Assignment 6: Text Classification using Apache Spark and TF-IDF Representation

phuonglh@gmail.com

May 9, 2023

1 Problem

In this assignment, you will implement a text classifier using Apache Spark. Your program should be very similar to a guided example that your professor demonstrated in class. Some differences include:

- You need to put all the code to a source file instead of using the shell;
- You will use the TF-IDF vector representations of given texts instead of the count vectors;
- You will use the logistic regression classification method rather than the naive Bayes model;
- You will arrange all processing stages into a pipeline so that it needs to call `fit()` and `transform()` methods once.

You use the `sample.txt` data file that was provided by your professor. You should split this file into two parts randomly: a training part of 80% and a test part of 20% of the dataset. The pipeline is fit on the training part, and applied (transformed) on both parts. Your program need to report the training accuracy and the test accuracy on these parts, respectively.

2 Submission

If you develop your program in Scala, Python, or Java, submit your source file *YourFullName.scala*, *YourFullName.py*, or *YourFullName.java*, respectively; for example *NguyenPhuTrong.scala*. If you have multiple source files then put them in a directory and zip it to *YourFullName.zip* and submit before due date.