Task A

1. aa %CE%92%CE%84_%CE%95%CF%80%CE%B9%CF%83%CF%84%CE%BF%CE%BB%CE%AE_%CE%99%CF%89%CE%AC%CE%BD%CE%BD%CE%B7/el/%CE%92 1 4854
   aa %CE%98%CE%B5%CF%8C%CE%B4%CF%89%CF%81%CE%BF%CF%82_%CE%91%CE%84_%CE%9B%CE%AC%CF%83%CE%BA%CE%B1%CF%81%CE%B7%CF%82/el/%CE%98%CE%B5%CF%8C%CE%B4%CF%89%CF%81%CE%BF%CF%82_%CE%91%27_%CE%9B%CE%AC%CF%83%CE%BA%CE%B1%CF%81%CE%B7%CF%82 1 4917
   aa %CE%9C%CF%89%CE%AC%CE%BC%CE%B5%CE%B8_%CE%95%CE%84/el/%CE%9C%CE%B5%CF%87%CE%BC%CE%AD%CF%84_%CE%95 1 4832
   aa %CE%A0%CE%B9%CE%B5%CF%81_%CE%9B%27_%CE%91%CE%BD%CF%86%CE%AC%CE%BD/el/%CE%A0%CE%B9%CE%B5%CF%81_%CE%9B 1 4828
   aa %CE%A3%CE%A4%CE%84_%CE%A3%CF%84%CE%B1%CF%85%CF%81%CE%BF%CF%86%CE%BF%CF%81%CE%AF%CE%B1/el/%CE%A3%CE%A4 1 4819
   aa %D0%A1%D0%BE%D0%BB%D0%B8_484_%D0%BF.%D0%BC 1 4750
   aa 271_a.C 1 4675
   aa Battaglia_di_Qade%C5%A1/it/Battaglia_dell%27Oronte 1 4765
   aa Category:User_th 1 4770
   aa Chiron_Elias_Krase 1 4694
   aa County_Laois/en/Queen%27s_County,_Ireland 1 4752
   aa Dassault_rafaele 2 9372
   aa Dyskusja_wikiprojektu:Formu%C5%82a_1/%22/pl/Polacy_w_Formule_1%22 1 4824
   aa E.Desv 1 4662
   aa Enclos-apier/fr/Enclos-Apiers_en_C%C3%B4te_d%27Azur 1 4772

2. total records: 5046226
3. min: 0 max: 141180155987 avg: 101423.92964801814
4. Log(en.mw,en,5466346,141180155987)
5. Log(en.mw,en,5466346,141180155987)
6. There are 6786 records match this criteria so listing them in the report was not a good idea.
7. There are 340585 records matching this criteria.
8. Nothing to report
9. (en.mw,5466346)
   (en,5310694)
   (es.mw,695531)
   (ja.mw,611443)
   (de.mw,572119)
   (fr.mw,536978)
   (ru.mw,466742)
   (ru,463437)
   (es,400632)
   (it.mw,400297)
10. 8951 titles are not part of an English project.
11. 79.38217590730181
12. 814444 unique terms
13. (of,120138)
    (the,70839)
    (in,41313)
    (de,39600)
    (list,27091)
    (and,21548)
    (a,9702)
    (user,9646)
    (la,8585)
    (by,8019)

Task B

The machine I am using has 4g ram and 2 cores. I created a cluster of 1 slave with 1 core and 1g ram (small cluster) which runs the job in 1.7 minutes. The big cluster has 2 workers with 1 core each and 1g of ram. It

executes the job in 1.3 minutes. In order to have a better understanding about clustering I created also a cluster with 3 workers, each with one core and 1g ram, which runs the job in 5.3 minutes and a cluster with 4 workers(1 core, 1g ram) which runs the job in 8.2. Based on the previous execution times we can safely make the following observation: as long as the pseudo-distributed cluster does not use more resources than the available the job's execution time reduces as we increase the workers.

1. The master is the program that the main job runs. It splits the program into tasks and distributes it to the workers. The workers receive the tasks from the master and execute them. When they finish their task they send the response back to the master.
2. As mention above the number of slaves affect the execution time, as long as the number of slaves does not succeed the available resources the job runs faster, but if they do succeed the resources the execution time increases.
3. The best approach would be to use 1 master with multiple slaves. Because the size of documents is about 10g each I would split the machine in 6 slaves with 10g of ram each and 2 or 3 cores each. This way the jobs would be equally distributed on every worker.