**CSCS 366 – Introduction to Natural Language Processing**

Assignment 3 Deadline: May 5th 11:59PM

**Part 1: POS Tagging**

*“The Cambridge Analytica scandal is more than a “breach,” as Facebook executives have defined it. 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, 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. Using NLTK, use pretrained POS tagger to identify tags for the above text using brown tagset. Comment on your result, highlighting inaccuracies in the results.
2. Train a HMM using NLTK on the Penn Treebank dataset and tag the above text. Comment on your results highlighting any improvements and inaccuracies in the results.

**Part 2: Named Entity Recognition**

*“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 for this task.
2. Use a pre-trained NER model to identify the named entities in the above sentence. Use Spacy for this task.
3. Use spacy to train NER model on CONLL-2003 NER shared task dataset. Use this custom trained model to identify the named entities in the above sentence. Evaluate the trained model in terms of Precision, Recall, and F1 score on the validation and test split of CONLL-2003 NER dataset and report these metrics.

**Submission Guidelines:**

* Submit Jupyter Notebook with all four outputs above.
* Show output of all above 5 tasks in the notebook.
* Submit on Moodle.
* Submit all the files so that I can run your code and reproduce the results.