Progress Report

1) Which tasks have been completed?

We have completed the code to scrape news content from "www.bbc.com/news" and "www.cnn.com". For each site, we created two separate files to store the information in each day. The first file stores the news headlines and news URLs in each single line. The second file stores the news headlines and their article contents. This facilitates the annotation work and provides the sources for users to trace where the news from. We also provide documentation for the data files.

2) Which tasks are pending?

We still haven't completed all the code to get information from other websites. We also need to manually annotate the news dataset we create.

3) Are you facing any challenges?

When getting the URL of each news, it is a challenge to filter out non-text news URLs, such as live streaming and video URLs depending on how that news websites are structured. Most news websites use dynamic structures and requires user authentication for accessing their contents, which increases more difficulty for scraping their news articles.