Bowling Green State University
CS 4800/5800 - Information Retrieval and Search Engines
Project 2 – Understanding the Vector Space Model (VSM)
Due Monday, August 10, 2020

The purpose of this project is to (i) give you an understanding of the Vector Space retrieval model, (ii) allow you to work with a modest to large amount of text, (iii) experiment by varying different parameters, and (iv) analyze the observations through standard retrieval metrics. In this project, you will gain a much better feel for the significantly experimental field that IR is. You will have to present *all* your work in a carefully written report that covers what you have done, what you have observed, why you did what you did (when applicable), how you tried to tune your parameters and influence your results, what results you got, what they mean, and so on. Consider the possibility that you plan to keep improving your program depending on the time and interest you have, as you learn and understand new techniques, or come up with an innovative idea yourself. In your report, also list/describe what kind of changes/improvements you might be interested in making to your program when you have more time, and what kind of work such improvements require.

You can do this using one of the following possible ways: (1) using C++ or other similar language (doing much programming from scratch), (2) using Python-NLTK, or (3) using an alternative approach that you should discuss with me. For the data, you may use any of the various test collections available; I particularly suggest taking a look at the **Glasgow IR test collections (http://ir.dcs.gla.ac.uk/resources/test_collections/)**. I suggest you use the LISA collection, which includes 6004 documents and 35 judged queries; it is easy to use but the results from our approach are not likely to be as good as one might expect. You may also choose additional/other collections. If you have trouble downloading, unzipping, or understanding the data formats, you might want to discuss with me. Since data are critical for this project, I suggest you download and understand the data, and get comfortable with it as your first task, to be completed by the end of the week the project is assigned.

You will start with several hundred documents and work with them as your document collection/corpus. This collection (or a subset in some specific experiments) will be the basis for your models. In other words, N (the total number of documents in the collection), n_i (the number of document containing a specific term t_i – also sometimes called the document frequency df_i for t_i), etc. will be derived from this base collection.

As you may recall, the vector space model is based on the notion of term weights, which may be computed in different ways, the most common method being the *tf-idf* method. You either need to develop or borrow tools to analyze the documents, and this would constitute a substantial part of the project. When you think about it, you will realize that a major data structure you would need is a postings list or a 2-d array that represents the weights of terms in documents. The kinds of tools you would find useful for this purpose are frequency counters, stemmers, stop word filters that use stop word lists, etc. Python's NLTK includes, or allows you to easily create, all these tools, and others (e.g., additional stop words lists) are available at various sites, but you need to verify their authenticity. I can supply a list of stop words and a C/C++ stemmer in C/C++ if you ask for them. Any data and tools supplied to you for use in this class are absolutely and strictly only intended for instructional purposes, and may not be used for commercial benefit.

Your project requires you to experiment with the vector space model. Once again, identify the data set you plan to use (I recommend LISA) early on, and discuss with me any questions you might have. Experiment along the following lines:

A. Work with tf-idf and another weighting scheme (e.g., plain term frequency or others indicated in Figure 6.15 on page 165) so you can identify what impact the term weighting scheme might have on performance.
B. Compare the four combinations of plain or stemmed words, combined with or without stop word filtering. [IMPORTANT: Do keep in mind that when your index is based on stemming or stop word filtering, queries must be treated consistently through the same process.]
C. Also, use your search engine for retrieving documents matching to a given document (i.e., a document is used as query in these cases)! Hypothesize your expectation and check how the results compare.
D. Evaluate top-k precision for at least 3 values of k. Think about possibly creating precision-recall curves.
E. All in all, you will minimally have all the following combinations to plan for: 2 weighting schemes, 2 stemming alternatives (plain and stemmed), 2 word filters (unfiltered and stop-word-filtered), A number of queries (if not all 35 in the case of LISA), 3 values in top-k! Is there a way to present your results in a simpler way by some kind of averaging of your performance metrics?
F. If you run into any difficulty with any of the details for any reason, let me know, so we can identify some options.

Minimum work to be completed:

0. **Start immediately and plan your work carefully**. Read the entire project description and estimate the time you would need to spend to complete the project, especially considering the shortness of this summer session. Prepare a time table and as you complete each task, check it off, but also revise your estimates as needed. ***Keep in mind that there may be other items you need to work on BEFORE the due date of this project!***
1. Make sure to identify the data set you would like to use ASAP. It would be hard to change it later, due to changes needed in preprocessing the data and other factors specific to the dataset. Note that the more useful datasets all come with a set of queries and prejudged relevance assessments of documents against those queries, so you can benchmark your search engine against them.
2. You need to write your own code and/or use tools to separate the documents from each other, identify words, process them and count them (both term frequency in each document and document frequency for each word) word), and maintain the necessary data structures. From these counts, you will derive the necessary weights and apply the formulas.
3. Approach your experimentation methodically. You should make sure to include some interesting queries (some queries clearly would have a very small or a very large number of relevant documents), and queries can be short or long (the latter case including that of using an entire document as a query). Use the same queries in the different approaches so you can readily compare them. (In general, you should change no more than one parameter at a time in your experimentation.)
4. Do you think you can create precision-recall curves? How / why not?
5. Have you used the model to retrieve documents similar to a given document? In the ranked result set, where (at what position) does the "query" document itself appear? At what rank? What other documents can you expect to appear in the results, and why; do they?
6. Write a report that describes exactly what you have done, what results / observations you got, how they compare / contrast with each other, an analysis and interpretation of the results and

your conclusions. How easy would it be to work with a different corpus (off the Internet or a new one)?

7. Have you done any additional work out of curiosity or passion/fun? If so, make sure to devote a separate section in your paper to describe the what, why, how, etc., as well as the results, and whether you found the results interesting or disappointing.

Report / Paper:

In writing your paper (a minimum of about 5-10 double-spaced pages) paper, you are expected to have understood the spirit of experimentation and to demonstrate that understanding in the paper. Some of the points to keep in mind are: Why is experimentation important? What purpose does it serve? Is the experimental methodology reasonable (any biases)? Are all the major steps in the experiments sensible? You should also address broad but important questions such as:

- What was expected of this project? (Goals?)
- What did you end up with?
- What did you learn from the project?
- How are any discrepancies between expected and experimental results to be explained?

Choose a suitable title, and organize the paper into meaningful sections, such as:

- Abstract
- Introduction and Goals
- What exactly you have done (including the experimental setup)?
- What results did you get? Present them in useful, easily understandable form.
- Analysis and interpretation of results
- Any additional work you have completed
- Conclusion (including what you have learned from the project)
- References

BOTH in this report, and in the other papers you will be writing, **absolutely make sure** to follow academic honesty practices:

- **Avoid plagiarism (read this part carefully and also refer to the course syllabus):**
    - You must submit your own work. You may collaborate with others on ideas, but what you present (your programs, your papers, etc.) must be your own, independent work. A key point is to avoid copying without attribution and quoting (in your reports, papers, or code).
    - Make sure to cite your source (e.g., [Salton, et al. 1975a]) and list the cited source completely in your references (e.g., [Salton, et al., 1975a] Salton, G., A. Wong, C.S. Yang: A Vector Space Model for Automatic Indexing, Communications of the ACM 18, pp. 613-620. Also reprinted in, Sparck Jones, K. and P. Willett (Eds.), Readings in Information Retrieval, Morgan Kaufmann Publishers, 1997, pp. 273-280.)
    - Note that citing your sources as indicated above is important to avoid a charge of plagiarism. If several paragraphs are based on a source, cite the source once per paragraph. Failing to cite when your material is based on some other source is dishonest, since it amounts to claiming (if only implicitly) that it is your own work!
    - Avoid self-plagiarism, i.e., presenting **your own** past work as original now.
    - Make sure to use quotation marks ("…") when you use sentences from your source verbatim; this must be done *in addition to* citing your source after the quotation. Failing to quote when you have taken an author's words verbatim (even if you cite

your source) amounts to an implicit claim that the words are your own. For academic purposes, you are expected to read your sources, put them aside, and paraphrase your understanding of what you have read (while making sure to cite your sources).
    - Work including mostly quotations from your sources indicates little (if any) work on your own; such work generally might not even merit a grade of C.
    - Make sure to understand the difference between citing and quoting, as well as the need for and importance of either kind of practice.
- Avoid reference padding and other kinds of dishonesty.
- Your paper will be processed through ***turnitin***[1] for plagiarism detection.
- Ask me if any of the above is unclear. A misunderstanding here could result in a low or failing grade and a record of academic dishonesty in the student's file, since all instances of academic dishonesty **are required** to be reported to the dean. Academic dishonesty can adversely affect a graduate student's chances of assistantship, as well.

Important Note: If your work is systematically performed and analyzed, and a good report is written, it could result in a student paper that can be submitted to some forum.

[1] Visit www.turnitin.com to get an idea, if you are not familiar with it.