Q1: Download the complete works of ‘Shakespeare’ from project gutenberg.org <http://www.gutenberg.org/cache/epub/100/pg100.txt>

* Write a MR program to Output the **TOP N words** from this.
* You have to consider /n/r/t/s to tokenize, can get rid of all the numbers and punctuations, don’t need to get rid of stop words.
* Submit the Code + Output as a “out.txt” file + Screenshots of execution state in a zip file

To build a search engine we need to build an ‘Inverted Index’ <http://en.wikipedia.org/wiki/Inverted_index>. Now Map Reduce was [first conceived in Google to build web-scale Inverted Index](http://research.google.com/archive/mapreduce-osdi04-slides/index-auto-0005.html).

* Build an **Inverted Index** using MR. Find out the in which line the word is present in the document and produce an output like the one below. The words should be alphabetically sorted.
  + Hamlet,[ 1060.hamlet,2070.hamlet,5680.hamlet,4560.hamlet]
  + Love,[300.romeo&Juliet,500.hamlet]

where the integer values(1060,2070) are the **line numbers where the word is present in the book and the string is the book name.**

* Use at least 5 different books of Shakespeare from <http://www.gutenberg.org/ebooks/author/65>
* You have to consider /t/s to tokenize, can get rid of all the numbers and punctuations, don’t need to get rid of stop words.
* Submit the Code + Output as a “out.txt” file + Screenshots of execution state in a zip file.

There are two ways to go about it Java/Python.

Use the python way if you have constraints of installing Hadoop Cluster/Or if you are more comfortable with it. We are fine with code in both python/ java.

1. Python :
   1. Requirement
      1. Python needs to be installed in machine.
   2. Download the library <https://github.com/michaelfairley/mincemeatpy> . There is a word count example attached in the git source

mincemeat.py is a Python implementation of the [MapReduce](http://en.wikipedia.org/wiki/Mapreduce) distributed computing framework.

* 1. [You should import mincemeat and write your MR code.](http://en.wikipedia.org/wiki/Mapreduce)