CS380 Information Retrieval and Web Search

Lab 1

Building your own IR dataset

Monday, Jan 27

In this lab you will formulate, document and implement a plan for building your own information retrieval dataset. You will use your dataset for at least the next next two assignments and labs. So choose wisely. Many of you will use your dataset in later assignments as well.

The first task is to formulate a plan for your dataset. That is, come up with an idea for a (at least semi) cohesive set of text that you would like to use as your dataset. The text should have most, and preferentially all, of the following properties:

1. Size — number of documents. There should be at least 100-200 documents in your dataset. Bigger is better.
2. Size — number of words per document. Each of the documents should be fairly large; averaging a minimum of 1000 words. Bigger is better; up to a point. For instance, full length novels are probably too big, but if you can break them up into smaller natural chunks then that could be OK.
3. Size — total number of words. The dataset should have a minimum of 500,000 words.
4. Diversity. The documents should have a number of different authors. Ideally the authors should be identifiable. If they are identifiable, the number of authors should be significantly smaller than the number of documents. For example, if you have 100 documents, then you should have at least 10 and no more than 40 authors.
5. Cohesiveness. The documents should not be random but rather they should share a common theme. For example, (and no you cannot use this idea) novels by 19th century British writers. Nor can you use plays by 16th century British writers. So no Dickens and no Shakespeare.

6. Availability. You should be able to get the documents; legally. You will not be re-publishing the documents; but be legal.

7. Interest. The documents should be on a topic about which you have some interest. For instance, the spread of and attempts to control invasive insects — the spotted lantern fly, emerald ash borer,.

8. The documents should be in a single language. Other than English is OK. However, I encourage you to choose a language for which there are a lot of documents and which follow patterns like English; i.e. the words have removable suffixes that can be recognized and removed without understanding the word or the surrounding text. It will be much easier on me (and in many ways easier on you) if you use English, but I am willing to be flexible.

During this lab I want you to come up with an idea, then research that idea to verify that is is feasible with respect to the points above. Discuss your ideas with other students. I find that some of my best ideas come from otherwise meaningless discussions with people. Also, your fellow students might have novel (possibly even of borderline legality) ideas about where to get documents on your topic area.

Some thoughts about document sources: google scholar (the major problem here is that the documents are often in PDF so collecting the documents in plain text will be slow … doable but slow); Project Gutenberg; the US government (have you read the tax code? or worse, the congressional record); any newspaper to which you have a subscription; fan fiction; and, or course, random web pages.

Write up a one paragraph summary of your idea. Find a fellow student who has also completed their summary. Review each others plan. Comment critically (if you can) and make suggestions. Adjust the writeup per your discussions. Then give me a copy of the plan and get my approval. Be sure to acknowledge your reviewer. Do not leave lab without handing in a reviewed and approved plan.

The first part of you homework for the week will be implement you plan. You might find the unix command `wget` is helpful for the task. (Or you might not). If you decide for some reason that you really hate your plan and want to change, that is OK. Write up a new plan and find me to discuss it (actually start by just emailing the new plan to me).