Assignment 1 – 20%

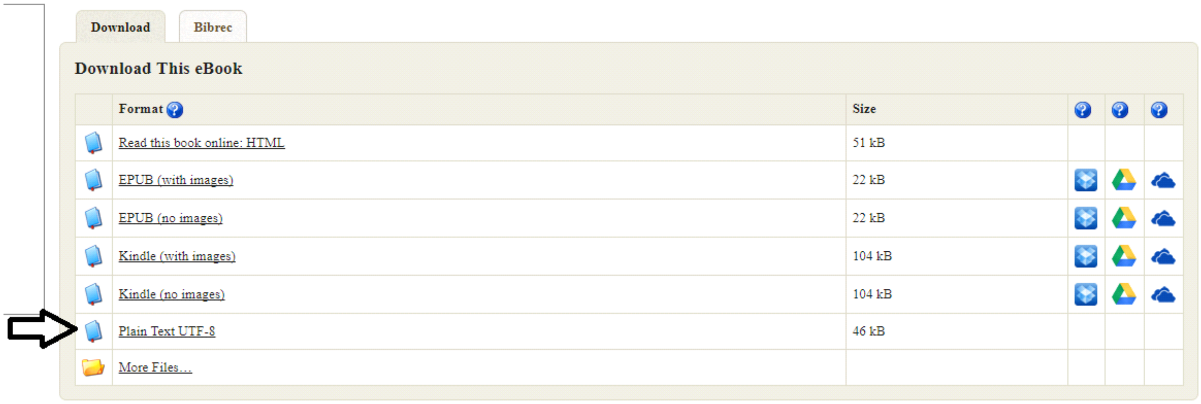
Description:

This assignment is divided into three parts and will test your ability to query a Hadoop file system using a python script. You have been provided with the mapper and reducer Python code. You will apply the map reduce methodology to text analytics within the Hortonworks Hadoop sandbox.

For this assignment, you will be tasked to modify the code to create two types of filters. One filter to detect stop words and another filter to find popular words starting or ending in the letter ‘e’.

Setup Instructions:

- Navigate to the Gutenburg website (http://www.gutenberg.org/ebooks)
- Select and download an English eBook. To do so, you should click on the title of the selected book and for the purpose of this assignment, download the “Plain Text UTF-8 format”.



Download This eBook

Format	Size	?	?	?
Read this book online: HTML	51 kB			
EPUB (with images)	22 kB			
EPUB (no images)	22 kB			
Kindle (with images)	104 kB			
Kindle (no images)	104 kB			
Plain Text UTF-8	46 kB			
More Files…				

- You can either move it to Hadoop cluster through the Ambari interface or use the Hadoop terminal to copy it from your computer to the sandbox.
- Complete the following questions by modifying the provided mapper.py code. Be sure to test your code as you have been taught before running the actual MapReduce job.

Part 1:

One of the preliminary tasks in statistical text processing is to get rid of “stop words”. Stop words (e.g. the, a, as of, there, is, are) do not convey as much information as the other key words in the text book and so we would like to remove them.

You have been provided with the mapReduce code to perform a word count. Modify the mapper.py in wordcount code to remove the following stopwords from your text:

- a
- about
- above
- after
- again
- against
- all
- am
- an
- and
- any
- are
- aren't
- as
- at
- be

(4 marks)

Part 2:

Using the wordcount code as your template, create a new filter to find the 50 most popular words in your text which begin or end in with the letter ‘e’.

Hint; to extract the most popular words out of dictionary you should sort them by value. For example, using a dictionary data structure, the sorted keys could be extracted through following code:

```
for w in sorted(d, key=d.get, reverse=True):
        print w, d[w]
```

Where d is a dictionary

(4 marks)

Part 3:

Explain why we would prefer to embed our conditions and filters in the mapper method as opposed to the reducer method.

(2 marks)

Submission Instructions:

In your submission include the following:

- Your book source text file
- Python code for parts 1 and 2
- The output file exported from the sandbox for parts 1 and 2
- Screenshot of the Hadoop stream for parts 1 and 2
- Text document explaining the question for part 3

Total Marks: 10