Project D: Twitter Data Analysis

* Full name: Justin Figueroa
* E-mail: jfigu042@ucr.edu
* UCR NetID: jfigu042
* Student ID: 862136079
*** Did Task 1**

* Full name: Anand Mahadevan
* E-mail: amaha018@ucr.edu
* UCR NetID: amaha018
* Student ID: 862132182
* **Did Task 2**

* Full name: Justin Do
* E-mail: jdo062@ucr.edu
* UCR NetID: jdo062
* Student ID: 862248675
*** Did Task 3**

Introduction:

This project consists of cleaning tweet data for training and test data for the building of a machine learning model that assigns a particular topic to a tweet. Precision and recall is then computed to determine the efficacy of the model.

For Task 1, **Spark Dataframe** was used to facilitate the use of multiple SQL queries, as well as Spark's built-in explode function, which reduced the code needed to count the top 20 hashtags.
For Task 2, **Spark Dataframe** was used because the array_intersect method belongs to it, which simplifies the process of filtering out hashtags immensely.
For Task 3, **Spark Dataframe** was chosen because it has a machine learning library that is simple and efficient to use.

QUESTIONS:

Task 1:
In the report, include the top 20 keywords you found for the 10k dataset.
ANSWER:
"ALDUBxEBLoveis","FurkanPalalı","no309","LalOn","chien","job","Hiring","sbhawks","Top3Apps","perdu","trouvé","CareerArc","Job","trumprussia","trndnl","Jobs","ShowtimeLetsCelebr8","hiring","impeachtrumppence","music"

Task 2:
In the report, include the total number of records in the tweets_topic dataset for the 10k dataset.
ANSWER:
269

Task 3:
Compute the precision and recall of the result you found and include them in the report for the 10k dataset.
ANSWER:
0.9023458896417154