Course Project

CS6370: Natural Language Processing Release Date: 09/04/2021

Having completed the assignment on implementing a simple Vector Space Model (VSM) based search engine, you are now set to explore improvisations over it. In part 2 of the assignment, questions 16 and 17 were as given below:

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

The goal of this project is to improve your search engine by addressing its current limitations. Based on the factual record of actual retrieval failures that you have reported in the assignment, you can come up with hypotheses that could address these retrieval failures. To realize the improvements, you can use any method(s) including hybrid methods that combine knowledge from linguistic, background and introspective sources to represent documents. Some examples taught in class are LSA and ESA. You can also explore ways in which a search engine could be improved like its efficiency of retrieval, robustness to spelling errors, ability to autocomplete queries, etc. You are also expected to test these hypotheses rigorously using appropriate hypothesis testing methods. Note that unlike the assignment, the scope of the project is open-ended and not restricted to the ideas mentioned here.

For each method, the final report must include critical analysis of results; methods can be combined to come up with improvisations. It is advised that such hybrid methods are well founded on principles, and not just adhoc combinations (an example of an adhoc approach is a simple convex combination of three methods with parameters tuned to give desired improvements). Teams can ask for resources they need over email (not Moodle), we hope to provide each team with specific resources you ask for.

You are required to submit a 1-2 page proposal with the following details by 20/04/2021 :

- What limitation(s) of the Vector Space Model you are trying to address.
- State your hypotheses for addressing the above limitation(s).
- Describe how you would realize the above hypotheses in your search engine.
- Describe how you would evaluate your system.

You could either build on the template code given earlier for the assignment or develop from scratch as demanded by your approach. A leaderboard will be created to let you gauge the relative performance of your method. Note that while you are free to use any datasets to experiment with; the Cranfield dataset will be used for evaluation. The project will be evaluated based on the competitiveness of your system, the rigour in methodology and depth of understanding, in addition to the quality of report and your performance in viva.