# 1 Data Cleaning

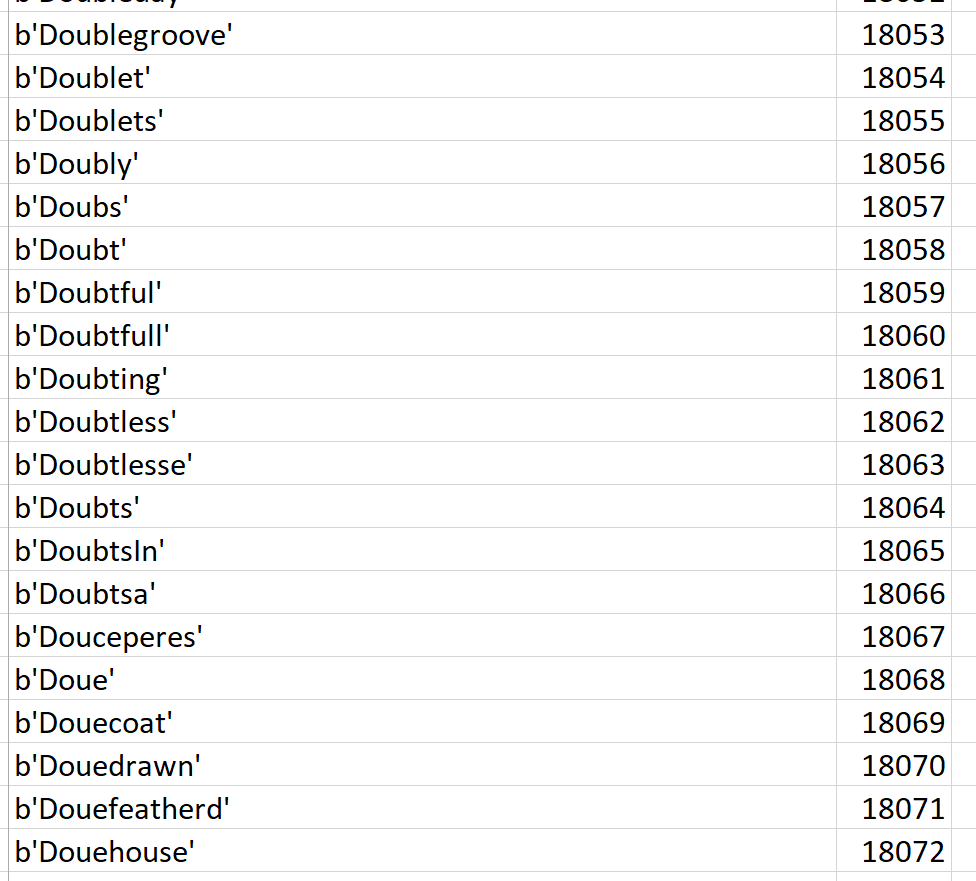
Environment: Python 3.7, Jupyter Notebook

Packages: pandas 1.1.0

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

## Generate a word dictionary

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

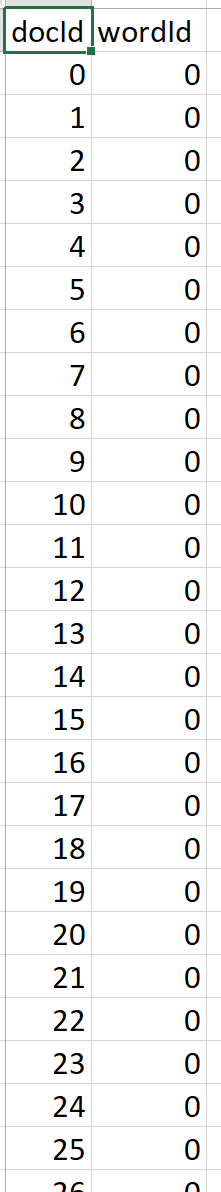


## Generate pair (wordId, docId)

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



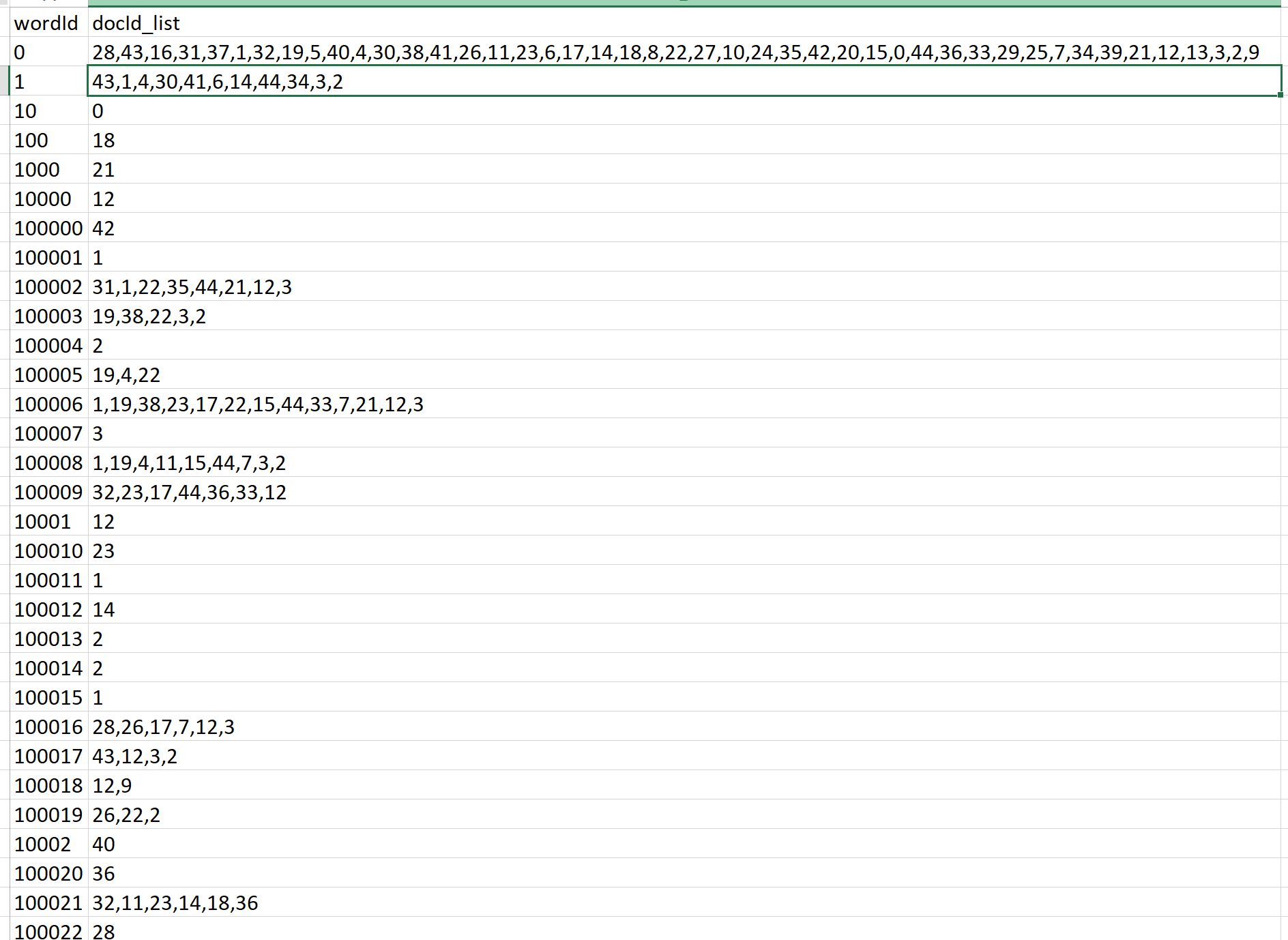
# Submit a spark job

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.



# 3 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