1 Data Cleaning

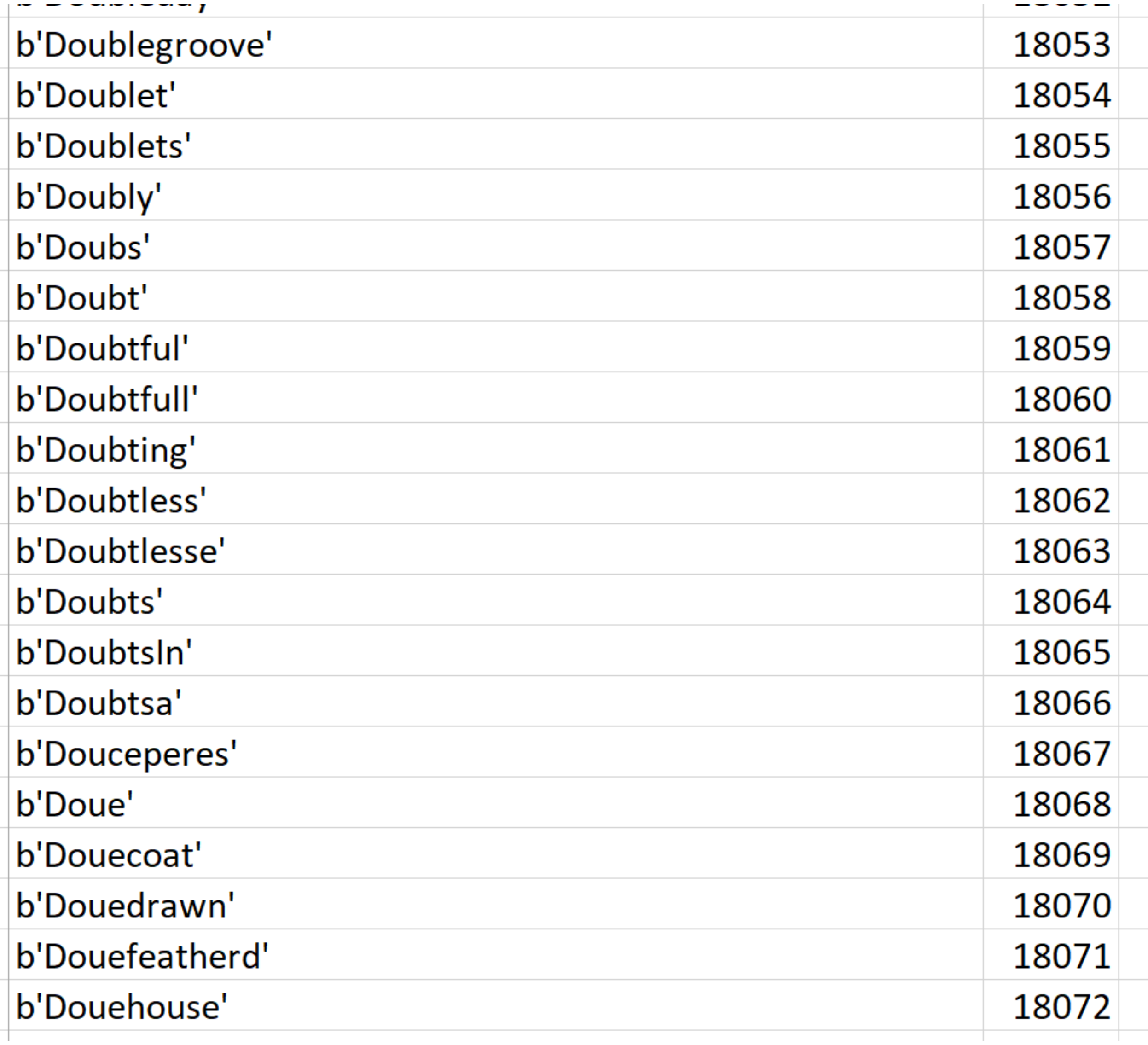
Environment: Python 3.7, Jupyter Notebook

Packages: pandas 1.1.0

Path of script: ./scripts/1-pre-process.py

1. Generate a word dictionary

The first column is the word, the second column shows a unique id for each word.



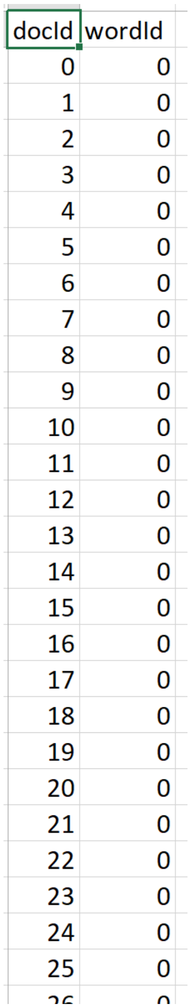
b'Doubleday'	18052
b'Doublegroove'	18053
b'Doublet'	18054
b'Doublets'	18055
b'Doubly'	18056
b'Doubs'	18057
b'Doubt'	18058
b'Doubtful'	18059
b'Doubtfull'	18060
b'Doubting'	18061
b'Doubtless'	18062
b'Doubtlesse'	18063
b'Doubts'	18064
b'DoubtsIn'	18065
b'Doubtsa'	18066
b'Douceperes'	18067
b'Doue'	18068
b'Douecoat'	18069
b'Douedrawn'	18070
b'Douefeatherd'	18071
b'Douehouse'	18072

2. Generate pair (wordId, docId)

The pair shows as below after some operations. The result is sorted by wordId and docId.

docId: the name of each file

wordId: an unique number from the dictionary



docId	wordId
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0

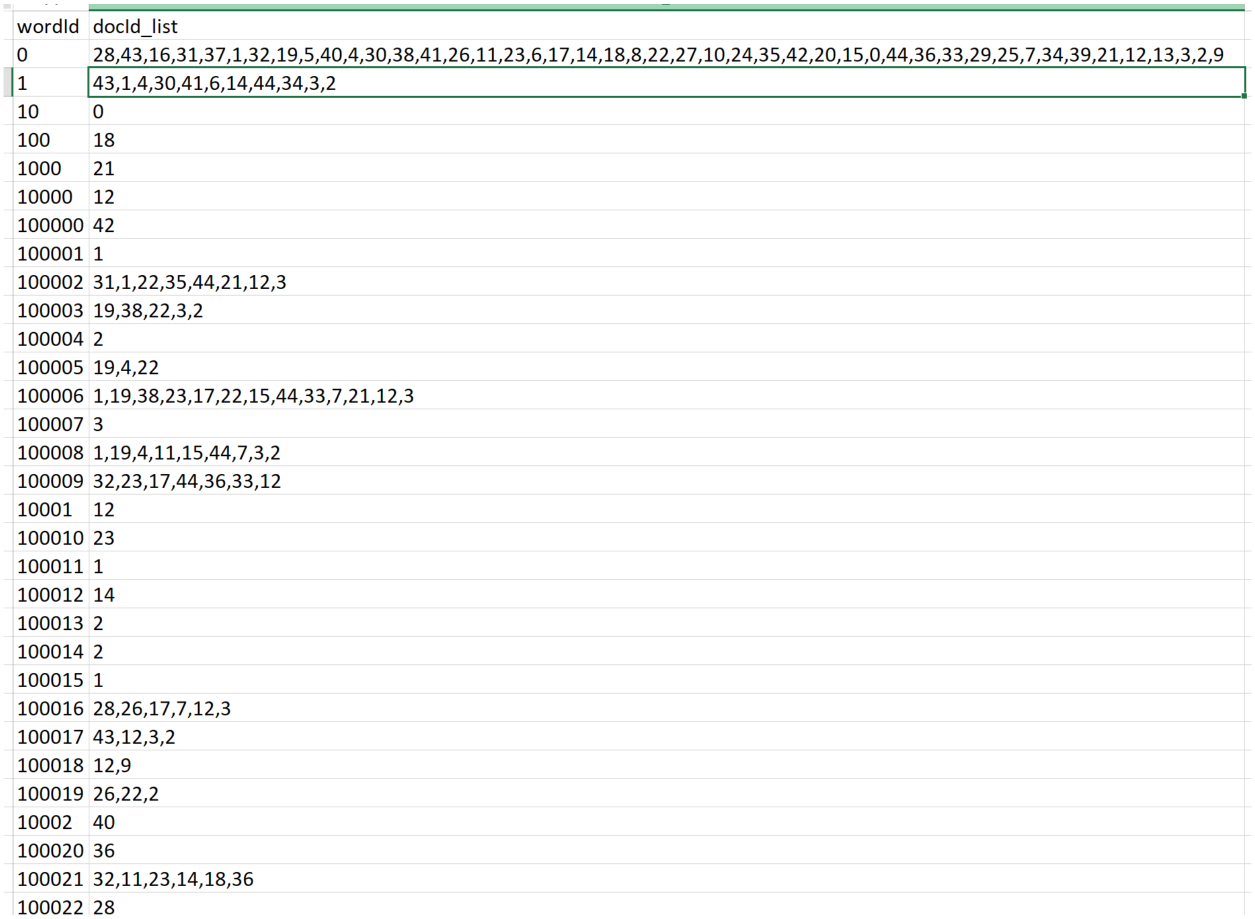
2. Submit a spark job

Environment: Spark 2.3.2

Path of script: ./scripts/2-inverted_index.py

Due to a memory issue in Python standalone environment, generating the invert index in Spark Cluster.

The final result shows as below. The first column is a unique wordId, the second column is a list of document Id for each word.



wordId	docId_list
0	28,43,16,31,37,1,32,19,5,40,4,30,38,41,26,11,23,6,17,14,18,8,22,27,10,24,35,42,20,15,0,44,36,33,29,25,7,34,39,21,12,13,3,2,9
1	43,1,4,30,41,6,14,44,34,3,2
10	0
100	18
1000	21
10000	12
100000	42
100001	1
100002	31,1,22,35,44,21,12,3
100003	19,38,22,3,2
100004	2
100005	19,4,22
100006	1,19,38,23,17,22,15,44,33,7,21,12,3
100007	3
100008	1,19,4,11,15,44,7,3,2
100009	32,23,17,44,36,33,12
10001	12
100010	23
100011	1
100012	14
100013	2
100014	2
100015	1
100016	28,26,17,7,12,3
100017	43,12,3,2
100018	12,9
100019	26,22,2
10002	40
100020	36
100021	32,11,23,14,18,36
100022	28

3 How to execute script

I assume you've already pull the whole directory to the local. The only thing you need to change is the path of data loading and saving.

1. Runs the first script in python3 environment and pandas must be installed.

command: **python /<your_path>/scripts/1-pre-process.py**

After this, we will have a list of pairs (wordId, docId).

2. Submit a spark job with command shown below.

Spark-submit —deploy-mode client —num-executor 2 —executor-cores 3 —executor-memory 2G —driver-memory 1G /<your_path>/scripts/2-inverted_index.py

After the script finished its job, we will have the inverted index result in a CSV file.

4 Optimizations

This task has no issue in algorithms, but some issues in data cleaning. To do further data cleaning, I think need to discuss with Data Scientist to see how to process data issues like mix type and similar words.

Case sensitive or not: Daughter vs daughter

Digit number: 0000157

Mix type: 00ws110zip

Similar words: beset, besets, besetting