Romain Bui - 80975422

**Big Data Case Study in Social Media**

**“Smile and Tweet – The usage of Smileys to train a Sentiment Analysis on tweets”**

**WORKING DEMO**

NB. All the necessary source codes, jar files and text documents can be found in the source folder

1. Install Mahout

sudo yum install mahout

vi .bash\_profile

MAVEN\_HOME=/usr/lib/apache-maven-3.2.1

export MAVEN\_HOME

MAHOUT\_HOME=/usr/lib/mahout

export MAHOUT\_HOME

HADOOP\_HOME=/usr/lib/hadoop

export HADOOP\_HOME

HADOOP\_CONF\_DIR = /usr/lib/hadoop-0.20-mapreduce/conf

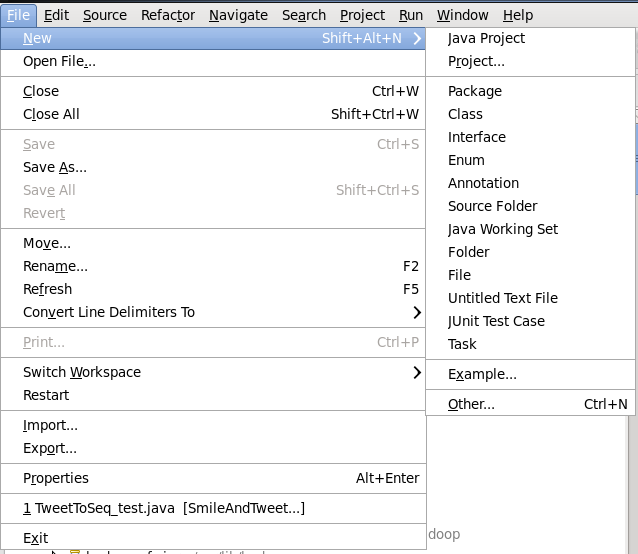
export HADOOP\_CONF\_DIR

PATH=$JAVA\_HOME/bin:$PATH:$HOME/bin:$MAVEN\_HOME/bin:$MAHOUT\_HOME

export PATH

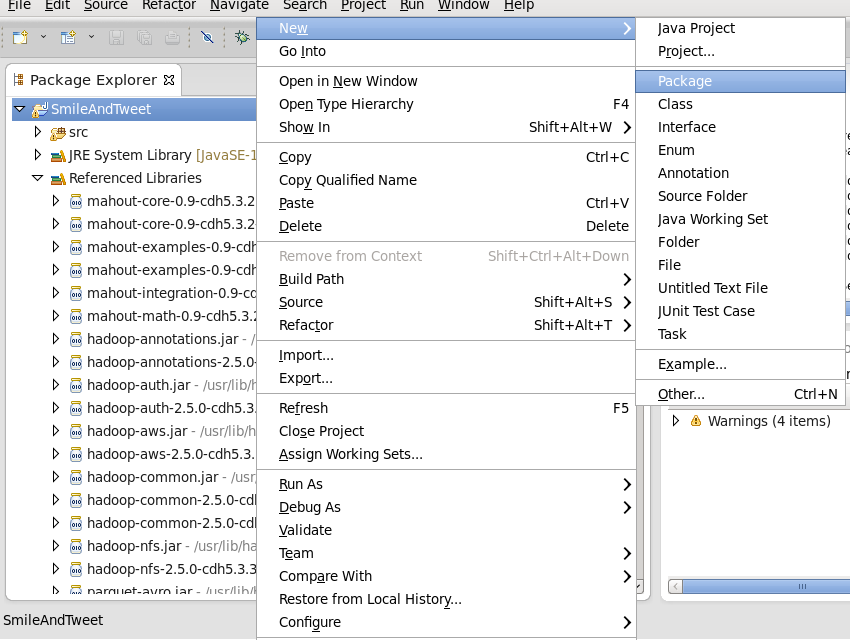
1. Create the Java Project

Eclipse – File New Java Project



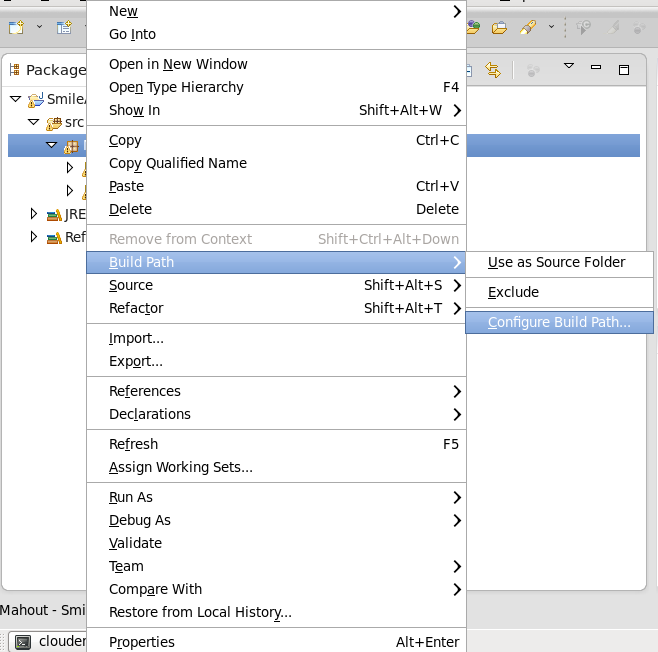
Project Name: “SmileAndTweet”

In the Project Explorer – Right Click on the Project Folder – New Package

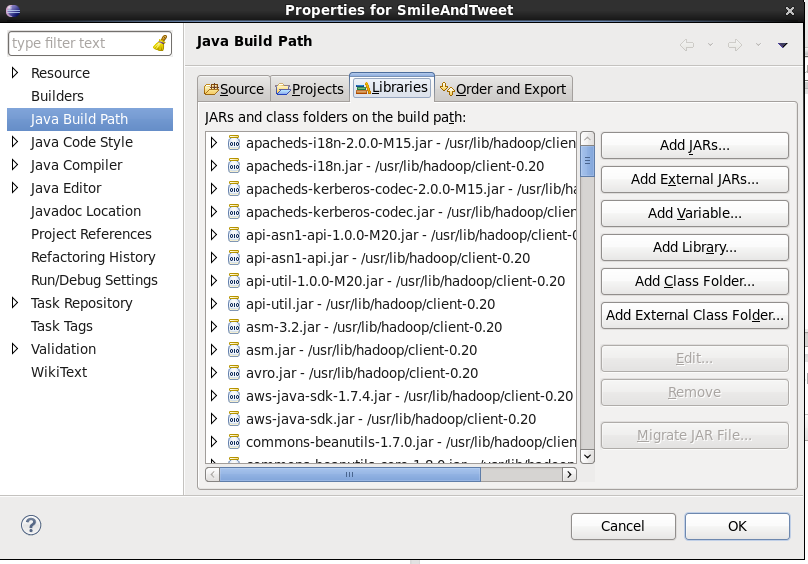


Package Name : Mahout

Right Click on the Package (“Mahout”) – Build Path – Configure Build Path



Sheet Libraries, Add External Jars

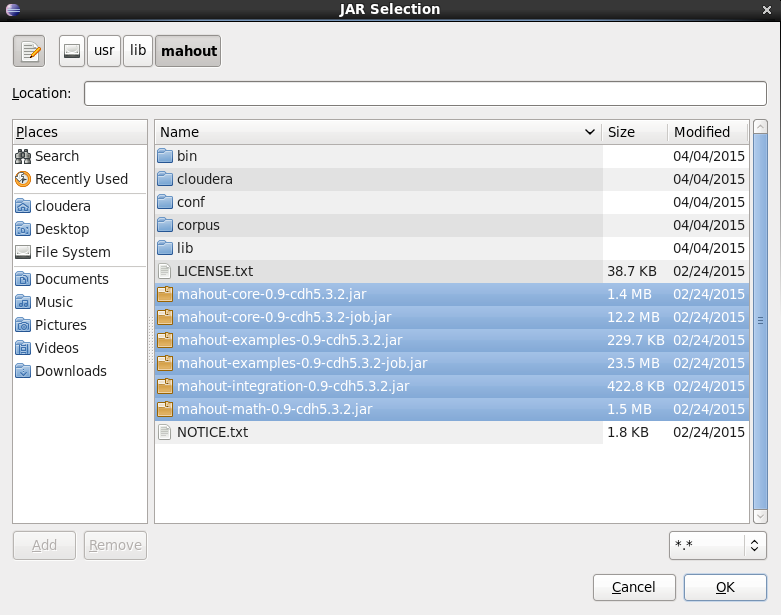


Add The Hadoop and Mahout Jars in the folders :

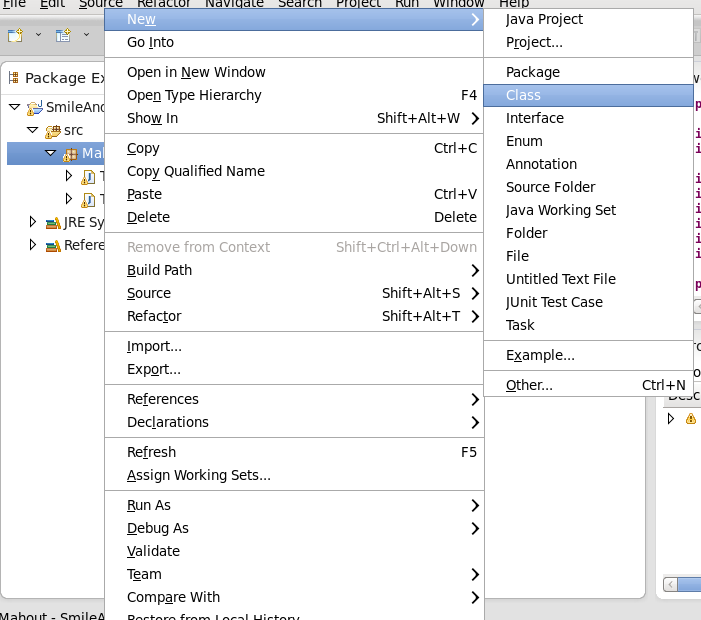
Usr/Lib/Mahout

Usr/lib/hadoop

Usr/lib/hadoop/client-0.20

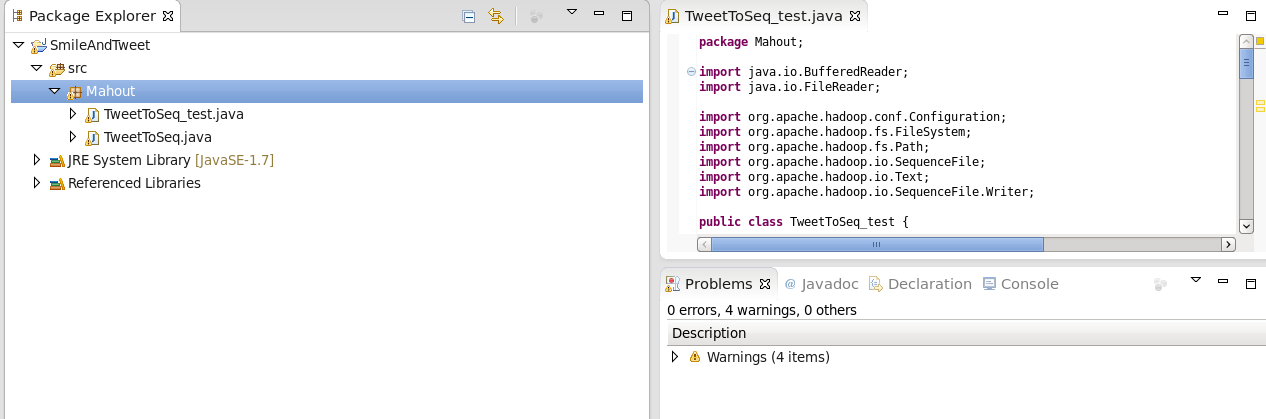


Right Click on the Package (“Mahout”) – New – Class



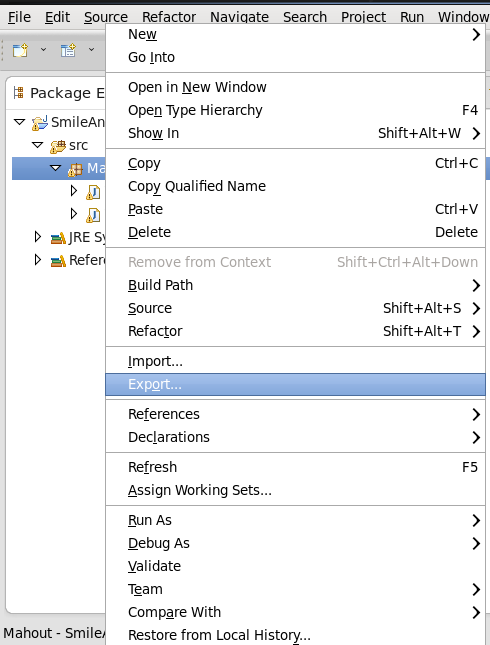
Name of the Class : “TweetToSeq” inser the code you can find in the assignment package under “Java Classes”

Do the same for TweetToSeq\_test

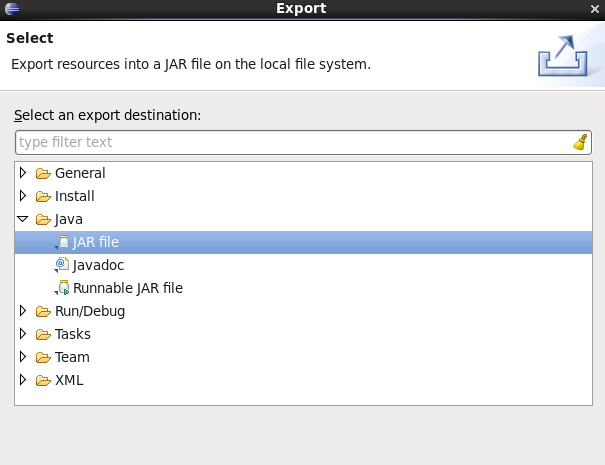


1. Create the Jar

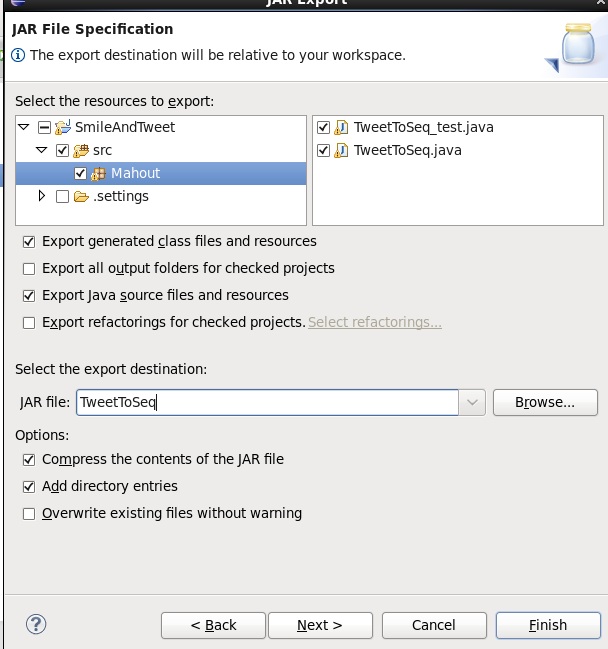
Right click on the project (“Mahout”) – Export …



Select Java/Jar



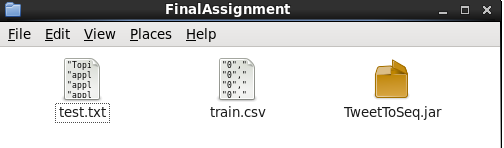
Select the options, as in the print screen below, call the jar file : TweetToSeq – Finish



The Jar can be found in the folder “JAR”

1. Prepare the folder

Create a new local folder on the root user folder (“Cloudera” for eg.) call it : “FinalAssignment”, put the two text files (you can find them in the Data folder), put the jar created in the previous point.



1. Run the Command Lines :

hadoop fs -mkdir FinalAssignment

hadoop jar FinalAssignment/TweetToSeq.jar Mahout.TweetToSeq FinalAssignment/train.csv FinalAssignment/train-seq

hadoop jar FinalAssignment/TweetToSeq.jar Mahout.TweetToSeq\_test FinalAssignment/test.txt FinalAssignment/test-seq

mahout seq2sparse -i FinalAssignment/train-seq -o FinalAssignment/train-seq-vectors -lnorm -nv -wt tfidf

mahout trainnb -i FinalAssignment/train-seq-vectors/tfidf-vectors -el -o FinalAssignment/model -li labelindex -ow

mahout seq2sparse -i FinalAssignment/test-seq -o FinalAssignment/test-seq-vectors -lnorm -nv -wt tfidf

mahout testnb -i FinalAssignment/test-seq-vectors/tfidf-vectors -m model -l labelindex -ow -o FinalAssignment/test-res