IDC4022C Module 9 Project: Named Entity Recognition (NER) Using NLP

The “idc4022cMod9.ipynb” notebook in this repository contains code which demonstrates Named Entity Recognition (NER) using the SpaCy NLP library. The data file “idc4022cMod9Data.csv” contains approximately 500 lines of news article data extracted from the “articles” series of datasets described in the NER section of the “Machine Learning for Finance” text’s required reading for this module. An additional dataset “npr.csv” is provided which is used for additional analysis in this assignment, refer to the instructions below for how to use it.

Upload the notebook to Google colab, then upload the CSV file. Depending on where you store the data file you may need to change the path in the notebook where the CSV file is read at the beginning of the notebook.

Examine and run each cell in the notebook and review the output. Write a review of the methods used and results obtained, covering the following topics:

- Briefly introduce the purpose of the notebook, including the primary objective of the analysis
- Describe the data preprocessing required, the analysis steps, and the visualizations rendered (note that these steps follow the associated NER section of the module’s required reading)
- Describe key takeaways of the analysis
- After executing the notebook as provided, add the npr.csv data to the original CSV dataset. Note that you will need to resequence the record number field (the unnamed field in the first column) in this data to continue the original dataset’s sequencing.
- Upload the revised dataset and rerun the notebook, making any changes necessary to complete execution and include the additional data. Describe changes you made and observations about any changes in the results.
- Using the “Machine Learning for Finance” text reading, suggest potential improvements or further analyses that could be conducted based on the notebook’s results (some suggestions can be found in the NER section)

Upload the executed notebook, revised dataset, and your review document to the GitHub classroom repo. Your review should be at least 2 pages long to provide sufficient detail necessary to earn full credit on this assignment.