Larry on Fire

(🐦🔛🔥)

Group Members

Núria Adell Raventós – nuriaadell

Sergio Olalla Ubierna – sergiou

Jonas Heim – jonasheim

Pavan Prathuru - pavanprathuru

Table of Contents

Project Requirements Compliance

Below is a quick description of how this project complies with every requirement:

- Two sources of data
  - ✓ Twitter API and CalFire
- Data analysis component
  - ✓ Four visualizations for the data analysis
- Specific components built by each team member
  - ✓ Refer to the relevant section
- Visual or textual output
  - ✓ Refer to the relevant section
- Each distinct component is a subpackage
  - ✓ Refer to the relevant section
- Virtual environment and requirements.txt
  - ✓ Refer to the relevant section
- README.md
  - ✓ Refer to the Git repository
- Project paper
  - ✓ Et voilà, here it is

Project Overview

The goal of this project is to give the user the ability to visualize when people tweet about the California wildfires (when they care about them) and what they tweet about (what their concerns are). Our application does this by allowing users to analyze the evolution over time (2015-2020) while offering several filters to interact with the data.

More precisely, it allows a user to analyze the social conversations taking place on Twitter around the California wildfires with several components that:

1. Map the location of the most significant wildfires
2. Compare the Twitter conversation volume with the wildfire intensity
3. Showcase the main topics of the conversations

In detail, we have produced the following interactive visualizations:

1. Map locating the wildfires over time
2. Line chart plotting the number of tweets and acres burned per week
3. Wordcloud with the most frequent words found in the tweets
4. Topic modeling from the tweets

For each of these components, the user can select a year between 2015 and 2020 and decide to take into account the entire year or only the fire season. For the latter two visualizations, one can also select whether to only analyze in-state/out-of-state or all tweets.

Project Structure

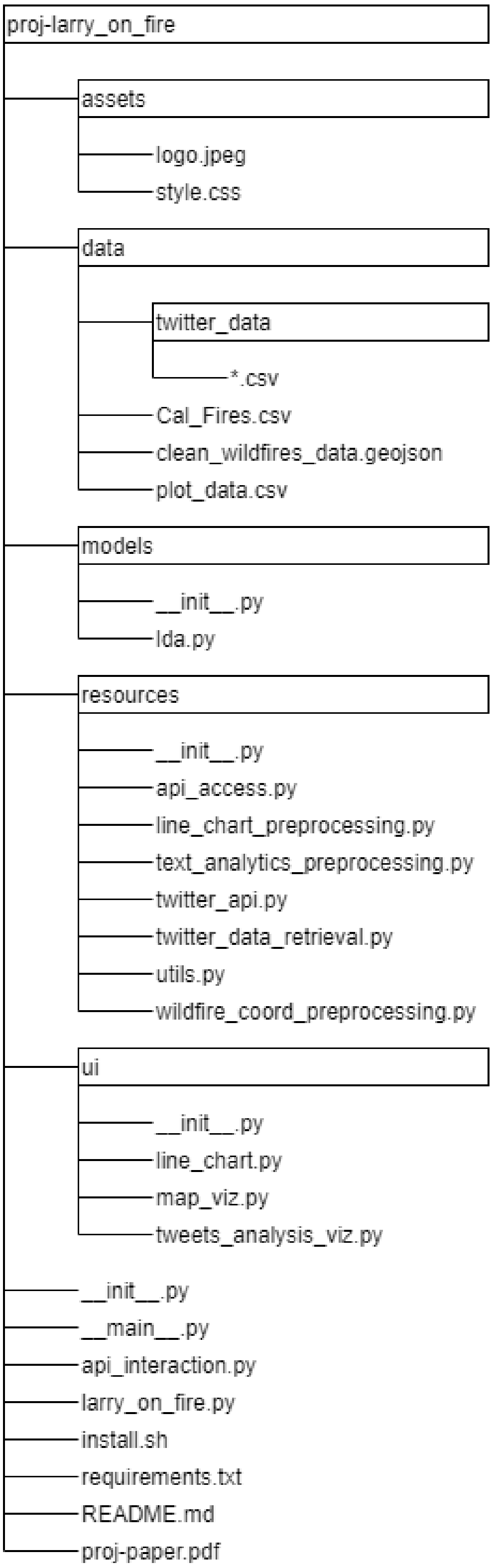
The application is organized in several submodules as seen on the diagram aside.

- ***assets:*** dash interface settings
- ***data:*** CSVs and GEOJSONs containing the data retrieved from the Twitter API as well as the wildfire data downloaded from the CalFire website. It also stores certain intermediate data so that the preprocessing does not have to take place several times
- ***models:*** contains the Latent Dirichlet Allocation (LDA) analysis
- ***resources:*** contains the helper functions to retrieve the data from the Twitter API and preprocess the data for the various visualizations
- ***ui:*** user interface module that contains the functions to create the visualizations

The top-level directory also contains several important files in addition to the ***README.md*** and ***proj-paper.pdf***:

- ***larry_on_fire.py***: the dash interface of the application
- ***install.sh***: bash script to set up the virtual environment
- ***requirements.txt***: containing the libraries required for the application

Behind the scene, ***larry_on_fire.py*** (dash), will call the components that are located in ***models*** and ***ui*** subpackages. Each of these components creates the according visualization by calling helper functions (mostly to pre-process the data) that are located in the ***resources*** folder.



Code Responsibilities

The responsibilities were distributed as follows:

- Data Retrieval and initial processing:
  - Twitter API: Jonas, Pavan, Sergio
  - Wildfire Data: Nuria
- Analysis:
  - Wildfire Map: Nuria
  - Line Chart: Jonas
  - Wordcloud: Nuria
  - Topic Analysis: Sergio
- Deployment:
  - Interactive UI (dash): Pavan

Overall, we do feel like everyone contributed in an equal manner. The Twitter API took longer than expected to figure out, which is why three of us ended up working on it. Nuria, Jonas, and Sergio each worked on separate visualizations while Pavan took on the task of putting it all together using Dash.

Interacting with the Software

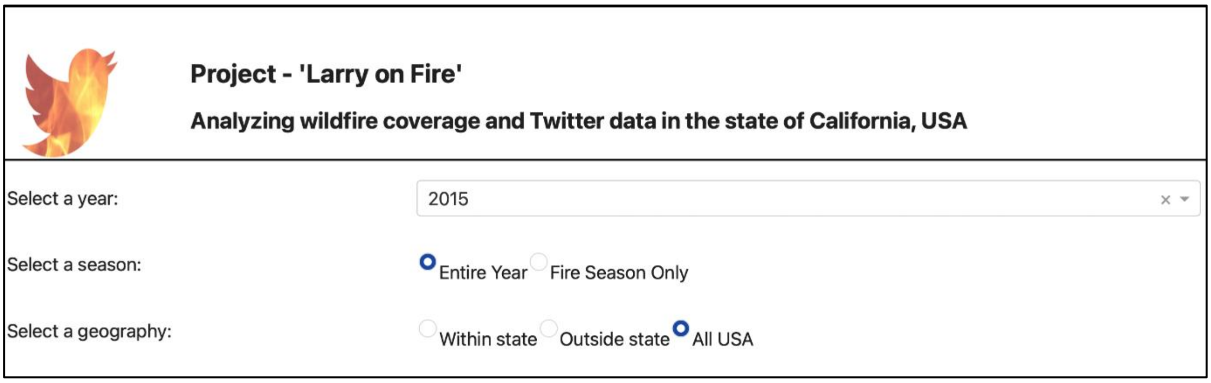
Interacting with the Application

To interact with the application, one first needs to create the virtual environment with ***`bash install.sh`*** in the top-level directory. Once the code has run, enter the virtual environment with ***`source virtual_larry/bin/activate`***. All that is then left to do is to launch the dash server with ***`ipython3 larry_on_fire.py`*** and click on the link displayed in the output. This will open a browser window to our application.

Once the three steps above are executed, one can interact with the application on the web browser by toggling three settings:

- Year (dropdown menu)
- Season (button)
- Geography (button)

The prior two dropdowns apply to all four visualizations, while the last filter only applies to the wordcloud and LDA. This allows the user to (1) locate the fires, (2) look at the tweet intensity compared to the wildfire intensity, and (3) get an overview of the discussed topics.



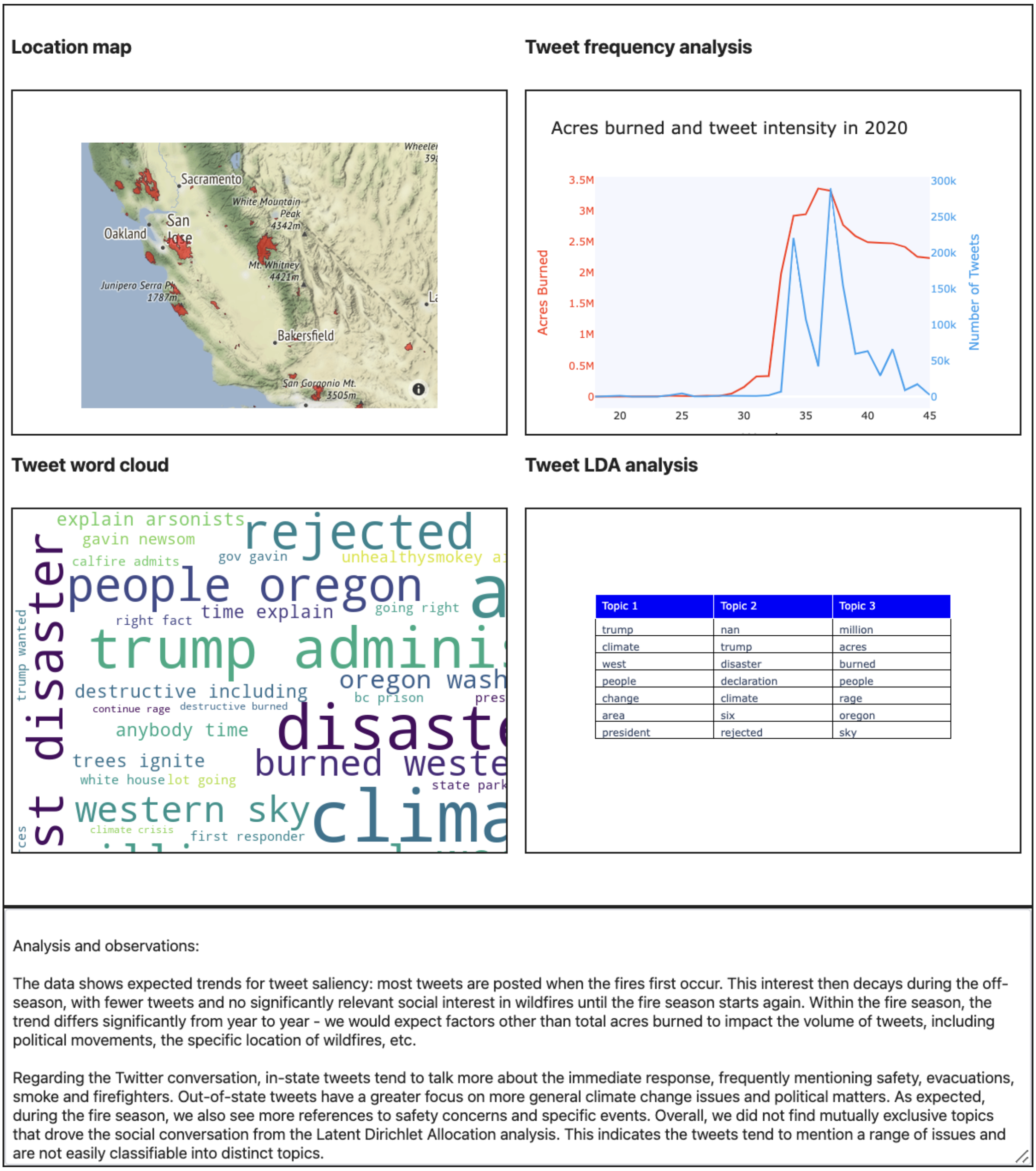
Project - 'Larry on Fire'

Analyzing wildfire coverage and Twitter data in the state of California, USA

Select a year: 2015

Select a season: ● Entire Year ○ Fire Season Only

Select a geography: ○ Within state ○ Outside state ● All USA



Location map

Tweet frequency analysis

Tweet word cloud

Tweet LDA analysis

Topic 1	Topic 2	Topic 3
trump	nan	million
climate	trump	acres
west	disaster	burned
people	declaration	people
change	climate	rage
area	six	oregon
president	rejected	sky

Analysis and observations:

The data shows expected trends for tweet saliency: most tweets are posted when the fires first occur. This interest then decays during the off-season, with fewer tweets and no significantly relevant social interest in wildfires until the fire season starts again. Within the fire season, the trend differs significantly from year to year - we would expect factors other than total acres burned to impact the volume of tweets, including political movements, the specific location of wildfires, etc.

Regarding the Twitter conversation, in-state tweets tend to talk more about the immediate response, frequently mentioning safety, evacuations, smoke and firefighters. Out-of-state tweets have a greater focus on more general climate change issues and political matters. As expected, during the fire season, we also see more references to safety concerns and specific events. Overall, we did not find mutually exclusive topics that drove the social conversation from the Latent Dirichlet Allocation analysis. This indicates the tweets tend to mention a range of issues and are not easily classifiable into distinct topics.

Interacting with the API

The current API call relies on one of our Twitter developer accounts. This account has Academic Access to the Twitter API, which gives us the ability to retrieve 10 million queries a month without any additional costs.

To interact with the API, run the command ***ipython3 api_interaction.py*** from the command line at the top-level directory. It will ask for two inputs; a ***start date*** and an ***end date***. This program will run a query on the Twitter API within the entered range. The inputs need to be in the following format; ***YYYY-MM-DD***.

The output is stored in a CSV-file at (overwriting previous queries stored in the same path) ***proj-larry_on_fire/data/twitter_data/simulation.csv***.

Please note that queries that include full months during the summer might take long since we see an increased Twitter activity during the wildfires season. As a suggestion, a ***2015-03-01*** to ***2015-04-01*** query will be completed in a few seconds.

Goals and Accomplishments of the Project

The project successfully provides a platform to better understand the conversations around the California wildfires on Twitter. However, there are two key objectives we had to reevaluate.

Initially, we wanted to visualize the fire location and the tweet density around California (two overlapping layers) on a map. However, we found that only very few tweets contained reliable, if any, localization. Thus, we had to abandon the idea of mapping the tweet density and instead made a line chart showing the tweet intensity compared to the wildfire intensity. We were still able to successfully map the wildfires on their own.

Additionally, we wanted to identify the main topics of conversation across all tweets to then classify every tweet with a certain topic. The goal was to see differences in topics between years and geographies. However, we did not find relevant key topics in the general model to use for classification. Therefore, instead of having a general LDA model to then apply it to specific tweets, we decided to run specific LDA models for each user query (filtering on year, fire season, and in/outside California). This way, the user can interactively see the differences in key topics across these areas of interest.