COMP4650/6490: Document Analysis

Assignment 2: NLP

Main details:

Maximum marks:	8
Programming language:	Java (only)
Assignment questions:	Post to the Wattle Discussion forum
Deadline:	Q1: Lab on Wed 29 August (no late days allowed, automatic 0 for Q1 if fail to attend/demonstrate) Q2-Q4: Friday 29 August, 5pm (online via Wattle)

Marking scheme:

- *Written:* Full marks given for a formulation that provides a well-reasoned and succinct response to the question that addresses all requested points. There may be more than one answer for each question that achieves full marks.
- *Code:* Full marks given for working, readable, reasonably efficient, commented code that performs well on the test case given in lab.
- *Academic Misconduct Policy:* All submitted written work and code must be your own (except for any provided Java starter code, of course) — submitting work other than your own will lead to both a failure on the assignment and a referral of the case to the ANU academic misconduct review procedures:

  ANU Academic Misconduct Procedures

Electronic submission (only):

All written questions should be in a file `ANSWERS.pdf`. MS Word or other document formats are not accepted. LaTeX formatting is preferred.

Please submit `ANSWERS.pdf` and Q1 source code zipped into a single file `assign2_yourname.zip` with a `README.txt` stating what each directory is for and how to run your code.

NLP Programming

Q1 [4 pts]. Keyphrase extraction (checked in lab).

Your task is to implement the keyphrase extraction algorithm [1] in Java and apply it to the *Assignment 2 Q1 Data* posted to Wattle. You should use the Stanford CoreNLP library and the stop list provided in the archive file.

In lab you will be asked to run your code on a *new* data set, similar to the *Assignment 2 Q1 Data*, but containing different documents. The code should display a list of extracted keyphrases ranked according to their C-Values / NC-Values in descending order. The grader will both inspect the quality of the extracted keyphrases (2 pts) as well as the code you write and your explanation of your system design (2 pts). If you fail to implement the NC-Value algorithm, you will get at most 3 pts.

Notes:

- The warm-up exercise in README will not be graded, though following the exercises and instructions will save you a great deal of time.
- Please do not submit your data directories and Stanford CoreNLP library.
- Code must be submitted with the assignment for purposes of plagiarism detection.

NLP Written

Q2 [1.5 pts]. Regular Expressions and Finite-State Automata.

Design an FSA that recognizes simple date expressions like 5 March 2012. You should include all such "absolute" dates (e.g. *not* the ones relative to the current day like *the day after tomorrow*) (1 pts). In order to avoid drawing too many arcs, you should use a word class for all month names (e.g. MON → Jaunary|Feburary|...). The day and year should be characterised by using single digit or character classes of digits such as [0-9]. Therefore, you are not allowed to use one symbol to represent all possible years or all possible days. For simplification, consider only the date format DAY MONTH YEAR, DAY MONTH, MONTH YEAR and YEAR. Write also an equivalent regular expression in your answer sheet (0.5 pts). Please use the notations in http://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html. You could test your regular expression with the following text by using any regular expression tool.

Anthony John "Tony" Abbott (born 4 November 1957) is an Australian politician. Since 2013, he has served as the 28th and current Prime Minister of Australia, and has been the Leader of the Liberal Party since 2009. Abbott is the Member of Parliament representing the Sydney-based Division of Warringah, having first been elected at a 1994 by-election.

The first funding for Google was an August 1998 contribution of US$100,000 from Andy Bechtolsheim, co-founder of Sun Microsystems, given before Google was incorporated.[51]

In 2012, Fortune ranked IBM the No. 2 largest U.S. firm in terms of number of employees,[7] the No. 4 largest in terms of market capitalization, [8] the No. 9 most profitable,[9] and the No. 19 largest firm in terms of revenue.[10]

"One Thing I Know", an unreleased track co-written by Bunton for Free Me, was recorded by another 19 Management act, S Club 8, for their album Sundown.

The musical opened at the West End's Piccadilly Theatre on 11 December.

Q3 [1 pts]. Context-Free Grammars

Noun phrases (*NPs*) like *my uncle's bicycle* or *Companies' workers* are called **possessive** noun phrases. A possessive noun phrase can be modelled by treating the sub-NP like *uncle's* as a determiner of the following head noun. Add grammar rules for English possessives to the rule set given in the lecture slide 10 of *constituency Parsing.* You may treat *'s* and *'* as if they are separate words. Use *PNP* to denote the nonterminal of possessive noun phrase. You should modify the rules in the slides to distinguish between singular and plural nouns. Your grammatical rules should at least cover the following cases:

- *my uncle's bicycle*
- *Companies' workers*
- *a car*
- *his books*
- *the bus stop*

Q4 [0.5 pts]. CKY Parsing
Although the CKY algorithm given in the lecture slides requires that the context-free grammars are converted into Chomsky Normal Form (CNF), in practice, people often use a generalised CKY algorithms which handles the unit productions directly rather than converting them to CNF. (Recall that unit productions take the form of $A \to B$, where both A and B are nonterminal symbols.) Please modify the pseudocode of the CKY algorithm given in the lecture slides (constituency parsing, P. 48) so that it can accept grammars that contain unit productions.

Q5 [1 pts]. Transition-based Dependency Parsing.

We have learned in the lecture that Nivre's parsing algorithm has four parsing actions (**Left-Arc**, **Right-Arc**, **Reduce**, **Shift**).

- Which parsing actions ensure that the resulting dependency graphs do not contain cycles? Please explain the reasons. (0.5 pts)
- Why is the time complexity of Nivre's algorithm linear $O(n)$? (0.5 pts)

References

[1] Frantzi, KaterinaT. and Ananiadou, Sophia and Tsujii, Junichi. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. *International Journal on Digital Libraries 3.2 (2000): 115-130.*