Team Members:

1. احمد محمد توفيق محمد

2. عمر عبدالرحمن فتحي

3. محمد ابراهيم محمد احمد

4. اسلام عصام عبد العاطى سالم

5. شهاب ايهاب احمد طلعت

6. زياد محمد حسين خبيز

[Topic Modeling]

[Team 23]

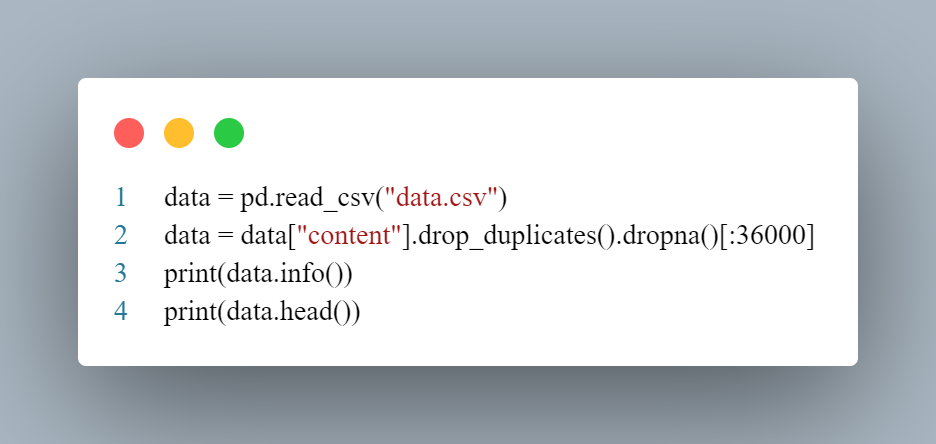
Topic Modeling

Introduction:

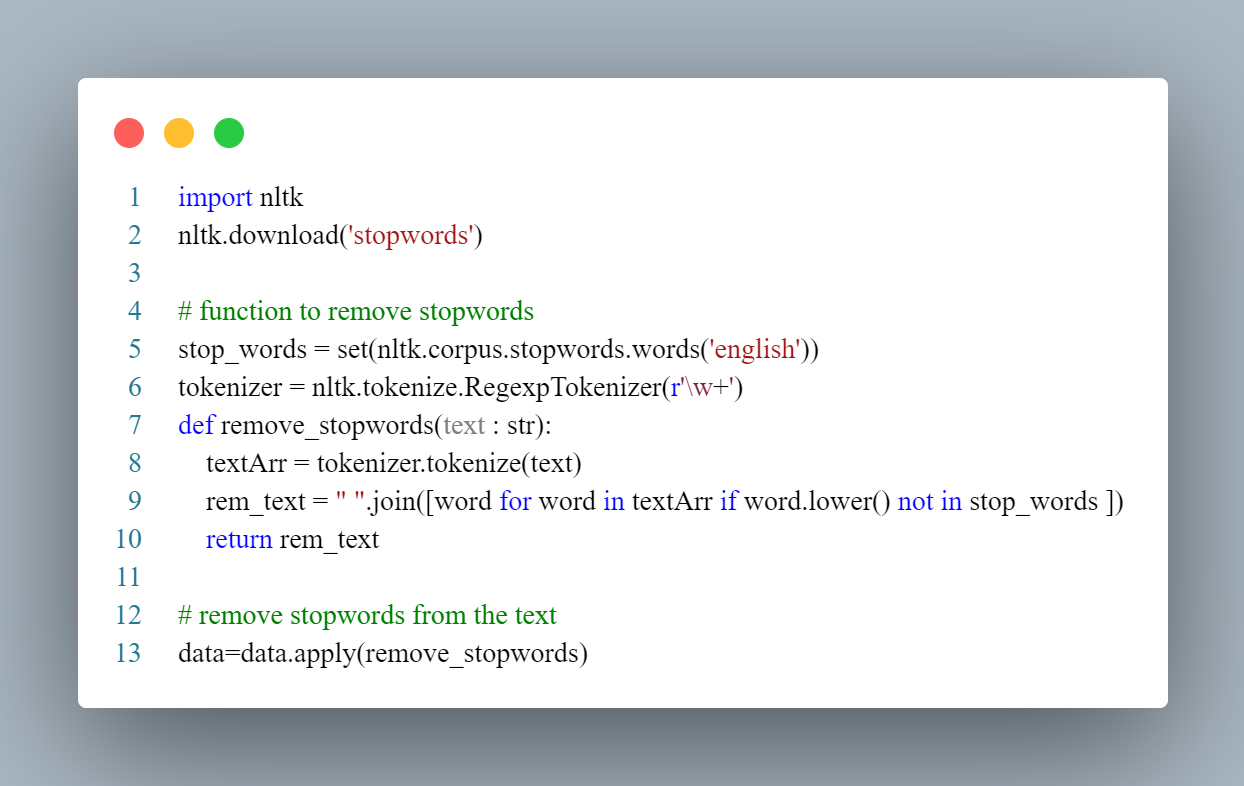
Topic modeling is an unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

Methodology:

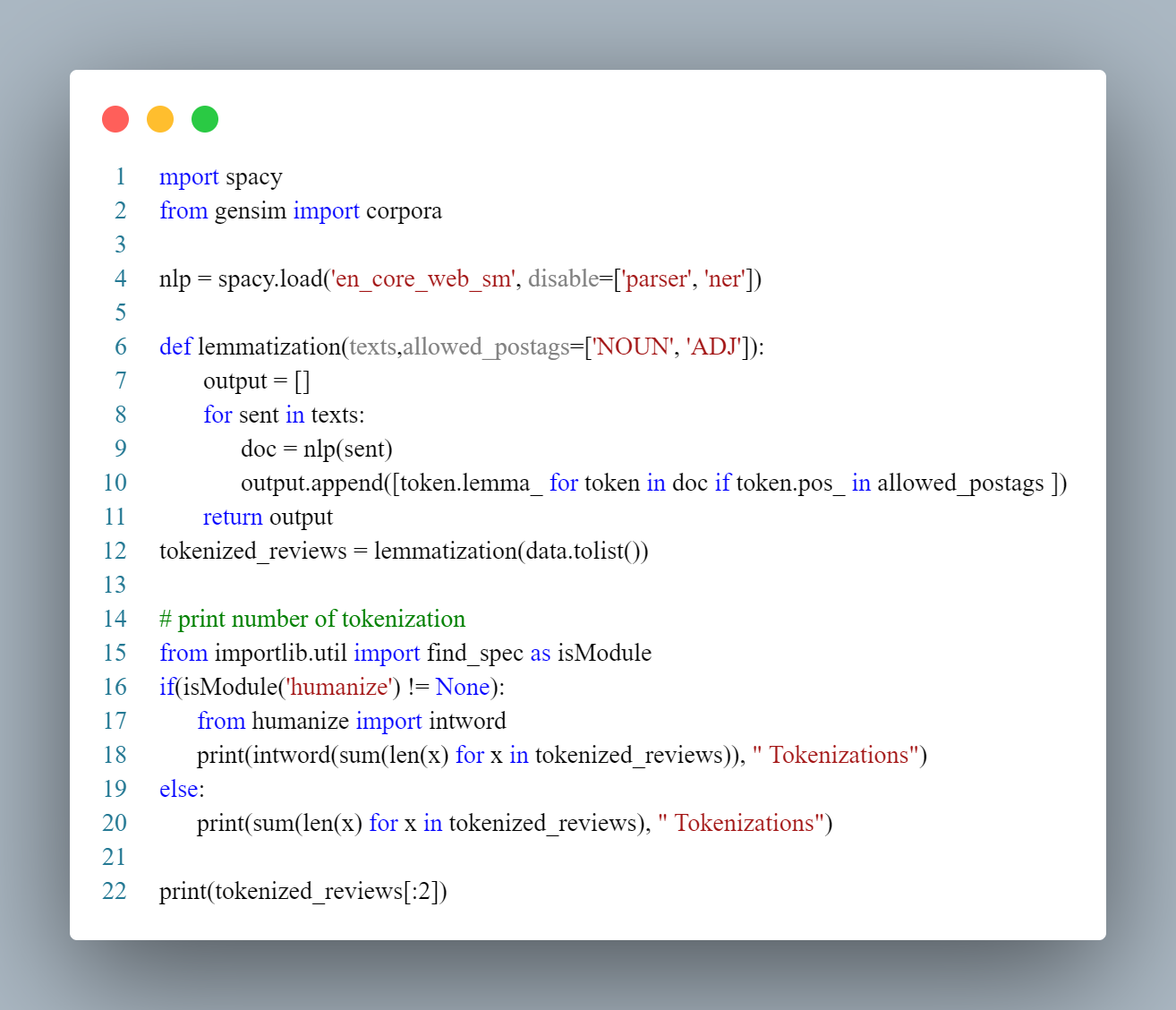
To cluster the topics, we found out that we only need the content row, so we dropped the rest of the features. Also, we dropped null and duplicated values in content column.



Then we remove the stop words in the articles.

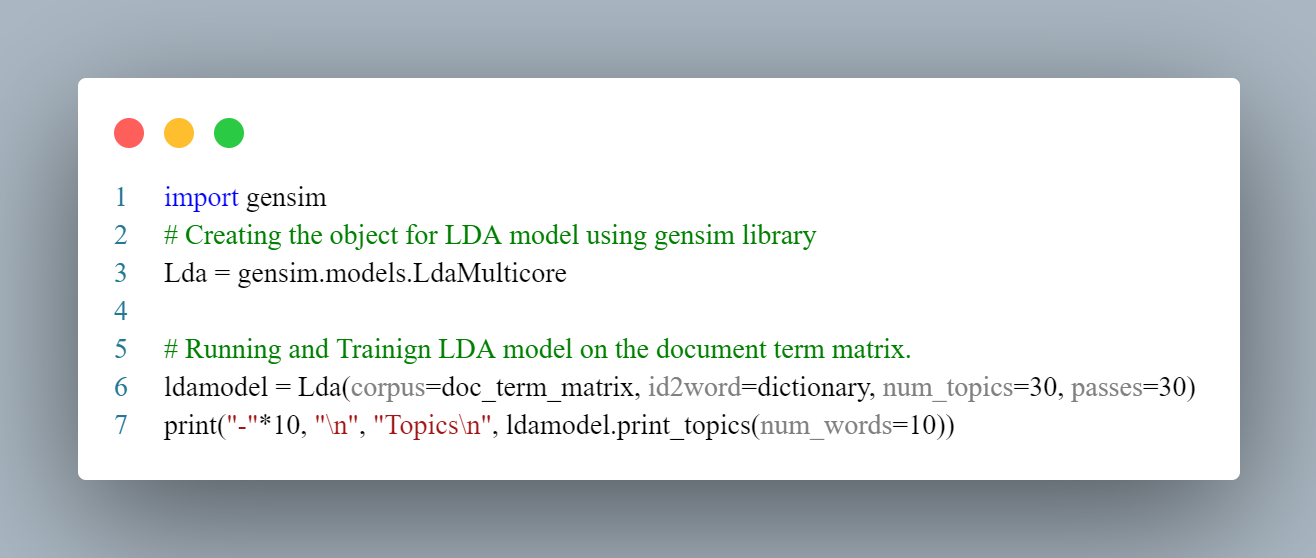


And we Lemmatize.



The simplest way to find a group of words that represents the topics of the document is that we find the words with the most frequency. But it’s not the most accurate way to do it, so we use LDA.

 the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

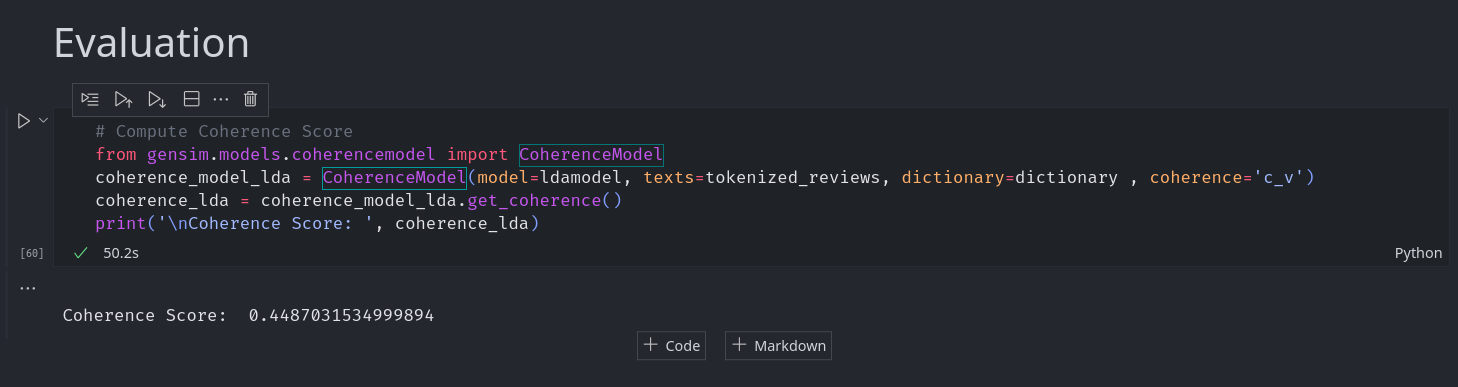
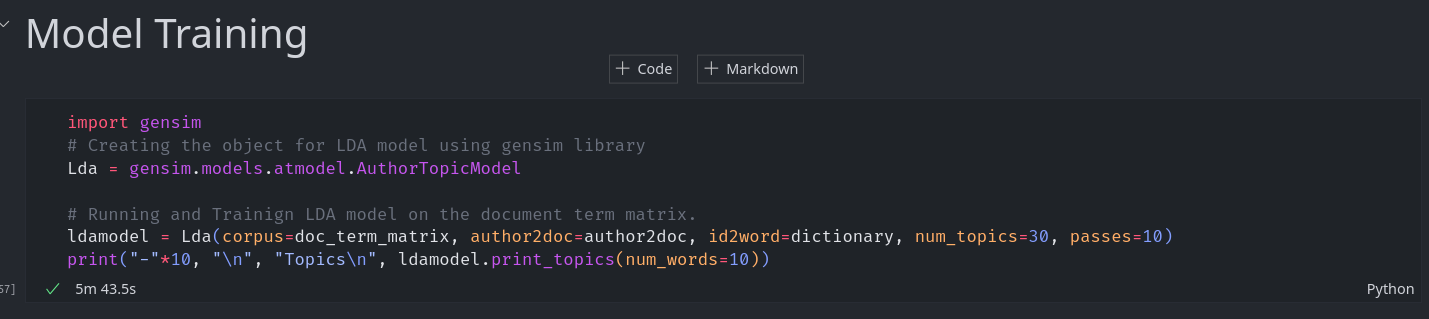


Data Set Summary:

1-What is the data set used?

We used the given data set, which named: articales1.csv

2- What is the summary of the dataset columns?

1. Id: A unique column and doesn’t relate to model training.
2. Title: it has small number of words so it doesn’t affect frequency.
3. Author: We tried to use the author, but it decreased the coherence, so we discard it   
     
   
4. Publication: a newspaper that doesn’t relate to the model training.
5. Date, year, and month: The time that the article released in and doesn’t relate to the model

3- Visualize the dataset statistics

Some topics with their probability.

Text, letter

Description automatically generated

Results:

