ibimbing

Sentiment Analysis

by : Ivan Manuel Wicaksono

Ivan Manuel W.

Education

S1 Teknik Elektro - Institut Teknologi Sepuluh Nopember (2018-2022)

Working

Business Analyst

Overview Project

- **Sentiment Analysis**

  Mencari trend sentimen yang terjadi melalui kanal berita website

Project Background

This project is to develop an automated web scraping from financial news (CNN and New York Times) to do a sentiment analysis (positive or negative) for each article to indicate trends

The project can help traders or investor to analyze US news sentiment (is the news give negative sentiment or positive) to decide or manage their investment portfolio

Problem Statement

In the financial market, there is a vast amount of data available, including stock exchange information, economic indicators, global events, and financial news.

Investors can use sentiment analysis to make informed investment decisions.

The rise and fall of financial asset prices are heavily influenced by market sentiment, political news, policies, inflation data, interest rates, bond yields, and more.

To gather this information, investors need to search for news across multiple websites and look up stock and commodity prices, as this data is often scattered across the internet.

Traditionally, analysis relies on historical data, which causes delays in report generation as it takes a long time to manually read the news one by one.

With a data engineering pipeline, large market data can be processed, sentiment toward a stock can be identified, and valuable insights can be generated.

The dashboard displays:

- Sentiment analysis for each article
- Comparison with SP500

Data Platform Understanding

Timeline

Data Extraction

Scrape financial news website

Data Processing

Cleaning null values and duplicates values with spark

Data Storage

Load the data to Bigquery

Data Visualization

Visualize data in Tableau

Data Understanding

Datasets :

- Websites (CNN, New York Times 250 articles in the last 7 days)
- CSV (SP500 price in the last 7 days)

Data Orchestration :

- Using airflow for scheduling and monitoring batch processing

Extract :

- Web scraping the data with python and selenium

Transform :

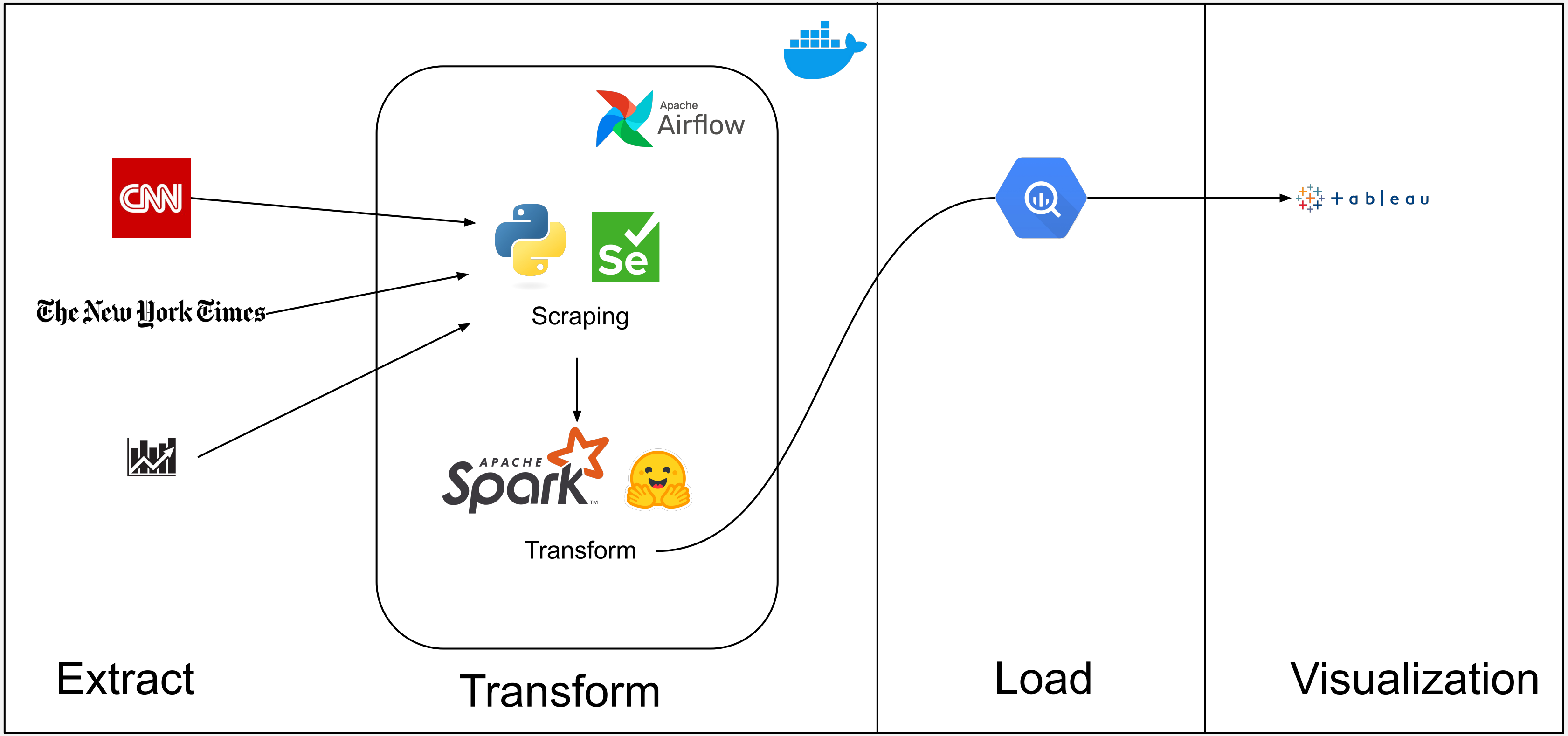
- Clean the data, duplicate values, and missing values using pyspark
- Do a sentiment analysis with ML model hugging face

Load :

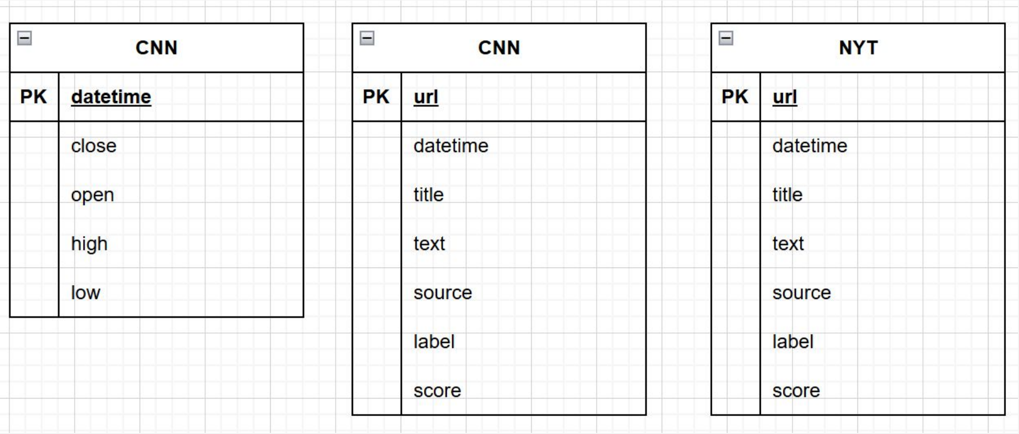
- Load the data to BigQuery

Transformation & Consideration

Architecture



Data Modeling (Business)



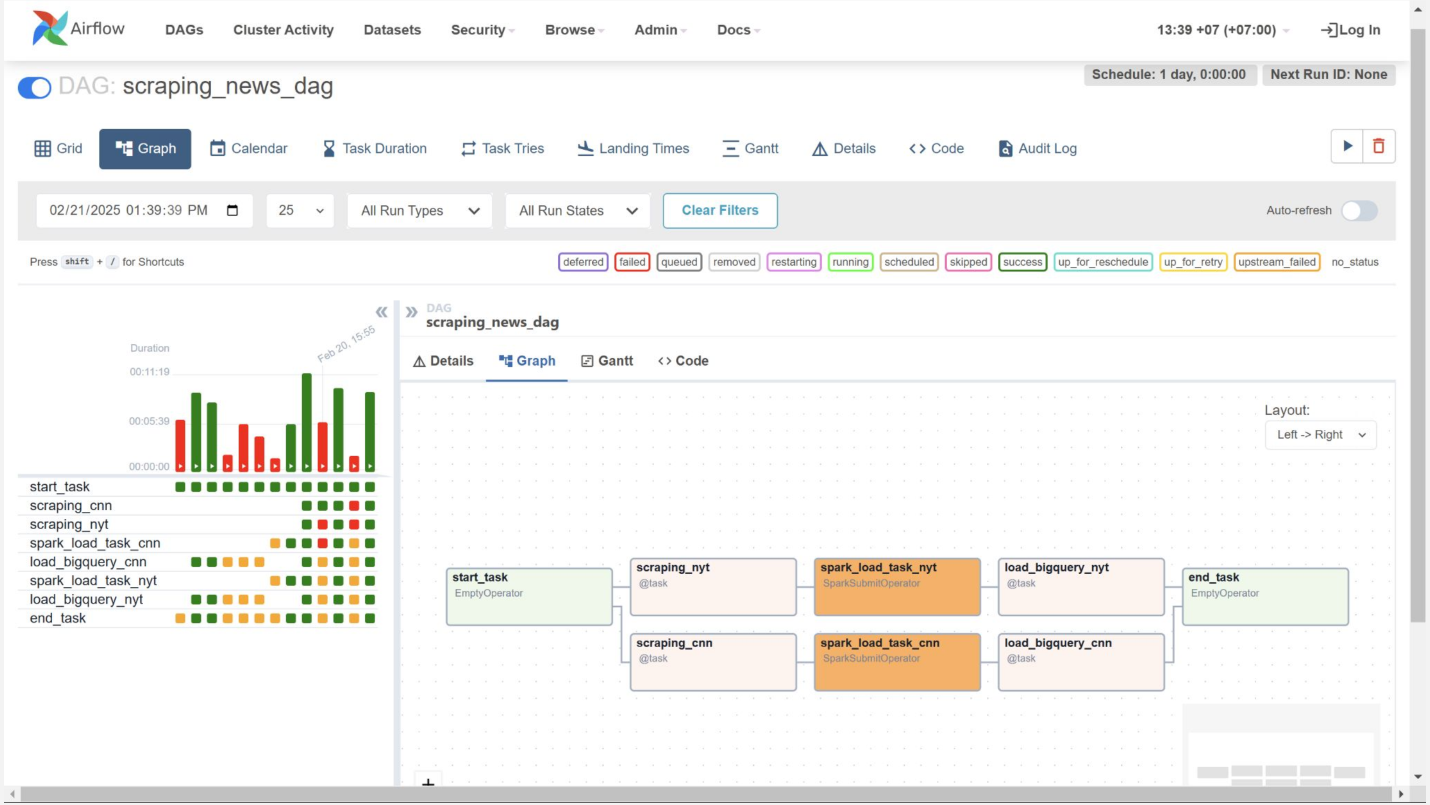
	CNN
PK	<u>datetime</u>
	close
	open
	high
	low

	CNN
PK	<u>url</u>
	datetime
	title
	text
	source
	label
	score

	NYT
PK	<u>url</u>
	datetime
	title
	text
	source
	label
	score

Conclusion & Recommendation

DAG Airflow - News



DAG: scraping_news_dag

Schedule: 1 day, 0:00:00

Next Run ID: None

02/21/2025 01:39:39 PM

25

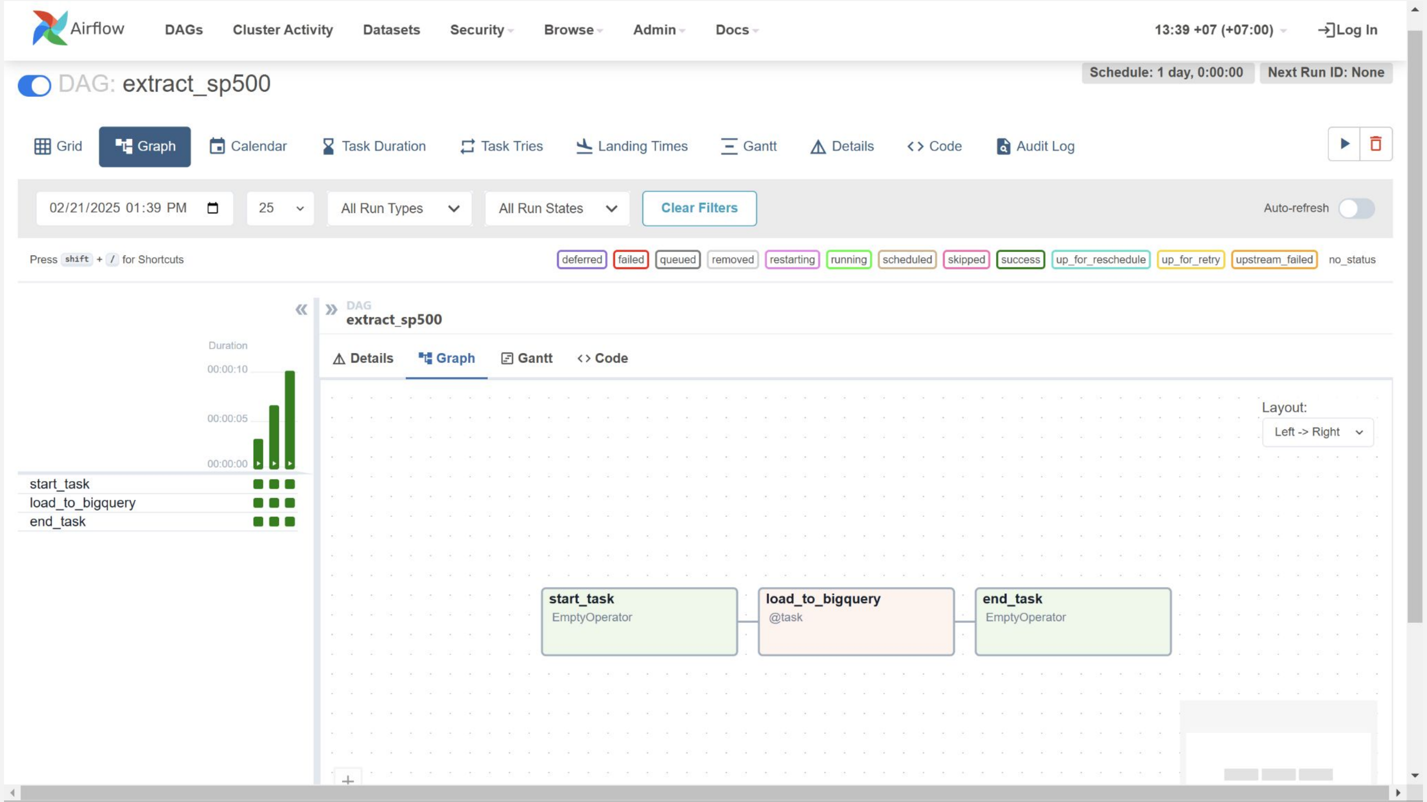
All Run Types

All Run States

Auto-refresh

Press shift + / for Shortcuts

DAG Airflow - SP500



DAG: extract_sp500

Schedule: 1 day, 0:00:00

Next Run ID: None

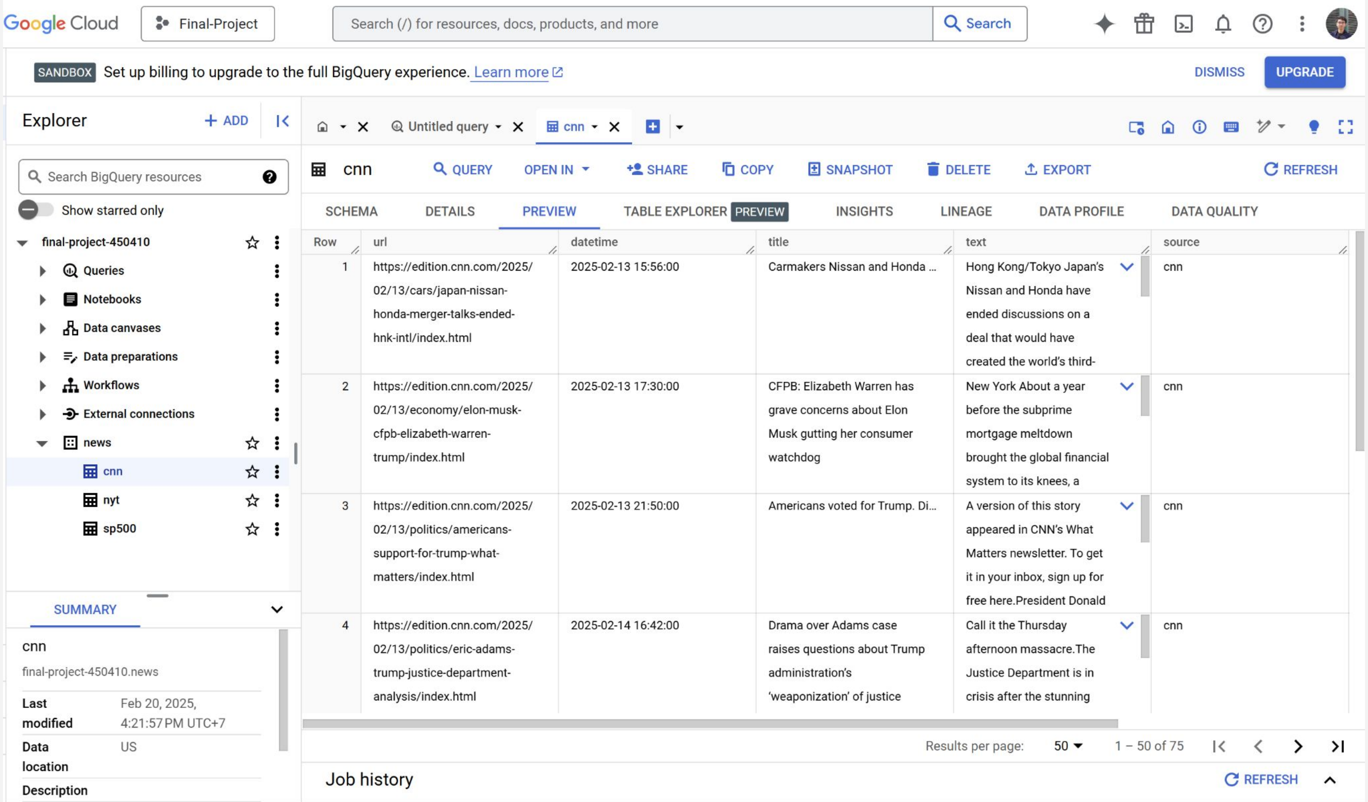
02/21/2025 01:39 PM

25

All Run Types

All Run States

Load BigQuery



Set up billing to upgrade to the full BigQuery experience. Learn more

cnn

Row	url	datetime	title	text	source
1	https://edition.cnn.com/2025/02/13/cars/japan-nissan-honda-merger-talks-ended-hnk-intl/index.html	2025-02-13 15:56:00	Carmakers Nissan and Honda ...	Hong Kong/Tokyo Japan's Nissan and Honda have ended discussions on a deal that would have created the world's third-	cnn
2	https://edition.cnn.com/2025/02/13/economy/elon-musk-cfpb-elizabeth-warren-trump/index.html	2025-02-13 17:30:00	CFPB: Elizabeth Warren has grave concerns about Elon Musk gutting her consumer watchdog	New York About a year before the subprime mortgage meltdown brought the global financial system to its knees, a	cnn
3	https://edition.cnn.com/2025/02/13/politics/americans-support-for-trump-what-matters/index.html	2025-02-13 21:50:00	Americans voted for Trump. Di...	A version of this story appeared in CNN's What Matters newsletter. To get it in your inbox, sign up for free here.President Donald	cnn
4	https://edition.cnn.com/2025/02/13/politics/eric-adams-trump-justice-department-analysis/index.html	2025-02-14 16:42:00	Drama over Adams case raises questions about Trump administration's 'weaponization' of justice	Call it the Thursday afternoon massacre.The Justice Department is in crisis after the stunning	cnn

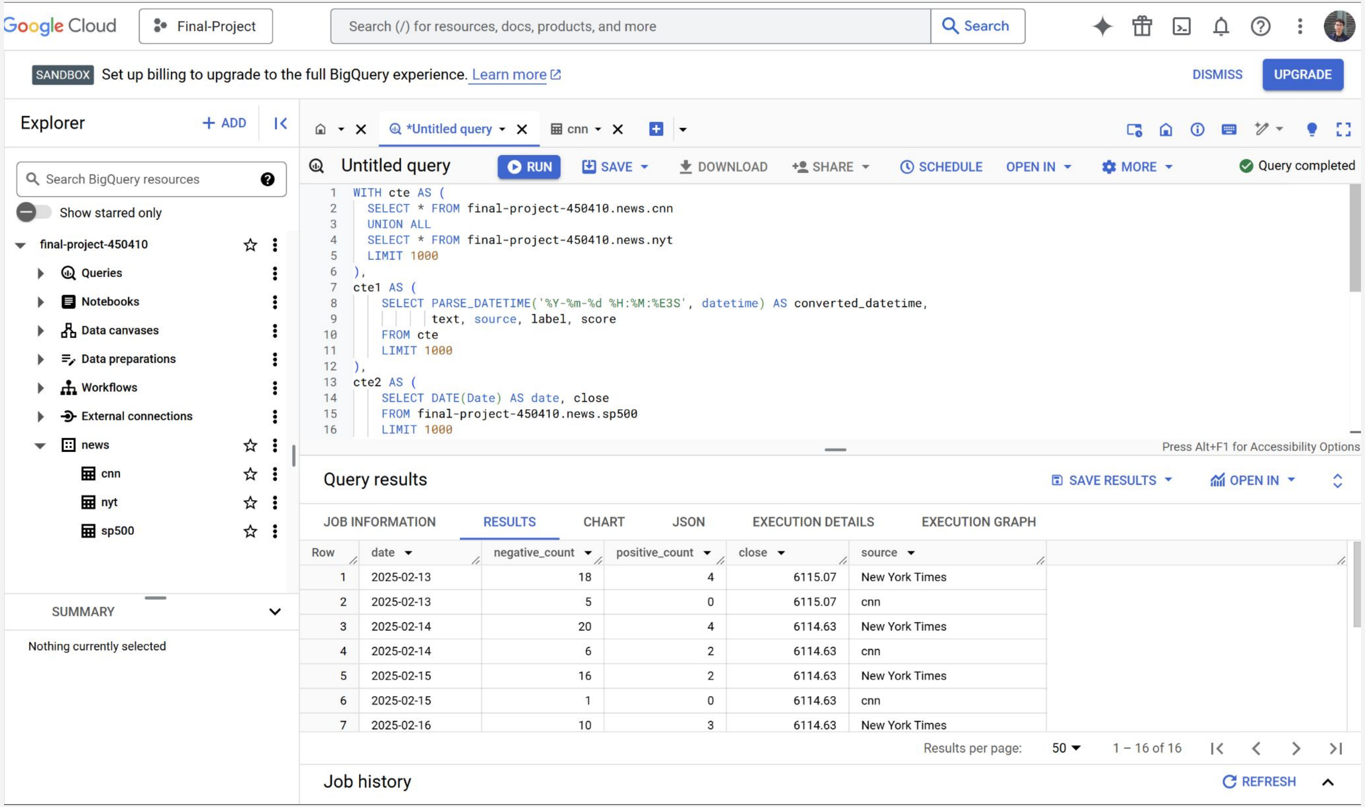
Results per page: 50 1 – 50 of 75

cnn

final-project-450410.news

Last modified	Feb 20, 2025, 4:21:57 PM UTC+7
Data location	US
Description	

Perform Query



Set up billing to upgrade to the full BigQuery experience. Learn more

Nothing currently selected

Untitled query

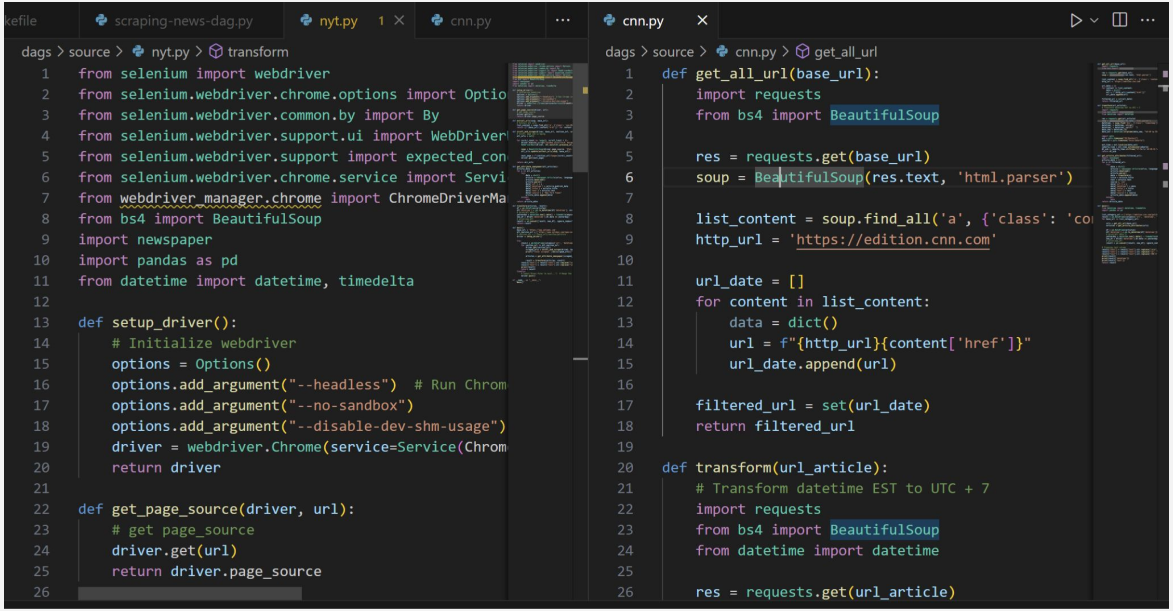
```
WITH cte AS (
  SELECT * FROM final-project-450410.news.cnn
  UNION ALL
  SELECT * FROM final-project-450410.news.nyt
  LIMIT 1000
),
cte1 AS (
    SELECT PARSE_DATETIME('%Y-%m-%d %H:%M:%E3S', datetime) AS converted_datetime,
          text, source, label, score
    FROM cte
    LIMIT 1000
),
cte2 AS (
    SELECT DATE(Date) AS date, close
    FROM final-project-450410.news.sp500
    LIMIT 1000
```

Query results

Row	date	negative_count	positive_count	close	source
1	2025-02-13	18	4	6115.07	New York Times
2	2025-02-13	5	0	6115.07	cnn
3	2025-02-14	20	4	6114.63	New York Times
4	2025-02-14	6	2	6114.63	cnn
5	2025-02-15	16	2	6114.63	New York Times
6	2025-02-15	1	0	6114.63	cnn
7	2025-02-16	10	3	6114.63	New York Times

Job history

Code Overview - Web Scraping

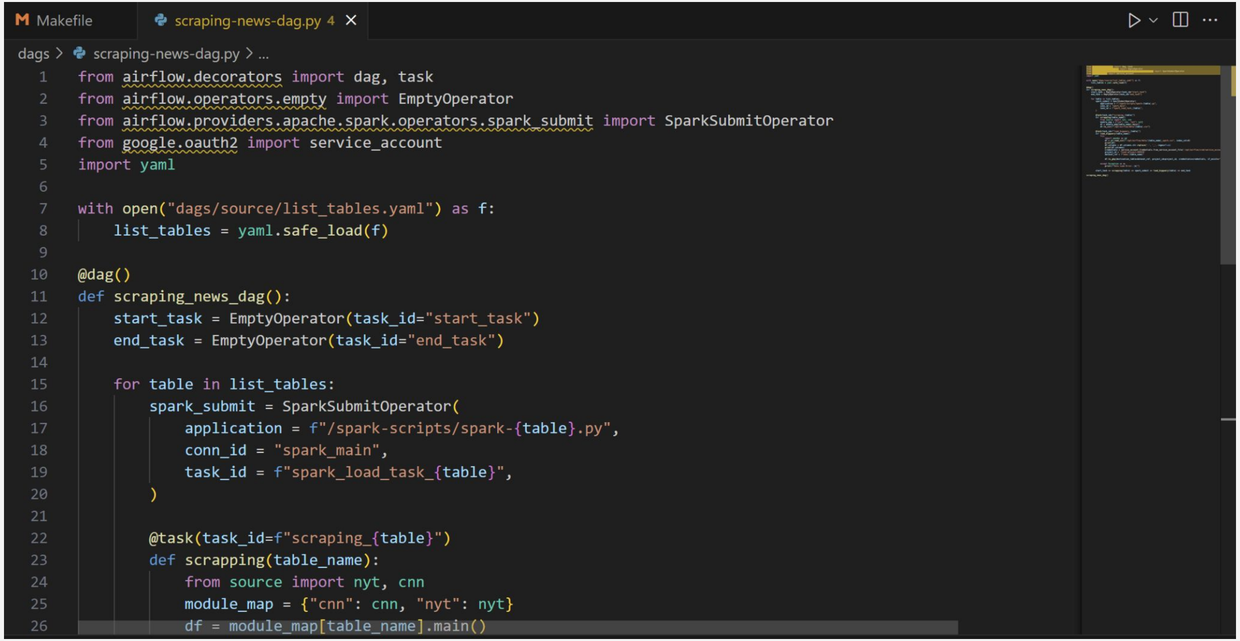

kefile
scraping-news-dag.py
nyt.py 1 ×
cnn.py
dags > source > nyt.py > transform
cnn.py ×
dags > source > cnn.py > get_all_url

```python
from selenium import webdriver
from selenium.webdriver.chrome.options import Optio
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriver
from selenium.webdriver.support import expected_con
from selenium.webdriver.chrome.service import Servi
from webdriver_manager.chrome import ChromeDriverMa
from bs4 import BeautifulSoup
import newspaper
import pandas as pd
from datetime import datetime, timedelta

def setup_driver():
    # Initialize webdriver
    options = Options()
    options.add_argument("--headless")  # Run Chrom
    options.add_argument("--no-sandbox")
    options.add_argument("--disable-dev-shm-usage")
    driver = webdriver.Chrome(service=Service(Chrom
    return driver

def get_page_source(driver, url):
    # get page_source
    driver.get(url)
    return driver.page_source

```

```python
def get_all_url(base_url):
    import requests
    from bs4 import BeautifulSoup

    res = requests.get(base_url)
    soup = BeautifulSoup(res.text, 'html.parser')

    list_content = soup.find_all('a', {'class': 'co
    http_url = 'https://edition.cnn.com'

    url_date = []
    for content in list_content:
        data = dict()
        url = f"{http_url}{content['href']}"
        url_date.append(url)

    filtered_url = set(url_date)
    return filtered_url

def transform(url_article):
    # Transform datetime EST to UTC + 7
    import requests
    from bs4 import BeautifulSoup
    from datetime import datetime

    res = requests.get(url_article)
```

These are the codes for web scraping through CNN and NYT. Because NYT is a dynamic website, I use selenium to scrap dynamic website

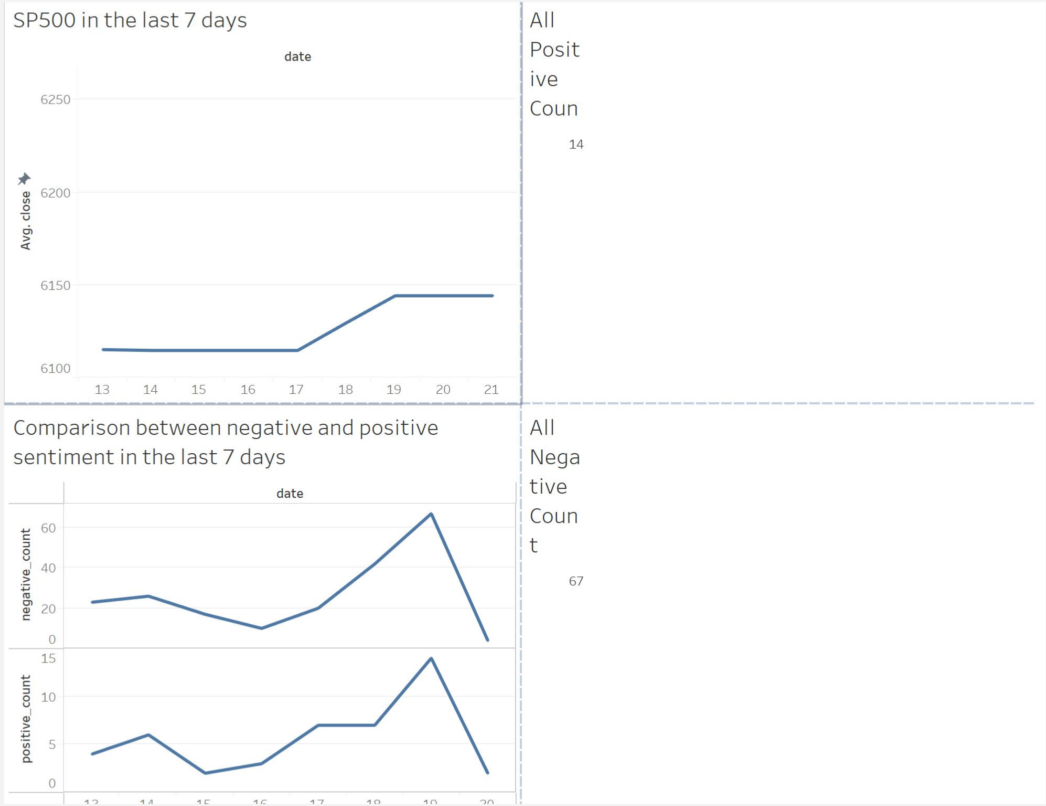
Code Overview - DAG



```python
from airflow.decorators import dag, task
from airflow.operators.empty import EmptyOperator
from airflow.providers.apache.spark.operators.spark_submit import SparkSubmitOperator
from google.oauth2 import service_account
import yaml

with open("dags/source/list_tables.yaml") as f:
    list_tables = yaml.safe_load(f)

@dag()
def scraping_news_dag():
    start_task = EmptyOperator(task_id="start_task")
    end_task = EmptyOperator(task_id="end_task")

    for table in list_tables:
        spark_submit = SparkSubmitOperator(
            application = f"/spark-scripts/spark-{table}.py",
            conn_id = "spark_main",
            task_id = f"spark_load_task_{table}",
        )

        @task(task_id=f"scraping_{table}")
        def scrapping(table_name):
            from source import nyt, cnn
            module_map = {"cnn": cnn, "nyt": nyt}
            df = module_map[table_name].main()
```

This is dynamic dag to scrap multiple website. You can add more if you want to

Visualization



SP500 in the last 7 days

All Positive Count

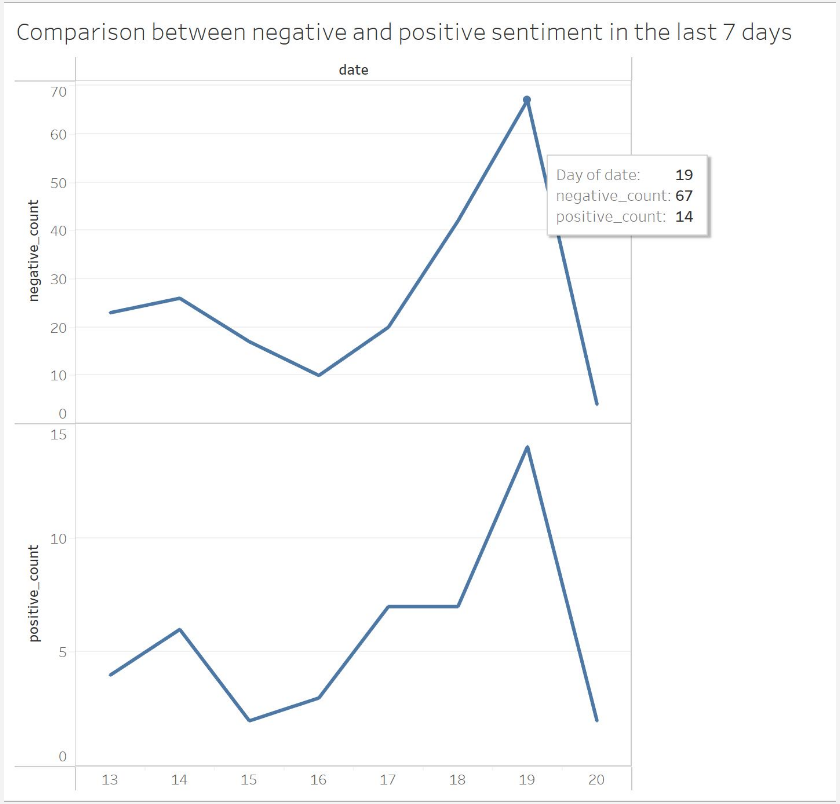
14

Comparison between negative and positive sentiment in the last 7 days

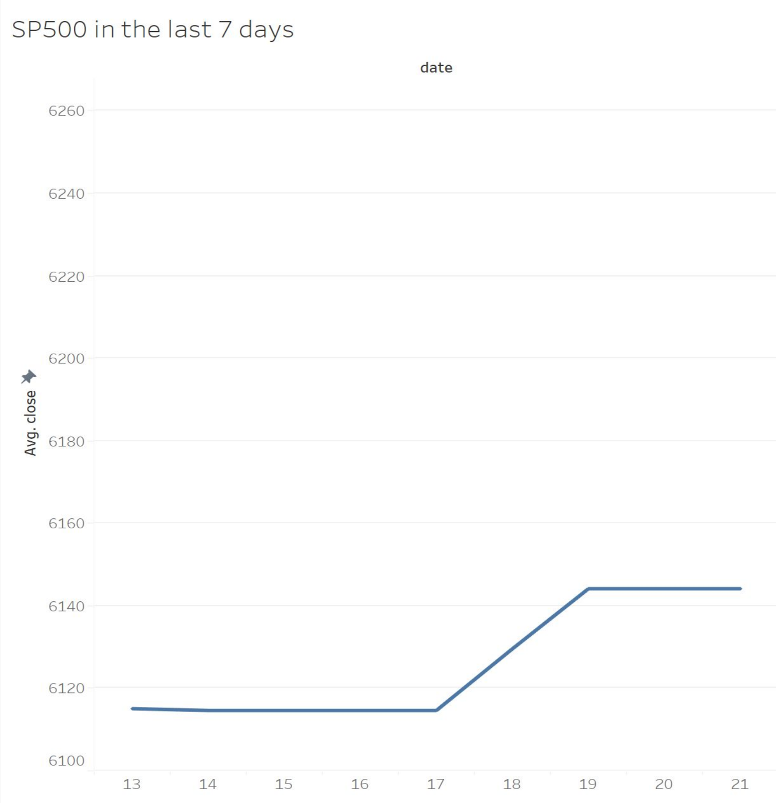
All Negative Count

67

Visualization



Comparison between negative and positive sentiment in the last 7 days



SP500 in the last 7 days

Conclusion

We can see from the dashboard that negative sentiment dominate the news. The ratio between positive and negative news are 0.1 to 0.3. So, we can conclude that the news in last 7 days are heavily negative. But if we compare with sp500, the index has increased from 6115 to 6140 or 0.4% gains at 17 February until 19 February.

Date	13	14	15	16	17	18	19	20
Positive	4	6	2	3	7	7	14	2
Negative	23	26	17	10	20	42	67	4
Ratio	0.174	0.231	0.118	0.3	0.35	0.167	0.209	0.5

Conclusion

From this analysis we can conclude that we can’t predict the stock price using only one metric. This project only one of supporting tools to help analyze US sentiment

The limitation of this project that it’s not sufficient using only 2 website sources. More data can help to generalize the news and get broader insight. But it is hard to find news website that can be scrape. Because I am using built-in model machine learning library to give sentiment analysis, there is a limitation for tokenize words.

Terima Kasih.