Homework 7: Spark Programming

**Word Count**

**Output of wc.txt**

took 0.203317 s

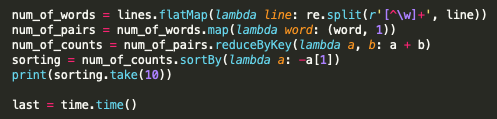
[('', 197060), ('the', 23455), ('I', 22225), ('and', 18715), ('to', 16433), ('of', 15830), ('a', 12851), ('you', 12236), ('my', 10840), ('in', 10074)]

Total program time: 4.02 seconds

The code below was researched and looked from different resources and slightly modified according to our implementation.

The official <https://spark.apache.org/examples.html> was a big help I just wanted to break down the code into variables so it was easier to follow other than that it was pretty much really similar apart from maybe changing **line.split(" "))-> re.split(r'[^\w]+', line))**

sorting the number of counts and taking the top 10.



**Page Rank**

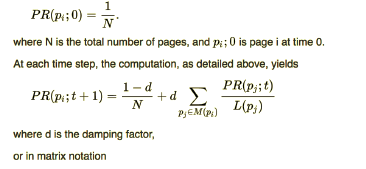
**Output of full.txt**

took 0.148354 s

[(263, 0.0020202911815182184), (537, 0.0019433415714531497), (965, 0.0019254478071662631), (243, 0.001852634016241731), (285, 0.0018273721700645144)]

Total program time: 23.11 seconds

**In this PagePank implementation I have also looked at several links and took pieces from from different sources to implement this code. The main difference was that I just used the formula given within the assignment to make it more relevant to the assignment.**

****

The implementation screenshot is shown below and it has been commented out for relevant explanation: Last part is sort and take the first 5. Similar idea to word count implementation.

