CIS5560 Term Project Tutorial

Authors: Tom Cruise; Mel Gibson

Instructor: Jongwook Woo

Date: 05/18/2017

Lab Tutorial

yourname (yourname@calstatela.edu)

06/10/2016

Yelp Data Analysis using Spark (your Title)

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually using REST API
- Create Spark cluster
- Train NLP system
- SQL commands to perform the analysis.
- Visualization

Platform Spec

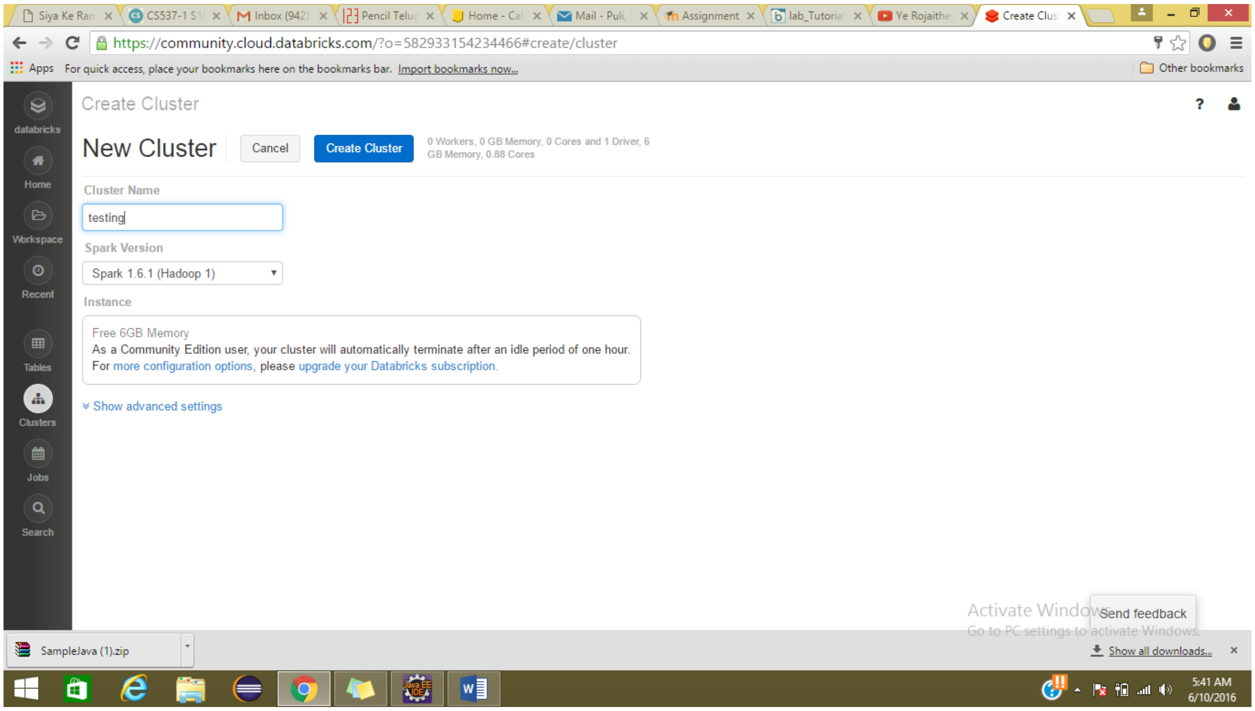
- IBM Bluemix BigInsights
- CPU Speed: ?

- # of CPU cores: ?
- # of nodes: ?
- Total Memory Size: ?

Step 1: Get data manually using REST API

Explain what this step is for. This step is to get data manually….

1. Create Google API keys at https://develop:
2. Sign into your databricks account.
3. Go to Clusters option on the left and click on create cluster.
4. Give the cluster name and click create cluster.



https://community.cloud.databricks.com/?o=582933154234466#create/cluster

Create Cluster

New Cluster

0 Workers, 0 GB Memory, 0 Cores and 1 Driver, 6 GB Memory, 0.88 Cores

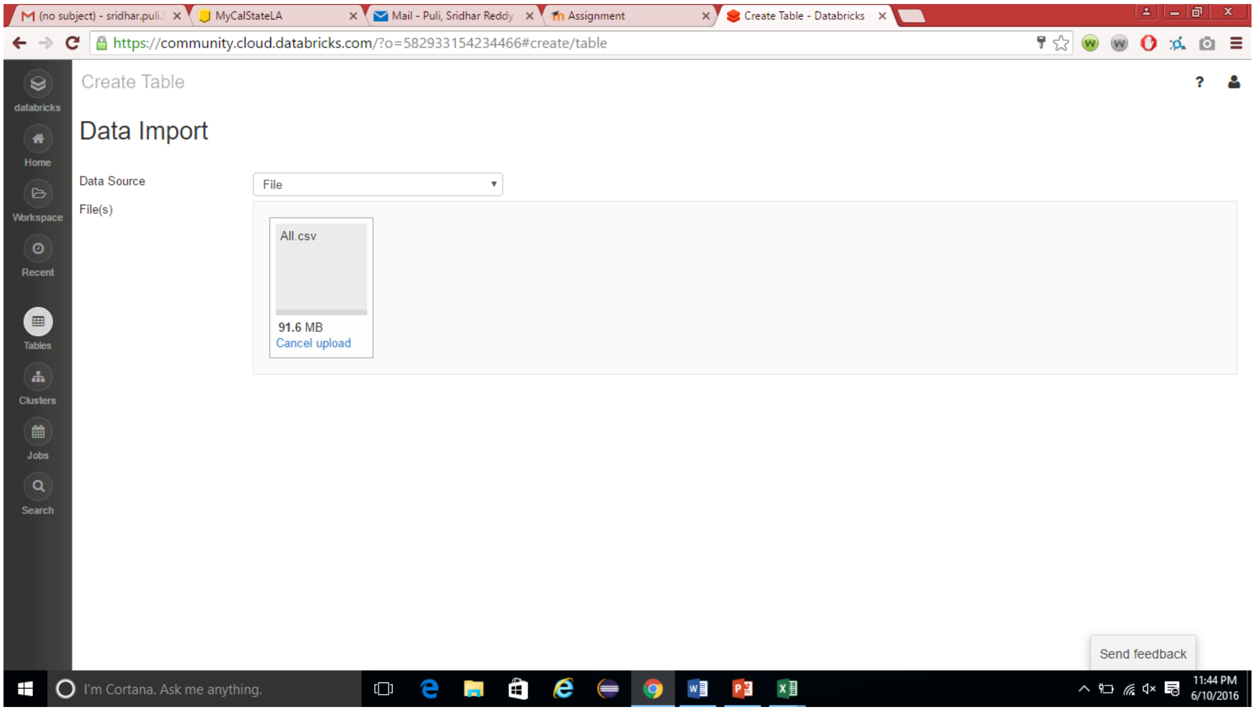
Cluster Name: testing

Spark Version: Spark 1.6.1 (Hadoop 1)

Instance: Free 6GB Memory

As a Community Edition user, your cluster will automatically terminate after an idle period of one hour. For more configuration options, please upgrade your Databricks subscription.

5. Under tables section click on create table and select the file to upload.


(no subject) - sridhar.puli.
MyCalStateLA
Mail - Puli, Sridhar Reddy
Assignment
Create Table - Databricks
https://community.cloud.databricks.com/?o=582933154234466#create/table
databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
?
Cancel upload
Send feedback
I'm Cortana. Ask me anything.
11:44 PM
6/10/2016

Create Table

Data Import

Data Source: File

File(s): All.csv, 91.6 MB

Step 2: Train NLP

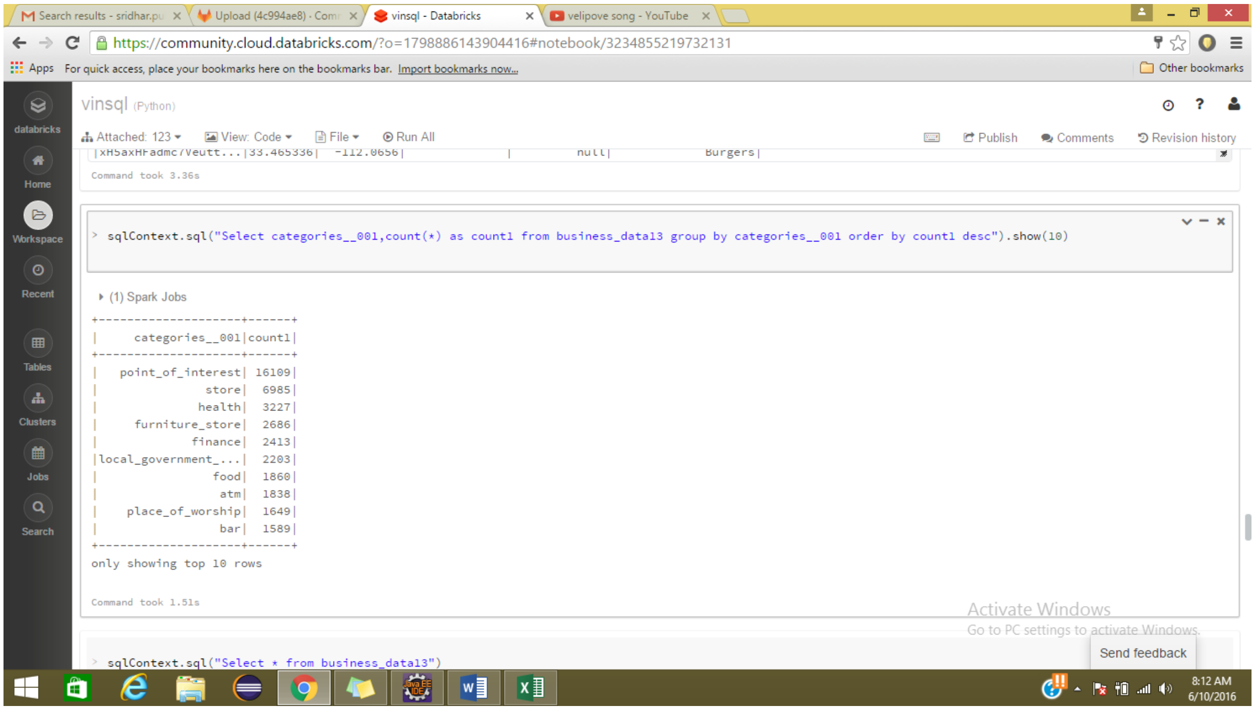
Explain what this step is for. This step is to …
Code should be in the following format and indent:

```
import org.apache.spark.ml.feature.RegexTokenizer
val tokenizer = new RegexTokenizer()
  .setPattern("\\p{L}+").setMinTokenLength(3)
.setGaps(false)
  .setInputCol("text")
  .setOutputCol("words")

val tokenized_df=tokenizer.transform(splits(0))

vi) Use the below code to remove stop words
Run them in separate cells for better understanding

%sh wget
http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words -O
/tmp/stopwords
%fs cp file:/tmp/stopwords dbfs:/tmp/stopwords
val stopwords = sc.textFile("/tmp/stopwords").collect()

import org.apache.spark.ml.feature.StopWordsRemover
// Set params for StopWordsRemover
val remover = new StopWordsRemover()
```

```
    .setStopWords(stopwords) // This parameter is optional
    .setInputCol("words")
    .setOutputCol("filtered")

  // Create new DF with Stopwords removed
  val filtered_df = remover.transform(tokenized_df)
```

1. To show top ten categories

```
sqlContext.sql("Select categories__001,count(*) as count1 from

business_data13 group by categories__001 order by count1

desc").show(10)
```



```
|xH5axHFadmc7Veut...|33.465336|  -112.0656|                |          null|                 Burgers|
Command took 3.36s
```

```
> sqlContext.sql("Select categories__001,count(*) as count1 from business_data13 group by categories__001 order by count1 desc").show(10)
```

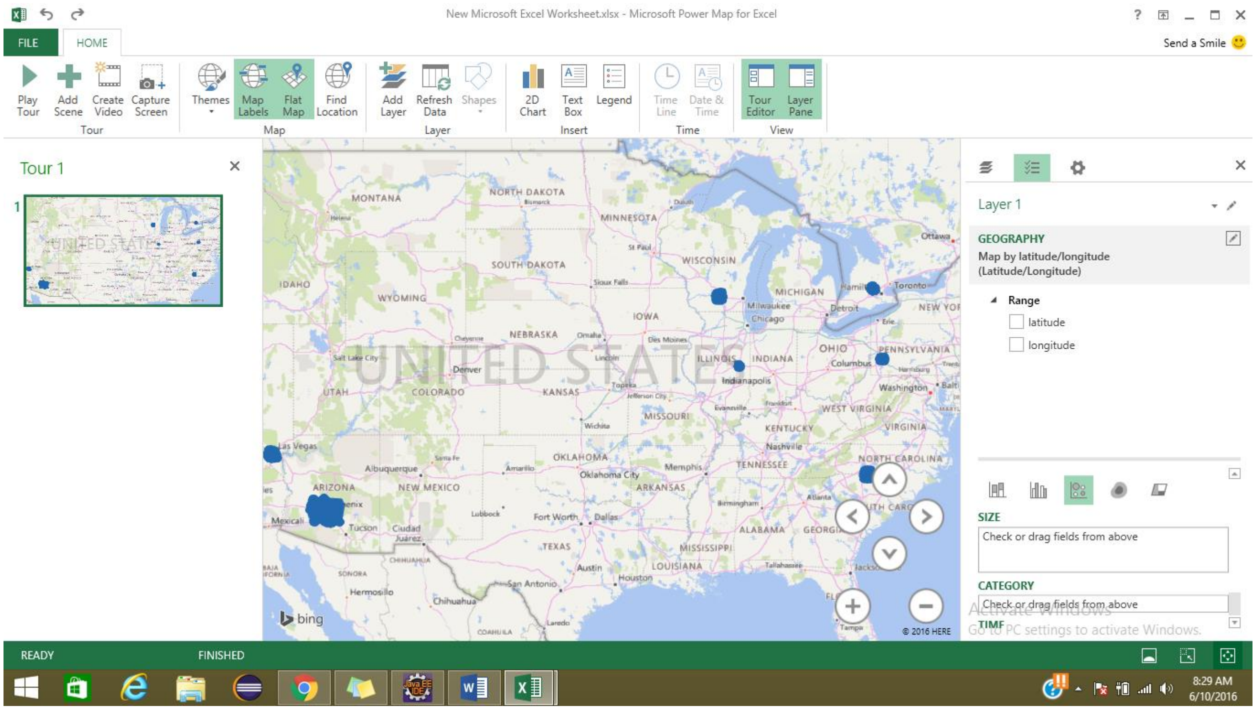
▸ (1) Spark Jobs

```
+--------------------+------+
|      categories__001|count1|
+--------------------+------+
|   point_of_interest| 16109|
|               store|  6985|
|              health|  3227|
|     furniture_store|  2686|
|             finance|  2413|
|local_government_...|  2203|
|                food|  1860|
|                 atm|  1838|
|     place_of_worship|  1649|
|                 bar|  1589|
+--------------------+------+
only showing top 10 rows

Command took 1.51s
```

```
> sqlContext.sql("Select * from business_data13")
```

Step 3: Visualization

Explain what this step is for. This step is to…



Layer 1

GEOGRAPHY
Map by latitude/longitude
(Latitude/Longitude)

Range
□ latitude
□ longitude

SIZE
Check or drag fields from above

CATEGORY
Check or drag fields from above

TIME

1. To visualize location type of results on map, convert csv file to excel and click on map button under insert tab.

References

1. URL of Data Source, http://www.calstatela.edu

2. URL of your Github

3. URL of References

4.