IR Assignment 2

Group 57

Meenal Jain (MT21050) Prakhar Kumar (MT21066)

DATA

Humor, Hist, Media, and Food.
Total Documents = 1134

PREPROCESSING

1. Converted the text to lower case.
2. Perform word tokenization using word_tokenize
3. Remove stopwords from tokens.
4. Remove punctuation marks from tokens
5. Remove blank space tokens

Q1

Jaccard Coefficient

METHODOLOGY

Calculated the Jaccard coefficient for each query with respect to the given documents. And retrieving the top 5 documents using the Jaccard coefficient

Formula

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

J = Jaccard distance
A = set 1
B = set 2

Firstly, applied the preprocessing steps to the input query and the documents. Then find the length of the Intersections and unions of queries and documents And calculate the Jaccard coefficient using the above formula.
Then sort the documents in decreasing order and produce the top 5 documents

TF-IDF Matrix

Using the same Data and Preprocessing steps.

The Term Frequency is calculated using 5 different variants:

Weighting Scheme	TF Weight
Binary	0,1
Raw count	f(t,d)
Term frequency	$f(t,d)/\sum f(t', d)$
Log normalization	log(1+f(t,d))
Double normalization	0.5+0.5*(f(t,d)/ max(f(t',d))

The IDF value of each word is calculated using the formula as mention below:

Using smoothing:-

IDF(word)=log(total no. of documents/document frequency(word)+1)

METHODOLOGY

1. Firstly, read the data files and preprocess the data into postings.
2. Calculated the document frequency of each word using postings. The total vocab size comes out as 84028
3. Calculated the IDF value for each word using the above-mentioned formula.
4. Calculated the term frequency TF with 5 different variants and then find the tf_idf matrix to retrieve the top 5 documents.

Q2

METHODOLOGY

1. took only qid:4 files as asked in the question. Converted to dataframe for easier manipulation.
2. Maximum DCG is obtained when documents are retrieved in descending order of rankings. So we sorted the documents in decreasing order which gave us a ranking. A number of such rankings is calculated in a notebook.

3. NDCG is calculated as (DCG of documents)/(max DCG that can be obtained). We created a function for this.
4. The precision-Recall curve in the notebook.

Q3

METHODOLOGY

1. Preprocessed the dataset by removing everything except alphabets and numbers.
2. Text is converted to lower case.
3. Tokens of length <= are removed.
4. Tokens are lemmatized to bring them to the root word.
5. Randomly split data.
6. Top k features for each class are calculated using TFICF and then merged to form a new vocabulary.
7. Vectorization is done using TFIDF.
8. Gaussian Naive Bayes is implemented for each class.
9. Results like accuracy and confusion matrix are shown in the notebook.
10. Inferences about the results are also shown in the document.