Introduction to Multimedia Big Data

Final Project Report



Amazon Product Data Analysis and Machine Learning

Michael Tjitra

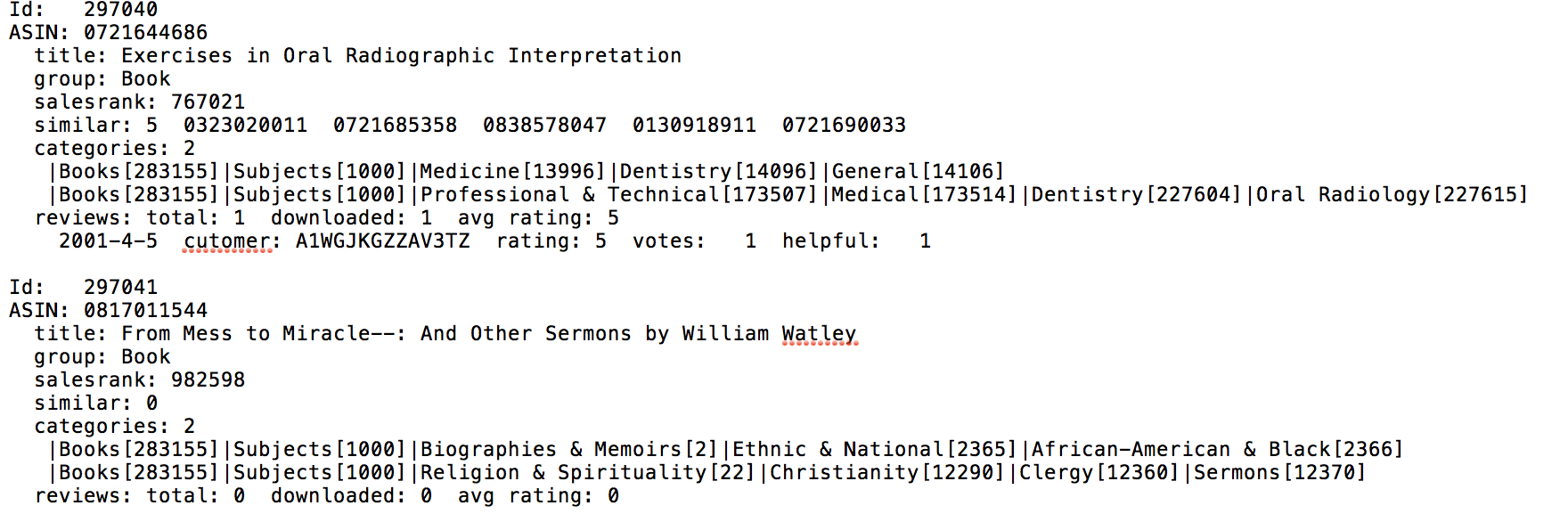
2018130172

1. Background

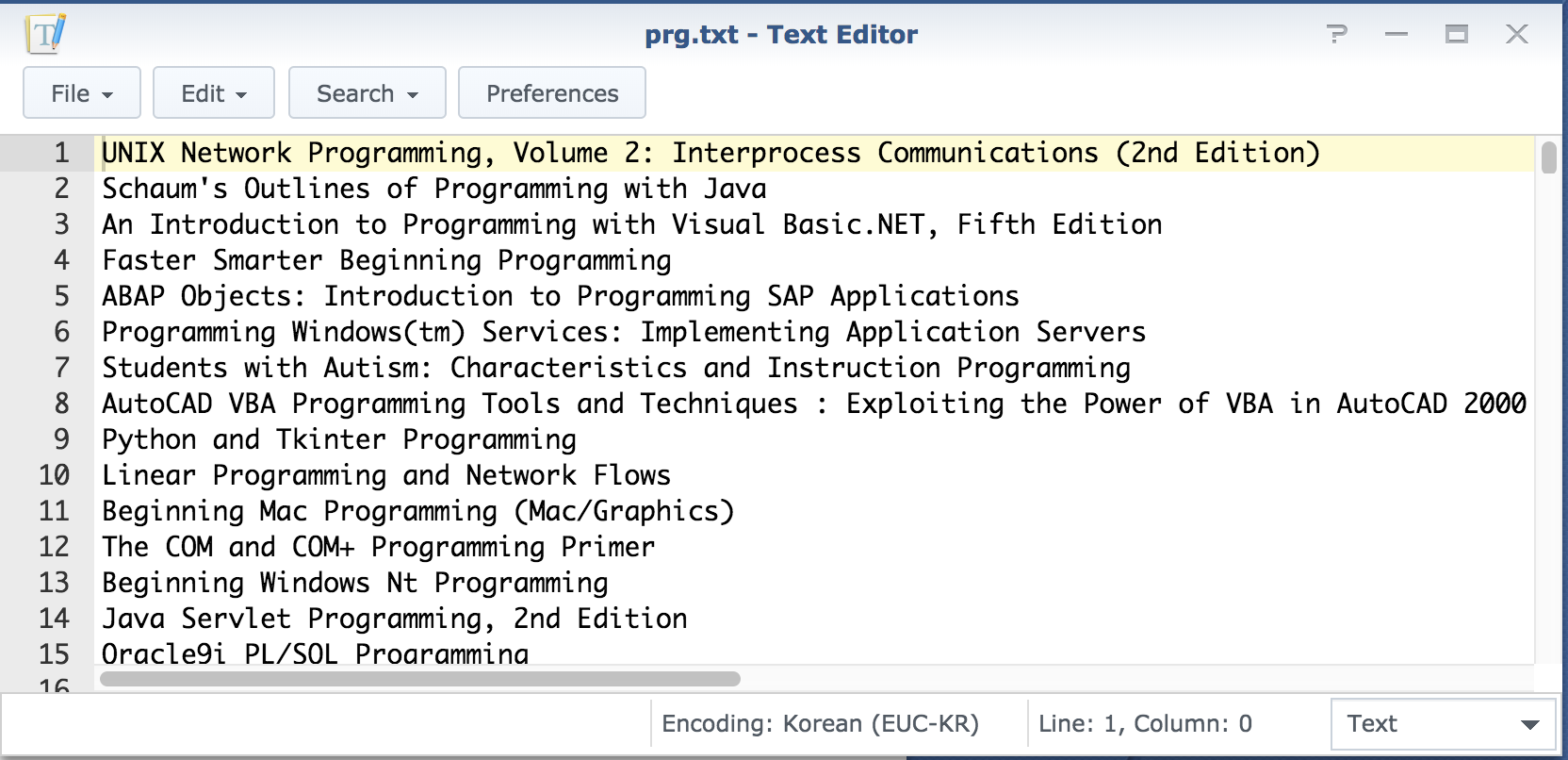
With the increase of technology usage in the business world right now there are a need to use Big Data analytics technology. Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

In this project I try to get some valuable data from the amazon product dataset and do a machine learning to predict if the product title is about programming or not.

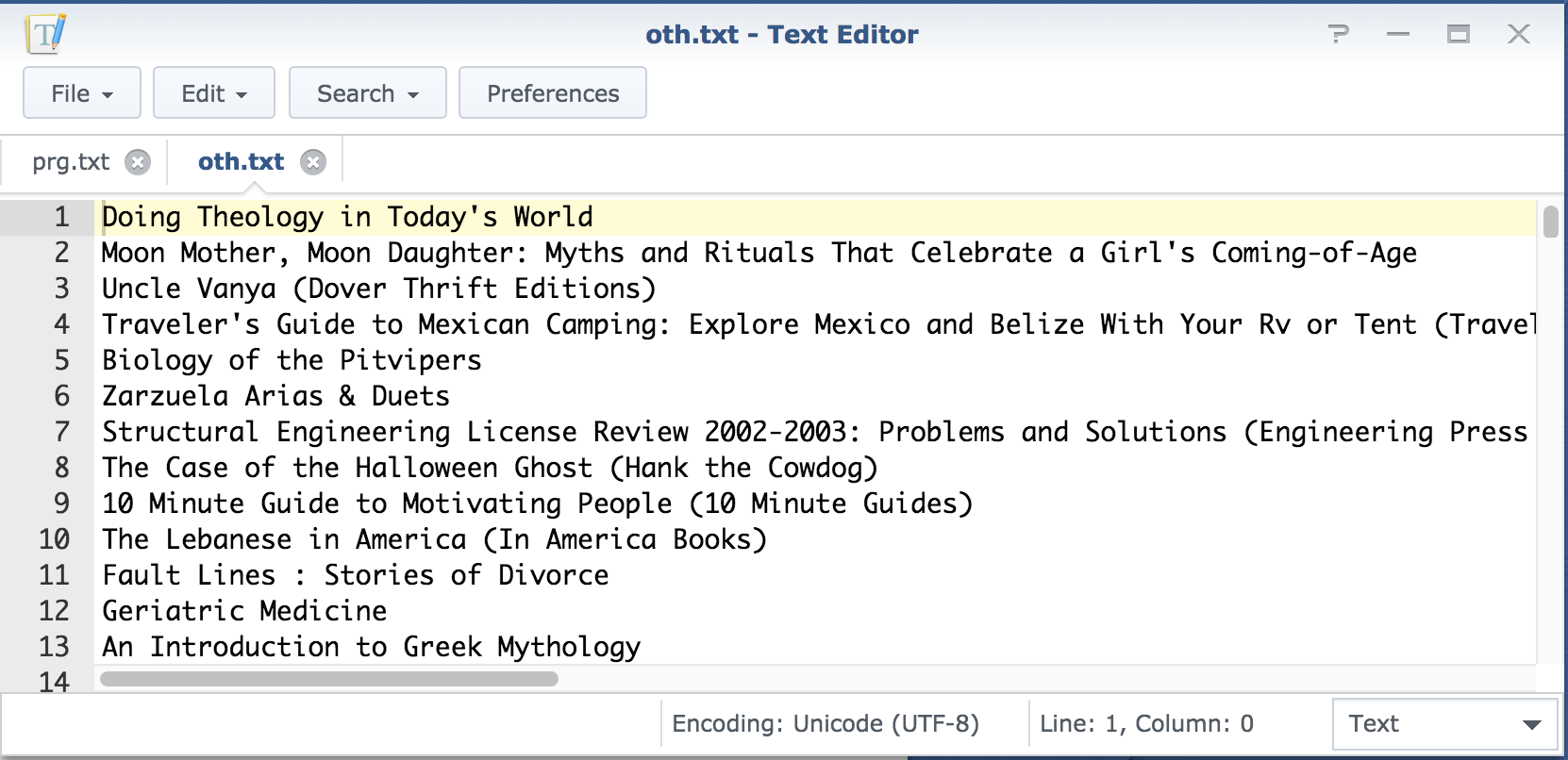
1. Data that want to be analyze
   1. Finding how much product in the dataset
   2. Finding how much product that get discontinued
   3. Finding the amount of product based on each group
   4. Finding top 10 customer that review a lot of product
   5. Do machine learning classification to classify its programming product or not based on the title
2. Equipment
   1. Raspberry pi
   2. 16G SD card
   3. Heatsink
   4. 5V/2A smartphone charger
   5. Raspberry pi 7-inch touchscreen display
   6. Internet connection
3. Dataset
   1. Amazon product dataset



* 1. Programming title dataset

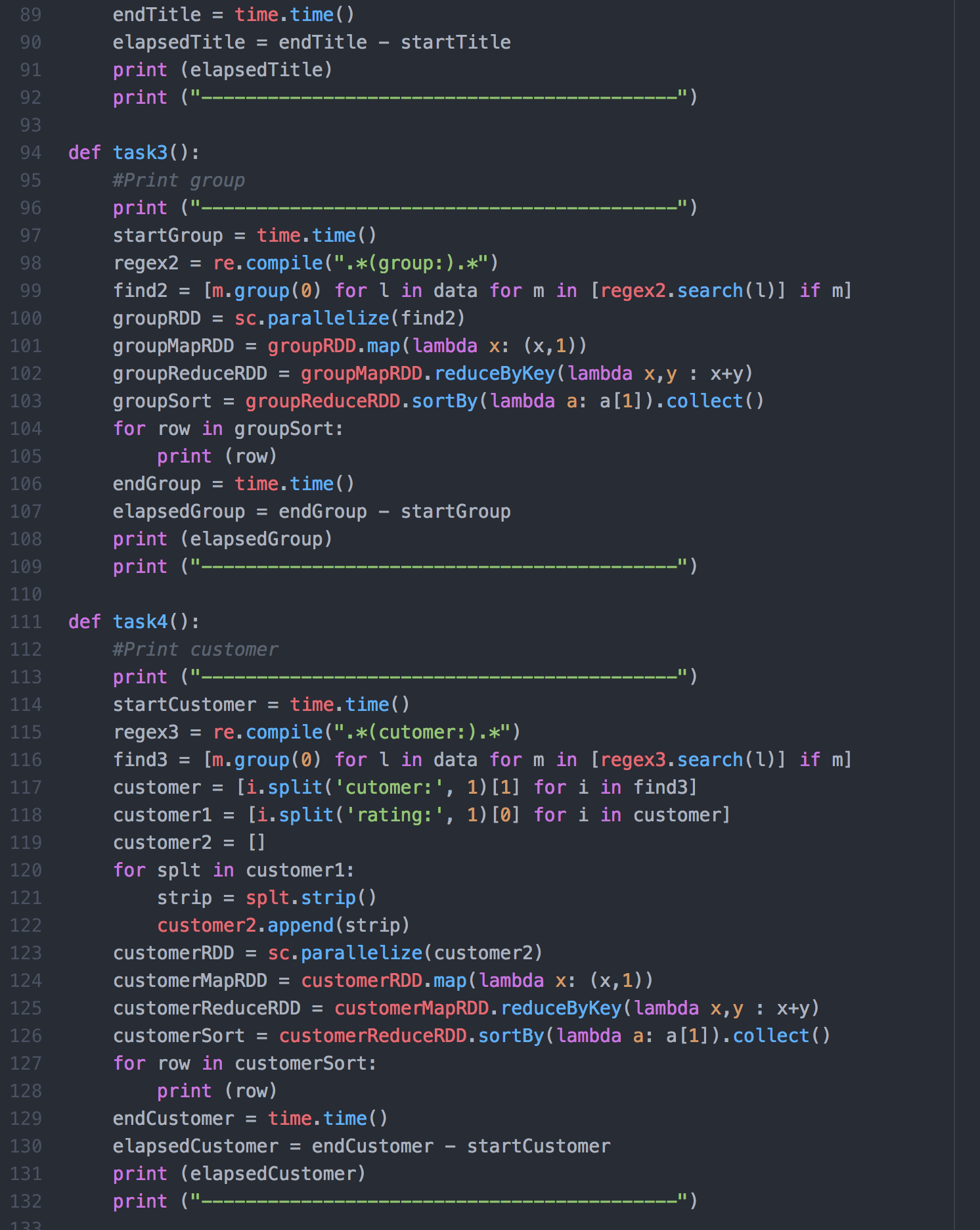
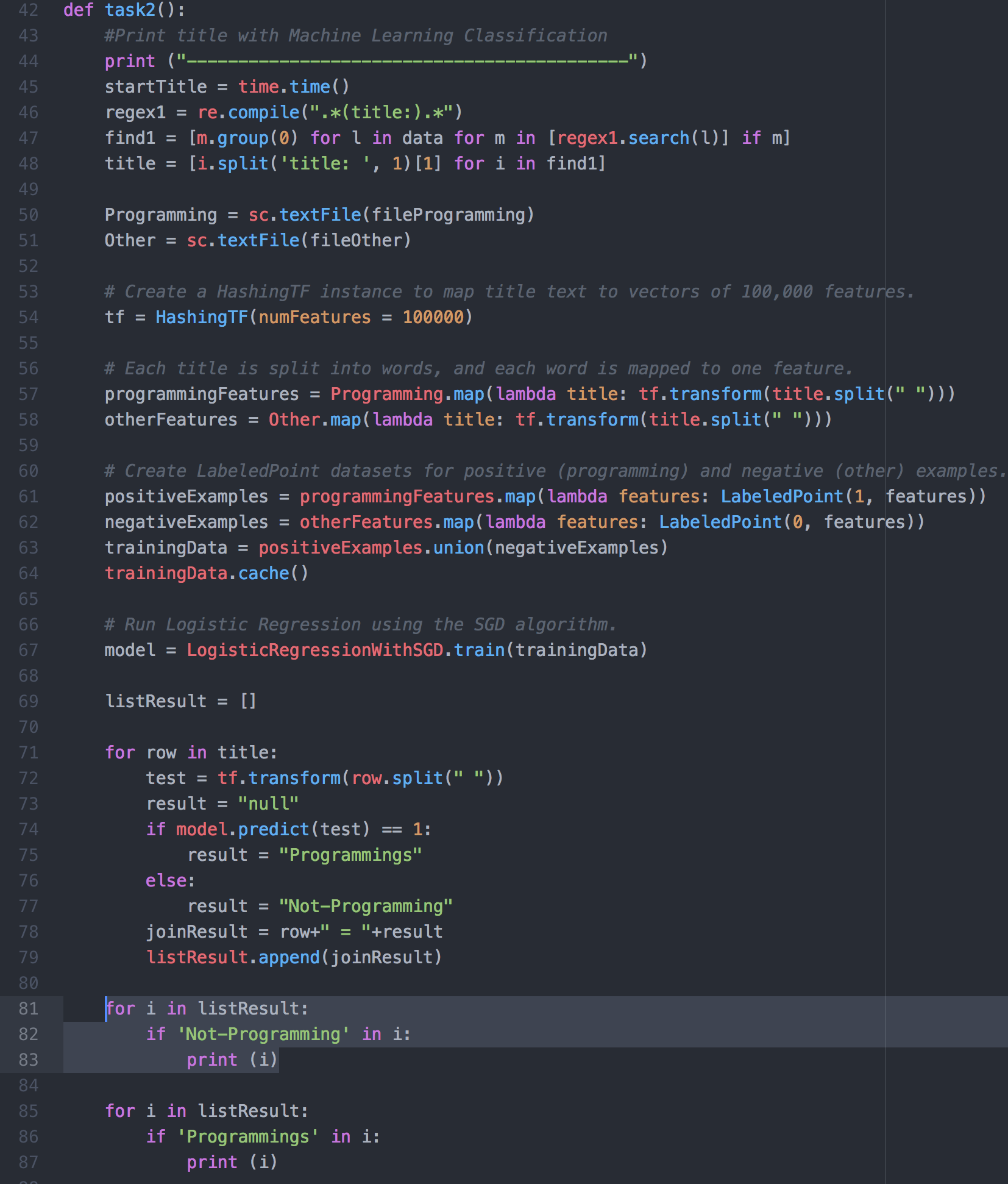


* 1. Other title dataset

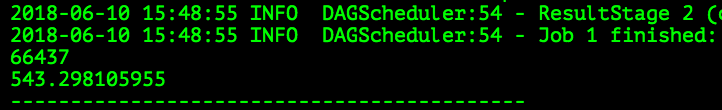


1. Source code



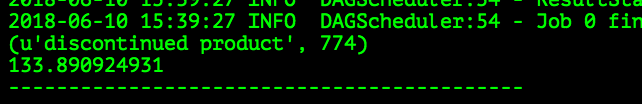


1. Step to run the program
   1. Turn on raspberry pi machine
      1. Username: pi
      2. Password: m1ch43lz
   2. Open terminal
   3. Type cd Document and hit enter
   4. Type spark-submit finalProject.py and hit enter
2. Outcome/Result
   1. Finding how much product in the dataset



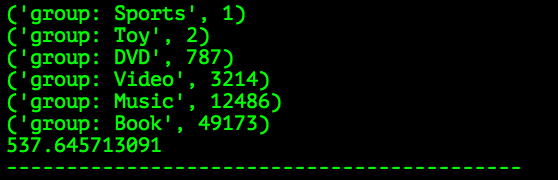
Time to completion = 9 minutes

* 1. Finding how much product that get discontinued



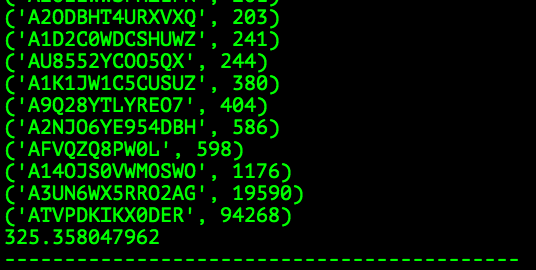
Time to completion = 2.2 minutes

* 1. Finding the amount of product based on each group



Time to completion = 8.96 minutes

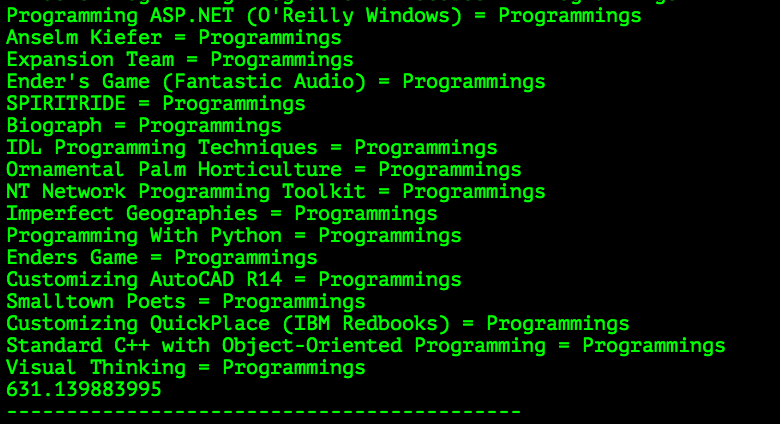
* 1. Finding top 10 customer that review a lot of product



Time to completion = 5.4 minutes

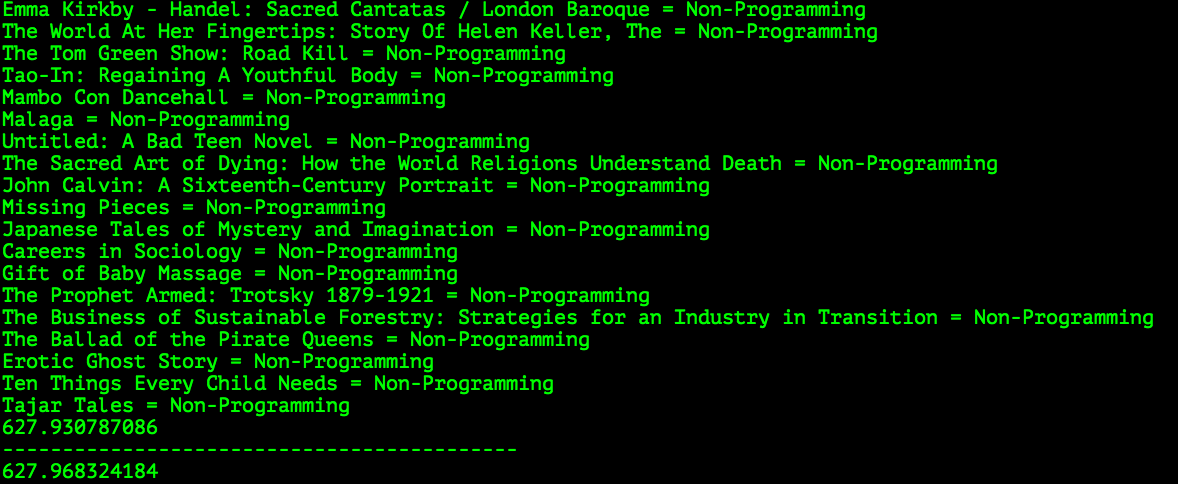
* 1. Do machine learning classification to classify its programming product or not based on the title

Programming Title:



Time to completion = 10.5 minutes

Non-programming Title:



Time to completion = 10.4 minutes

1. Interesting finding

When raspberry pi uses multi-processing function time reduce from ±50 minutes to only ±20 minutes but the drawback of using that function is Raspberry Pi easily get overheat because that function use all cores in the same time. When not using multi-processing function the Raspberry pi temperature only 50°C but when it gets use Raspberry Pi temperature get more than ±80°C

1. Limitation when building the project
   1. Raspberry Pi low memory size

Problem solving:

1. Reduce dataset from 1Gb to 100Mb
2. Increase swap size in Raspberry Pi to 2Gb
   1. Raspberry Pi low computing power
   2. Raspberry Pi stop working when overheat

Problem solving:

Open Raspberry Pi back cover and put it directly in front of the AC

1. References

https://www.sas.com/en\_us/insights/analytics/big-data-analytics.html