CSE 4334/5334 – Data Mining

Fall 2013 - Project 1
Due: 11:59pm Central Time, Friday, Nov 1, 2013

REQUIREMENTS:

You need to submit your project document and source code.

- The project document describes how you tackle the problem, and how you evaluate your solution.
  - No limitation on the number of pages, or the format of the document.
  - The document should clearly describe how you design, implement, and evaluate your classifier.
- Source code must be submitted in a .zip file
  - You have to implement the whole project by yourself. No software package allowed, except standard libraries (such as C/C++ library), stemming and stop-words removal tools, if necessary;
  - You can use any language, though I recommend C++, Java, Python;
  - Your source code must pass compilation. Any non-executable submission is not acceptable.
- Compilation and execution
  - Compile and test your program on omega.uta.edu before you submit
  - Please strictly follow following compilation instruction, your submission may be rejected if compilation fails:
    - First, go to the folder holds your source code;
    - For C++, compile your program using: *g++ *.cpp*
    - For Java, use *find . -type f -name "*.java" -print | xargs javac*
  - Please strictly follow following execution instruction, your submission may be rejected if execution fails:
    - For C++, run your program using: *./main train_file test_file*
    - For Java, use: *java Main train_file test_file*
    - For Python, use: *python main.py train_file test_file*
  - For other languages, you may have to demonstrate to the TA how to compile and run your program.
  - The output of the program should be a file named *result.txt*. The file should contain a header and have the following format:

```
Id,Tags
1,"c++ javaScript"
2,"php python mysql"
3,"django"
```

PROBLEM SCENARIO:

We adopt a Facebook Recruiting competition as our project: http://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction. Thus you're encouraged to submit your solution to the competition if you're looking for a position in Facebook.

Your task is to **predict the tags for questions** from Stack Exchange sites (e.g. stackoverflow.com), **given only the question text and its title**.

Training data is available at: http://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction/download/Train.zip (compressed ~2.2gb, uncompressed ~6.5gb). As described on the competition page, the training data is given in the Train.csv file, which contains 4 columns: Id, Title, Body, Tags.

- Id - Unique identifier for each question
- Title - The question's title
- Body - The body of the question
    - Note: the body is in HTML format
- Tags - The tags associated with the question (all lowercase, should not contain tabs '\t' or ampersands '&')
    - Note: a questions may have multiple tags. Tags are separated by space. A multi-word tag is connected by a hyphen "-"

The first example in the training set:

*"1","How to check if an uploaded file is an image without mime type?"," I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg, gif, bmp) or another file. The problem is that I'm using Uploadify to upload the files, which changes the mime type and gives a 'text/octal' or something as the mime type , no matter which file type you upload. *

* Is there a way to check if the uploaded file is an image apart from checking the file extension using PHP? ", "php image-processing file-upload upload mime-types"*

So the information about this question are:

- Id – "1"
- Title - *How to check if an uploaded file is an image without mime type*
- Body – the third quoted text
- Tags - *php, image processing, file upload, upload, mime-types*

Rules stated on the competition page also apply to our projects. For more detailed information about the competition, please refer to the competition page and its competition forum: http://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction