Project 1: Review

Information Retrieval and Interation

Datalogisk institut, Copenhagen University (DIKU)

Oleksandr Shturmov
oleks@diku.dk

December 7, 2013.

The below is a review of an anonymous student submission of Project 1. The name of the file reviewed is `GR5-PR1.pdf` and its `md5sum` is `e22b47db4ae3f52eab0d6f7c7742819a`.

1

First, we'll state some general notes on this exercise, derived from the below.

- There is no overview of what has been done (exercise 1 and 3), or how far below the appendices are.
- There is little discussion of how the results were attained outside the appendices, which hurts the credibility of the answers. It is of interest to the reader whether e.g. the dataset was normalized in any way prior to answering each question.
- Some subtasks were rephrased, sometimes resulting in a different meaning. It is better to copy the exact wording of an exercise text, and then later discuss an interpretation of it, if necessary.

Now we'll look at each question in turn, excluding the appendices.

- *What is the mean number of queries per user id?*

  There is no discussion of how this was solved and, given that it is unclear how far below the appendix is, the reader is left astray looking for how this answer came about.

- *Analyse the variability of query length (i.e., in words or in characters).*

  This subtask was reformulated without further notice: the author has prematurely chosen a method of analysis, without justifying it. The student should state why it was chosen to count word lengths, why it was chosen to visually analyse a plot rather than a numeric method, and what sort of "regular expressions" were used.

- *What percentage of queries are mixed case? Upper case? Lower case?*

  This subtask was also rephrased, but unlike the previous subtask, the same meaning was retained. (See general notes above.)

- *Count the number of questions (look for patterns such as starting with Wh-words, or ending with a '?' symbol). What percentage of queries do questions make up? What is the most common type of question?*

  Again, the phrasing was changed.

- *What are the k most common queries issued?*

  We're not sure what's going on with subtask five, it looks like the student counted the most common keywords, excluding stopwords.

- *What percentage of queries contain stopwords like "and", "the", "of", 'in", "for"?*

  No comment.

- *What are the k most common non-stopwords appearing in queries that contain the word download?*

  The word "download" appears multiple times, presumably due to non-normalized spacing and punctuation. It is a good idea to normalize the dataset for every question, ignoring those aspects which are irrelevant.

- *What are the k most common non-stopwords appearing in queries?*

  No comment.

- *What percentage of queries were asked by only one user?*

  No comment.

- *How often is a consecutive query a reformulation of the previous one? (Not the same query to greater depth.)*

  Not completed.

- *What kind of spelling mistakes do users seem to make in general?*

  Consider an automated method.

- *What percentage of queries contain a person's name?*

  Not completed.

- *How often do URLs appear in queries?*

  Again, this is a question where some explanation is certainly adequate given the generally complicated nature of URLs.

- *Is it likely that this web query log puts anyone's privacy at risk?*

  It is not justified why the occurrence of phone numbers in a query log puts anyone's privacy at risk.

- *Can you find addresses, phone numbers, or other identifiers in the log file?*

  This was just barely mentioned, and under a different question. Consider to elaborate on how the American phone numbers and birthdays were searched for.

- *Is query volume constant throughout the day?*

  Not completed.

Now we turn to the appendix:

- Consider ordering it as in the question outside the appendix, for each question, describe the steps you took to attain the result.
- Remember to mention all your steps. Your results should be reproducible in general.
- Titles like "Shell script" are not very useful to the reader.
- The approach for counting the number of queries containing the different numbers of words is highly inefficient. It's hard to read and maintain, and incurs unnecessary computational cost. Consider this method instead: `http://unix.stackexchange.com/questions/18736/how-to-count-the-number-of-a-specific-character-in-each-line`.
- The approach for counting the various casing of the queries containing only two words, is statistically justifiable, but you can do this for the entire query log by looking for those queries that are composed of lower-cased words or upper-cased words (regardless of the number of words).

- The assignment asked us to use the handed out stopword file, instead of compiling our own lists.
- Consider justifying your regular expressions when dealing with URLs, phone numbers, etc.

3

It is probably a very good idea to show the algorithm itself. You could summarise the formulas in the different cells in your report.