# CS6370: Natural Language Processing
# Course Project

Release Date: 15/03/2023 | Date of submission of codes and report: 07/05/2023
Proposal submission: 19/03/2023 | Viva + Presesntation + Demo: 13/05/2023 or 14/05/2023

Having completed the assignment on implementing a simple Vector Space Model (VSM) based search engine, you are now set to explore improvisations over it. The following are two questions from assignment 2:

- 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 auto-complete queries, etc. You are also expected to test these hypotheses rigorously using appropriate hypothesis testing methods. 60% of your project credit is reserved for the soundness of the experimental methodology and the rigor of your result analysis. 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 ad hoc combinations (an example of an ad hoc approach is a simple convex combination of three methods with parameters tuned to give desired improvements).

You are required to submit a 1-2 page proposal with the following details by 19/03/2023:

- 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. 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 effectiveness and efficiency of your IR system in comparison to that of the systems developed by other groups, the rigor in methodology and depth of understanding, in addition to the quality of report and your performance in viva that will be scheduled on either 13/05/2023 or 14/05/2023. Teams on the top of the leader board might get an opportunity to present their work to the class.

Note: The teams found to have plagiarized components in their project will be awarded 'U' grade.