Forest Fire Analysis Project

CMPT 732 Project Proposal

Aidan Vickars (301255806)

Rishabh Kaushal (301443177)

Anant Sunilam Awasthy (301467798)

Description

Due to global warming, forest fires are becoming an increasingly common and dangerous norm in today's world. This was particularly evident in British Columbia where temperatures reached over 40^oC that kicked off one of the worst fire seasons on record. This resulted in hundred of fires across the province that displayed thousands of people, causes millions of dollars in property damage, and destroyed huge swarths of forest. While this problem is particularly local to us it is not limited to just British Columbia and is a major problem around the world. Therefore, the aim of this project is to analyze factors in relation to forest fires over time. This includes air quality, forest density and identification of high-risk areas. Furthermore, if time permits, we are aiming to leverage our analysis and ETL code into a live pipeline to develop a machine learning model that identifies areas that may be or are very likely to experience a forest fire.

Deliverables

- Develop a scalable pipeline using PySpark. Some of the datasets that will be used are:
  - Air quality Monitoring Data. This dataset contains hourly air quality data from across the province. https://catalogue.data.gov.bc.ca/dataset/air-quality-monitoring-verified-hourly-data\
  - Historical forest fires and perimeter coordinates from 1929 – 2020. https://catalogue.data.gov.bc.ca/dataset/fire-perimeters-historical
  - Forest fires and perimeter coordinates for 2021. https://catalogue.data.gov.bc.ca/dataset/fire-perimeters-current
  - API containing weather data both historical, current, and forecasted. Has a free student bracket that should be sufficient for our needs. https://openweathermap.org/price#history
  - Data from the Global Historical Climatology Network as used in class. https://www.ncei.noaa.gov/products/land-based-station.
  - Forest information for the entire province. https://catalogue.data.gov.bc.ca/dataset/results-forest-cover-inventory
- Analysis (Note: this is purposely vague, as we expect this to change once we get into the data)
  - Analysis of high-risk areas for forest fires.
  - Analysis of air quality changes by region over time with respect to forest fires.
  - Analysis of forest fire frequency in relation to forest characteristics.
- Dashboard/Final Visualization
  - Using either PowerBI or Dash, depending on how well each technology fits to the visualization.

Nice-to-haves

- Implement the above deliverables in AWS in a live system. This will depend on the ability to leverage the APIs as well as cost having the system running as well.
  - Using Kafka / Cassandra for storing and updating live hourly Air Monitoring and Weather Data.
- Develop a machine learning model that uses the above pipeline as input to predict the risk of a forest fire in a specific region.

Application to Big-Data

The application to big data for this project is that we are aiming to develop it in such a way that if we had all forest fire data, weather data etc. for all regions that are effected by forest fires and if the granularity of the data increases (i.e. instead of hourly data, we have minute by minute data), our pipeline and model could easily be scaled up to process this large amount of data.

Expected Challenges

- Figuring out how to partition British Columbia into regions. This will require some investigation into the locations of the weather/air stations where the data is coming from. Using this we can develop a methodology to partition British Columbia into regions. Attention will be paid to develop a scalable methodology that can automatically be applied other regions.
- Figuring out how to visualize the results. Plotting custom regions may be very finicky and could require significant development time.