1

1.1

For the main part of this exercise, please see our groups code.

After visualizing the tf_idf ranking, we decide to use 8000 stems.

1.2

The following figures plot the perplexity scores as a function of the samples we draw.

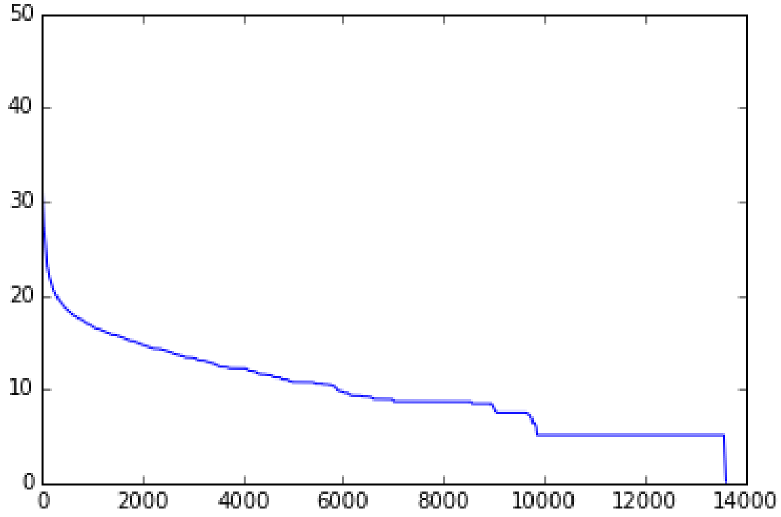
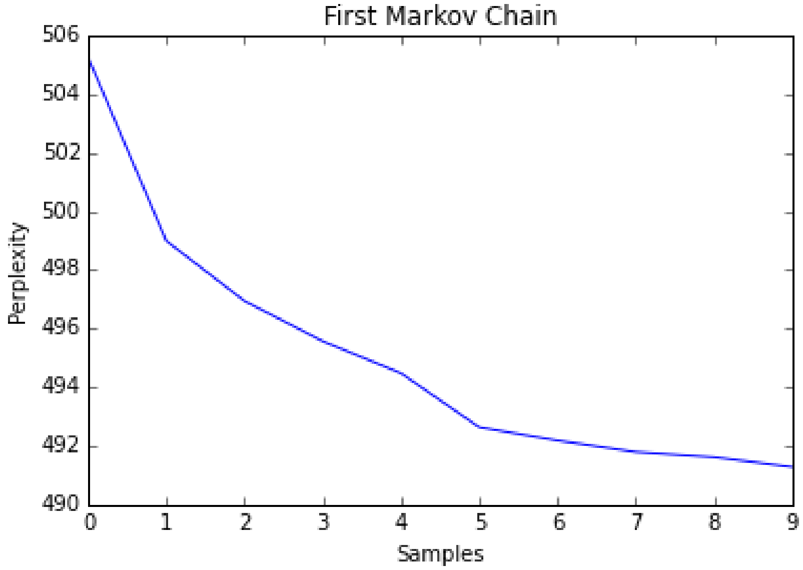
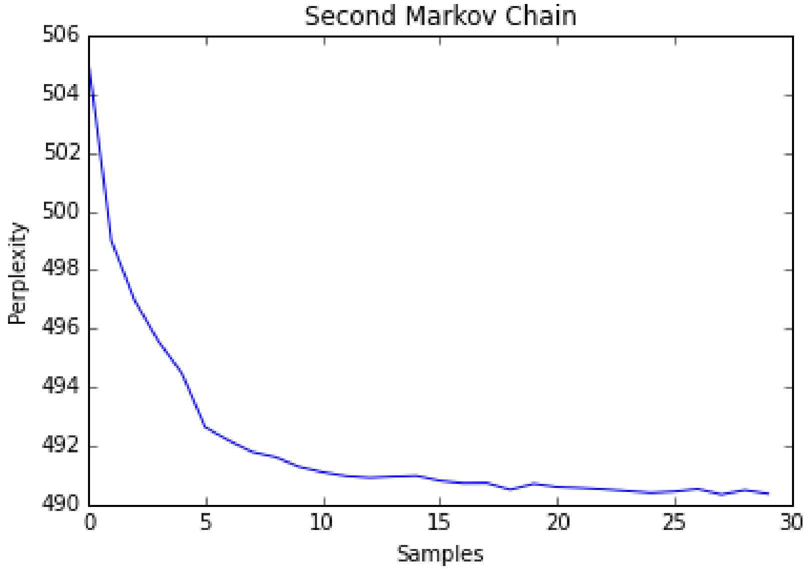


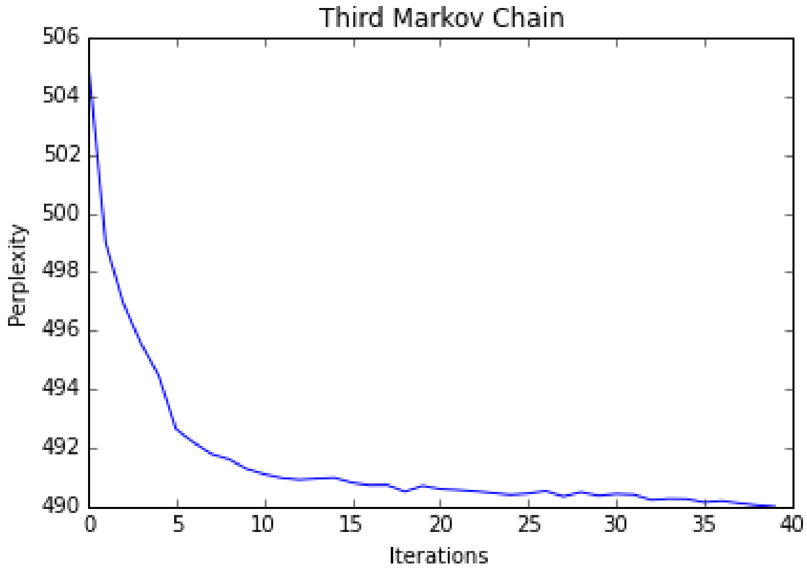
Figure 1: TF-IDF Ranking of Stems

1.3

The following figures plot the perplexity scores as a function of the samples we draw.


First Markov Chain
Perplexity
Samples
506
504
502
500
498
496
494
492
490
0
1
2
3
4
5
6
7
8
9


Second Markov Chain
Perplexity
Samples
506
504
502
500
498
496
494
492
490
0
5
10
15
20
25
30



We see that they stabilize at about the last 20 draws in the third markov chain. Since the exercise text restricts us to ten, we choose to keep the last ten samples. After having generatied the topics and document-topic distributions (see code), let's examine the topics.

```
topic0,oper,financ,support,econom,countri,region,project,
    agreement,cooper,provis
topic1,eupol,copp,rafah,head,action,bam,oper,polic,staff,
    eusr
topic2,product,materi,origin,use,ex,manufactur,work,valu,
    declar,head
topic3,budget,payment,research,commit,programm,amount,
    revenu,titl,financi,will
topic4,remarksthi,reserv,basiscouncil,year,action,third,
    basisregul,servic,project,train
topic5,commit,expenditur,payment,programm,area,polici,
    appropri,chapter,action,countri
topic6,appropri,financi,revenu,year,remark,chapter,
    expenditur,staff,assign,servic
topic7,agreement,price,flower,rose,carnat,protocol,origin
    ,fix,strip,bank
topic8,al,birth,syrian,regim,syria,minist,repress,respons
    ,civilian,popul
topic9,appropri,outturn,payment,differenti,commit,financi
    ,expenditur,research,legal,programm
```

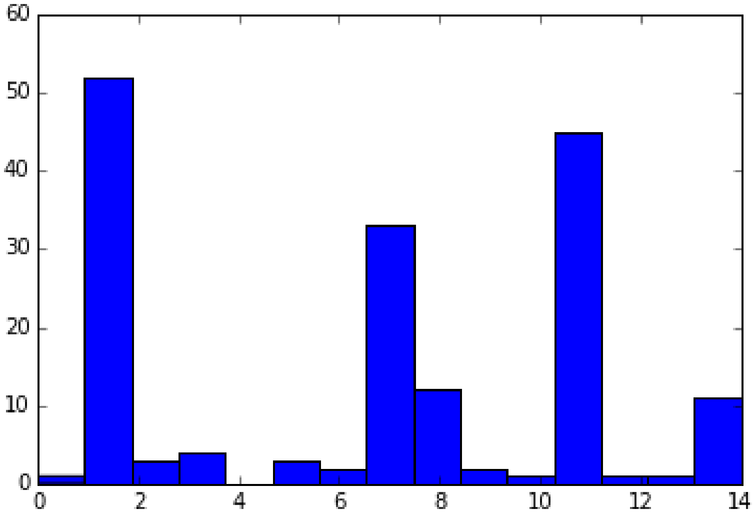
```
topic10,fund,parliament,project,pilot,innov,payment,
    commit,remark,programm,financi
topic11,al,born,takfir,hijra,card,ident,activist,algier,
    aka,hofstadgroep
topic12,row,part,credit,crr,exposur,asset,amount,financi,
    institut,column
topic13,syn,var,product,brassica,leav,remov,bean,citru,
    seed,root
topic14,bio,third,product,countri,number,code,categori,
    point,permit,month
```

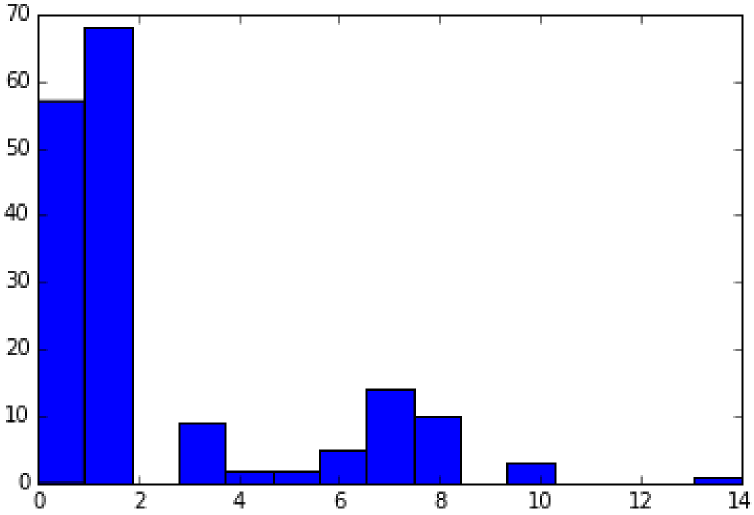
Here we see some interesting patterns. For example, topic11 seems to be a djihadism-terrorism related topic (since it talks about the hofstadgroep, a Dutch terrorist organization influenced by Takfir wal-Hijra), wereas topic1 might be related to police, and topic6 is definitely economy-related.

1.4

To generate out of sample distributions, we use the titles of our data set.

We think that it would be interesting to see if the estimated topic of documents and titles match. This is the case for 69.6% of the documents. This sounds reasonable at first. Looking at the histograms of documents (first) and titles (second) we see that the picture is quite different.. This is probably due to the short length of the titles which makes it a lot harder to do valid inference on them.





1.5

We spot check documents 1 and 2 (0 and 1 in Python Indexing) which match their topic (1) (police / border control missions). 6 and 7, also belonging to the same topic (7) seem also to talk about a similar topic, regarding import / customs regulations.