CS6350

Big data Management Analytics and Management

Spring 2019

Homework 3

Submission Deadline: 7th April 2019

In this homework, you need to extract features and build a classifier over a stream of news articles. The task is to gather real-time news articles using stream tool provided by Guardian API. This part of the code (stream_producer.py) has been provided by us for your implementation. To use this API, you will need to sign up for API key here:

https://open-platform.theguardian.com/access/

Stream_producer.py is the script that generates the Kafka streaming data from Guardian API every 1 second. The topic created has been named 'guardian2'. Run this from your command prompt window. You should see the news getting printed on the screen in the following format:

Label index (news category) || headline + bodyText

Each news article will have one of the 32 categories, such as Australia news, US news, Football, World news, Sport, Television & radio, Environment, Science, Media, News, Opinion, Politics, Business, UK news, Society, Life and style, Inequality, Art and design, Books, Stage, Film, Music, Global, Food, Culture, Community, Money, Technology, Travel, From the Observer, Fashion, Crosswords, Law. This above-mentioned script can capture a category for each news article and assign a label index for that category.

Simply type the following command on your command prompt to run this script:

python3 stream_producer.py API-key fromDate toDate

Ex: python3 stream_producer.py API-key 2018-11-3 2018-12-24

You need to set up standalone Spark and Kafka on your own system.

In Spark context, you need to create a Pipeline model with Tokenizer, Stopword remover, Labelizer, TF-IDF vectorizer, and a Classifier. You can use different classification techniques (two classifiers by your choice) and present performance results (i.e., Accuracy, Recall etc.).

Please note that, first, you need to train your model with offline data. To collect the offline data, you can save some news article on your machine by using 'Stream_producer.py'. Then, apply this model on the stream of data (window) with different 'fromDate' and 'toDate' of news articles. Each window of the stream of data will constitute a batch.

The label (news category) can be solely used in the training process. Furthermore, during testing, the labels can be utilized to report accuracy, recall, and precision for each batch.

You are allowed to use NLP processing tools, python Scikit-learn and MLlib in Spark packages.

Submission::

You have to upload your submission via e-learning before the due date.

Please upload the following to eLearning:

1. Source files

2. Readme file

2. Some screenshots of the output of your program and classifiers' performance