Assignment 2

Jan-May 2022 CS6370: Natural Language Processing
Release Date: 1/03/2022 (Tuesday)
Deadline: 18/03/2022 (Friday) 11:59 pm

The goal of the assignment is to build a search engine from scratch, which is an example of an information retrieval system. In the class, we have seen the various modules that serve as the building blocks of a search engine. The first part of the assignment involved building a basic text processing module that implements sentence segmentation, tokenization, stemming/lemmatization and stopword removal. This module involves implementing an Information Retrieval system using the Vector Space Model. The same dataset as in Part 1 (Cranfield dataset) will be used for this purpose.

1. Now that the Cranfield documents are pre-processed, our search engine needs a data structure to facilitate the ’matching’ process of a query to its relevant documents. Let’s work out a simple example. Consider the following three sentences:

   S1 Herbivores are typically plant eaters and not meat eaters

   S2 Carnivores are typically meat eaters and not plant eaters

   S3 Deers eat grass and leaves

   Assuming *{are, and, not}* as stop words, arrive at an inverted index representation for the above documents (treat each sentence as a separate document).

2. Next, we must proceed on to finding a representation for the text documents. In the class, we saw about the TF-IDF measure. What would be the TF-IDF vector representations for the documents in the above table? State the formula used.

3. Suppose the query is "plant eaters", which documents would be retrieved based on the inverted index constructed before?

4. Find the cosine similarity between the query and each of the retrieved documents. Rank them in descending order.

5. Is the ranking given above the best?

6. **Now, you are set to build a real-world retrieval system. Implement an Information Retrieval System for the Cranfield Dataset using the Vector Space Model**.

7. (a) What is the IDF of a term that occurs in every document?

   (b) Is the IDF of a term always finite? If not, how can the formula for IDF be modified to make it finite?

8. Can you think of any other similarity/distance measure that can be used to compare vectors other than cosine similarity. Justify why it is a better or worse choice than cosine similarity for IR.

9. Why is accuracy not used as a metric to evaluate information retrieval systems?

10. For what values of α does the F_α -measure give more weightage to recall than to precision?

11. What is a shortcoming of Precision @ k metric that is addressed by Average Precision @ k?

12. What is Mean Average Precision (MAP) @ k? How is it different from Average Precision (AP) @ k?

13. For Cranfield dataset, which of the following two evaluation measures is more appropriate and why? (a) AP (b) nDCG

14. **Implement the following evaluation metrics for the IR system**:

    (a) Precision @ k

    (b) Recall @ k

    (c) F-Score @ k

    (d) Average Precision @ k

    (e) nDCG @ k

15. Assume that for a given query, the set of relevant documents is as listed in *cran_qrels.json*. Any document with a relevance score of 1 to 4 is considered as relevant. For each query in the Cranfield dataset, find the Precision, Recall, F-score, Average Precision and nDCG scores for k = 1 to 10. Average each measure over all queries and plot it as function of k. **Code for plotting is part of the given template**. You are expected to use the same. **Report the graph with your observations based on it**.

16. Analyse the results of your search engine. Are there some queries for which the search engine's performance is not as expected? Report your observations.

17. Do you find any shortcoming(s) in using a Vector Space Model for IR? If yes, report them.

18. While working with the Cranfield dataset, we ignored the titles of the documents. But, titles can sometimes be extremely informative in information retrieval, sometimes even more than the body. State a way to include the title while representing the document as a vector. What if we want to weigh the contribution of the title three times that of the document?

19. Suppose we use bigrams instead of unigrams to index the documents, what would be its advantage(s) and/or disadvantage(s)?

20. In the Cranfield dataset, we have relevance judgements given by the domain experts. In the absence of such relevance judgements, can you think of a way in which we can get relevance feedback from the user himself/herself? Ideally, we would like to keep the feedback process to be non-intrusive to the user. Hence, think of an ’implicit’ way of recording feedback from the users.

Submission Instructions:

1. **The template for the code (in python) is provided** in a separate zip file and you are expected to fill in the template wherever instructed to. Note that any python library, such as nltk, stanfordcorenlp, spacy, etc can be used.

2. A folder named ’Team_number.zip’ that contains a zip of the code folder (named ’code’) and a PDF of the answers to the above questions **must be uploaded on Moodle by 18/03/2022.**

3. Please include the names and roll numbers of each member of the team in the document.

4. Please make sure that the runtime of your code should not go beyond **10 minutes.**

5. All sources of material must be cited. The institute’s academic code of conduct will be strictly enforced.
