**Big Data Integration and Storage**

**PROG8451 - Winter 2024 - Section 1**

**NewsAPI ETL Pipeline**

**Marc Hansel John Thomas: 8849344**

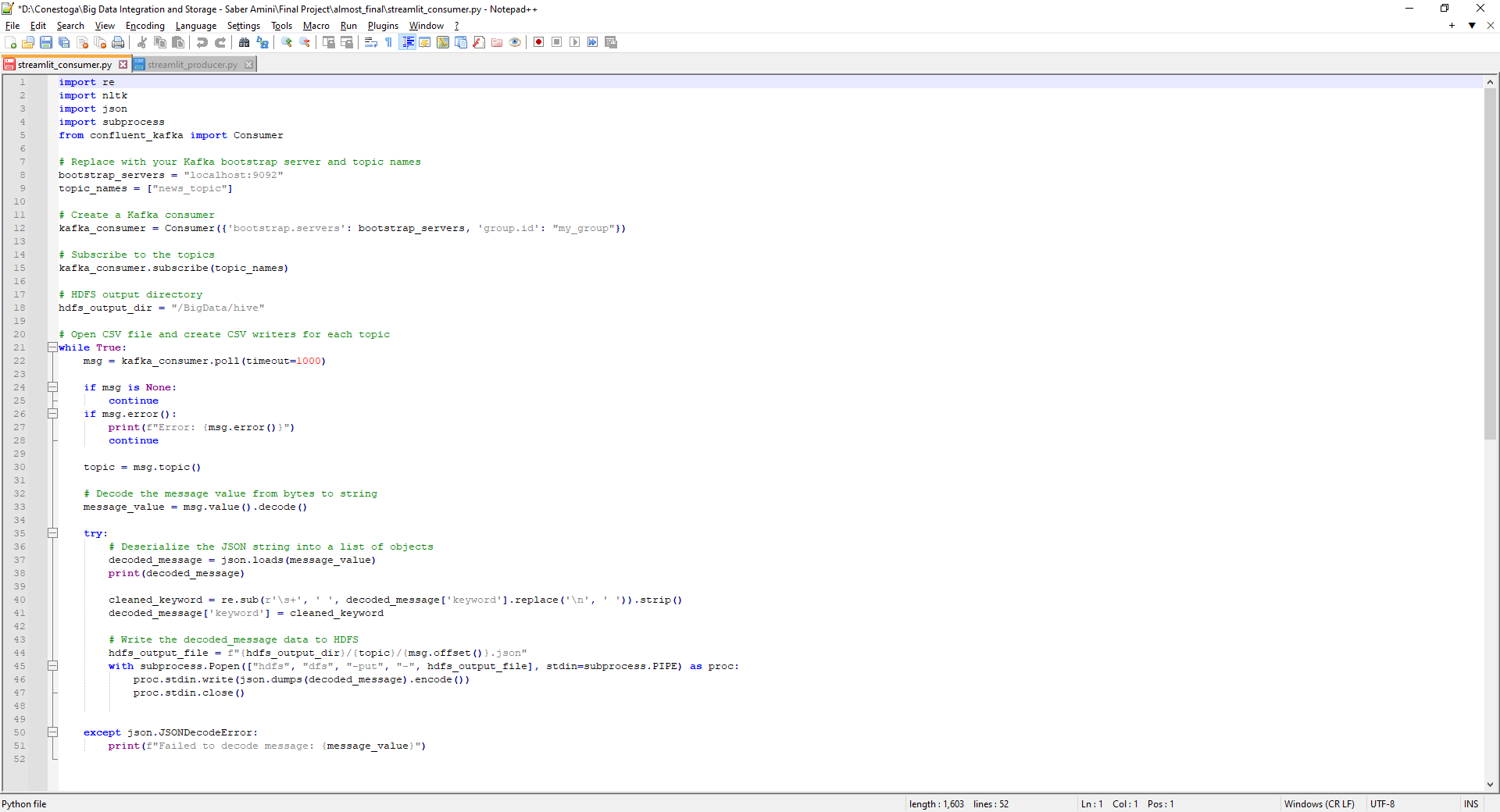
**Richard Andrey Biscazzi: 8903530**

To: Technical Manager

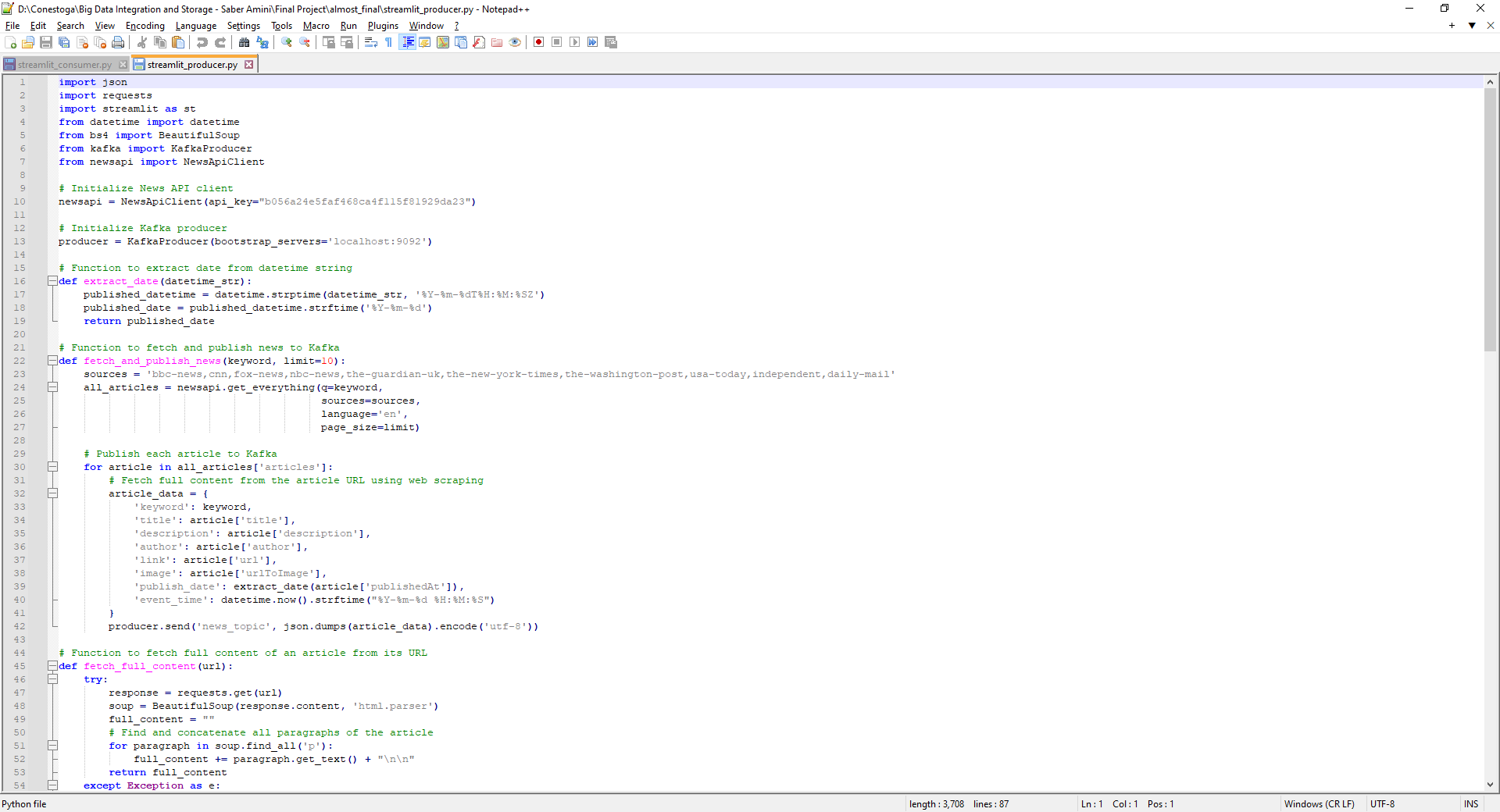
From: Big Data Solution Architect

Our system consists of an ETL process running using python to fetch data from the NewsAPI.

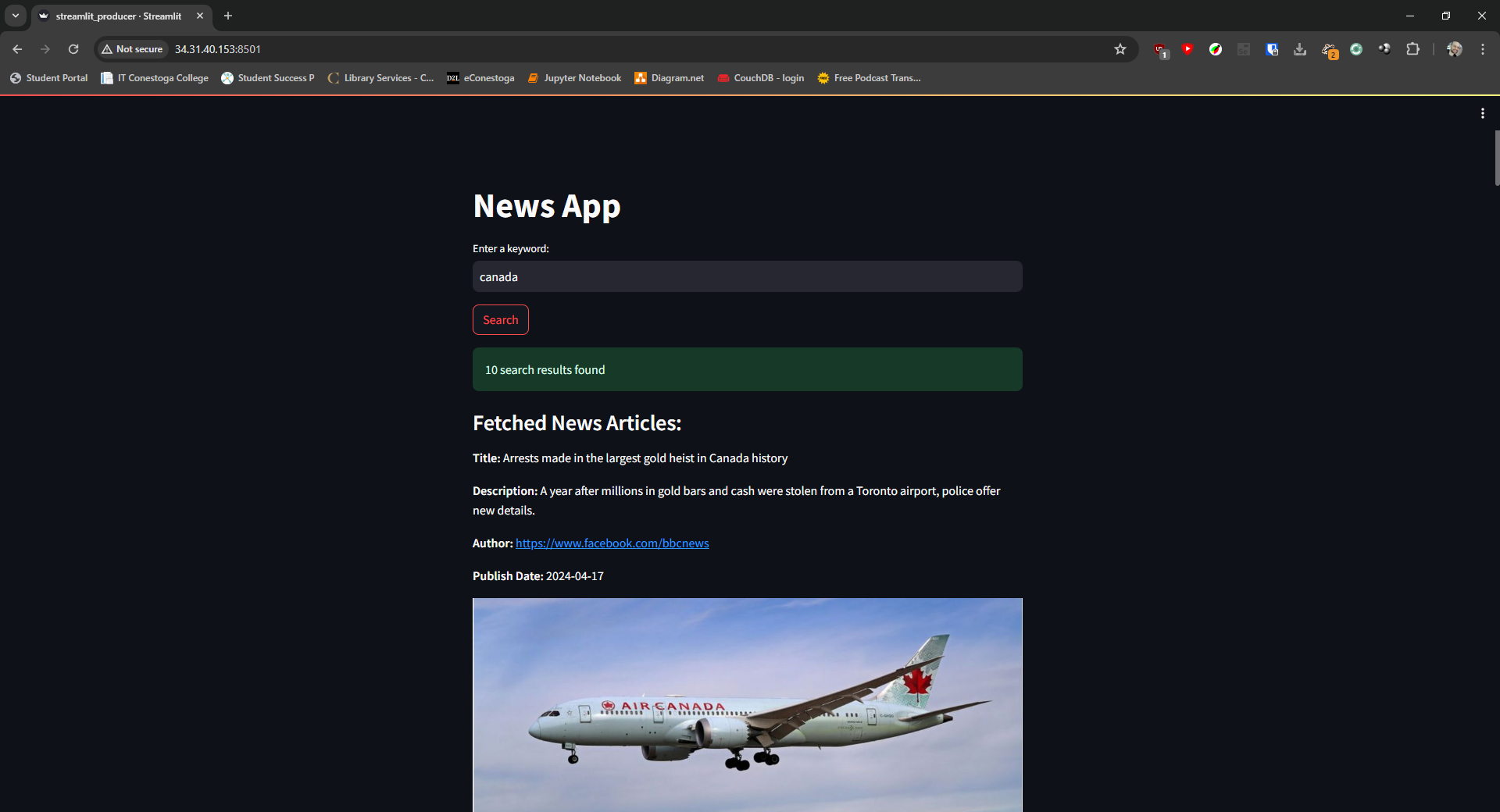
Firstly, we run the streamlit\_consumer.py code written in Python to consume the data the comes from the producer.



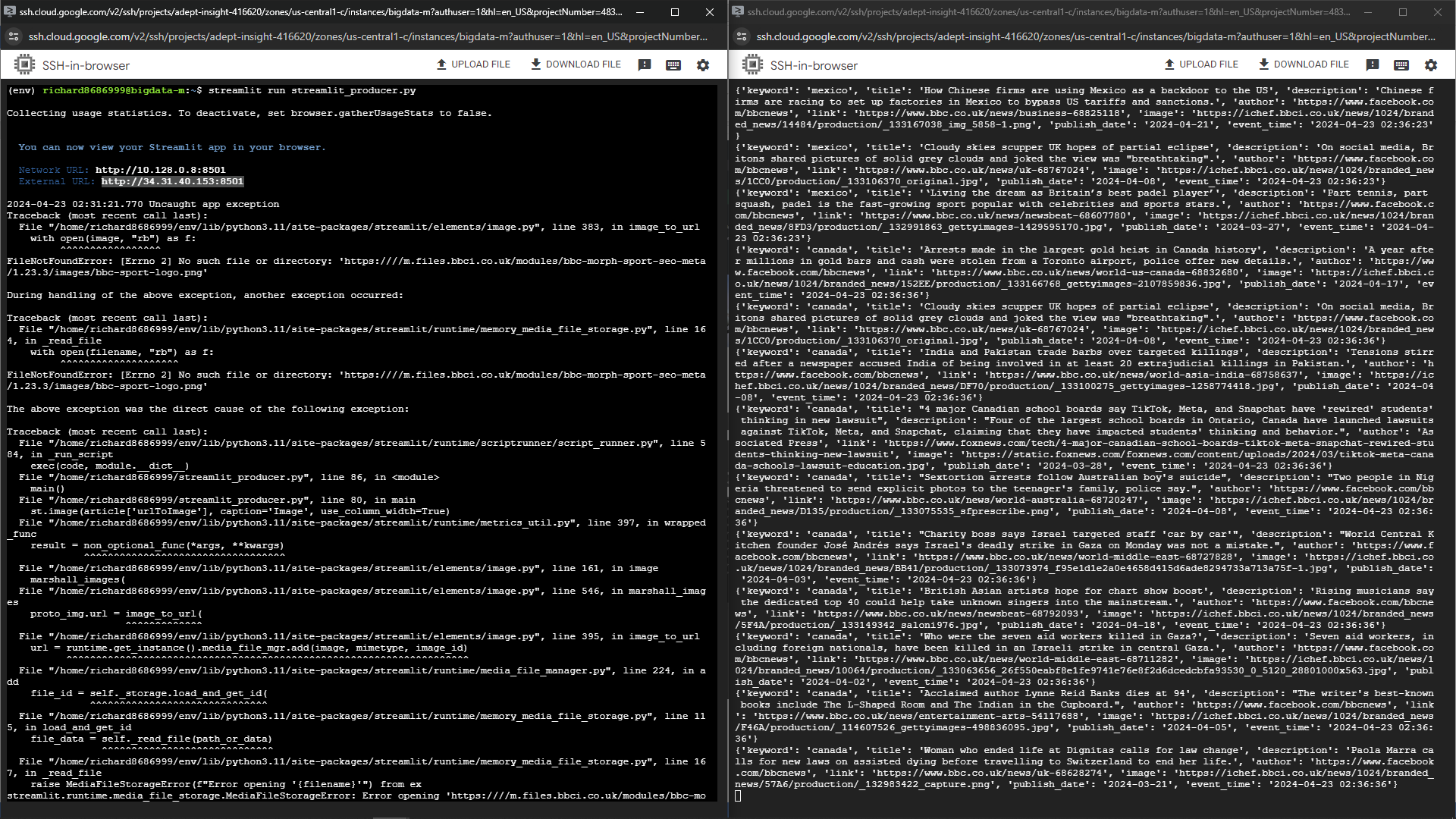
Secondly, we run the streamlit\_producer.py code written in Python to fetch the data from API. In this case we will use an webpage to insert the keyword related to the search the we want to make among the articles.



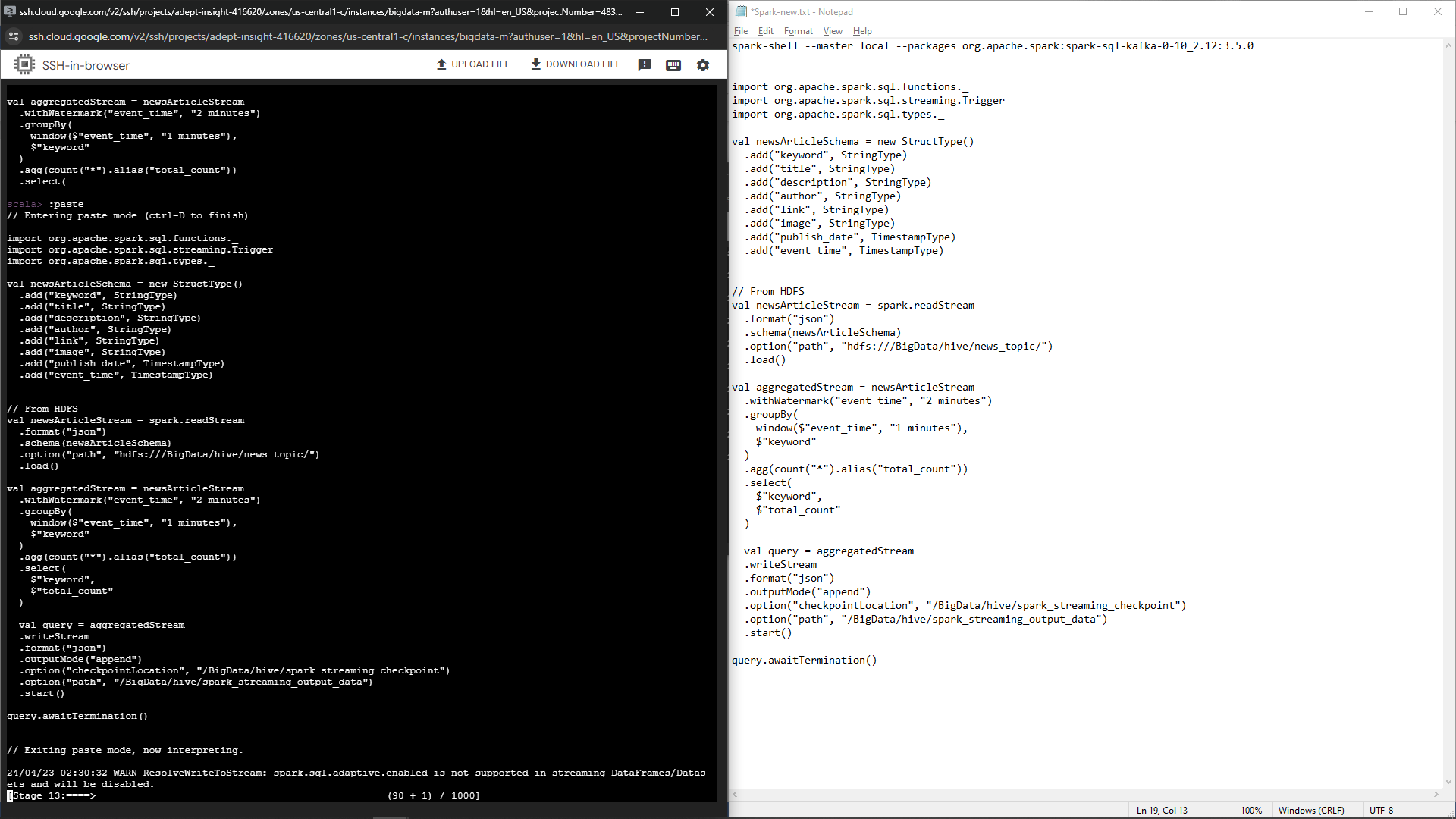
Searching for keywords in the webpage.



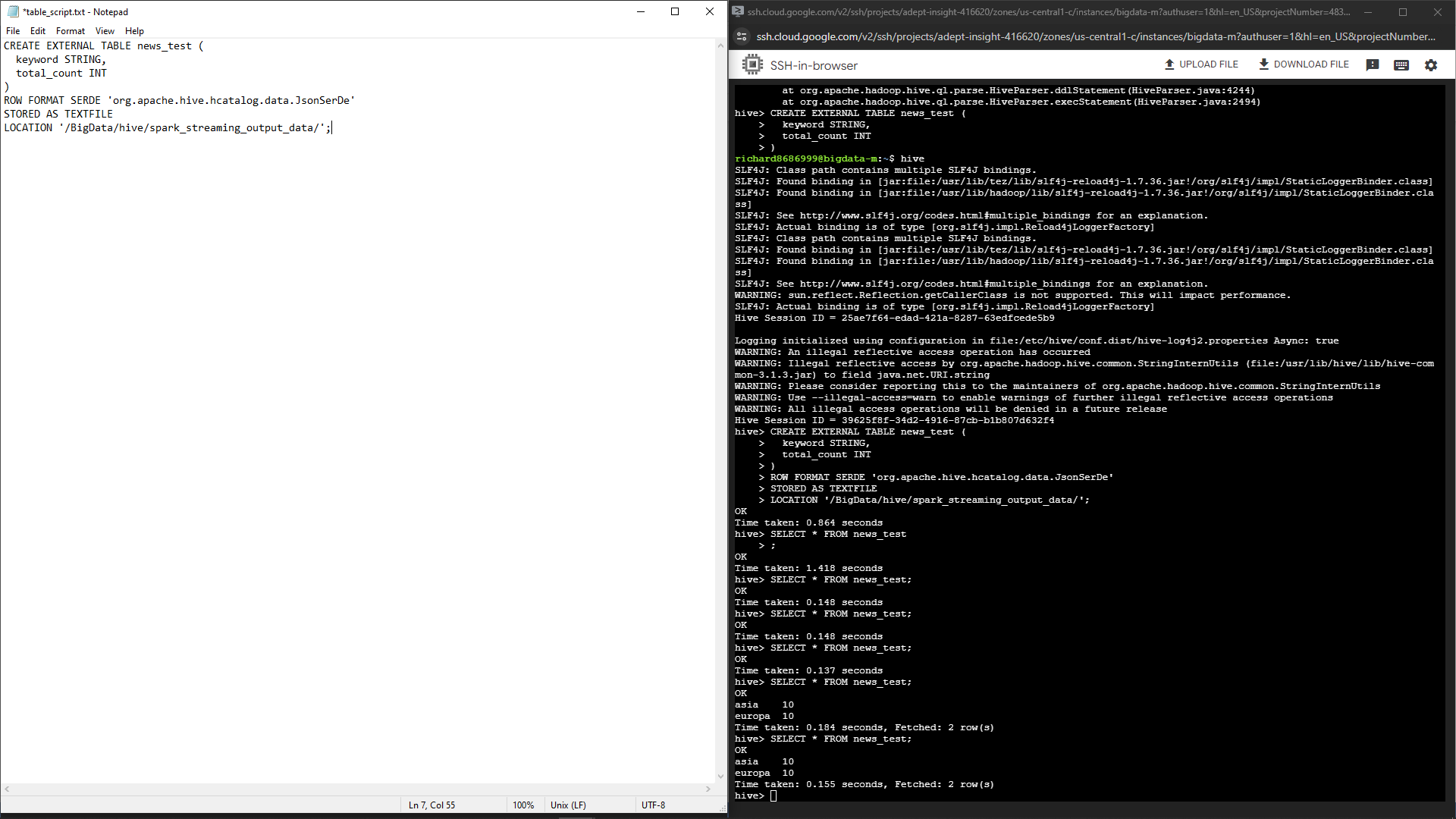
Here is our consumer and producer running.



Next our Spark Streaming read the information in the json file that is created with the consumer and then aggregate to have some insights of information. Here is the Spark Streaming running and the code used in it.



After the count of articles and the aggregation, we build an table in Hive to could use this data to create an dashboard or something else.



This is the final result of our system running everything together.

