DS501
Information Retrieval
Assignment 2
Deadline: 26.3.2023

Question 1.

Part 1: Consider the dataset provided for Assignment 1 part 2 i.e. https://docs.google.com/document/d/1FxfmZ0tji8FKBmOJDoTsUMRBEN4ABtJTjAWYvZChceM/edit?usp=sharing. :

a) Randomly sample 100 documents out of all the documents in the dataset; also sample any 10 of the 20 queries you have used for evaluating the different dictionary compression schemes in Assignment 1 part 2. For each of these 10 queries, label the documents as relevant or non-relevant using three annotators (the group members) and calculate the average kappa value. Use the label as per majority voting.

b) Develop an inverted index - dictionary and postings list using standard data structures in Python (Dictionary, Json Formats, List…). Compression is not required here. You need to tokenize and lemmatize/normalize the data. Use NLTK libraries (http://www.nltk.org/install.html). Develop the vector space (TF-IDF) scoring module. Run on the 10 queries selected in part (a) and tabulate the speed of execution. Also calculate precision, recall and MAP for the given query set for top 10 results. Metrics should be calculated as the average for at least 10 runs for each query (this is valid for the rest of the assignment too).

c) Build an inverted index using python based Elasticsearch - https://pypi.python.org/pypi/elasticsearch. Now again tabulate the speed as well as Precision/Recall/MAP for top 10 and compare with the previous approach.

Part 2: Use the index and model of part 1-b. Apply pseudo relevance feedback assuming top 5 documents to be relevant. In the query updation equation for pseudo relevance feedback, assume beta=1-alpha. Vary alpha with a step of 0.2 between [0,1], and for each alpha value, perform the pseudo relevance feedback. Report the alpha which maximizes the MAP. Does this improve the performance of the IR engine?

Part 3: Use query expansion by adding maximum two synonyms of the nouns and verbs present in the query. Search the synonyms from NLTK Wordnet. Compare the performance with the models in Part 1-b and Part 2.

Part 3: Using the same dataset now to show the comparative behaviors of four different document ranking models - a) TF-IDF b) BM25 c) Language Model and d) LSI (use 50 latent dimensions for terms and documents). For each case, mention the model you are using and corresponding assumptions. Use precision/Recall/MAP for performance measurements and tabulate the speed of execution for each model.

Part 4: Apply K-Means clustering on the chosen 100 documents in part 1-a. Define the similarity between two documents as the jaccard coefficient of their terms (lemmatized and case-folded). Vary the K value, plot the corresponding RSS and find the optimal number of clusters.

Part 5: Apply bottom up hierarchical clustering on the chosen 100 documents with the same similarity function defined in part 4. Use single linkage, complete linkage and centroid based linkage. Cut the dendrogram where the number of clusters is the same as the best K found in the last step.

a. Assume K-means clusters as the ideal ground truth clusters. Calculate the purity and NMI values for single linkage, complete linkage and centroid linkage.
b. Assume that the documents relevant to each query should ideally belong to a separate cluster. With that respect, which of the four aforementioned methods (K-means and three bottom up schemes) gives the best result?

For all questions submit the codes along with a document containing the relevant assumptions and comparison results.

Question 2.

Build an index for this Video Game Sales with Ratings dataset:
https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

Download the csv dataset and index it on your local elasticsearch setup or on the elastic cloud. Do appropriate pre-processing as needed to run the queries. You can use the Bulk API for this :
https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-bulk.html

Write queries to extract the following information from the index :

1. Which game is the best selling game overall ?
2. Which is the best selling genre overall ?
3. Which publisher has the highest number of global sales ?

4. What is the average sales in NorthAmerica? How does this compare with the average global sales?
5. Find the top 5 games with the most number of global sales. What are the top 5 genres for each of them?
6. Find all the games which can be categorized as “Action” or “Fighting”

Submit the queries and corresponding results.