Verslag BDA opdracht 1

Sil Vaes en Maarten Evenepoel

October 11, 2021

In deze opdracht was de bedoeling dat we een maximal frequent itemset algoritme ging implementeren. Hier hebben we het a-priori algoritme met hashing, dus het PCY algoritme.

Het inlezen van de dataset in het XML-formaat wordt gedaan met behulp van regexes, dus de XML library in de standard library van Python is niet gebruikt. We hebben bewust deze keuze gemaakt omdat dit ons eenvoudiger leek, want we moesten enkel de entries tussen bepaalde tags matchen en dan enkel the authors. Dus we zijn tot de conclusie gekomen dat een hele XML parser wat overkill was.

De dataset wordt ook niet in een keer ingelezen, maar in chunks van x megabytes, waar x een commandline argument is.
Deze werkt op subsets van de grote dataset. We hebben dit algoritme geprobeerd te runnen op de gehele dataset, maar dit duurde veel te lang.
Output 150000 entries:

```
Frequent pairs: {frozenset({'Christoph Meinel', 'Harald Sack'}): 6, frozenset({'Christo


3-tuple itemset: {frozenset({'Ophir Frieder', 'Steven M. Beitzel', 'Eric C. Jensen'}):


4-tuple itemset: {frozenset({'C. C. Bordeianu', 'Ioan Valeriu Grossu', 'C. Besliu', 'Al


5-tuple itemset: {frozenset({'C. Besliu', 'Al. Jipa', 'C. C. Bordeianu', 'T. Esanu', 'I


Max frequent itemsets: {frozenset({'C. Besliu', 'Al. Jipa', 'C. C. Bordeianu', 'T. Esan
```