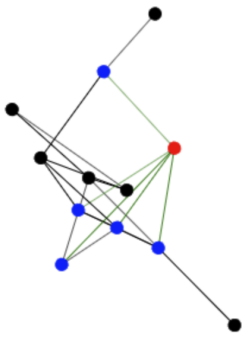
Massive Graph Data Extraction

CentralSupelec

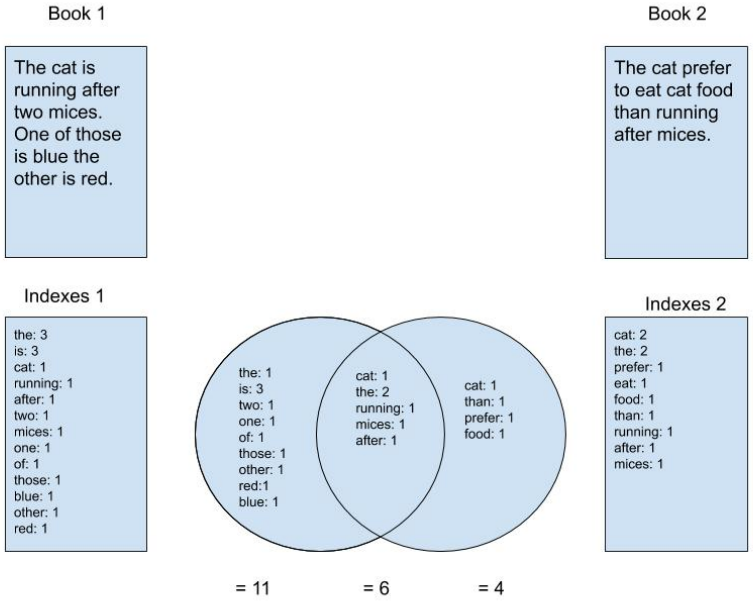
8 novembre 2019



Our goal today: extract data from a graph. **About the graph:** It come from http://www.gutenberg.org/ which provide us thousands of books freely in different format.

The graph have been created using Jaccard similarity index of the number of apparition of the words for each books.

sans titre.jpg



Jaccar = 6 / (11+4) - 6 = 6/9 = 0.6666666

The Adjacency matrix have already be computed for you and is available on my github.

It contains indexes between books you can find directly on the gutenberg website :

This index **9975.txt.utf-8.json** refere to

http://www.gutenberg.org/cache/epub/9975/pg9975.txt