# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS5560 Term Project Tutorial

#### Authors:  Atharva Vikas Borole, Daniel Ozorio Delfin, Ryuichi Matsushita

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 04/29/2025

**Lab Tutorial**

Ryuichi Matsushita ([rmatsus@calstatela.edu](mailto:rmatsus@calstatela.edu))

04/29/2025

**Soccer Match Outcome Prediction Using Historical Betting Data with Spark**

**Objectives**

In this hands-on lab, you will learn how to:

* Acquire historical soccer data with betting odds
* Clean and prepare the data for analysis using Python and Spark
* Train a classification model to predict match outcomes (Win/Draw/Loss)
* Use SQL and DataFrame operations to perform feature engineering
* Visualize prediction results

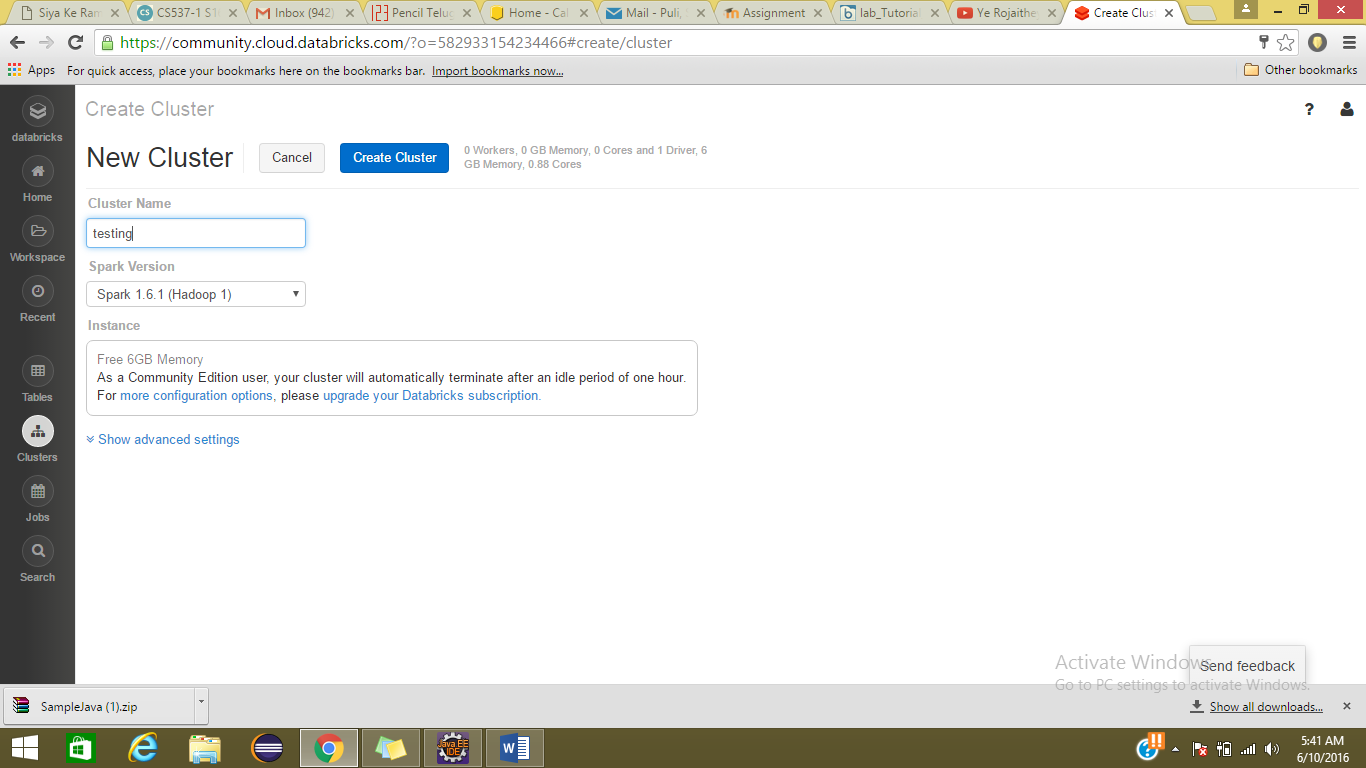
**Platform Spec**

* Databricks on AWS
* CPU Speed: ?
* # of CPU cores: ?
* # of nodes: ?
* Total Memory Size: ?

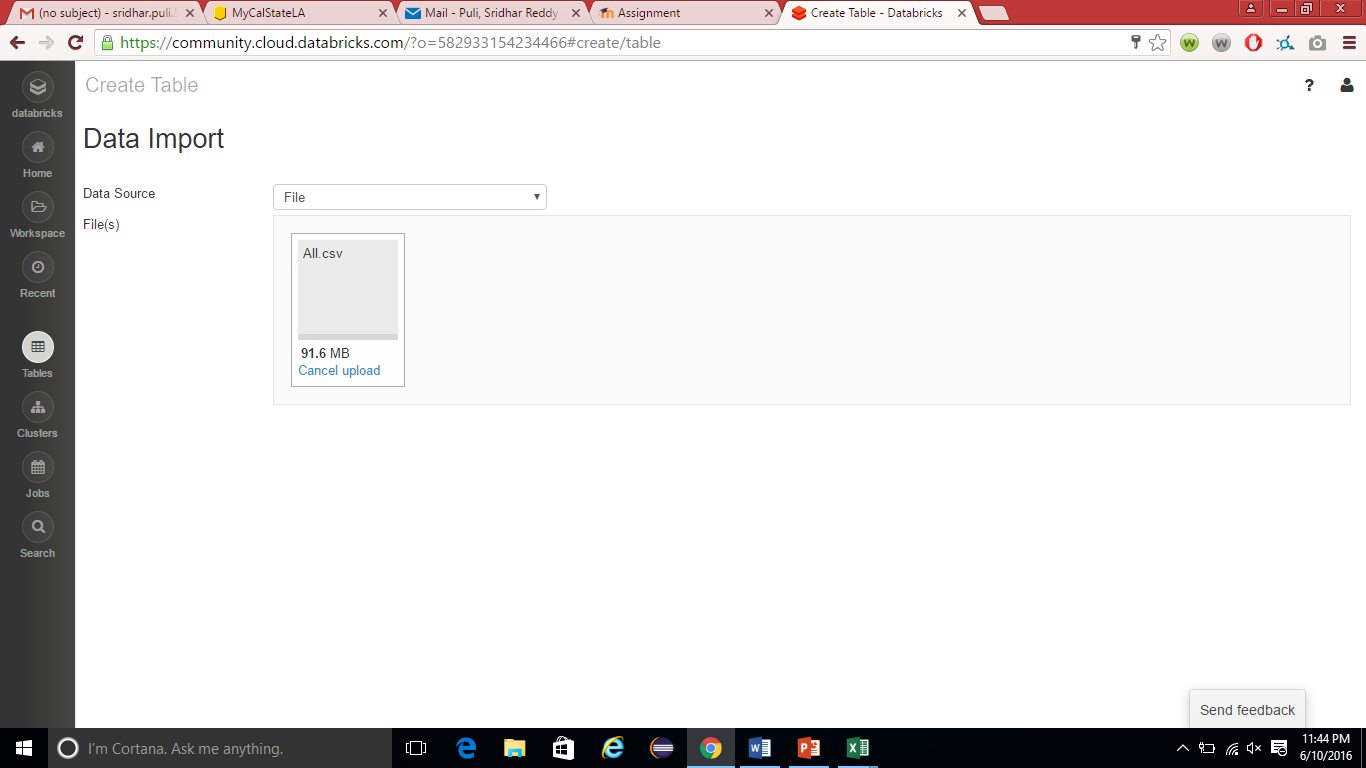
Step 1: Get data manually using REST API

**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.



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



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")

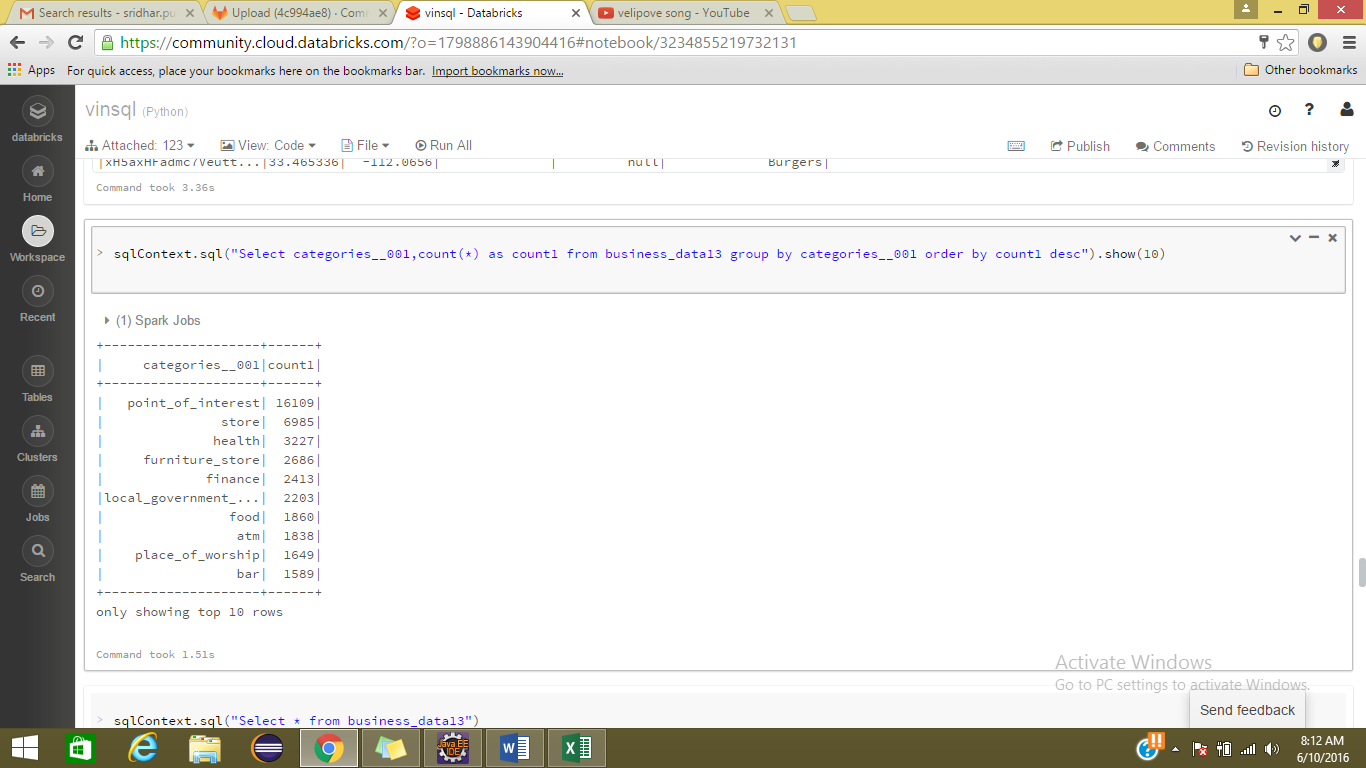
.setOutputCol("filtered")

// Create new DF with Stopwords removed

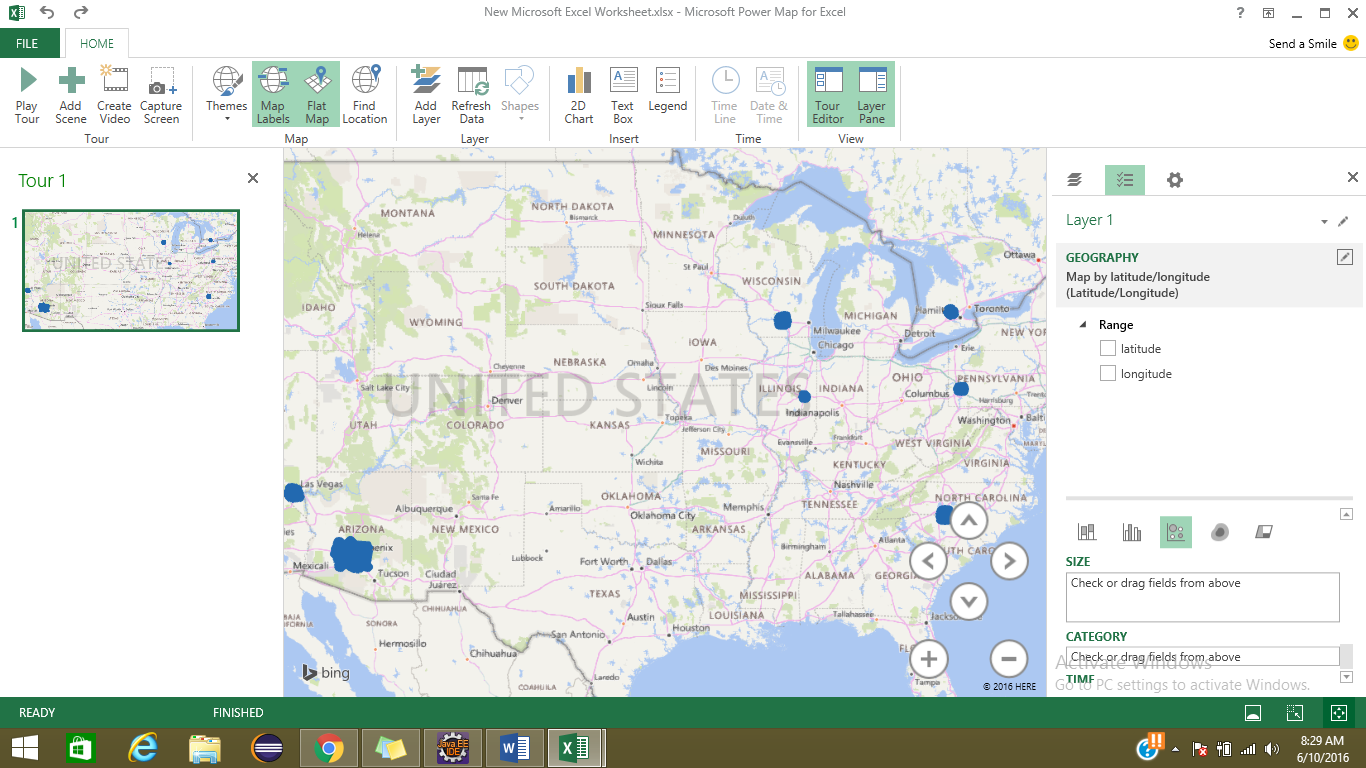
val filtered\_df = remover.transform(tokenized\_df)

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)



Step 3: Visualization

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

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