Kyle Cielencki

Luis Cerrilla

Will Pumphey

News ETL Project

The purpose of this project is to **E**xtract information from three main news websites, **T**ransform the scraped data into a structured data frame and **L**oad this data frame into a SQL data base, to be able to have the latest news from this tree main websites on one single table.

Extract

Jupyter Notebook’s, Splinter and Beautiful Soup were our main tools to be able to extract the latest news, including headlines, date and URL’s from CNN, NPR and CNBC. We chose these websites because we think they are some of the most reliable, while providing different takes on potentially similar stories.

We also made sure that all the websites had the same information that we were going to extract like an article header, date and URL. This prevented having NaN’s on the database and having all the information of each article.

We used similar methods for the scrapping, but encountered some differences in the layouts of the websites.

Transform

We only transformed all the data that we scraped into a data frame to make sure that we could load this data into the data base in a proper way. Also to be sure that all the data frames had the same headers as the data base and being able to push all the extracted data into the SQL data base. We also made some sanity checks querying the data base to make sure it was correctly pushed.

Load

The data base is a SQL cloud based that consist in four different tables, connected between different ID’s. The first table is the only dynamic table, this means, the information will constantly be updating, it contains all the articles headers, URL’s, date published, and tag/site ID’s. The second table is more focused on the website, it has the site ID, main URL and site name. The other table has the tag number and tag ID. And the last table is referring all the keywords to categorize the articles. All table have a column indicating the last time a row was updated.

Future work and limitations.

The first thing we would do after this work is expose the database on a front-end table, like a website. We spent a lot of time working on getting the articles URLs

Our first limitation was the number of keywords that we could assign into each type of news, we had to limit it into a few words. That is something we can improve and reduce the amount of NaN’s values on the type of news.

One of the limitations we had, was that one of the websites did not had a daytime on the same format as the other two. Because of this issue, we had to change the table date column from a date type into a varchar type.

One topic for future works was the differences in t=website design. Each of the three websites had different styles, labels, bins, and layouts. Some had information that was more hidden than others, while some had constraints on what information we could use in general.