**Lab 2 (Cont. Live session)- Exercises about Information Retrieval -- You need to submit your script (python package together with dataset, if any, as a zip file, or you can also use Jupiter Notebook)**

**Information Retrieval**

1. Consider an academic journal of your own choice and collect 30 abstracts using a method of your own (can be a simple manual copy-and-past operation or webcrawling) in a single file using the same query term (E.g., Go to webscience or sciencedirect or Springer and input a query T of your choice). Save the first 30 results as documents, where each document has four metadata: Title, List of Authors, Abstract text, List of keywords).
2. We want to test how much the keyword metadata content matches the document and query. Concatenate the list of keywords (in metadata) of all documents and construct an inverted file, stating for each keyword the document number it matches. Let KK be this indexed file of keywords (ordered in alphabetical order). Write a script that computes the edit similarity between each keyword in the metadata and the query T (If T contains more than one query term, then the similarity is understood as the maximum similarity score (Edit metric) among all terms of query, e.g., if T = T1, T2, T3, then for keyword K, we have Sim(K,T) = max(Sim(K,T1), Sim(K,T2), Sim(K,T2)). Save the result as an array X whose size is size of all keywords (in metadata) in all documents (excluding repetition).
3. Repeat 2) when considering the Edit similarity between each keyword K in the metadata and the title of the document containing K (considering the same rule applies when the title is made of several tokens so that the maximum similarity over all token is computed, and also when the keyword is found in more than one document). Save the result as new array Y.
4. Use appropriate script to compute the Person correlation coefficient between X and Y and the associated p-value.
5. Now we want to exploit the content of the abstract text. Initially, we want to test the extent to which the keywords of the metadata are part of the tokens of the abstract. Write a script that calculates for each keyword K, the frequency of K in the corresponding abstract. Save the result as a matrix M (n x m) where n stands for the number of documents and m the number of keywords.
6. Consider the query T used originally to construct the database, write a script that uses Boolean model to find out relevant document, utilizing matrix M (we assume the vocabulary is made only of keywords in KK).
7. Repeat 6) when you use tf-idf model.
8. We want to test the consistency between title of each document and the abstract text. Write a script that computes for each title, the FuzzyWuzzy score between the title and the corresponding abstract. Save the result in an array Z. We assume that the matching is accepted if the Fuzzy-wuzzy score is greater than 80%. Comment on the results.