

InvertedIndex - Using RDD

1. View the contents of the file `hortonworks.txt`. It represents semi-structured content where the first string in each line is a Web page URL, followed by a comma-separated list of terms that appear on that particular Web page.
2. Write a Spark application using Rdd that takes a set of <key,value> pairs (key is Web page URL and Value is list of terms on the web page) and inverts the index, so that each value(term) becomes a key, and the list of Web page URLs where that term is appearing as a list of values. You can see from the output file how this information could be used as the beginnings of a web search engine.
3. Here is a subset of what the output should look like

```
download http://hortonworks.com/community/, http://hortonworks.com
/,http://hortonworks.com/download/
downloads http://hortonworks.com/get-started/,http:
//hortonworks.com/events/
enterprise http://hortonworks.com/products
/hortonworksdataplatfrom/, http://hortonworks.com/about-
us/contact- us/,http://hortonworks.com/
feed http://hortonworks.com/products/, http://hortonworks.com
/kb,http://hortonworks.com/resources/
founders http://hortonworks.com/about-us/,http://hortonworks.com
/get-started/
hadoop http://hortonworks.com/,http://hortonworks.com/download/,
http://hortonworks.com/community/,http://hortonworks.com/kb,http:
//horto nworks.com/about-us/,http://hortonworks.com/resources/,
http://hortonworks.com/webinars/,http://hortonworks.com/resources
/,http: //hortonworks.com/hadoop-rooting/
```

Note: The inverted index problem is one of the earliest and most common use case of MapReduce: invert the data so that the words on the Web page become the key, and the Web page URLs become the value.