**Natural Language Processing: COMS 4705**

**Hirschberg, Fall 2010**

**Due: October 1, midnight**

**Homework 1: Stock Market QA System (100 points)**

Please post any questions concerning this assignment to the Courseworks (*http://courseworks.columbia.edu*) discussion board, under the HW1 topic.

**1 General Instructions**

The main goal of Homework 1 is to produce a simple question answering (QA) system using regular expressions to retrieve information from a single news article related to the stock market.

Your system should take as input a file of financial news and a file of questions, one question per line. Your system should produce as output answers to each of those questions when run on the news file. Some questions may have multiple answers and some may have no answers. Your output should include each question in the question file together with each answer to that question on a separate line. See the examples below.

We will provide sample files of news text and questions as input. You can use these files as `development data' as you build your QA system. These files will be linked to the course syllabus page.

You must write the code yourself. Don't use publicly available code (please refer to the Academic Integrity policy if you have questions in this regard). Unless you have discussed it with the TA’s beforehand, your assignment should be written in Perl (recommended) or Java **and should compile and run on the CS cluster machines**. Please check this explicitly before submitting your final version. We do not want to have to debug your code. If you wish to use any additional tools please check with the TA first by posting a question to Courseworks.

Your submission must include a README file as specified in Section 2.2 below. Also include code for your program and any supporting data files you use, as well as the required compilation and execution scripts as described in Section 3 below.

**2 Grading**

You will be graded on the following elements:

**2.1 Functionality (80 points total)**

**Functionality (42 points)**

Your system should be able to correctly answer ***at least*** the following questions. It should also be able to handle paraphrases of the questions for indices such as the *Dow Jones* (ex: *Dow Jones industrials*, *the industrial* *average*) and verbs (e.g. *rise*, *climb*). You should always give the source for the answer (i.e., the sentence where the answer came from). Do not worry about time. If the input file indicates any rise or fall throughout the day or in the past, you should mention it.

1. Did <index> rise or fall?
2. Did <company stock> rise or fall?
3. How much did <index> rise/fall?
4. How much did <company stock> rise/fall?
5. How much did <index> close/open at?
6. How much did <company stock> close/open at?

**Examples:**

Q: How much did the Dow Jones close at?

A: 2569.6

Source: “The Dow Jones industrials closed at 2659.26” (line 10)

Q: How much did the Industrials Average close at?

A: 2569.6

Source: “The Dow Jones industrials closed at 2659.26” (line 10)

Q: Did Delta Airlines rise or fall?

A: It fell.

Source: For example, their selling caused trading halts to be declared in USAir Group, which closed down 3 7/8 to 41 1/2, Delta Air Lines, which fell 7 3/4 to 69 1/4, and Philips Industries, which sank 3 to 21 1/2. (line 43)

Q; How much did Delta Airlines drop?

A: 7 3/4

Source: For example, their selling caused trading halts to be declared in USAir Group, which closed down 3 7/8 to 41 1/2, Delta Air Lines, which fell 7 3/4 to 69 1/4, and Philips Industries, which sank 3 to 21 1/2. (line 43)

**Regular Expression templates (8 points)**

You should have good regular expression templates for the questions and answers. Quality is more important than quantity. More general regular expressions, that can match multiple questions or multiple kinds of sentences for the answer, are better than rigid regular expressions that match only one string. With more general regular expressions, you will need to create fewer overall, but of course they may also over-generalize. Try for a happy medium.

**Preciseness of answer (10 points)**

Your answer should be as specific as possible. For example, for the first question above, the answer should be **“2569.6”** not **“closed at 2569.6”**.

**Multiple answers (10 points)**

List all possible answers where applicable. You will be penalized for missed answers.

**Example:**

Q: How much did the Dow Jones fall?

A1: 55

Source 1: The Dow industrials were down 55 points at 3 p.m. (line 10)

A2: 114.76

Source 2: At 3:30 p.m., at the end of the "cooling off" period, the average was down 114.76 points. (line 15)

**No answer available (10 points)**

Correctly identify when there is no answer available.

**Example:**

Q: How much did the S&P market rise?

A: No information available.

**Incorrect Answer**

You will be penalized for an incorrect answer

**Example:**

Q: How much did the Dow fall?

A1: Dow Richardson

Source: -- Dow Richardson. (line 5)

**2.2 Software Engineering (includes documentation) (20 pts.)**

**Your README file must include the following:**

* Your name and email address.
* Homework number
* A description of every file in your solution, the programming language used, supporting files, any additional resources used, etc.
* How your QA system operates, in detail.
* A description of special features (or limitations) of your QA system.
* How to run your system, with a sample command line.

**Within Code Documentation:**

* All environmental variables should be set appropriately within the program.
* Methods/functions/procedures should be documented in a meaningful way. This can mean expressive function/variable names as well as explicit documentation.
* Informative method/procedure/function/variable names.
* Efficient implementation
* Programmer, Memory, and Processor efficiency. Don't sacrifice one unless another is improved
* Don't hardcode variable values, etc.

**3 Submission instructions**

If you use a language that requires compilation, you must include a shell script that automatically compiles your code **on the CS cluster machines**. You should provide a script that takes 2 inputs, a text file of news and a text file of questions (1 question per line) and returns a text file of answers to the questions in the form specified above in the examples (i.e. each question on a single line followed by every answer to the question you find, or the string “No information available.”).

When you have completed your system, you will submit your solution electronically using instructions under the “submission” link at the top of the course syllabus or at http://www.cs.columbia.edu/~julia/courses/CS4705/submission.html.

**4 Academic Integrity**

Copying or paraphrasing someone's work (code included), or permitting your own work to be copied or paraphrased, even if only in part, is not allowed, and will result in an automatic grade of 0 for the entire assignment or exam in which the copying or paraphrasing was done. Your grade should reflect your own work. If you believe you are going to have trouble completing an assignment, please talk to the instructor or TA in advance of the due date.