COMP4650/6490: Document Analysis

Assignment 4: NLP

Main details:

Maximum marks:	10
Programming language:	Python (only)
Assignment questions:	Post to the Wattle Discussion forum
Grading lab of Q1 :	Week of 9th October in your respective labs (No late days allow
Submission deadline of Q2 - Q5:	**Friday 13 October, 11:59pm** (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 Python 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 `assign3_yourname.zip` with a `README.txt` stating what each directory is for and how to run your code.

NLP Programming

Q1 [5 pts]. News Title Classification (checked in lab).

Your task is to implement the fastText algorithm for text classification [1] and apply it to the *Assignment 3 Q1 Data* posted to Wattle. You do not need to implement the *Hierarchical softmax* (Section 2.1) and N-gram features (Section 2.2). You should use Tensorflow and NLTK libraries.

In lab you will be asked to run your code on the data sets included in your assignment. The code should provide a list of sentence categories for all sentences in the test set. The grader will both inspect the quality of the assigned categories (2.5 pts) as well as the code you write and your explanation of your system design (2.5 pts).

Notes:

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

NLP Written

Q2 [1 pts].
We are given the following corpus:
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green apples and Sam </s>

If we use Kneser Ney smoothing for Bigrams (slide 25), what is P(Sam|am)? Include <s> (begin of a sentence) and </s> (end of a sentence) in your counts just like any other tokens.

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. Find all errors in the following context-free grammar rules for English possessives and correct them.

$PRP\$ \rightarrow my|his|her|its$
$PNP \rightarrow nounEndWithS\ '$
$Nominal \rightarrow PNP$
$Det\ Nominal \rightarrow Det\ Noun$
$Nominal \rightarrow PRP\$\ Nominal$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Noun$

All grammatical rules should at least cover the following cases:

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

Q4 [1 pts]. Word Embeddings
Suppose we are given a training corpus for learning word embeddings and a test dataset for evaluating these embeddings with the word analogy task. There are m words that appear only in the test dataset, but not in the training corpus. Could you

provide a solution to deal with the unseen words while learning word embeddings on the training corpus?

Q5 [2 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**).

- (1pts) Explain the reasons why
  - **Left-Arc** needs to remove the topmost element from the stack.
  - **Right-Arc** needs to add the leftmost element of the queue to the stack.
- (1 pts) We have learned that the time complexity of Nivre's algorithm is $O(n)$. What is then its space complexity? Please explain the reasons.

References

[1] Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759.*, 2016