Tools

* Scraping and Saving Data:
* Puppeteer => is a Node library which provides a high-level API to control Chrome or Chromium over the DevTools Protocol. One of its main purposes is to scrap and crawl web pages and generate pre-rendered content.
* Firebase – Realtime DB => is a cloud-hosted database where data is stored as JSON and synchronized in realtime to every connected client.
* Node-Geocoder => is a node.js library for geocoding and reverse geocoding for extracting latitude & longitude values.
* BeautifulSoup =>
* NLTK – NLP =>
* Web Platform:
* EJS => Stands for Embedded JavaScript. It’s a simple template engine that lets you generate HTML markup with plain javascript.
* Node.js => is a cross-platform, JavaScript runtime environment that executes JavaScript code outside a web browser.
* Express.js => is Fast, unopinionated and modular web framework for Node.js that lets you handle multiple different http requests and is used mainly to build RESTful APIs for web and mobile applications.
* Mapbox => is an open source mapping platform for custom designed maps powered by vector tiles and WebGL.
* Bootstrap => is an open source CSS framework directed at responsive, mobile-first front-end web development. It contains CSS and JavaScript-based design templates.
* Firebase – Authentication => is a firebase service that provide backend SDK to authenticate users to your app.
* Morgan => is HTTP request logger middleware for node.js application.
* Socket.io => is a JavaScript library for realtime web applications. It enables realtime, bi-directional communication between web clients and servers.
* Axios => is a promise-based HTTP client for the browser and node.js.
* Web Speech API => provide speech-input and text-to-speech output features in a browser that are typically not available when using standard speech-recognition or screen-reader software.
* Font Awesome & Feather Icons.

Main Phases (Tasks)

* News Extraction – Scraping:

In this step, we first gathered news articles from several trusted news providing websites such as Globalvoices – BBC – CNN using software tools such as puppeteer and beautiful-soup that help us extracting content from web pages. We are extracting content data we are concerned about such as article title, summary, description, category, location if available, tags and imageURL. Then we structure the scrapped data in a similar way in order to be parsed to the following step.

* Localization:

When the scraped data are ready, we are parsing them to various localization algorithms. So each news article will be tagged with its own list of location, latitude, longitude and country-code values. We use the title and description of each article to figure out the related location using NLTK tool that is based on NLP(Natural Language Processing) technology. Then after locations have been extracted and inserted, we use Node Geo-Coder library to resolve the coordinates (longitude - latitude) and country-code values based on each article’s location.

* Saving Date:

Once the location and coordinates related information have been linked to each article, we start using firebase node.js SDK to save the news we scrap in the firebase realtime-Database to be used in web and mobile application interfaces. The news articles are saved in JSON format in firebase, since firebase provide a RESTful API for the saved data so we can hit this API endpoints and retrieve the data we want based on the request parameters we provide. Also we use the firebase-admin node.js library to filter data using the provided methods and limiting the retrieved articles to a specific number.

* Updating Data:

We set the scraping system to update the database periodically every news release.

* Map-Based Searching: