**####### Process word frequency based on all essay list in opendata\_essay.csv ####**

1. **pre-processing essay raw data :**

**#######replace all ", " to "##@@##" inorder to process as seperated column #####**

**sed -i 's/,\ /##@@##/g' opendata\_essays.log**

**#######replace all "^M" to "" inorder to process as seperated column #####**

**sed -i 's/\r//g' opendata\_essays.csv**

**sed -i 's/\\n//g' opendata\_essays.csv**

**sed -i 's/"""//g' opendata\_essays.csv**

**sed -i '1d' opendata\_essays.csv**

**2.**

**CREATE TABLE opendata\_essays**

**(projectid string, teacher\_acctid string, title string, short\_description string, need\_statement string, essay string, paragraph1 string, paragraph2 string, paragraph3 string, paragraph4 string, thankyou\_note string, impact\_letter string)**

**ROW FORMAT DELIMITED FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE;**

**load data local inpath "/home/cloudera/workspace/big\_data\_project/database/opendata\_essays.csv" into table opendata\_essays;**

1. **select essay column from opendata\_essays table**

**INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/workspace/big\_data\_project/database' select \* from opendata\_essays;**

**hadoop jar Tokenized\_and\_Stem.jar cw2189/NLP/essay/ cw2189/NLP/token\_stem**

**java -jar Stop\_Word.jar token\_stem /home/cw2189/database/stop\_word.list Stop\_List**

1. **ordering the word count result**

**create table rank ( word string, count int ) STORED AS TEXTFILE;**

**load data local inpath “ /home/cw2189/NLP/Stop\_List" into table opendata\_essays;**

**create table final\_rank**

**as select \* from rank order by count desc limit 100;**

**select \* from final\_rank;**

Got the result

1. normalization by 1000

(word count / total sum of count) \* 1000

1. duplication words by the number of word count

java -jar word\_count\_duplicate.jar normalization.txt output.txt

##############################################################

1.Finding the essay from all the funded project

select essays.essay from essays JOIN fundness ON essays.projectid = fundness.projectid AND fundness.fully\_funded = 1

2.######## Same process as Step 2~8 First Part #########

########## Create all word classification by all project focused area ######

Applied Learning

Health & Sports

History & Civics

Literacy & Language

Math & Science

Music & The Arts

INSERT OVERWRITE LOCAL DIRECTORY '/home/cw2189/database/Applied\_Learning' select essays.essay from essays JOIN projects\_funding\_status ON essays.projectid = projects\_funding\_status.projectid AND projects\_funding\_status.fully\_funded = 1 AND projects\_funding\_status.primary\_focus\_area = 'Applied Learning' ;

2.######## Same process as Step 2~8 First Part #########