Information Retrieval

Program 1

Due: September 20, 2016 by 8:00AM.

The goal of this program is:

1. Extract index terms from a simple collection
2. Prepare the input for the classification program

Given the following reviews for a vacuum cleaner:

1. The following were assigned the class **excellent**:
2. I love this vacuum
3. The best vacuum for the money
4. Excellent suction, with a few minor annoyances
5. Amazing. I love this
6. The one to buy
7. I am blown away by how awesome this sucks
8. Wow. Tiny but mighty.
9. This thing is worth the money
10. A great buy
11. The following were assigned the class **good**:
12. Better than I thought but not perfect
13. Impressive vacuum, but feels a little cheap
14. Works well so far
15. I would recommend for the price
16. The following reviews were assigned the class **bad**:
17. It lasted two weeks
18. Stay away. Works well for a short period
19. Extremely disappointing. Works well but not durable.
20. Does not last

Treat each review as a document, extract its terms and store in lexicographic order in a table with four columns. Column1 will contain the terms, and the next 3 columns the number of times each term occurs in its class.

Words extracted from the collection will become a terms if:

1. They are converted to lower case.
2. They are not one of the following stop words: and, an, by, from, of , the, with, a, in
3. They do not include a comma, a period, question mark, colon, semicolon, or an exclamation mark followed by space. (leave them in the token if they are not followed by a space)
4. They do not contain “’s ” or “s’ “. So car’s, and cars’, will be changed to cars and then stemmed to car.
5. Do the following minimal stemming that deals with plurals and third person:
   1. if word ends in “ies” but not “eies” or “aies” then “ies”->“y”;
   2. else in “es” but not “aes”, “ees” or “oes” then “es”->e;
   3. else in “s” but not “us” or “ss” then “s”->NULL endif

Example

The following reviews are assigned the class **excellent**

Love this vacuum.

The one to buy.

The following reviews are assigned the class **good**

Works well so far.

Works well, but not durable

The following review is assigned the class **bad**

Does not last

|  |  |  |  |
| --- | --- | --- | --- |
| Term | E | G | B |
| but |  | 1 |  |
| buy | 1 |  |  |
| does |  |  | 1 |
| durable |  | 1 |  |
| far |  | 1 |  |
| last |  |  | 1 |
| love | 1 |  |  |
| not |  | 1 | 1 |
| one | 1 |  |  |
| so |  | 1 |  |
| vacuum | 1 |  |  |
| well |  | 2 |  |
| work |  | 2 |  |

**Submission directions:**

You may write the code using C, C++ or Java. Your program should compile on **remote.cs.binghamton.edu**. No exceptions.

It should be purely a command line program. NO GUI will be accepted. This enables the TAs to run your code using test scripts. DO NOT assume that since your program compiled and ran correctly on your laptop it will also compile and run correctly on **remote.cs.binghamton.edu**

Please drop a .tar.gz file to the digital drop box of the assignment in blackboard.

The tar.gz file should be named (lower case) as <bmail userid>\_nb1.tar.gz

Example: if your name is pankaj saha and your bmail id is **psha4**, then your package name should be **psaha4\_nb1.tar.gz**

When the file is unzipped it should contain a directory with the same name (like psaha4\_nb1) as the zip file. Please stay updated with blackboard posts regarding assignments and for any query email TA or meet him during his office hours.

**The directory should contain the following:**

1. A “make” file.
2. A read-me file named **readme.txt** which should contain the exact compilation and execution steps. TA should be able to follow the readme.txt steps to evaluate the project. Failure to write a proper readme.txt and make file will deduct 5 points for each.
3. The source code directory.

**Plagiarism Policy:**

All your code will be checked for similarity to other submissions using Moss. Programmers have an uncanny knack of reproducing the same code that they have seen before. So you are advised not to look at each other's code.