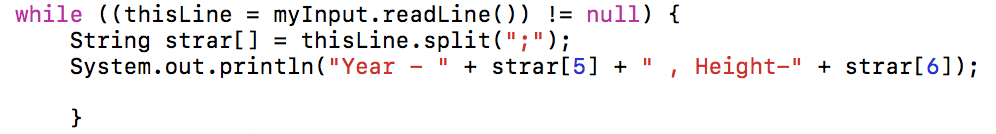
Big Data Algorithm & Platform

Lab 1 - Yijing RONG

**Question 2.7 Displaying the content of a CSV file**

The objective of this exercise is to display the the year and the height of each tree of the file *arbres.csv*.

Year = ANNEE PLANTATION, Height = HAUTEUR

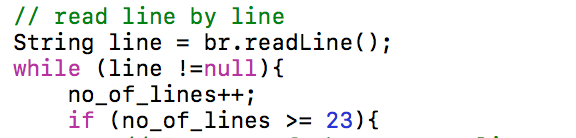


Total lines = 98

**Question 2.8 Displaying the content of a compact file**

Your job is to write a program that will display the USAF code, the name, the country (FIPS country ID) and the elevation of each station.

* The first 22 lines of the file have to be ignored



* The name of each station begins at the character 13 of a line and its size is 29 characters long. *-> the name of station 13-42*



* The FIPS begins at the character 43 of a line and its size is 2 characters long.



* The altitude begins at the character 74 of a line and its size is 7 characters long.

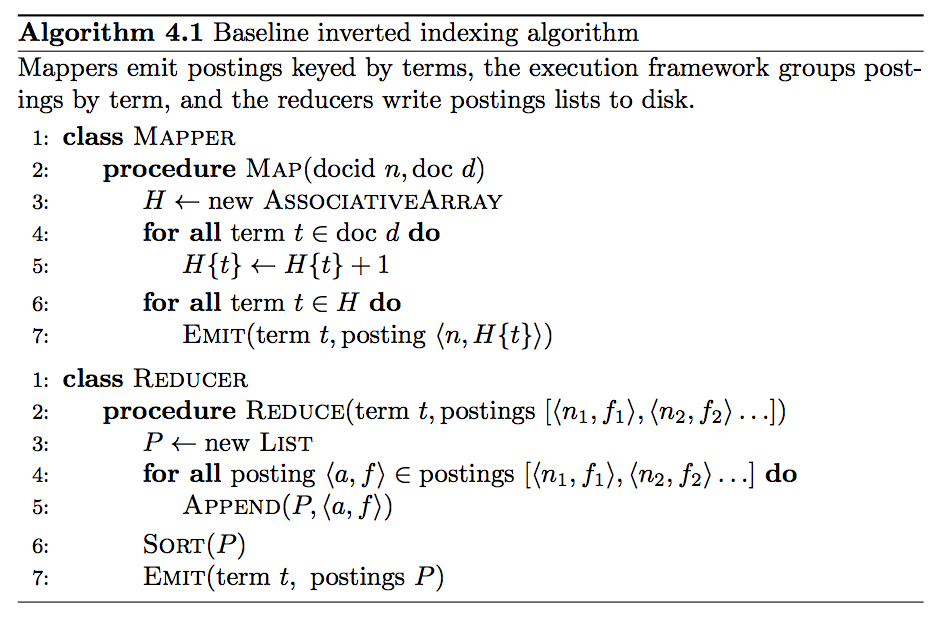


**Question 4.1 TF-IDF (Term frequency – inverse document frequency)**

Term frequency ****

Inverse document frequency 

TF\*IDF 



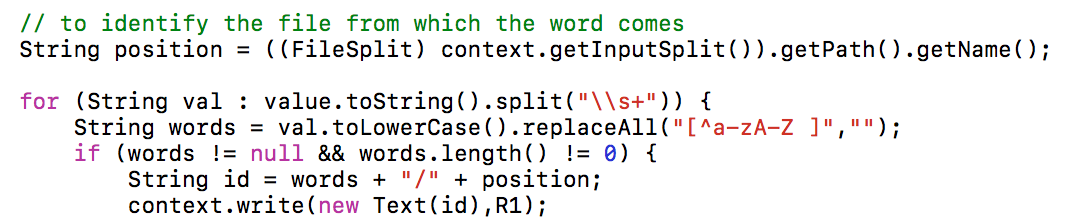
<http://www.dcs.bbk.ac.uk/~dell/teaching/cc/book/ditp/ditp_ch4.pdf>

Doc1 – <http://www.textfiles.com/etext/FICTION/defoe-robinson-103.txt>

Doc2 - <http://www.textfiles.com/etext/FICTION/callwild>

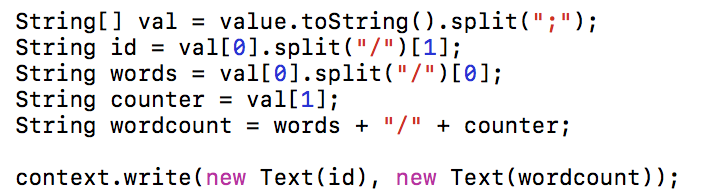
Round 1 – Wordcount

Map phase

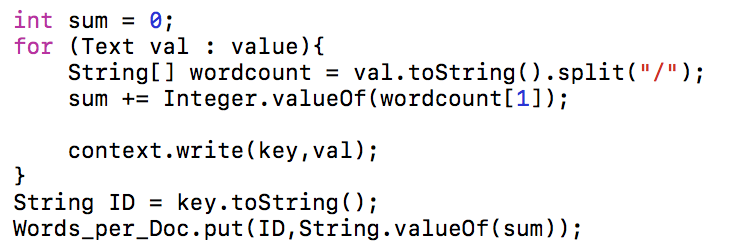


Round 2 – Wordcount per doc

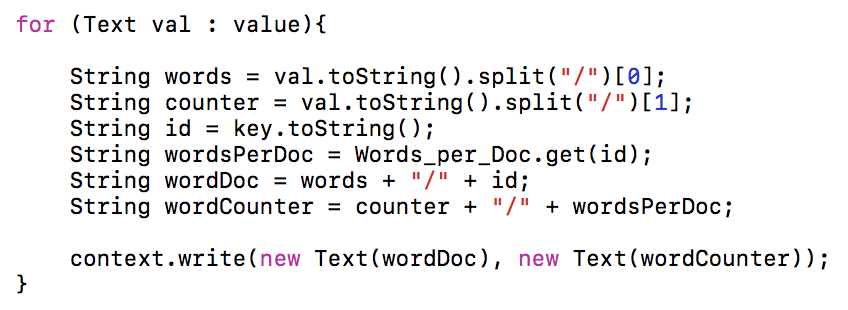
Mapper



HashMap

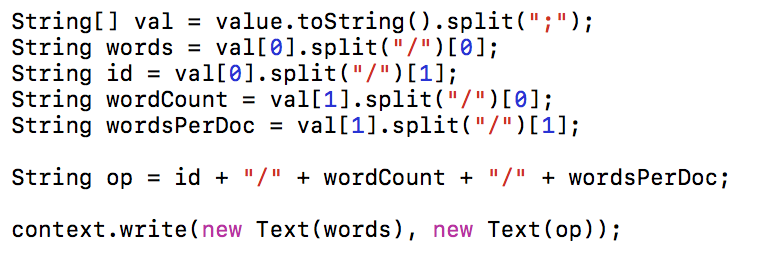


Reducer

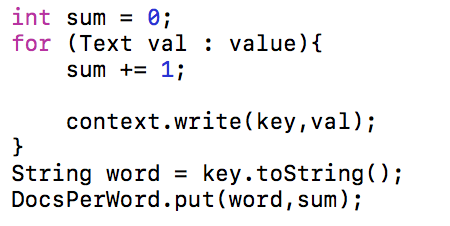


Round 3 – Doc count per word

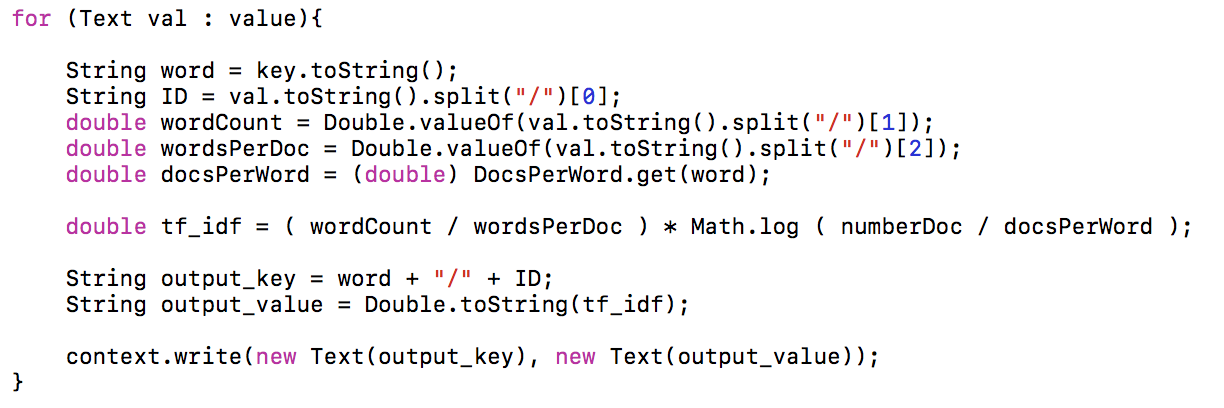
Mapper



HashMap Combine



Reducer



Which 20 words have the highest TF-IDF scores in these documents? List them in descending order.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Call Wild** | | **Dafoe Robinson** | |
| **Word** | **TF-IDF score** | **Word** | **TF-IDF score** |
| 1 | the | 0.049691254 | the | 0.033684536 |
| 2 | the | 0.033684536 | i | 0.029398972 |
| 3 | and | 0.033396536 | and | 0.027501895 |
| 4 | i | 0.029398972 | to | 0.024604854 |
| 5 | and | 0.027501895 | of | 0.020136438 |
| 6 | to | 0.024604854 | a | 0.012919548 |
| 7 | of | 0.020136438 | my | 0.012239572 |
| 8 | of | 0.018999597 | was | 0.011519597 |
| 9 | he | 0.017712598 | in | 0.011068184 |
| 10 | was | 0.015160413 | that | 0.010799622 |
| 11 | to | 0.014724142 | it | 0.010542488 |
| 12 | a | 0.014309685 | as | 0.008913974 |
| 13 | a | 0.012919548 | had | 0.008713981 |
| 14 | my | 0.012239572 | for | 0.007514023 |
| 15 | his | 0.012237398 | me | 0.007142607 |
| 16 | in | 0.01169206 | but | 0.006348349 |
| 17 | was | 0.011519597 | with | 0.006159785 |
| 18 | in | 0.011068184 | not | 0.005599804 |
| 19 | that | 0.010799622 | which | 0.005136963 |
| 20 | it | 0.010542488 | he | 0.005039824 |

**Question 4.2 Page Rank**

Which 10 users have the highest PageRank scores in this social network? List them in descending order.

|  |  |  |
| --- | --- | --- |
| 10624 | 9.885209721700881E-5 | 41718 |
| 10939 | 9.885209721700881E-5 | 11240 |
| 11164 | 9.885209721700881E-5 | 4544 |
| 14107 | 9.885209721700881E-5 | 22016 |
| 1541 | 9.885209721700881E-5 | 669 |
| 1654 | 9.885209721700881E-5 | 1399 |
| 1890 | 9.885209721700881E-5 | 61 |
| 2006 | 9.885209721700881E-5 | 2861 |
| 2024 | 9.885209721700881E-5 | 6170 |
| 2041 | 9.885209721700881E-5 | 19 |