Data Preparation Document

sudo cp -r ~/Downloads/ProjectData/Test/* .
sudo chown -R mysql mysql

Load into HDFS:
$ sudo java -jar Sql2Seq-0.0.1-jar-with-dependencies.jar
*** Welcome to SQL2Seq ***
MySQL username: root
Password:
Table to import from: sentences
Field(s) to import: sentence
HDFS Output directory: hdfs://localhost:8020/user/michael/seqs/
Scanning working directory : /var/lib/mysql

Load into Mahout:
mahout seq2sparse -i /user/michael/seqs/* -o vects -ow

- Uses Lucene StandardAnalyzer (by default) to tokenize the document(s), storing individual words in the tokenized-documents/ folder.
- setting the '-ng' flag to n will denote the maximum size of n-grams to be selected from the collection of documents.
- '-a' flag allows specification of a different Lucene analyzer.
- '-seq' flag specifies that the output Vectors should be SequentialAccessSparseVectors instead of RandomAccessSparseVectors, which work better with KMeans and SVD.

mahout kmeans -i /user/michael/allvect/tfidf-vectors/ -c initial-clusters -o kmeans-clusters -dm org.apache.mahout.common.distance.CosineDistanceMeasure -cd 1.0 -k 20 -x 20 -cl

- cd -> convergence delta, default is 0.5
- cl -> if present, run clustering after the iterations have taken place.
- k -> number of clusters.
- x -> max number of iterations.

mahout clusterdump -dt sequencefile -d hdfs://localhost:8020/user/michael/vects/dictionary.file-0 -i hdfs://localhost:8020/user/michael/kmeans-clusters/clusters-1-final -o results -b 10 -n 10

Successful results (by language):
$ cat results
:VL-129942{
Top Terms:

he => 0.4321995088141447
his => 0.35200716884713684
said => 0.3464244438552207
from => 0.3382853389657676
have => 0.2981334591790388
has => 0.2903694849275929
i => 0.2664428856072751
were => 0.21967396370770398
had => 0.2193561644182676
who => 0.21803578885692346
:VL-129943{
Top Terms:
de => 1.935401108220715
la => 1.41327124929813
le => 1.2486838799570223
et => 1.1165280979879373
à => 1.0691059800650915
les => 1.0340855076165913
des => 0.926662732820834
en => 0.9255334632971403
du => 0.7806759012191344
un => 0.7049445195880997

Project Gutenberg:

```
rsync -av --del ftp@ftp.ibiblio.org::gutenberg /var/www/gutenberg
```