filter top 50 rows from the DF
2. Create a new DF with sorted relevance
3. Calculate DCG for for DF and SortedDF
4. Calculated NDCG by DCG/DCGMax

Steps to Find NDGC for whole document

1. Load qid:4 data into DF
2. Create a new DF with sorted relevance
3. Calculate DCG for for DF and SortedDF
4. Calculated NDCG by DCG/DCGMax

Precision vs Recall Curve

1. Filter qid:4 data
2. Sort in Descending order of value in Col75
3. Calculate Precision @ 1 , Precision @ 2 ….
4. Calculate recall@1 , recall@2….
5. Plot the graph between Precision and Recall



Question-3: TF-ICF and Naive Bayes

Preprocessing steps

1. Removal of punctuation marks.
2. lowercasing of the characters.
3. word tokenization.
4. Removal of stop words.
5. Applying Lemmatization.

Steps followed :

- Load Data
- Perform Train Test Split
- Create word list for all train data
- Create word list for all test data
- Create global list
- Create TF , CF and ICF dictionaries
- Iterate over Train Data and create TF-ICF dataframe for train data
- Iterate over Train Data and create TF-ICF dataframe for test data
- On train data perform column wise summation of values
- Pick top k features from each class having the highest value of TF-ICF score
- Create df from train data having the k features
- Apply Naive Bayes Classifier
- Predict the Labels on Test data
- Measure Accuracy