Lab 2 Documentation

Michael Klein & Anthony Introne

Our activities are packaged into a Python web scraper and AWS mapper/reducer. First steps involve running the Python scripts for Twitter, NYT, etc. These scripts scrape a large amount of tweets and NYT article links that are then scraped for the articles themselves. The scripts run for 5 different kinds of data based broadly on sports. These texts are saved locally and cleaned of HTML tags using Beautiful Soup. With the data cleaned, the local files are sent to our AWS server where the Mappers/Reducers are hosted that reduce our files to all words that occur in the article and the number of times they occur. These files are transferred back locally and sorted by the most common occurring words for clarity. These files are then used to create word clouds deployed on our local server through a webpage displaying the most common occurring words in a common word cloud format. Word clouds are broken into sections for Twitter and NYT as well as the different subtopics under the general sports term (esports, NBA, NCAA, NHL, & NFL).

Our process can very easily be repeated locally for different results. By changing the search terms in the scraping scripts, different data instead of sports can be obtained. From here, the process is generally the same where one could use the cleaning script to remove HTML tags, run the mapper/reducer scripts on AWS, and then add the sorted and reduced files to the website code to produce the word clouds.

Block Diagram:

