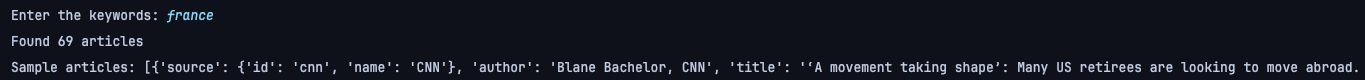
**Midterm Project**

**Getting the news from News API**

****

For this pipeline, we received the articles found for the inserted keywords. To get the news, we use the unofficial library of News API to fetch any news related to the keyword and parameters configured. In this case, using the keyword “France” we obtained 69 results.

**Sending news through the producer**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

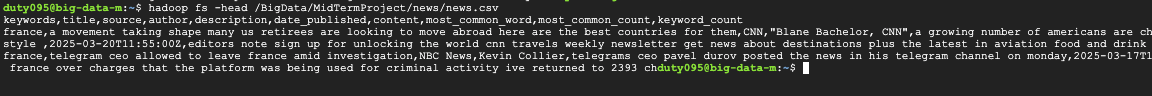
After retrieving the news from the API, we send it through the producer. Before sending it, to send the less useless data without losing significance, we perform some transformations in the title, content, and description to remove all non-alphanumeric characters, making the data sent less heavy. After performing that text processing operation, we proceed to send the article through the pipeline.

**Receiving the news to the consumer and saving it into HDFS**

**A black screen with white text

AI-generated content may be incorrect.**

We can check here that the consumer is receiving the news sent by the producer. After receiving the information, the consumer performs some operations to extract insightful observations from the article, including the most common word, its number of occurrences, and the number of occurrences of the keyword. After processing, the data is saved in an HDFS file.

****

The image shows that the data is stored in a CSV file (news.csv). The image also shows part of the data stored in the CSV file. This data will be loaded into the hive table to gain some insights.

**Setting up the Hive database and table**

**A screenshot of a computer program

AI-generated content may be incorrect.**

The schema of the table was created to load the news. This table and database store all the streamed data produced and saved by the consumer.

**Loading the data generated**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

The data is loaded into the created table from the file the consumer generates to save the news. In this example, the data is already stored in the table. With this data, insights can be generated to check for useful information.

**Insights**

**Top 5 sources by news count**

**A screenshot of a computer

AI-generated content may be incorrect.**

The first insight helps us understand which sources or media outlets contribute more to the information. For example, in this insight, NBC News is by far the one that contributes the most to this data.

**Top 5 used words in the news**

**A screen shot of a computer

AI-generated content may be incorrect.**

The second insight reveals the most used word in media, revealing potential biases to one ideology. In this example, the most common words found in the articles are common words like prepositions.

**Top 15 used words per source**

**A screenshot of a computer

AI-generated content may be incorrect.**

A more advanced insight of the second insight can get the most frequent words by source. For example, BBC and Fox News use the pronoun “the” as the most frequent word in their news.