Assignment 2

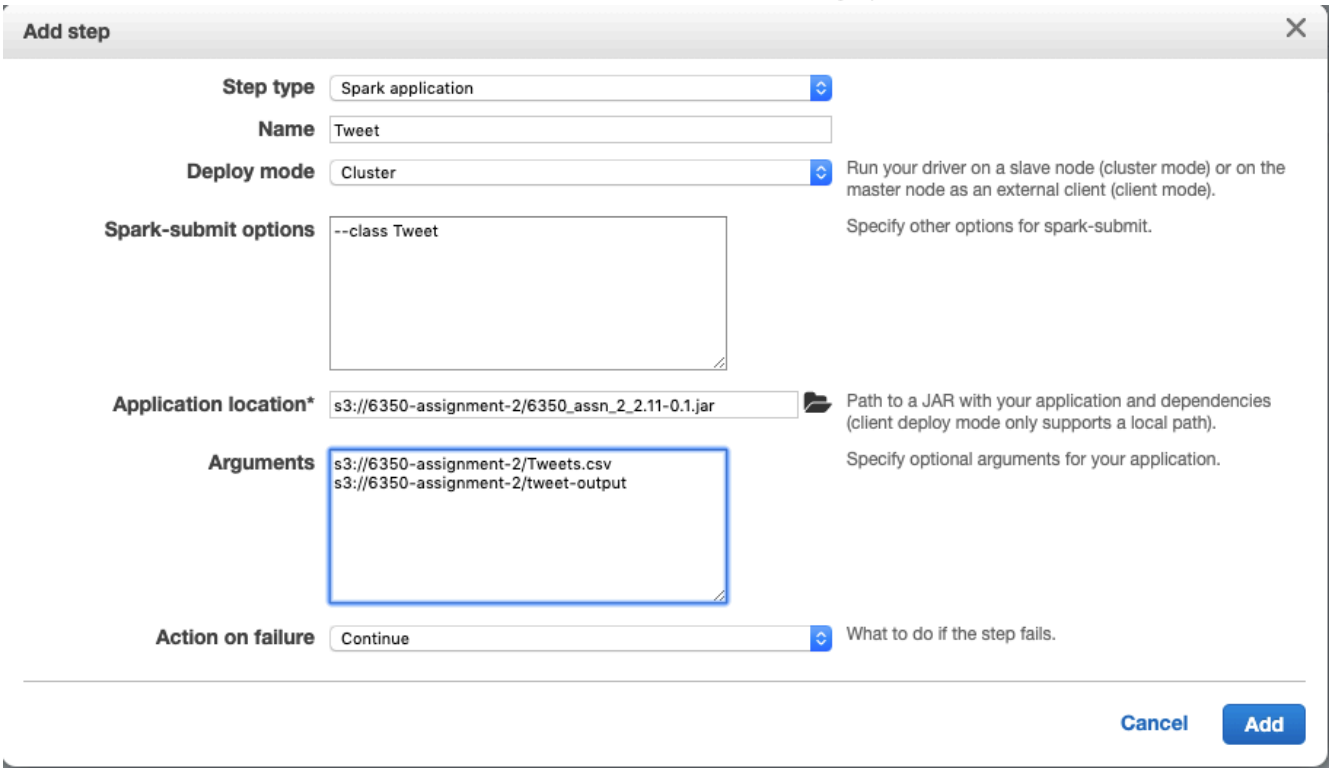
Language: Scala
IDE: Intellij
Platform: AWS

Part 1

Tweet Processing & Classification using Pipelines

How to run:

1. Build and package the project in Intellij to get a jar file named ***6350_assn_2_2.11-0.1.jar*** under the path /target/scala-2.11/.
2. Upload the jar file and the tweets.csv to AWS S3.
3. Run it on a cluster with the following parameters.



Result:

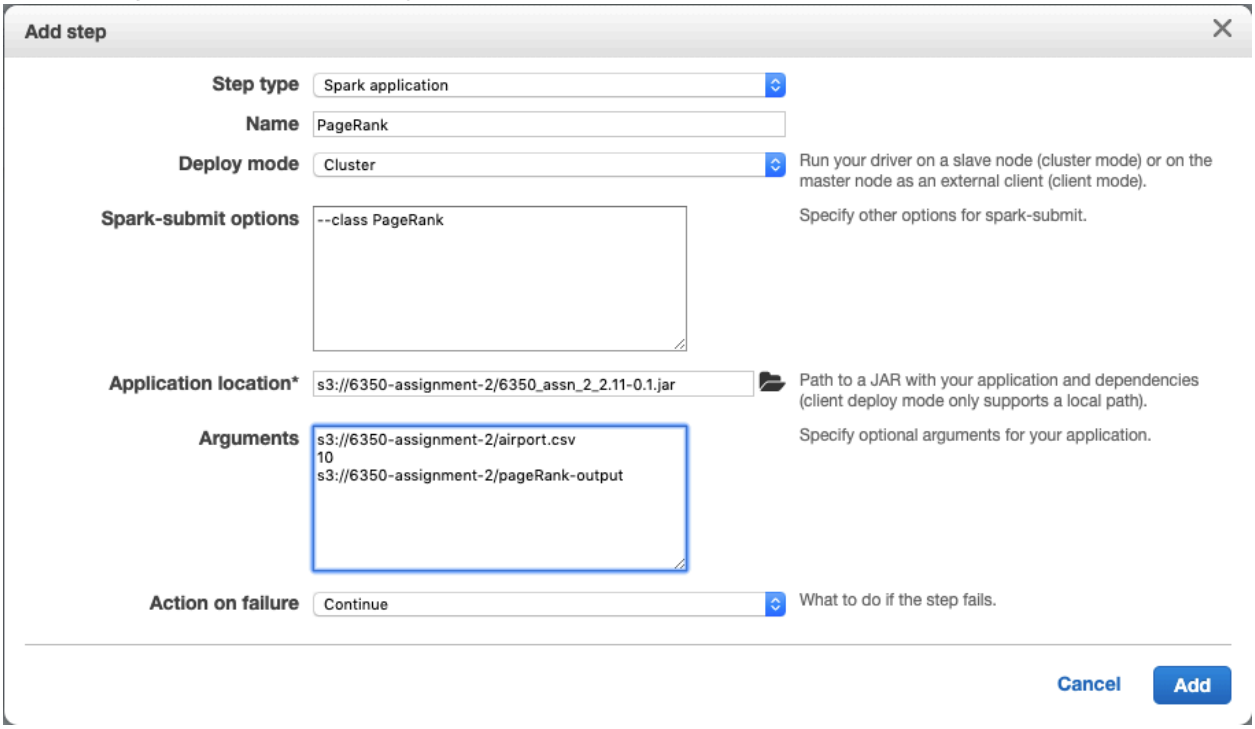
```
1  Accuracy: 0.7703381319049213
2
```

Part 2

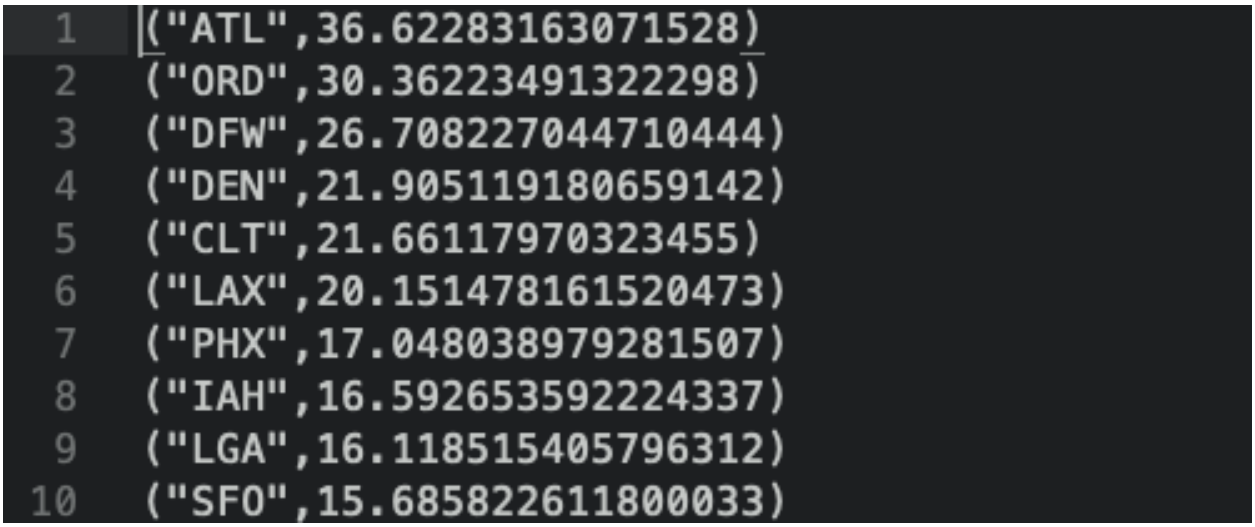
PageRank for Airports

How to run:

1. Upload the airport.csv to AWS S3.
2. Run it on a cluster with the following parameters using the jar file from part 1.



Part of Result:

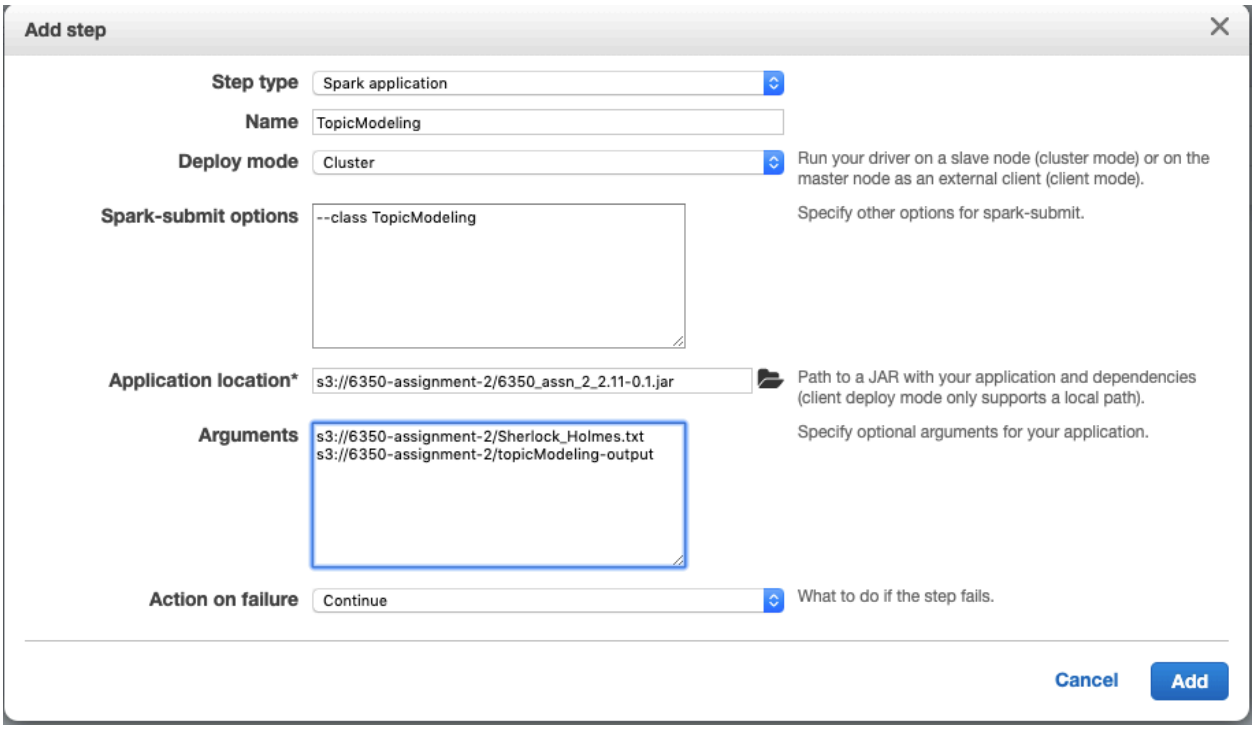


Part 3

Topic Modeling from Classic Books

How to run:

1. Download a book from the Gutenberg project http://www.gutenberg.org. Here we choose Sherlock holmes.
2. Upload the Sherlock_Holmes.txt to AWS S3.
3. Run it on a cluster with the following parameters using the jar file from part 1.



Result:

```
TOPIC 1:
took    0.003401641634888759
project 0.0033752398170424675
quite   0.0033725033017256356
without 0.003264842725623772
miss    0.0031649037539665556
good    0.003133985623192229
away    0.0031218365118898268
nothing 0.0031134844081667987
left    0.0030178158007912563
matter  0.002959855383350144

TOPIC 2:
quite   0.003434815321012206
nothing 0.0033264954557114717
good    0.0032908088048380977
tell    0.0032369444189662825
left    0.003206091796323911
every   0.0031140244169175196
face    0.0030624477458456715
make    0.0030245676409928718
took    0.0030059774811405674
away    0.0029483253710657726

```

```
TOPIC 3:
project 0.003408121720498146
quite   0.0032265720550807976
nothing 0.003146731978841166
every   0.0030595409592994558
tell    0.003008278458386203
face    0.002998439562566579
took    0.0029355547609897685
make    0.002899894972985773
good    0.0028683344047137405
last    0.0028535202974561413

TOPIC 4:
good    0.0035014259434134657
tell    0.0034594319567720815
project 0.0033519010169094096
every   0.0032213507641849983
quite   0.0031526040888385344
nothing 0.0031320610148372866
without 0.0031275376910154936
away    0.003124805150349058
face    0.003100970400044747
took    0.0030980735324592384

TOPIC 5:
miss    0.0035025775089876365
away    0.0033239803644040736
good    0.003259036216534128
take    0.003194759599812901
nothing 0.0031525390894828077
tell    0.0031449903465934057
young   0.0030756540109242716
quite   0.003049677855757757
door    0.003014718779274901
case    0.0029073308618538813

```