6350 Final Project readme

We have two modules:

Original data analysis

Machine learning

1 Original data ayalysis

1-1 Pig Part

File location: Source code/analysis_part1/

Task list:

1-1 analysis on task list:

1. overall sentiment(negative, neutral, positive percentage)

2. perfectage of tweets per airline

3. Proportion of negative/neutral positive sentiment tweets per airline

4. Reasons for negative sentiment tweets

5. Reasons for negative sentiment per airline

6. Tweet location exploration

7. Tweet timezone study

8. Location of tweets: Visualization on maps

Running method: You need to run the pig script on Hadoop interactively, or you can run it as pig script.

1-2 MapReduce Part

File location: Source code/ analysis_part2/

task list:

1 Find top 300 negative words in the negative comments in decreasing order
2 Find top 300 positive words in the positive comments in decreasing order
3 Choose 5-10 from top negative words, find their top 100 co-occurrence words in negative comments.

4 Choose 5-10 from top positive words, find their top 100 co-occurrence words in the positive words

Running method:

1 first you need to upload the tweets_Update.csv(it is in the input_data folder) to Hadoop cluster

2 copy the Final6350-0.0.1-SNAPSHOT.jar to CS6360

3 M3.java is for task 1, M4.java is for task2, Co1.java is for task3, Co2.java is for task4

4 run every java class with 2 parameters, the first parameter is the input file which is tweets.csv, the second parameter is the destination direction.

5 example:

hadoop jar Final6350-0.0.1-SNAPSHOT.jar Final6350.M3 Tweets_Update.csv final/part1/v1

hadoop jar Final6350-0.0.1-SNAPSHOT.jar Final6350.M4 Tweets_Update.csv final/part1/v2

hadoop jar Final6350-0.0.1-SNAPSHOT.jar Final6350.Co1 Tweets_Update.csv final/part1/v3

hadoop jar Final6350-0.0.1-SNAPSHOT.jar Final6350.Co2 Tweets_Uptae.csv final/part1/v4

notice: please make sure tweets.csv is located at Hadoop cluster with the same name

Machine Part

File location:

source code/machine learning_part1

Source code/machine learning_part2

The task list and running method are listed below:

Machine learning analysis part 1 task list:

//convert css to libsvm file(convert.py) (for Decision tree, Logistic Regression, Random Forest) (**not big data approach**)

eg: python convert.py sample_train.csv train.csv

you could directly use 1.csv and 2.csv in input data

//template.jar

//for 3 class label

Decision tree:

spark-submit --class template.template.dt --master yarn template.jar 2.csv 30 entropy 3

Logistic Regression

spark-submit --class template.template.lr --master yarn template.jar 2.csv 3

Naive Bayes:

spark-submit --class template.template.naive --master yarn template.jar 1000.csv

Random Forest:

spark-submit --class template.template.random --master yarn template.jar 2.csv 3

//for 12 class label

Decision tree:

spark-submit --class template.template.dt --master yarn template.jar 1.csv 30 entropy 12

Logistic Regression

spark-submit --class template.template.lr --master yarn template.jar 1.csv 12

Naive Bayes:

spark-submit --class template.template.naive --master yarn template.jar TweetsAll1000.csv

Random Forest:

spark-submit --class template.template.random --master yarn template.jar 1.csv 12

machine learning analysis part 2 task list:

First, make sure installed numpy, sklearn, re and naiveBayesClassifier package in Python 2.7.
Change the path to each python file.
Then, you need to add permission to each py file, using commend line, for example typing:
chmod +x code.py
Last, run the code, for example typing:
./code.py

3 Pre Processing of data (not bigdata approach)

File location: source code/pre process/data

Running method:

Please put the tweets.csv and stopWords at the same location with the java file

Javac PreProcessData.java

Java PreProcessData

It will generate four sample4 with a total of 12000 lines, please make sure the full file with tweets.csv is located here.

4 Output

We have already put all parts output at the folder

Ouput

please refer all parts in each folder under this directory