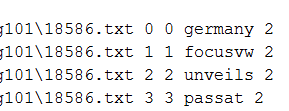
After implementing the LDA model, we obtain “.gz” file, which includes the every word assign to different topic.

Topic number



We use this file to create transactional databases for pattern mining.

In the Association rule mining.zip, you can see the introduction of the format of the transactional database in Using ARMSystem.txt. You can refer to the format of mushroom-bin.data and mushroom-bin.names. For each topic, we need one pair: .data and .names.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Document | *Z1* | | *Z2* | | *Z3* | |
|  | words |  |  |  |
| *d1* |  | *w1*,*w2*,*w3*,*w2*,*w1* |  | *w1,w9*,*w8* | *w7*,*w10*,*w10* |
| *d2* |  | *w2*,*w4 ,w4* |  | *w7*,*w8*,*w1*,*w8*, *w8* | *w1,w11*,*w12* |
| *d3* |  | *w2,w1*,*w7*,*w5* |  | *w7*,*w1*,*w3*,*w2* | *w4,w7*,*w10*,*w11* |
| *d4* |  | *w2*,*w7,w6* |  | *w9*,*w8*,*w1* | *w1*,*w11*,*w10* |

Suppose we have 5 documents, d1, d2, d3, d4, d5. Words assignments are listed in the figure. Then the three transactional databases for 3 topics are like below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| transaction | topic document transaction | transaction | topic document transaction | transaction | topic document transaction |
| 1 | {*w1*, *w2*, *w3* } | 1 | {*w1*, *w8*, *w9* } | 1 | {*w7*, *w10*} |
| 2 | { *w2*, *w4* } | 2 | { *w1*, *w7* , *w8* } | 2 | {*w1,* *w11*, *w12*} |
| 3 | {*w1*, *w2*,*w*5,*w7*} | 3 | {*w1*, *w2*,*w*3,*w7*} | 3 | {*w4*, *w7*, *w10*, *w11* } |
| 4 | {*w2*, *w*6, *w7*} | 4 | {*w1*, *w8*, *w9* } | 4 | {*w1*, *w11*, *w10*} |

Transactional datasets

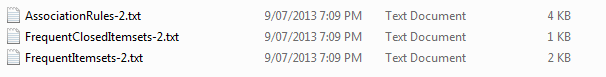
Note that when one word occurs in one document many times, we just count it for once. In each transactional database, a set of words are without any duplicates. After pattern mining, we get patterns in each topic. For example, the table below list the patterns in topic2. The threshold is 0.5.

|  |  |
| --- | --- |
| Patterns | *supp* |
| {},{,} | 0.75 |
| {},{,},{},{},{} | 0.5 |

Enter in “…\PS-System\Build”, command to call ARM is

java -Xms64M -Xmx1024M ARM “../[inputfilename.data]" 0.2 0.5 T F F F 1 2 2 "../../outFiles"

Notes: 0.2 is the minimum threshold, outFiles should be created in advance.



In outFiles, frequent pattern and closed pattern are obtained for the particular topic.