E4577 – Assignment 2: Data annotation

In this assignment, you will be running a data annotation task as well as a data analysis task.

1. Download the Twitter dataset from S3: https://aiops-2020-public.s3.us-east-2.amazonaws.com/twitter_stream_2019_05_01.tar
Un tar it and upload it to your S3 at the following path: s3://<you_bucket>/twitter/state=raw/
Make sure to keep the internal directory structure (./01/00 …)

2. Create a Glue crawler and run it on the top level "twitter" directory.
You should now have a Catalog database with your raw dataset schema.

3. Go in Athena and run a query to select only English tweets using the "user.lang" field.
The data should be stored under s3://<you_bucket>/twitter/state=selected as a new table using the Athena "CREATE TABLE" functionality

4. Select a subsample of 1000 samples of your selected dataset as a csv and store it in a new location.
Make sure to filter out non English and garbage tweets as much as you can, even by going through a quick manual run through.
Use this data to create an AWS SageMaker Ground Truth job to annotate the selected tweets for positive, neutral or negative sentiment.

5. Clone the repo https://github.com/pharnoux/columbia-aiops-glue-helper. You should find the Word Count ETL Job.
Use this code to create a Glue ETL job that produces an English Hashtag Word Count. Make sure to remove non-English tokens.
Don't forget to replace the tags in the code (<your_database>, <your_table> and <your_path_to_s3>)
Visualize your result as a Word Cloud.

To Help you developing this code, you will also find glue_dev_endpoint.py (instructions an in the README.md)

This script is going to spawn a spark cluster and allow you to connect to it.
While in there you should be able to develop your code interactively. You can simply copy/paste your code and see if it works.

BE SURE TO DELETE THE CLUSTER ONCE FINISHED !
This costs about $2/hour !!

Otherwise, you can simply run the Glue ETL Jobs but that’s going to take much longer.

6. Bonus question: Write an ETL job that produces the top trending English Hashtag every minute and find a nice way to visualize it.