CS 5316: ASSIGNMENT 4

* It is mandatory to provide a single word document with all the solutions and clear screenshots of code along with descriptive comments. It is your duty to explain what you have done as clearly as possible in your solution document.
* Ambiguous answers and unrelated submission documents will be heavily penalized.
* You also need to provide your code which shall be used for plagiarism checks. ( make sure its clean and readable.
* You will use python 2/3 for this assignment. (NLTK has most packages that you would need)
* You will be marked on clarity and accurate solutions. If required, a viva might be arranged.
* Submission deadline: 16th April

## Part 1: POS tagging

**Question 1:**

1. Write down the name of 3 POS tag annotated datasets for the English language.

*“The Cambridge Analytica scandal is more than a “breach,” as Facebook executives have*[*defined it*](http://www.bbc.com/news/world-us-canada-43494337)*. It exemplifies the possibility of using online data to algorithmically predict and influence human behavior in a manner that circumvents users’ awareness of such influence. Using an intermediary app, Cambridge Analytica was able to harvest large data volumes—over 50 million raw profiles—and use big data analytics to create psychographic profiles in order to subsequently target users with customized digital ads and other manipulative information. According to some*[*observers*](https://www.theguardian.com/uk-news/2018/mar/20/cambridge-analytica-execs-boast-of-role-in-getting-trump-elected)*, this massive data analytics tactic might have been used to purposively swing election campaigns around the world. The reports are still incomplete and more is likely to come to light in the next days.”*

1. Use a pre-trained pos tagger to identify tags for the above text. Which tagger did you use? Comment on your result highlighting inaccuracies and strategies for improving performance. (you will be marked accordingly)
2. Train a hmm using nltk on the penn-tree bank dataset and tag the above text. Comment on your result highlighting inaccuracies and strategies for improving performance.

**Question 2:**

1. Write briefly on the evaluation problem, decoding problem and the learning problem in HMMs. You should describe what the problem is and what the solution is. Mention the name of any algorithms used to solve the problem and provide an intuitive explanation of the algorithm. Use as less mathematical notation as possible and try writing in your own words.
2. Consider a Hidden Markov Model with three hidden states: N (noun), V (verb) and O (other). Let all transitions between states be equally probable. Consider the following possible outputs:

N: mimsy | borogoves

V: were | borogoves

O: All | mimsy | the

Let all these outputs be also equiprobable.

Consider the sentence "All mimsy were the borogoves”. Find the most probable tag sequence

**Question 3:**

1. Explain how you would use a deep learning model for POS tagging. Also comment on the input representation to the model.

## Part 2: Entity Recognition

**Question 4:**

1. Study the CONLL-2003 NER shared-task dataset. Provide a screenshot of the ‘head’ of the file and explain how it can be interpreted.

“Atlas Honda is expected to achieve sales of 1.1 million units by end of its financial year ending March 31, while it aims to hit sales of 1.3m bikes in its next financial year, a Honda dealer said. “

1. Use a pre-trained NER model to identify the named entities in the above sentence. (use NLTK)
2. Provide a brief summary (6 lines) of what a bidirectional LSTM-CRF model (paper <https://arxiv.org/pdf/1508.01991.pdf>) is.
3. Provide a detailed outline (point wise) of how to use such a model for NER tagging. More specifically, mention the dataset to use, the input representation of words, the ideal number of hyper parameters, the cost function and the overall architecture of the deep learning model.
4. Find a pre-trained deep learning model in python (any package but mention the package name in this document and provide screenshots) and use it to on the test sentence provided above.

The End ☺