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.

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.

Transform

We only transform 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.

Load

The data base is a SQL cloud based that consist in three 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 headers, articles URL’s, date published, and tag and site ID’s. The second table is more focused on the website, it has the site ID, main URL and site name. And the last table has the tag number and tag ID. All table have a column indicating the last time a row was updated.