# 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 annotated with part-of-speech 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 a sample file of news text annotated with part-of-speech tags and questions as input. You can use these 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. Do ***not*** use publicly available code or copy code from any other source and do ***not*** let anyone else look at or use your code (please refer to the Academic Integrity policy below and on the course syllabus or ask the instructor or TA 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 assignment. We will not 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*** questions like those below, on new text (i.e. with new names for indexes, stocks, etc. It should also be able to handle paraphrases of the questions for nouns such as the *Nikkei Index* (ex: *Tokyo's Nikkei Index, Nikkei Index*) and verbs such as *rose* (e.g. *rise, climb,*

*went up*). You should always give the source for the answer in the format shown in the example below. Do not worry about time of mentioned events. If the input file indicates any rise or fall throughout the day or in the past, you should mention it.

Some Sample Questions

1. What did <*index or company stock*> do?
2. Did <*index or company stock*> rise or fall?
3. How much did <*index or company stock*> rise/fall?
4. <What|How much> did <*index or company stock*> close/open at?
5. How much did <*index or company stock*> sell at?
6. What was the discount rate?
7. Did <currency> rise or fall against <currency>?
8. What <*indexes or company stocks*> rose/fell?

**Examples:**

Q: How much did the Nikkei Index close at?
A: 34795.05
Source: Tokyo's Nikkei Index of 225 issues, which fell 136.28 points Wednesday, closed at 34795.05, down 445.02. (line 3)

Q: How much did the Tokyo's Nikkei index close at?
A: 34795.05
Source: Tokyo's Nikkei Index of 225 issues, which fell 136.28 points Wednesday, closed at 34795.05, down 445.02. (line 3)

Q: Did Sumitomo Metal rise or fall?
A: It fell.
Source: Nippon Steel fell 10 to 698 yen ($4.83) a share, Sumitomo Metal lost 19 to 677, and Kobe Steel was down 9 at 690. (line 20)

Q: What did Sumitomo Metal do?
A: It lost.
Source: Nippon Steel fell 10 to 698 yen ($4.83) a share, Sumitomo Metal lost 19 to 677, and Kobe Steel was down 9 at 690. (line 20)

Q; How much did Nippon Steel drop?
A: 10
Source: Nippon Steel fell 10 to 698 yen ($4.83) a share, Sumitomo Metal lost 19 to 677, and Kobe Steel was down 9 at 690. (line 20)

**Regular Expression templates (8 points)**

You should create good regular expression templates to produce answers for the questions asked. 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. Note that, if there are multiple sources for the same answer, you should treat these as separate answers to the question.

**Example:**

Q: Did the Nikkei Index rise or fall?
A1: Fell.
Source 1: Tokyo's Nikkei Index of 225 issues, which fell 136.28 points Wednesday, closed at 34795.05, down 445.02. (line 3)
A2: Rose.
Source 2: In the first hour of trading in Tokyo Friday, the Nikkei Index rose 145.96 points to 34941.01. (line 4)

Q: What company stocks rose?
A1: Sharp
Source1: Sharp gained 20 to 1,550. (line 30)
A2: Nippon Shokubai
Source 2: Other gainers included Nippon Shokubai, which rose 70 to 2,270, Nikon, up 30 to 1,620, and Aiwa, which gained 120 to 2,000. (line 31)
A3: Nikon
Source 3: Other gainers included Nippon Shokubai, which rose 70 to 2,270, Nikon, up 30 to 1,620, and Aiwa, which gained 120 to 2,000. (line 31)
A4: Aiwa
Source 4: Other gainers included Nippon Shokubai, which rose 70 to 2,270, Nikon, up 30 to 1,620, and Aiwa, which gained 120 to 2,000. (line 31)
A5: BTR
Source 5: BTR ended 7 higher at 437 pence ($6.76) a share. (line 43)
A6: Ferranti International Signal
Source 6: Ferranti International Signal rose 1 1\2 to 58 on 8.5 million shares. (line 44)
A7: Argyll Group
Source 7: Other companies in the food sector also firmed on active volume, with Argyll Group gaining 8 to 223, Tesco up 1\2 to 194 1\2 and J.Sainsbury advancing 3 to 255. (line 48)
A8: Tesco
Source 8: Other companies in the food sector also firmed on active volume, with Argyll Group gaining 8 to 223, Tesco up 1\2 to 194 1\2 and J.Sainsbury advancing 3 to 255. (line 48)
A9: J.Sainsbury
Source 9: Other companies in the food sector also firmed on active volume, with Argyll Group gaining 8 to 223, Tesco up 1\2 to 194 1\2 and J.Sainsbury advancing 3 to 255. (line 48)
A10: British Steel
Source 10: British Steel edged 2 higher to 126 1\2 on 8.3 million shares, British Telecommuncations settled 6 higher at 268, and British Petroleum gained 1 to 307 1\2. (line 50)
A11: British Telecommuncations
Source 11: British Steel edged 2 higher to 126 1\2 on 8.3 million shares, British Telecommuncations settled 6 higher at 268, and British Petroleum gained 1 to 307 1\2. (line 50)

A12: British Petroleum
Source 12: British Steel edged 2 higher to 126 1\2 on 8.3 million shares, British Telecommuncations settled 6 higher at 268, and British Petroleum gained 1 to 307 1\2. (line 50)

**No answer available (10 points)**

Correctly identify when there is no answer available.

**Example:**
Q: How much did the Dow index rise?
A: No information available.

**Incorrect Answer**

You will be penalized for all incorrect answers.

**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 includes informative method/procedure/function/variable names as well as explicit documentation.
- Efficient implementation
- Programmer, Memory, and Processor efficiency. Don't sacrifice one unless another is improved
- Don't hardcode things that should be variables, 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 in the format of the training file (with part-of-speech tags) 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, please 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/submissions.docx.

## 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.