BLOG POST
December 10, 2021

California Wilfire Detection Data Science Project

In this blog post we are gonna look at historical wildfire data to analyze which places in California are under fire threat and how deadly they can get by areas. We will build a dataset for training from two different data sources: wildfire data and weather api. Then build a model to detect wildfires before they occur.

MOTIVATION – The problem we are trying to solve is the constant wildfires in California. As we are aware, there have been wildfires raging across California for several years now, and it has caused not just racked up massive cost in damages, but also destroyed several forests, and killed animals, and human beings. There is no question that this occurs due to Climate Change and rising temperatures which results in droughts that cause such fires to not only spread quickly but also make it hard to put out. Another example of this is Australia, which had such massive wildfires that persisted for weeks and even caused emergency actions from several nations around the world. Therefore, it would be vital in this day and age to determine if it is possible for a potential wildfire to take place so that the necessary precautions are taken to ensure the safety of the people, animals, and land.

Photo an Astronaut in space took of the California wildfires spewing smoke

DATA – The core of the data is coming from Kaggle. We will be downloading the dataset which lists all the wildfires from 2013 to 2019 in California. One of the best things about this dataset is it lists specifically which county the wildfire occurred and also how many acres it burned. These are very good details as they make it easier to get the weather information of these places much faster. There is also quite a lot of additional information such as injuries, structures damaged, personnel involved, etc. This makes this dataset quite useful for not just this research but several others. We will be using World Weather Online to get the weather information at the place on the day the fire started and the week before. We use this weather information data gathered to build, train, and test our model.

Kaggle dataset

World Weather Online API

CLEANING – We performed the cleaning process on the raw datasets which we have divided into wildfire data named **California_Fire_Cleaned.csv**, weather data named **weather_data.csv**, and merged dataset **cleaned.csv**. In order to create a reliable dataset, we have identified unnecessary columns, dropped them and removed missing and erroneous data. We believe that the downsized dataset in the name of **cleaned.csv** will help us generate more accuate results while training our machine learing model.

Firstly cleaning the fire dataset from Kaggle and generating the cleaned **California_Fire_Cleaned.csv** we used the **etl_california_fire.py** file to clean the data. This python file dropped unncessary columns such as *'SearchDescription', 'SearchKeywords', 'StructuresDamaged', 'StructuresDestroyed'*, etc. These columns although interesting for exploration purposes did not seems to be of great interest when it came to building the model.

Contributers & Writers

Farah Sultana
fsultan001@citymail.cuny.edu

Nishanth Prajith
nprajit000@citymail.cuny.edu