Big Data Mining: Term Project

COVID-19 Twitter Dataset:

In-Depth Analysis of Text Clustering and Sentiment Prediction

Group 10

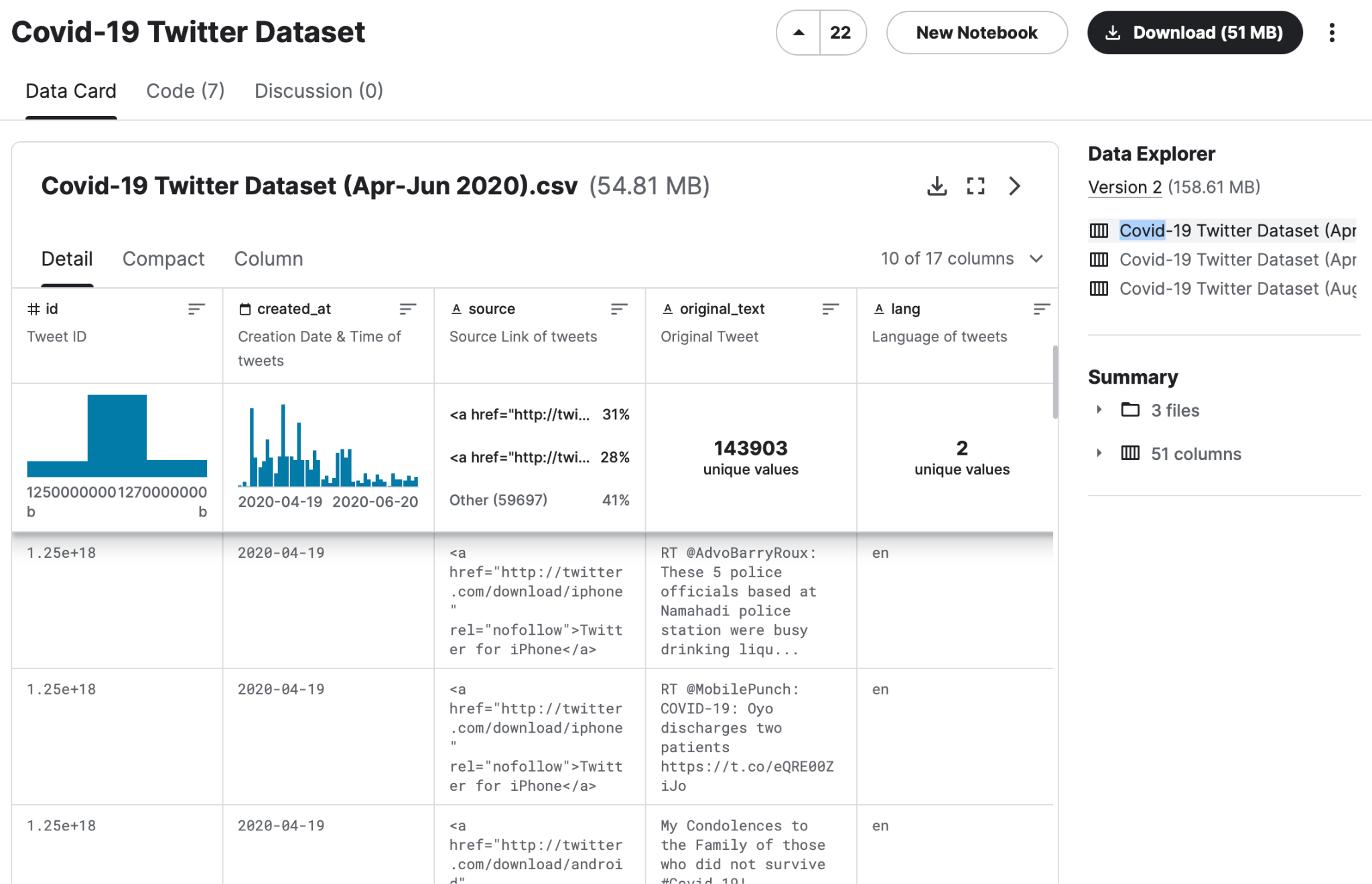
資工碩一 112598017 楊淨雯

資工碩一 112598028 余珮綺

**A. Background**

1.DataSet - Covid-19 Twitter Dataset

<https://www.kaggle.com/datasets/arunavakrchakraborty/covid19-twitter-dataset/data>



* The Covid-19 Twitter dataset, covering the period from April 2020 to June 2021.
* The dataset spans multiple phases of the pandemic and encompasses English tweets from around the globe.
* The dataset contains 3 files and 10 columns, has 1044825 tweets, and the data size is 51MB.

2. environment - using the PySpark, and T4 GPU in colab

3. Motivation

Amid the Covid-19 pandemic, social media has emerged as a primary platform for global information exchange. Our motivation lies in comprehensively understanding the dynamics and trends on social media during this crisis through extensive data analysis and exploration. Specifically, we aim to focus on global health-related topics to further investigate societal concerns regarding health and public safety. The objective of our research is to provide an in-depth analysis of health-related issues, offering insights into potential future similar situations.

4. Goal:

1. To comprehensively understand the dynamics and trends on social media during the Covid-19 pandemic through extensive data analysis and exploration.
2. Utilizing existing machine learning models to gain a deeper understanding of the topics people discuss, and further predicting the emotions associated with their tweets.

5. Methods and Functionality

1. Data Pre-Processing:

Processed data using NLTK-based function: Lowercasing, removing extraneous elements, converting 'covid' to 'covid19', and stemming.

1. Sentiment Analysis:

Utilized NLTK Sentiment Analyzer for polarity scores—categorized tweets as Positive, Negative, or Neutral based on compound sentiment scores.

1. K-means clustering:

is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

1. Principal component analysis (PCA):

is the process of computing the principal components and using them to perform a change of basis on the data,sometimes using only the first few principal components and ignoring the rest.

1. Convolutional Neural Network (CNN):

is a deep learning model that efficiently extracts image features through convolutional and pooling layers, widely used for image recognition and classification, but also can be used in sentiment prediction.

1. Long Short-Term Memory (LSTM):

is a type of recurrent neural network designed for processing and predicting time series data. Successful in language processing and time series analysis, it learns long-term dependencies.

**B.execution instructions**

We have a .py file: “BDM\_Final.py”, containing code primarily used for our report. The execution method is as follows:

**Environment: Multi-node on spark**

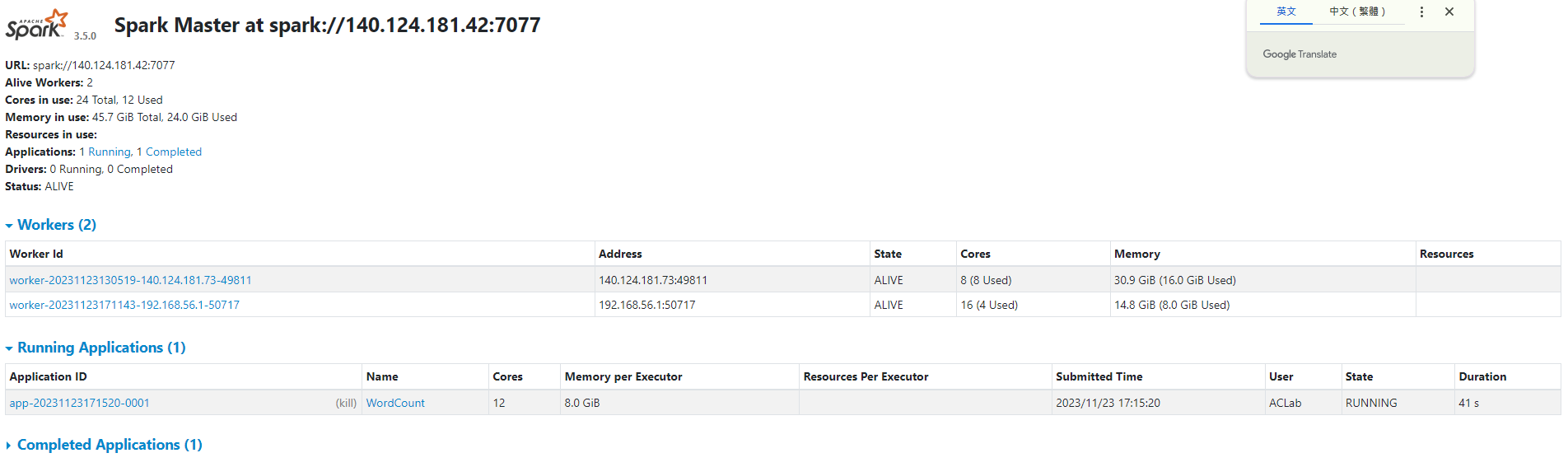
Both master and worker are running on Windows, and you need to follow the previous instructions to install Python, Spark, and Hadoop on both systems.

master: 140.124.181.42

worker 1: [192.168.56.1](http://192.168.56.1:8081/), 4 cores, 8GB Used

worker 2: [140.124.181.73](http://140.124.181.73:8081/), 8 cores, 16GB Used

And the UI like this, to check whether the connection succeed or not



**Execution: connection and execution**

* Master：

Open cmd as an administrator, navigate to the spark\bin folder, and enter the command:

spark-class org.apache.spark.deploy.master.Master

Wait until "Alive" appears.

* worker：

Open cmd as an administrator, navigate to the spark\bin folder, and enter the command:

spark-class org.apache.spark.deploy.worker.Worker spark://<master\_ip>:7077

* Ensure that the specified path contains the needed files.
* Open cmd, navigate to the spark\bin folder, and enter the command: spark-submit --master spark://http://<master\_ip>:7077 --executor-cores 4 --executor-memory 8g <code\_file\_name>

To see the detail of our project:

clustering: [clustering.ipynb](https://colab.research.google.com/drive/13zU6TnBXpCCqFXOINUR9xYgBtaiGB2I_?usp=share_link)

In this file, you can find details of Exploratory Data Analysis (EDA) on each column, as well as detailed programming information related to K-means and PCA.

BDM\_Final: [BDM\_Final.ipynb](https://colab.research.google.com/drive/1cK26D0F3_eyNHoa55OzoBQvnktGK5OB9?usp=share_link)

In this file, you can find details of EDA on each column.

BDM\_Final\_Model: [BDM\_Final\_Model.ipynb](https://colab.research.google.com/drive/1razaE_PKO6v1dcTE61fArBNCER5repMl?usp=share_link)

In this file, you can find details on building CNN, LSTM, and CNN+LSTM models.

**C.Task Responsibility**

楊淨雯 ： EDA on each columns, K-means with PCA

余珮綺 ： EDA on each columns, Sentiment Prediction