CS 419/519 Document Analysis
Lab and Assignment 3: Keyphrase Extraction

October 31, 2015

This lab and assignment involves extracting and ranking keyphrases from text data through the C-value algorithm and visually summarizing NSF abstract keyphrase trends over time.

- Maximum marks: 10
- Programing language: Java (for implementing C-value algorithm), Python(for visualizing)
- Assignment questions: CS419/519 Canvas Site
- Deadline: Monday Nov 9 @ 2pm (scheduling a meeting 3-5pm mandatory for discussing answers)

Marking scheme and requirements: Full marks will be given for (1) working, readable, reasonably efficient, documented code that achieves the assignment goals and (2) for providing appropriate answers to Q1-Q3 in a file `ANSWERS_lastname.pdf` submitted with your solution.

Please adhere to the collaboration policy on the course website – people you discussed the assignment solution with, or websites with source code you used should be listed in `ANSWERS_lastname.pdf`.

To verify that your code runs end-to-end, you may be asked to make some changes (to the code or data directory) during the evaluation lab and then re-evaluate results.

What/how to submit your work: Please submit `ANSWERS_lastname.pdf` and all your source code, zipped into a single file called `Assignment3_lastname.zip` and submit to **Canvas**. **Please do not submit data or other large binary files — Canvas will not accept large submissions.**

1 Preparation for the Introductory Lab and Assignment

In the last NER lab, you should have become familiar with the OpenNLP APIs and named entity extraction. This time, we are going to extract keyphrases from abstract of papers (which are not just proper nouns and where we are also concerned with ranking the phrases). To do this, we first need to extract some candidate phrases from the text. Then, we need to score each phrase to rank the importance of keyphrases w.r.t. the data.

To do this, please implement the C-value method from the following paper (the NC-value extension is not required):

`http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/ijodl2000.pdf`

Support code for implementing this algorithm is provided in the github repository that contains this assignment handout.

The code directory `code/ATR` includes a README file with instructions for getting started. It is suggested to use an IDE such as Eclipse, IntelliJ, etc. Make sure you can run the classes listed in the README and obtain error-free output. You are free to borrow code from these classes for your assignment (indeed you'll probably want to).

The dataset we are going to use this time is NSF Abstracts. This dataset contents 129,000 abstracts describing NSF awards for basic research in zipped text files. You will find it is arranged through year-named folders. Please download the data (Part1.zip,Part2.zip and Part3.zip) from:

`http://archive.ics.uci.edu/ml/machine-learning-databases/nsfabs-mld/`

An example text file is shown as follow. We are interested in Abstract only, which can be easily identified by the line starting with "Abstract".

```
Title       : CRB: Genetic Diversity of Endangered Populations of Mysticete Whales:
                Mitochondrial DNA and Historical Demography
Type        : Award
NSF Org     : DEB
Latest
Amendment
Date        : August 1,  1991
File        : a9000006

Award Number: 9000006
Award Instr.: Continuing grant
Prgm Manager: Scott Collins
              DEB  DIVISION OF ENVIRONMENTAL BIOLOGY
              BIO  DIRECT FOR BIOLOGICAL SCIENCES
Start Date  : June 1,  1990
Expires     : November 30,  1992   (Estimated)
Expected
Total Amt.  : $179720             (Estimated)
Investigator: Stephen R. Palumbi   (Principal Investigator current)
Sponsor     : U of Hawaii Manoa
              2530 Dole Street
              Honolulu, HI  968222225     808/956-7800

NSF Program : 1127       SYSTEMATIC & POPULATION BIOLO
Fld Applictn: 0000099    Other Applications NEC
              61         Life Science Biological
Program Ref : 9285,
Abstract    :

              Commercial exploitation over the past two hundred years drove
              the great Mysticete whales to near extinction.  Variation in
              the sizes of populations prior to exploitation, minimal
              population size during exploitation and current population
```

2 General Assignment Instructions

Our goals are to (a) extract term candidates from text files through regular expressions, (b) rank extracted candidates via the C-value algorithm and (c) visualize the evolution of extracted keyphrase frequencies over time to get a sense for trends in funded NSF abstracts.

For part (a), please do the following if not done previously:

- Clone support code from GitHub and import it into Eclipse (or another IDE) as an existing Project.
- Download NSF data from the UCI dataset website (provided previously).
- Modify and run code in `code/ATR/src/file/reader.java` to extract the abstracts.
- Write code to add POS tags to extracted sentence.
- Use regular expressions (see `code/ATR/src/extraction/RegEx.java`) to extract candidate phrases from the text. There are three possible regular expressions:
  - $Noun^{+}Noun$
  - $(Adj|Noun)^{+}Noun$
  - $((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)^{?})(Adj|Noun)*)Noun$

  Make sure you thoroughly test your code (there are online regex debuggers you can use, e.g., `regex101.com`).
- Remove POS tags from extracted candidate phrases (if necessary).

For part (b), using the reference paper, implement the C-value algorithm (including stopword pruning using a stoplist you develop yourself) and apply it to all of the NSF abstract data. Specific instructions on what to show are below.

The NSF dataset is large. If you run into memory issues, note that you can increase Java heap size with the `-Xmx` option.

Part (c) concerns visualization of the keyphrase data. To help with this, a Python script is provided in `vis/main.py` which requires a csv file in the following self-explanatory format:

```
Year,reliable mail system,computer science,nervous system,research project,molecular biology,
potential applications,information services,gene expression,molecular mechanisms,cell biology
1990,32.70344407,40.82404662,62.6,50.81809432,47.20085084,60.8,29.4,78.86685859,14.1,66.92190193
1991,34.71183749,33.67988118,62.1,51.46880537,47.22432481,60.8,28.7,78.99124597,14,66.24147465
1992,33.93165961,35.20235628,61,51.34974154,47.21939541,59.7,28.2,78.43518191,14.5,65.62245655
1993,34.94683208,35.77715877,60.2,51.12484404,47.63933161,58.7,28.5,77.26731199,14.9,65.73095014
1994,36.03267447,34.43353129,59.4,52.2462176,47.98392441,58.1,28.5,75.81493264,15.7,65.64197772
1995,36.84480747,36.06321839,59.2,52.59940342,48.57318101,58.8,27.5,75.12525621,16.2,65.93694921
1996,38.96977475,35.9264854,58.6,53.78988011,48.6473926,58.7,27.1,75.03519921,16.7,66.43777883
1997,40.68568483,35.10193413,58.7,54.99946903,48.56105033,60,26.8,75.1637013,17,66.78635548
1998,41.91240333,37.59854457,59.1,56.35124789,49.2585152,60,27,75.48616027,17.8,67.2554484
1999,42.88720191,38.63152919,59.2,58.22882288,49.81020815,61.2,28.1,75.83816206,18.6,67.82022113
2000,45.05776637,40.02358491,59.2,59.38985737,49.80361649,61.9,27.7,76.69214284,18.4,68.36599498
2001,45.86601517,40.69028156,59.4,60.71233149,50.27514494,63,27.6,77.37522931,19,68.57852029
2002,47.13465821,41.13295053,60.9,61.8951284,50.5523346,63.7,27,78.64424394,18.7,68.82995959
2003,47.93518721,42.75854266,61.1,62.1694558,50.34559774,64.6,25.1,78.54494815,18.8,68.89448726
```

The CSV file is used as input of the visualization code. Anaconda or Enthought Python installations are recommended for their ease of installing packages like *matplotlib* or *pandas*.

Mdify the given Python source code (`vis/main.p`) to specify the correct csv path and text label positions (as required for readability). When you run it, you should see a result like Figure 1.

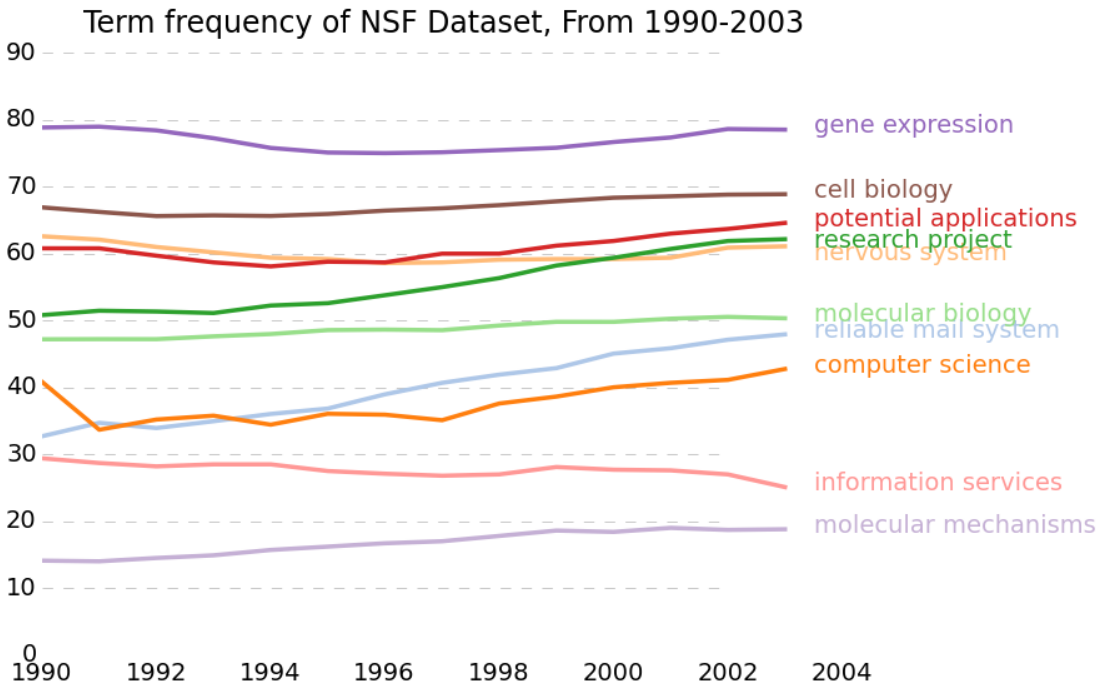

Figure 1: Sample visualization of keyphrase evolution over time for a sampling of 20 keyphrases.

3 Main Assignment (Assessed in Evaluation Meeting)

Q1. Keyphrase Template Matching

(a) Provide the regular expression you used in your Java code.

(b) Provide the top 20 words in the stoplist you used (you can provide any stoplist you think is appropriate). What was the rationale behind your stoplist? [1 sentence]

Q2. Keyphrase Ranking Methodology

(a) List the top-20 keyphrases by frequency (ignoring C-value), *displaying one keyphrase and corresponding frequency per line.*

(b) List the top-20 keyphrases by C-value, *displaying one keyphrase and corresponding C-value per line.*

(c) Qualitatively, which top-20 list do you think is better? Why? [1-2 sentences]

(d) Propose and implement one way to improve the baseline C-value results (the method can be as simple as introducing a threshold or adapting the stoplist specifically to the NSF data or altering where the stoplist is applied; it can also be as complex as the full NC-value method or some approximation of it). What was the improvement and what was the rationale for it?

(e) List the improved top-20 keyphrases. Are they qualitatively better? How?

Q3. NSF Abstract Visual Summarization and Trend Analysis

Here our goal is to obtain a keyphrase summary of NSF grant trends over the period of the data, which may require a small amount of manual pruning of poor keyphrases with high C-value. Consider that your code provides a very fast way to analyze trends and topics in 129,000 NSF Abstracts – something that would be extremely time-consuming and error-prone to do manually in its entirety.

(a) Show a frequency vs. time graph like Figure 1 for your top C-value keyphrases. You should manually prune out poor keyphrases among the top C-values to get a more informed view (you should be able to do this manual pruning in < 2 minutes). Do you note any interesting trends?

(b) An (up or down) trending keyphrase is one whose frequency is drastically changing over the time period of the data (e.g., you can look at the standard deviation over the years, the absolute range of frequencies over the years, or any other measure you think is appropriate for identifying trending keyphrases).

Choose (i) some measure of trendiness in keyphrases and (ii) some threshold on C-value. Extract the most trending keyphrases according to (i) that meet the threshold of (ii) and show them in a frequency vs. time graph like Figure 1.

Explain your choices for (i) and (ii). [1 sentence] Discuss one point of interesting behavior you observe about NSF grants over the time period in this new plot. [1 sentence]