Online Medical Data Clustering for Search Result Diversification

Objectives

Obtain experience in preprocessing & grouping online medical data and re-ranking output retrieval results for improving search diversification performance/aspect search performance.

What to do?

You are given a ranked passage retrieval result file “output-format-york07ga1.txt” and a file “top-passages-york07ga1.zip”. You will implement a clustering program for grouping these retrieved passages for each topic and re-rank these original retrieved passages for improving aspect search performance. You can implement the clustering program by yourself or you can use a tool like Weka (http://www.cs.waikato.ac.nz/ml/weka/) to group these retrieved passages for re-ranking purpose.

The Data Sets

The original ranked passage retrieval result file “output-format-york07ga1.txt” is formatted as follows:

```
200   12595615   1       43.022  3839    339     york07ga1
200   12595615   2       40.761  28426   295     york07ga1
200   10986300   3       35.756  45721   818     york07ga1
200   11809733   4       34.833  8556    1213    york07ga1
200   10662869   5       34.11   40348   1860    york07ga1
200   14963203   6       32.447  20397   1136    york07ga1
200   12595615   7       31.732  6534    570     york07ga1
```

where the first column is the topic number from 200 to 235; the second column is the document ID which is the official identifier for the document; the third column is the rank of the passage for the topic, starting with 1 for the top-ranked passage and preceding down to as high as 1,000; the fourth column shows the system-assigned score from the Okapi information retrieval system for the rank of the passage; the fifth column is the byte offset in the document ID file where the passage begins, where the first character of the file is offset 0; the sixth column is the length of the passage in bytes and the 7th column is your tag ID such as york07ga1 that should be distinct from the other retrieval results.

The file “top-passages-york07ga1.zip” contains up to 1,000 passages per topic that correspond to the passage retrieval file “output-format-york07ga1.txt”. Each record in the file “top-passages-york07ga1.zip” contains topic ID, passage ID such as 12595615_1_43.022_3839_339 and its passage content.

Steps

The project consists of three parts: (1) preprocessing the data, (2) implementing a clustering program and (3) re-ranking the original passage retrieval result for improving aspect search performance. Your re-ranked results should be formatted in the same way as the passage retrieval result file “output-format-york07ga1.txt”. Aspect retrieval performance will be measured using average precision for the aspects of a topic, averaged across all 36 topics. Your re-ranked results can be evaluated by the standard TREC Python evaluation program that is available on the course web site. The appropriate gold standard data files and the evaluation script are also available on the course web site.

What to submit?

You should submit the following items:

1. The assignment report that describes your medical data-clustering program and aspect search re-ranking program, the design of your programs and the analysis of your design and implementation.
2. The programs for preprocessing the raw data, for grouping original retrieved passages for each topic and re-rank these retrieved passages for improving aspect search performance.
3. A file called readme.txt where you give a tutorial on how to compile and run your programs.

How will you be graded?

The full mark for this assignment is 25. The following will play a crucial role in your grade for this assignment.

1. Correctness of programs for preprocessing raw data, grouping online retrieved passages and re-ranking passage retrieval results.
2. Your assignment report that should include introduction, description of your online medical data-clustering and re-ranking programs, description of your implementation, analysis of the results and conclusion. In particular, your report should focus on how to preprocess the raw data, why a specific clustering algorithm is chosen. If a few clustering algorithms have been chosen, which one can generate the best result in terms of aspect search performance and please justify in detail why.
3. Your class presentation. It accounts for 5 marks.
4. Clarity of your programs (comments!) and functionality of your programs.
5. Ease of using the README to test your programs and results.
6. Your group competition mark. Your solution will be compared with the solutions from other groups according to the search diversification performance/aspect search performance through using your clustering and re-ranking programs.
7. Student performance mark in your group. The student performance in your group is to evaluate your performance in the team work. The marks will come from the other members of your group. That is, at the end of the project, each of you will be asked to rate the performance of other members in your group. The ratings on you by the other members of your group peer will determine your “performance mark”. This is to encourage all the students in each group to get involved in the project. It accounts for 5 marks.