Project report: Australian Fires

Simon J. Binyamin <binyamin@kth.se>

& Ralfs Zangis <zangis@kth.se>

Problem description

Due to the hot weather and other causes, Australia has been a subject of fires over the last two decades. This project aims to utilize NASA satellite data to discern facts related to these unfortunate occurrences and make them conveniently available to the reader.

Tools

The following tools have been used to solve the problem:

- Language: Python
- Data processing- Spark
- Database- Firebase
- Prediction- MLlib
- Frontend- ASP.NET
- Deployment- Azure
- Virtualization- Docker

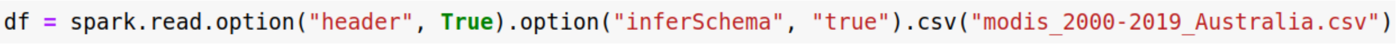
Dataset

The Australian fires dataset that has been used contains satellite information for years from 2000 to 2019. The data is available in a CSV formatted file with more than 5 million entries, accessible following the provided link:
https://www.kaggle.com/gabrielbgutierrez/satellite-data-on-australia-fires.
During the project the focus was placed on but was not limited to, the usage of the following columns and information therein: Latitude, Longitude, Brightness, Acquisition Date, Confidence.

Method

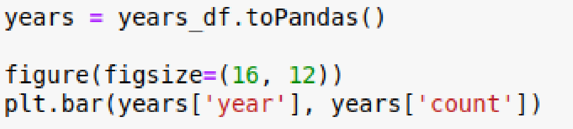
During this, assignment, data was loaded into the data processing solution as a Dataframe, with the schema being inferred from the contents of the file. The processing itself was done using spark, with results displayed as well as saved into variables, for the following plotting of graphs. Later, the generated images were base64 encoded and uploaded (as well as their description and filename) to the firebase, whereupon they could be viewed by anyone who has access to the webpage. Finally, in addition to the already described acts, predictions were made on the number of fires that could be expected each day of the year.

- Reading data into a Dataframe:



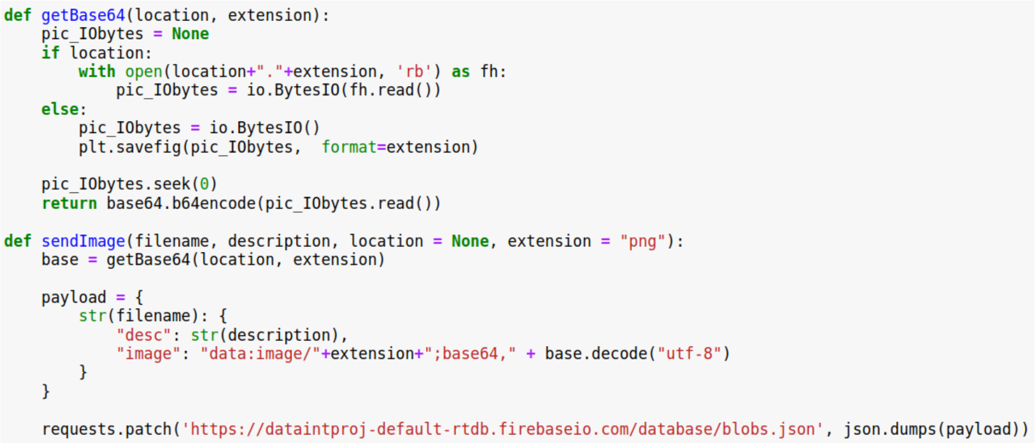
```
df = spark.read.option("header", True).option("inferSchema", "true").csv("modis_2000-2019_Australia.csv")
```

- Result plotting:



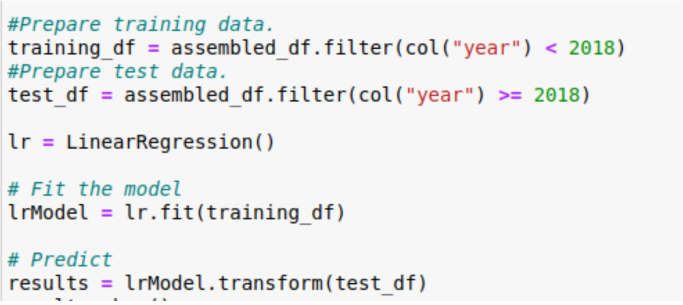
```
years = years_df.toPandas()

figure(figsize=(16, 12))
plt.bar(years['year'], years['count'])
```

- Uploading an image to firebase:



```
def getBase64(location, extension):
    pic_IObytes = None
    if location:
        with open(location+"."+extension, 'rb') as fh:
            pic_IObytes = io.BytesIO(fh.read())
    else:
        pic_IObytes = io.BytesIO()
        plt.savefig(pic_IObytes,  format=extension)

    pic_IObytes.seek(0)
    return base64.b64encode(pic_IObytes.read())

def sendImage(filename, description, location = None, extension = "png"):
    base = getBase64(location, extension)

    payload = {
        str(filename): {
            "desc": str(description),
            "image": "data:image/"+extension+";base64," + base.decode("utf-8")
        }
    }

    requests.patch('https://dataintproj-default-rtdb.firebaseio.com/database/blobs.json', json.dumps(payload))
```

- Basic prediction using linear regression:



```
#Prepare training data.
training_df = assembled_df.filter(col("year") < 2018)
#Prepare test data.
test_df = assembled_df.filter(col("year") >= 2018)

lr = LinearRegression()

# Fit the model
lrModel = lr.fit(training_df)

# Predict
results = lrModel.transform(test_df)
```

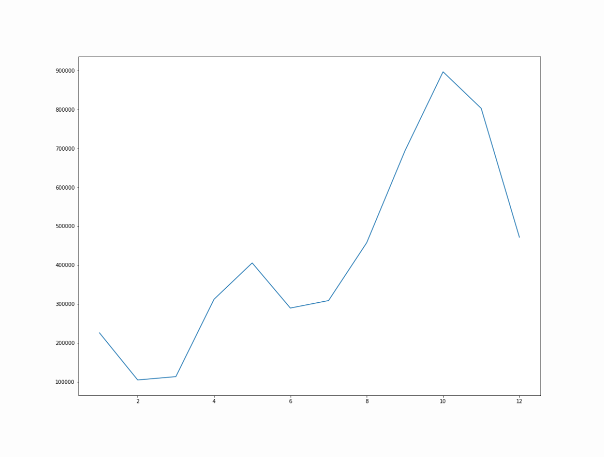
Results

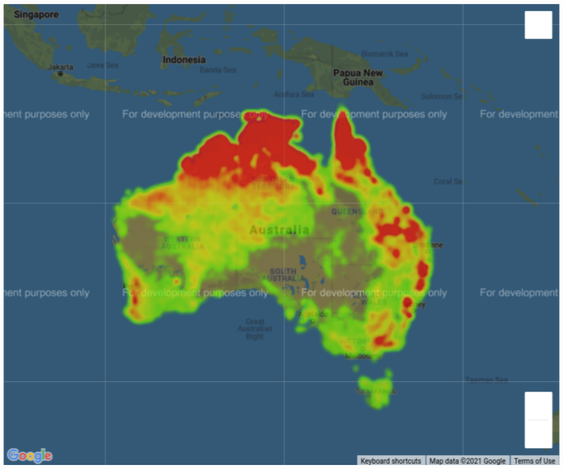
Results of the document will be briefly introduced in the following text. However, the code, as well as the deployed website and corresponding database, are freely available using the following links:

- Website showing results- https://mydataintproj.azurewebsites.net/
- Code repository- https://github.com/bubriks/ID2221_project
- Database- https://dataintproj-default-rtdb.firebaseio.com/database/blobs.json

Most interesting observations:

- It was noticed that Australia is most prone to fires around the time of October (which is spring in the southern hemisphere).
- The generated heat maps showed us that fires are most concentrated in the northern part of the continent (approximately corresponding to the tropical climate zone).





The machine learning algorithm performed as expected, being able to predict the number of fires that could be anticipated on any day of the year. However, due to its basic implementation and lack of information (other than the time when fires occur) the predictions are not of high precision. To improve this, in the future incorporation of data such as weather (temperature, cloud coverage, wing, etc.), human factors (proximity to settlement, holidays, etc.), and more should be considered.

Executing the code

The prerequisite for running the provided code is that the chosen environment offers pyspark (in the scenario of that not being the case the following link could come to be useful: https://python.plainenglish.io/apache-spark-using-jupyter-in-linux-installation-and-setup-b2cacc6c7701) as well as the presence of all python libraries being used (for this consult the provided Jupyter notebook).

```
gmaps.configure(api_key='AIza[redacted]')
```

Before starting to run the solution, ensure the “*modis_2000-2019_Australia.csv*” file is present in the same directory as the “*analysis.ipynb*” file. Furthermore, add an API key for google maps (this step could be skipped, but then gmaps would not provide any useful information). To get the google maps API key, follow the instructions provided in the following link: https://developers.google.com/maps/documentation/javascript/get-api-key.
Finally, after ensuring that the requirements are fulfilled you should be able to execute all data processing actions that have been provided in the notebook.