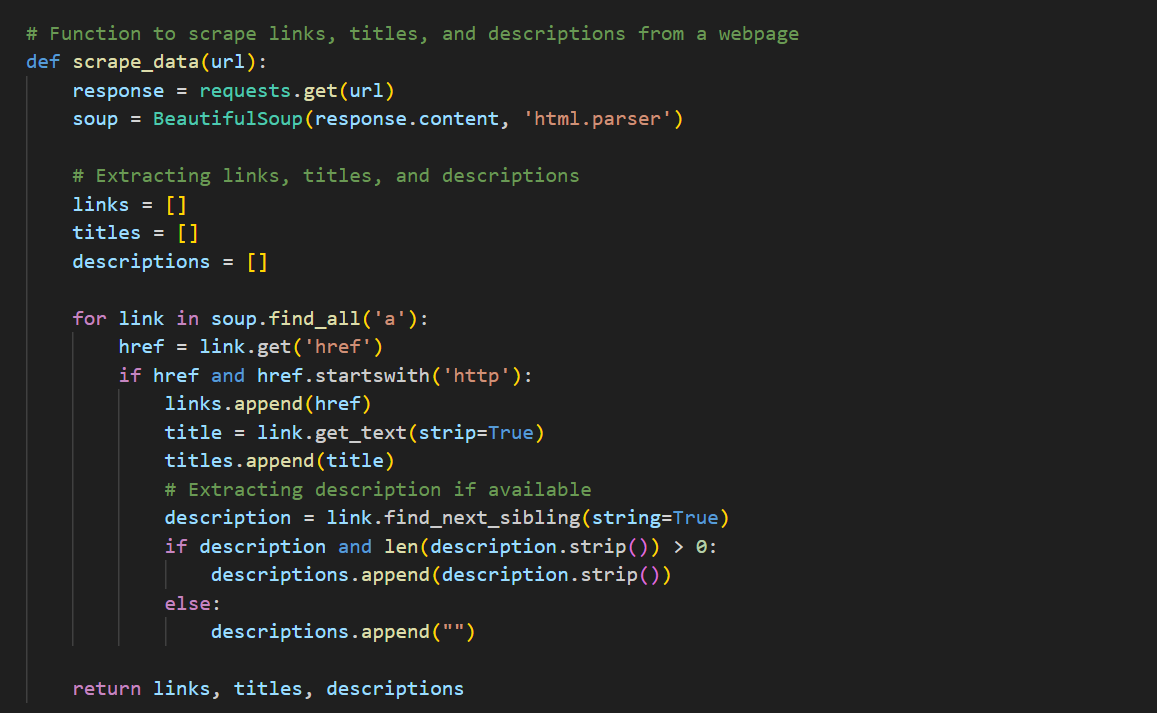
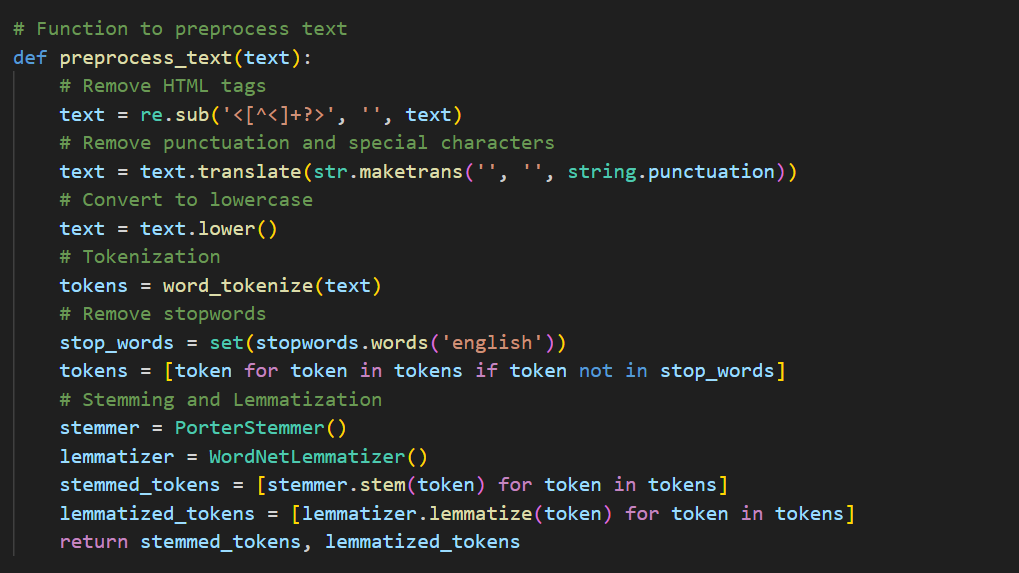
# Data Extraction

We used online web scrapping using beautiful soap from Dawn and BBC as these were the websites listed in the assignment description. We extracted links and their relative description on the landing page that were present and saved them to a csv after pre-processing.



# Data Transformation

Data Transformation basically refers to data pre processing which was done using NLTK libraries. Many more sub libraries were installed using NLTK to remove stop words and other non-useful characters that can be present in the scraped test from both websites. Below is the code attached for the Data transformation.

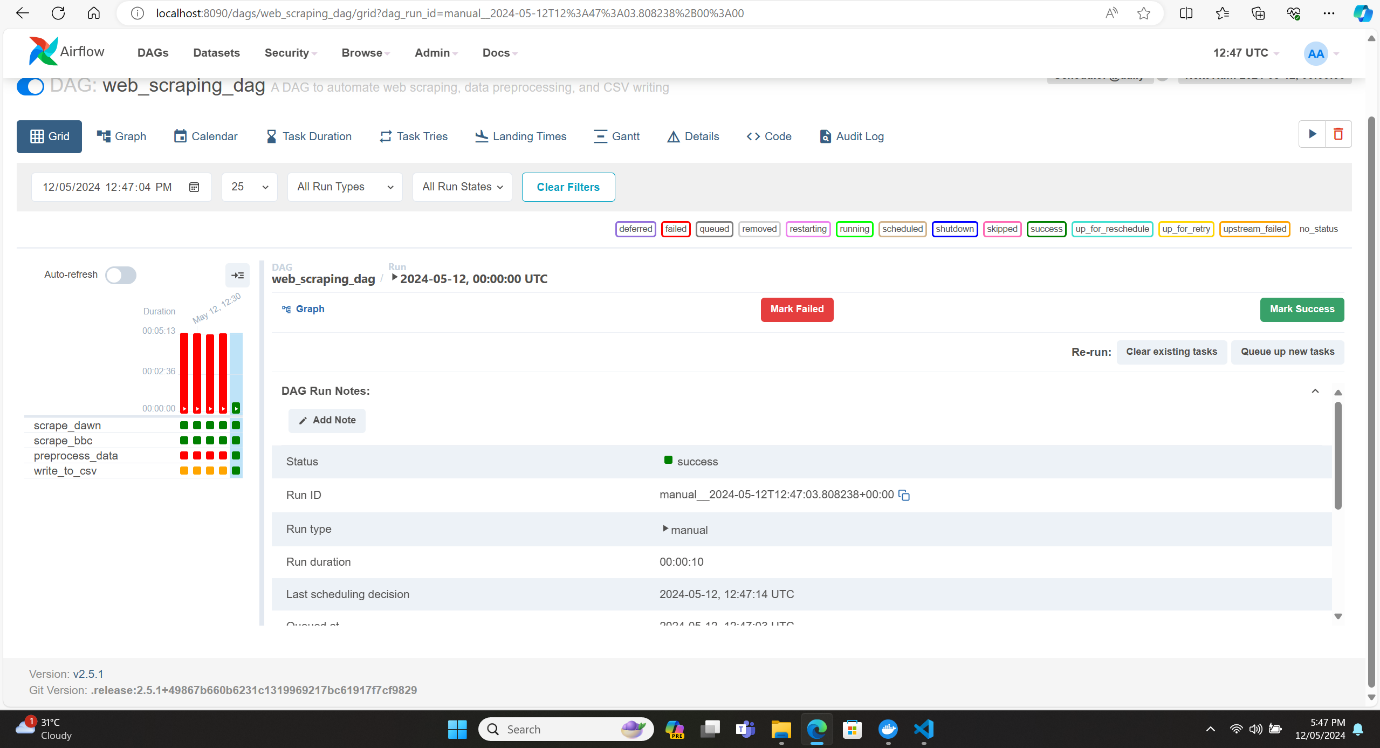


# Data Storage and Version Control

Data Storage and Version control was done using a tool named DVC. It is one of the most popular tool that is used. I first uploaded my dataset on some cloud platform named as google drive and then using DVC I versioned it and initialized it. This generated a file in my remote repository which was then integrated with GitHub.

# Apache Airflow DAG Development

First after installing Apache Airflow Using Docker Hub. I changed the port number to 8090 as 8080 was assigned to Jenkins. Then I wrote my own DAG file that were assigned to run tasks that could help me automate the process as its generally a pipeline. After 4-5 unsuccessful attempts and fixing error by eyeing log I was able to execute it successfully.



# Pushing Everything to GitHub Repo

Then after completing all the work including this work. Everything was pushed to my Github Repo. Which link is submitted as a proof for my hard work and commitment towards this assignment.