# CS-4420: Natural Language Processing

Assignment 1

## Web Scraping and Essential NLP Methods

1. **Gathering Data**
   - (a) Build a web scraper to aggregate at least 100 news articles about your selected topic (pick topics on the sheet shared on Classroom) over a time period of one month. These can come from a single source, or from a news aggregator/search engine like Google or MSN. Try to ensure the same article is not scraped multiple times. You are free to use any scraping libraries available, though the most commonly used libraries are `BeautifulSoup`, `Selenium`, and `Scrapy`. Store the title of the article, the link to the page, and the text scraped from the article in a `Pandas` DataFrame.
   - (b) Briefly explain how your scraper works, and provide reasoning for your design choices. (10)
2. **Essential NLP Tasks**
   - (a) Perform Part-of-Speech (POS) tagging on the text of each article using `Spacy`. Store the results in a new column in your DataFrame.
   - (b) Perform Named Entity Recognition (NER) on the text of each article using `Spacy`. Store the identified entities (e.g. persons, organizations, locations) in a new column in your DataFrame.
   - (c) Perform Coreference Resolution on the text of each article using Stanford's `CoreNLP` library or a similar tool (`Spacy` does not natively support this method). Store the resolved coreferences in a new column in your DataFrame.
   - (d) Identify and implement (on your own or using a library) an Entity Resolution algorithm to match entities across articles. Save the resolved entities in a new column in your DataFrame. In your report, discuss which Entity Resolution algorithm you chose and why. (10)
3. **Evaluation and Analysis**
   - (a) Manually annotate a small subset (30 documents) of your collection of articles to evaluate the performance of the **NER** model. You can use metrics such as accuracy, precision, recall, and F1-score. Discuss the results of your evaluation, and talk about what each evaluation metric indicates.
   - (b) Annotate 15 articles to evaluate the performance of the **Coreference Resolution** model. Discuss the results.
   - (c) Analyze the outputs from the NER model and extract meaningful insights from the identified entities and their distributions. What additional insights would you like to obtain that NER alone cannot provide? (10)

**Submission Instructions:** Save your DataFrame as a `csv`, and submit it in a **zip folder** along with your code and a PDF report with your answers to questions `1b`, `2d`, `3a`, `3b`, and `3c`.