NEWS APPS CRAWLER

Developed by:
Jaya Krishna Vadlamudi
Leela Chakravarthy

PROJECT IDEA

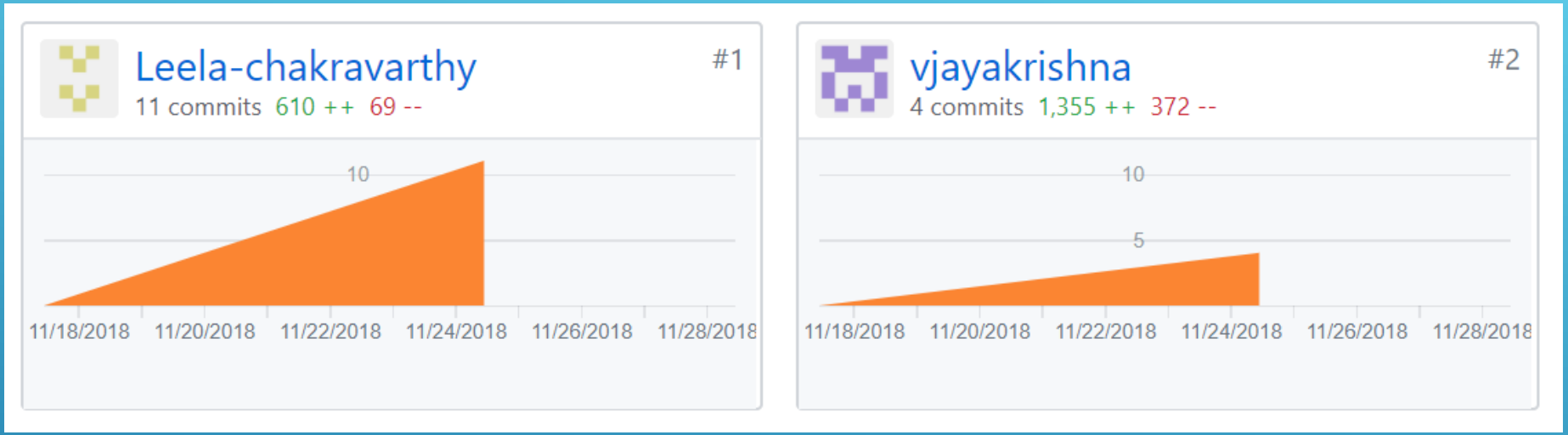
- Our project idea is to develop a web crawler that crawls through various news websites to extract content.
- Display the news content in a custom-designed GUI.
- Crawler was implemented using **HTMLParser**, from the python class library ‘html.parser’.
- Used four news websites for extracting news content - **CNN, NYTimes, NBC news and FOX news.**

HTML PARSER

- A basic web crawler class included in the python class library ‘html.parser’
- Has basic parsing functionality for html content.
- Has methods for handling tags and data of a html page which were used for extracting the news headlines in our project.

- Each news website had its own html format.
- Hence, we implemented separate crawlers for the news websites (like NewsParser_CNN, NewsParser_NBC etc.)
- For each website, we extracted news headlines of various types such as Politics, Entertainment, World, Business.
- Some of them had different html formats which were implemented in separate methods.

CHALLENGES FACED



GITHUB

Total lines of code: 1170

DEMO