Accounting for Data dependencies within a Hierarchical Dirichlet Process Mixture Model

CIKM 2011

Dongwoo Kim & Alice Oh
KAIST U&I Lab.

Background & Goal

- The goal of this work is to incorporate the dependencies of data into HDP framework to model the topic of corpus
- To do this we extend HDP to distance-dependent Chinese restaurant franchise
- From experiments we can capture the emergence of topics and disappearance of topics by incorporating time dependencies

Hierarchical Dirichlet Process

- is a non-parametric Bayesian prior for grouped clustering problems such as topic modeling
- Number of topics is not fixed a priori
  - Find adequate number of topics by model itself
- Called Chinese Restaurant Franchise (CRF) metaphorically

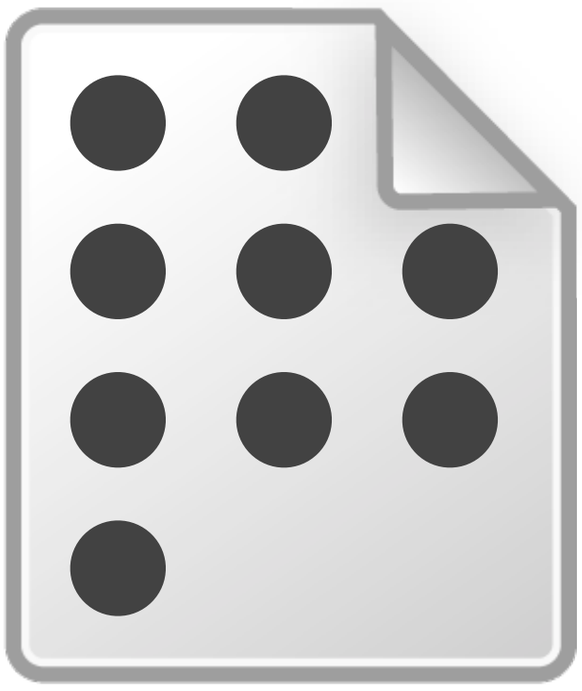
2 Processes of CRF

- CRF consists of two sequential processes for modeling topics from a corpus

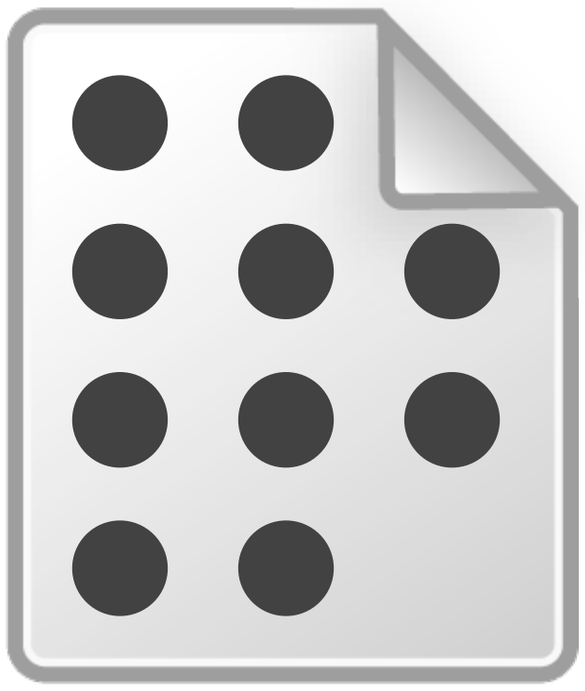
1. Partitioning words within a document
2. Grouping partitions across the corpus

1.Partitioning words within a document

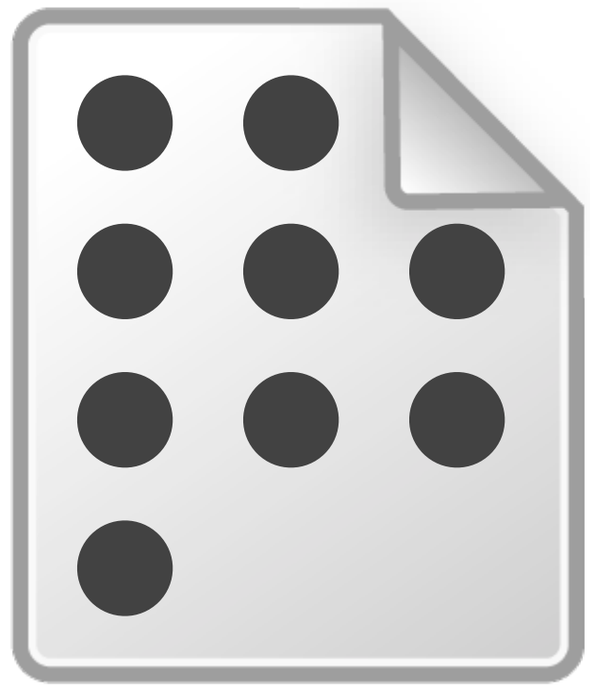
- By stochastic process of partitioning document



Document1



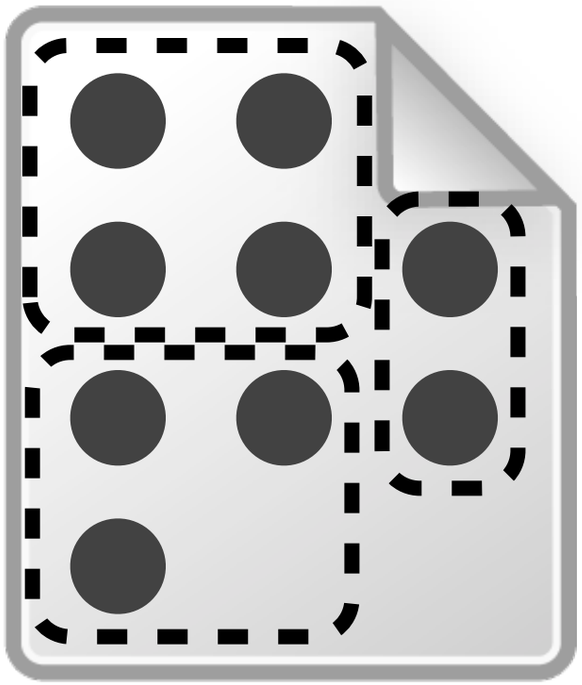
Document2



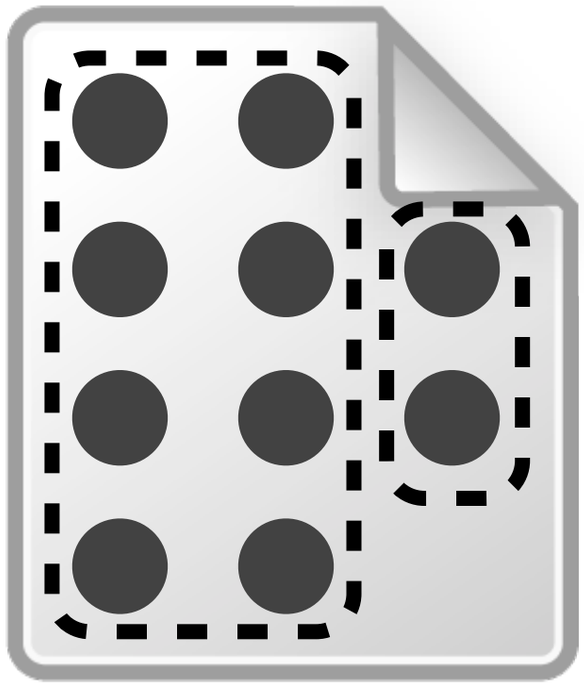
Document3

1.Partitioning words within a document

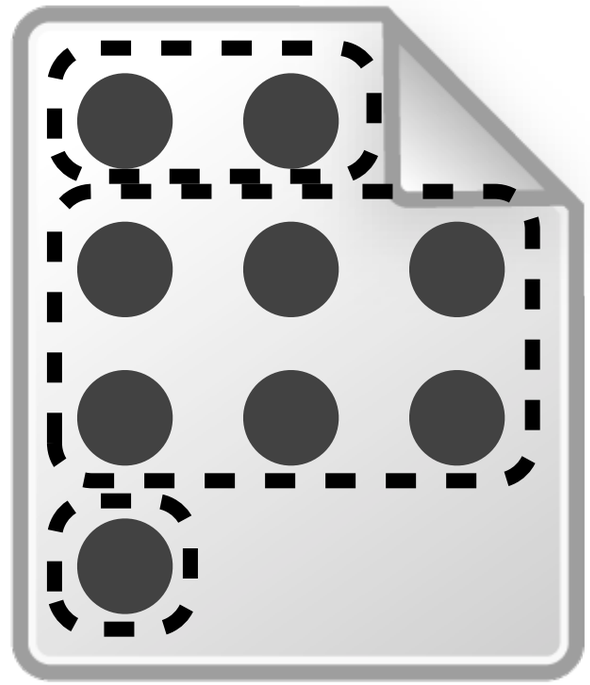
- Number of partition is not fixed a priori



Document1



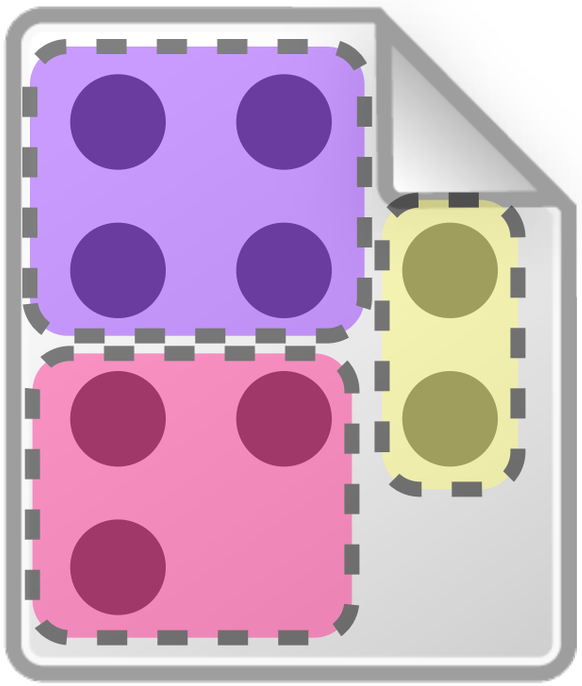
Document2



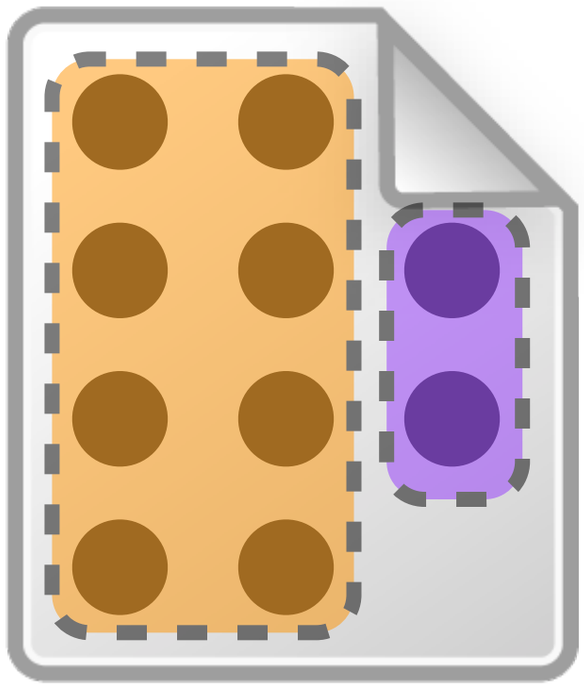
Document3

2.Grouping partitions across documents

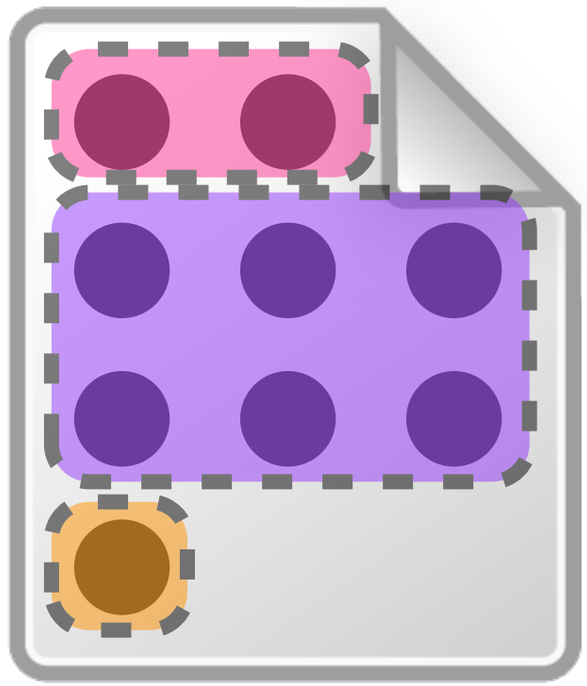
- By stochastic process of grouping partitions



Document1



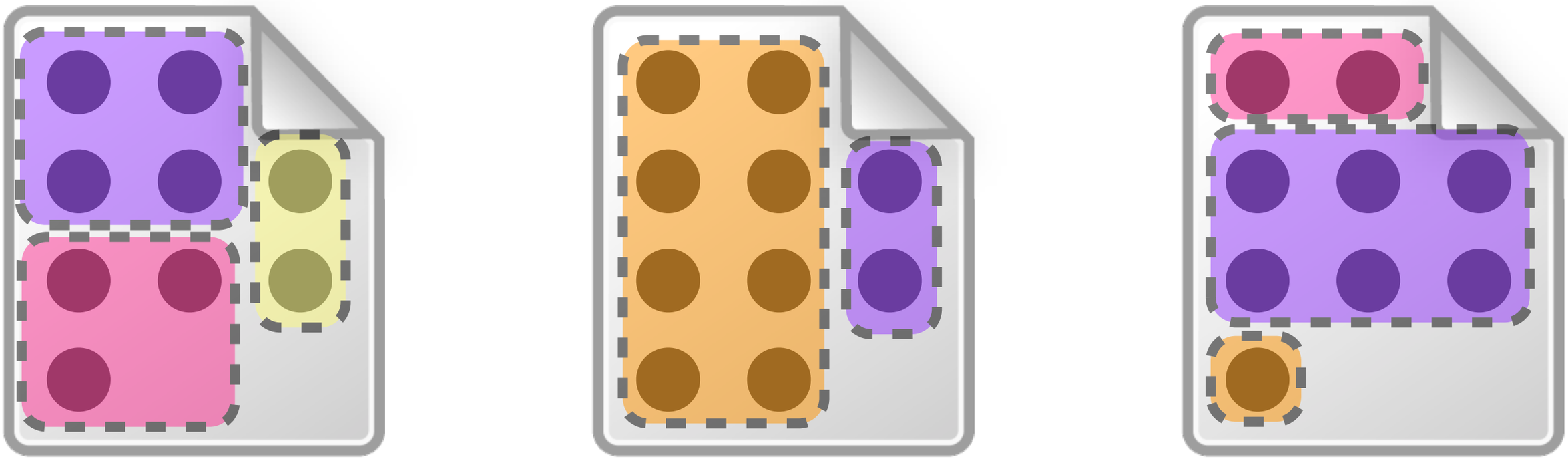
Document2



Document3

2.Grouping partitions across documents

- Also, number of group is not fixed a priori



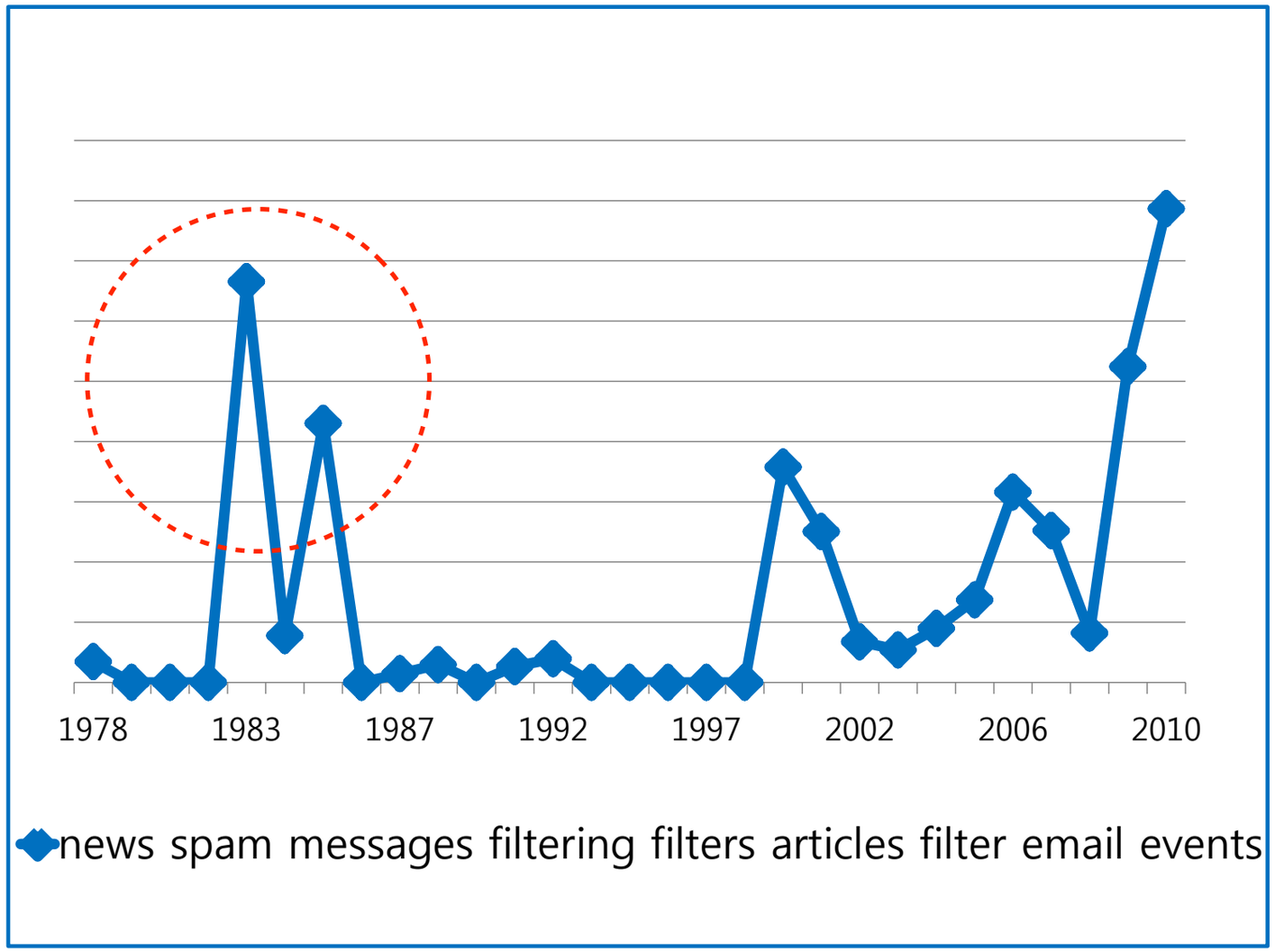
Document1 Document2 Document3

- Finally, the grouped partitions forms topics

Limitations with CRF

- CRF is an exchangeable prior

- CRF does not incorporate relationships (dependencies) between documents for modeling topics



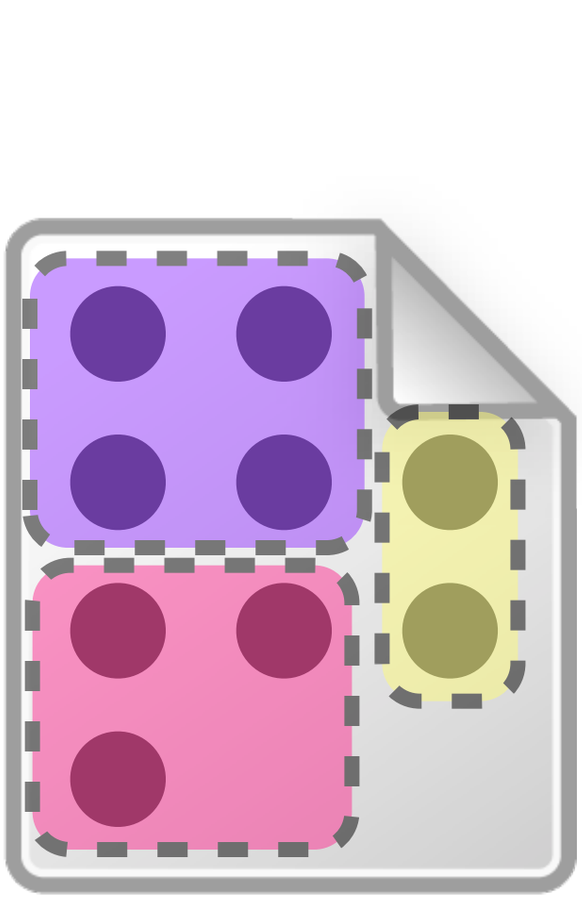
"Spam filtering" at 1980s ??

Distance Dependent Chinese Restaurant Franchise

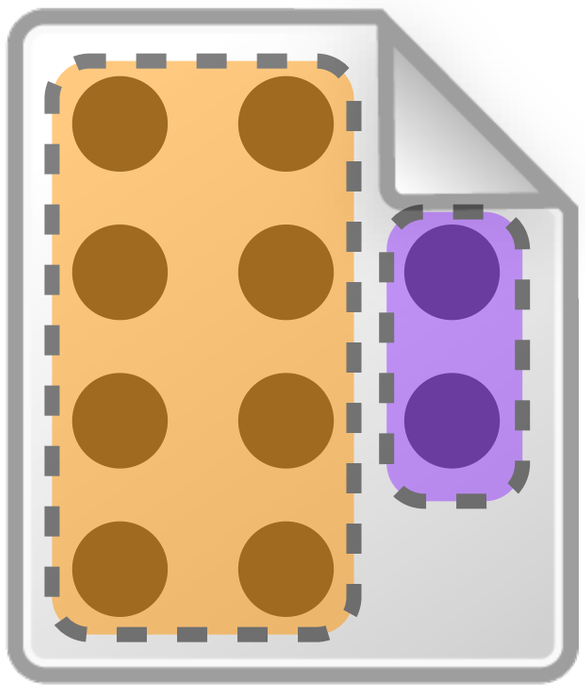
- Modified second process of CRF
- In CRF, deciding a group of new partition is proportional to the number of partitions already assigned that group only
- Grouping partitions by considering the relationship between documents

Assigning a group for new partition

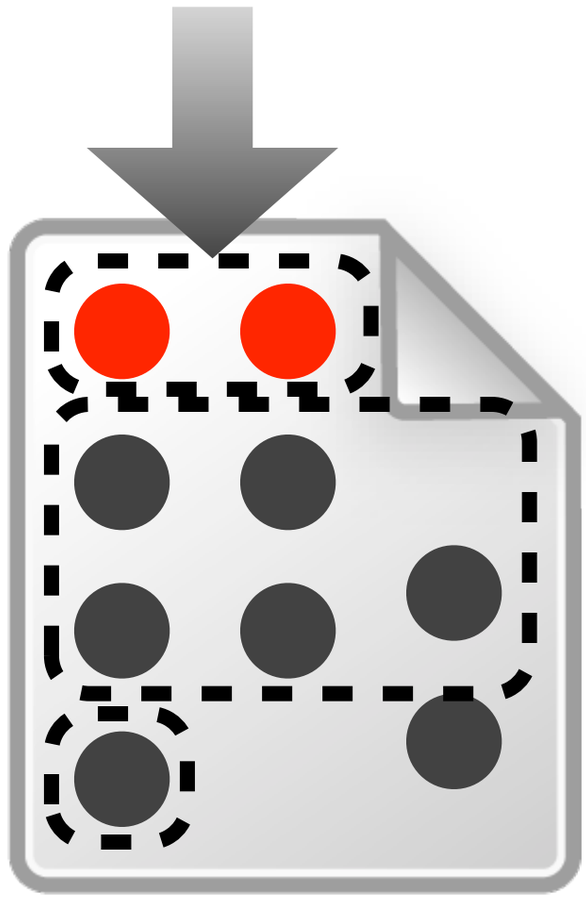
Let's consider about this partition



Document1

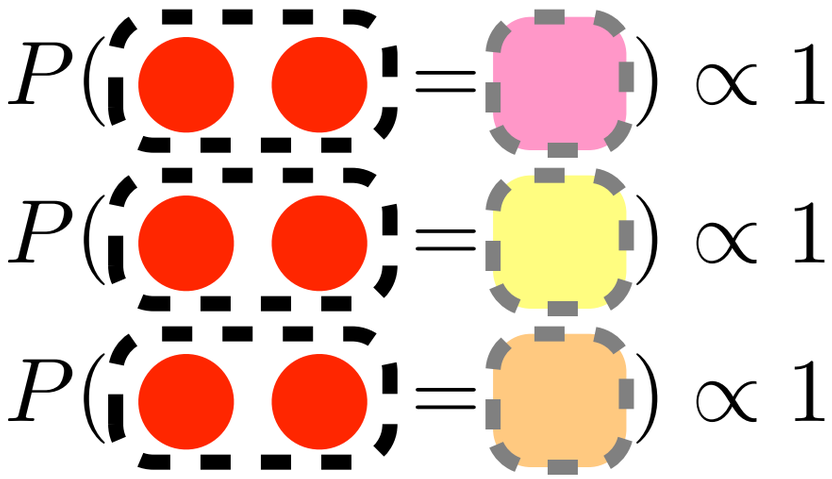


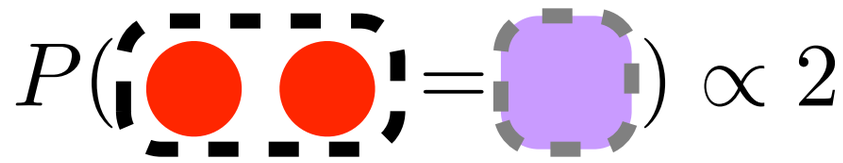
Document2

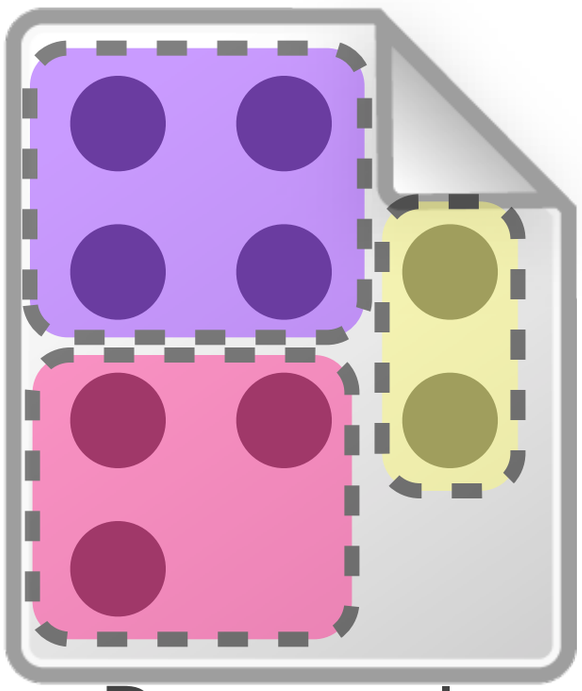


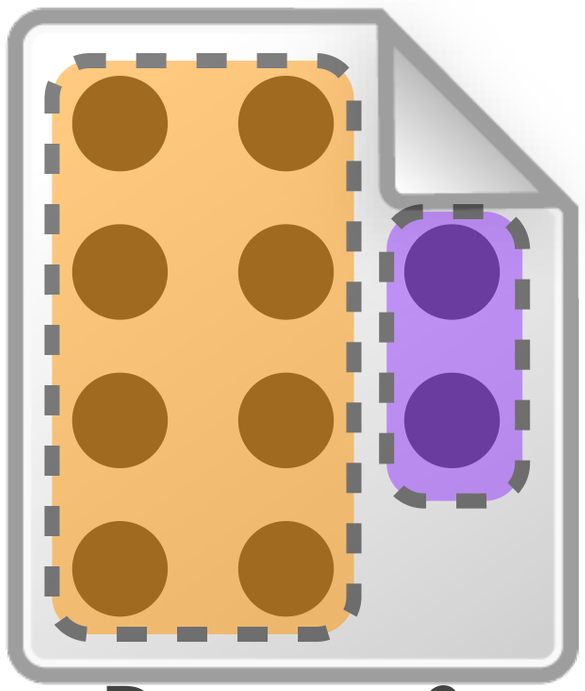
Document3

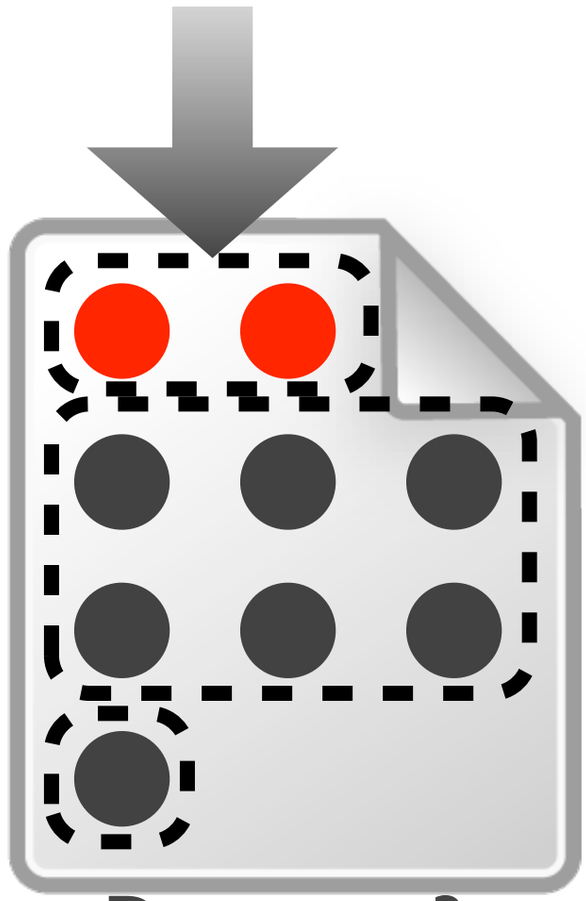
In Original CRF



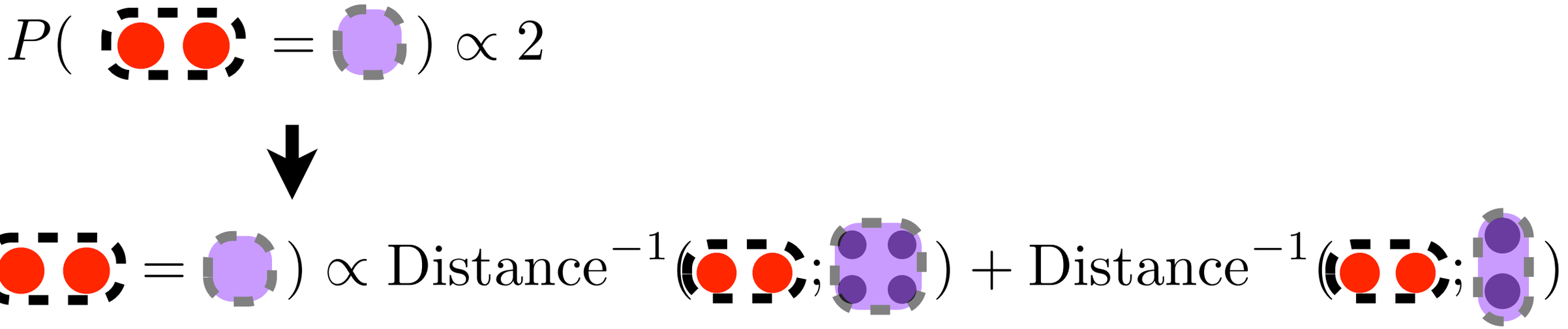


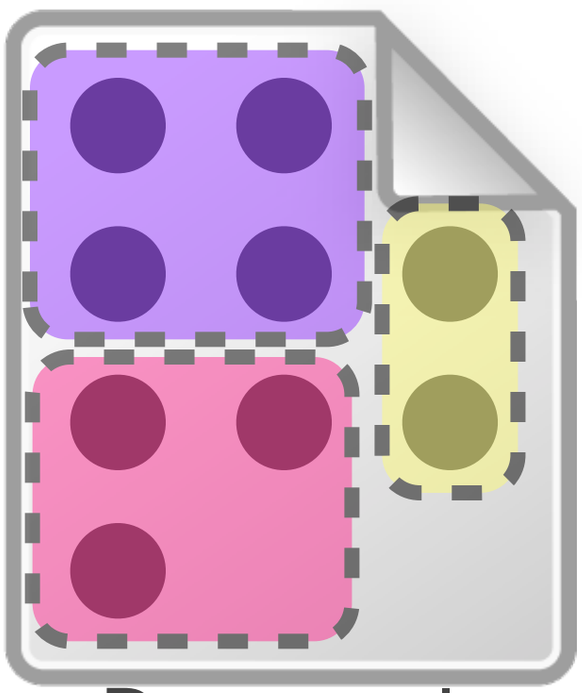

Document1

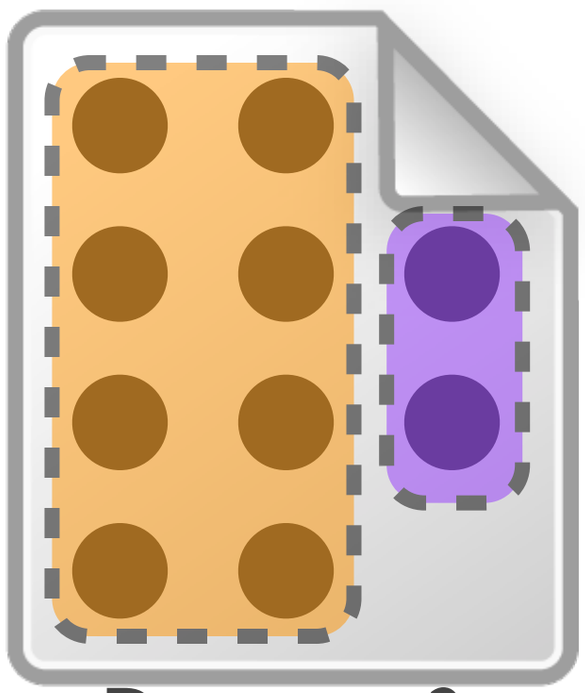

Document2

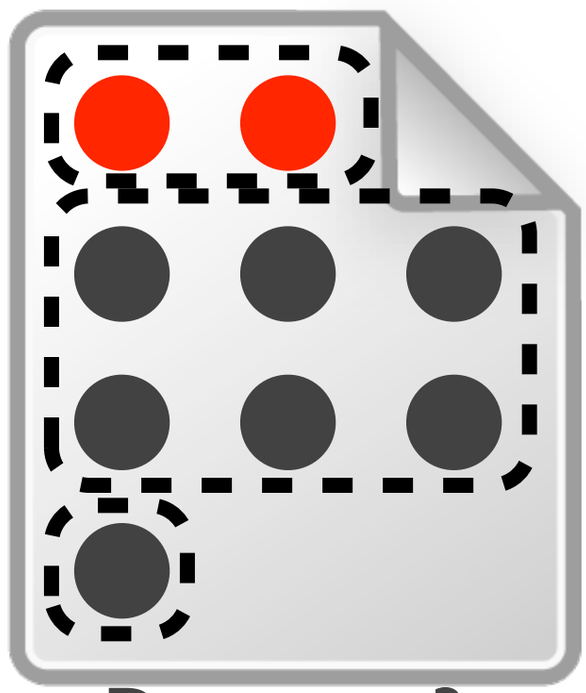

Document3

In ddCRF




Document1


Document2

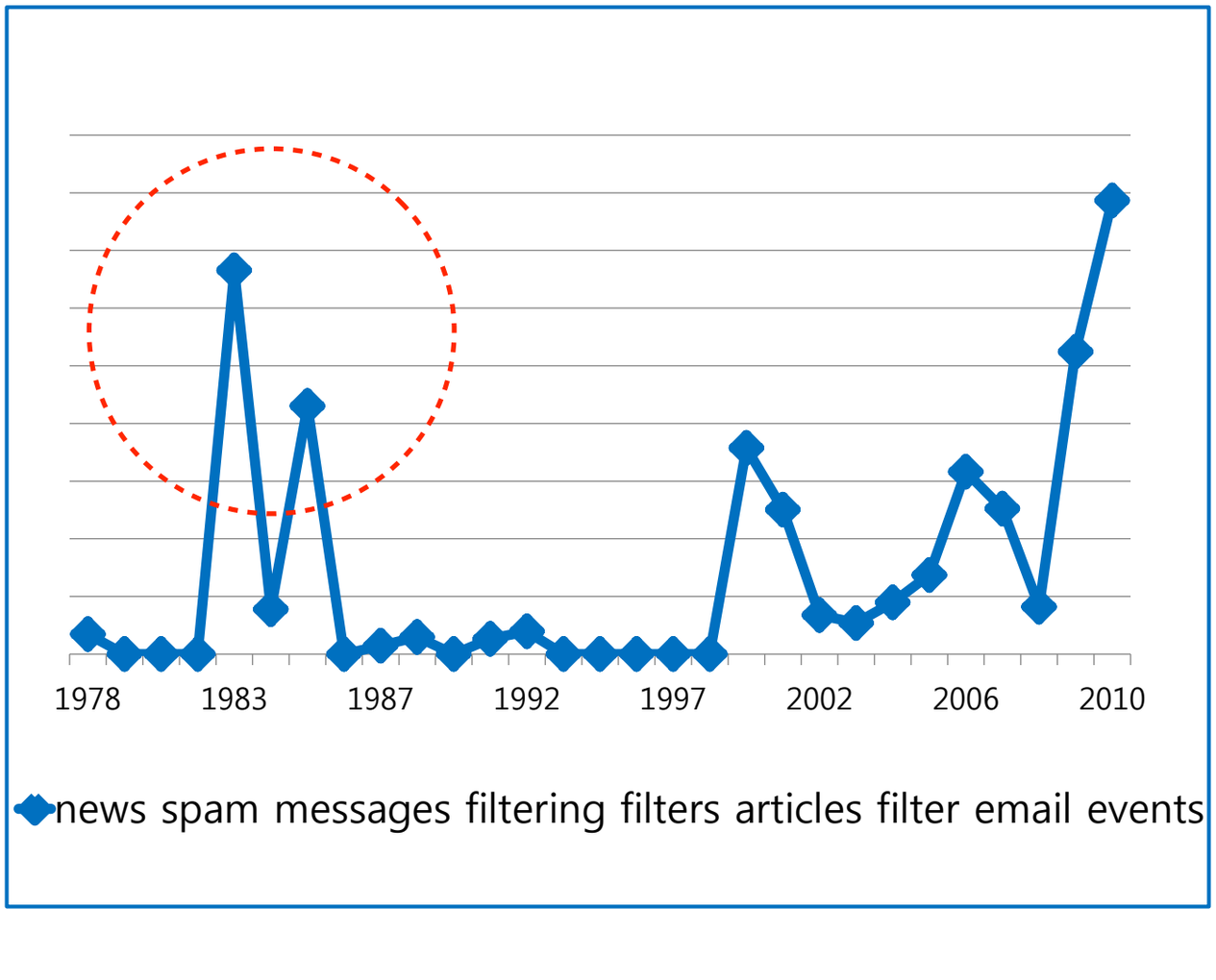

Document3

Experiments

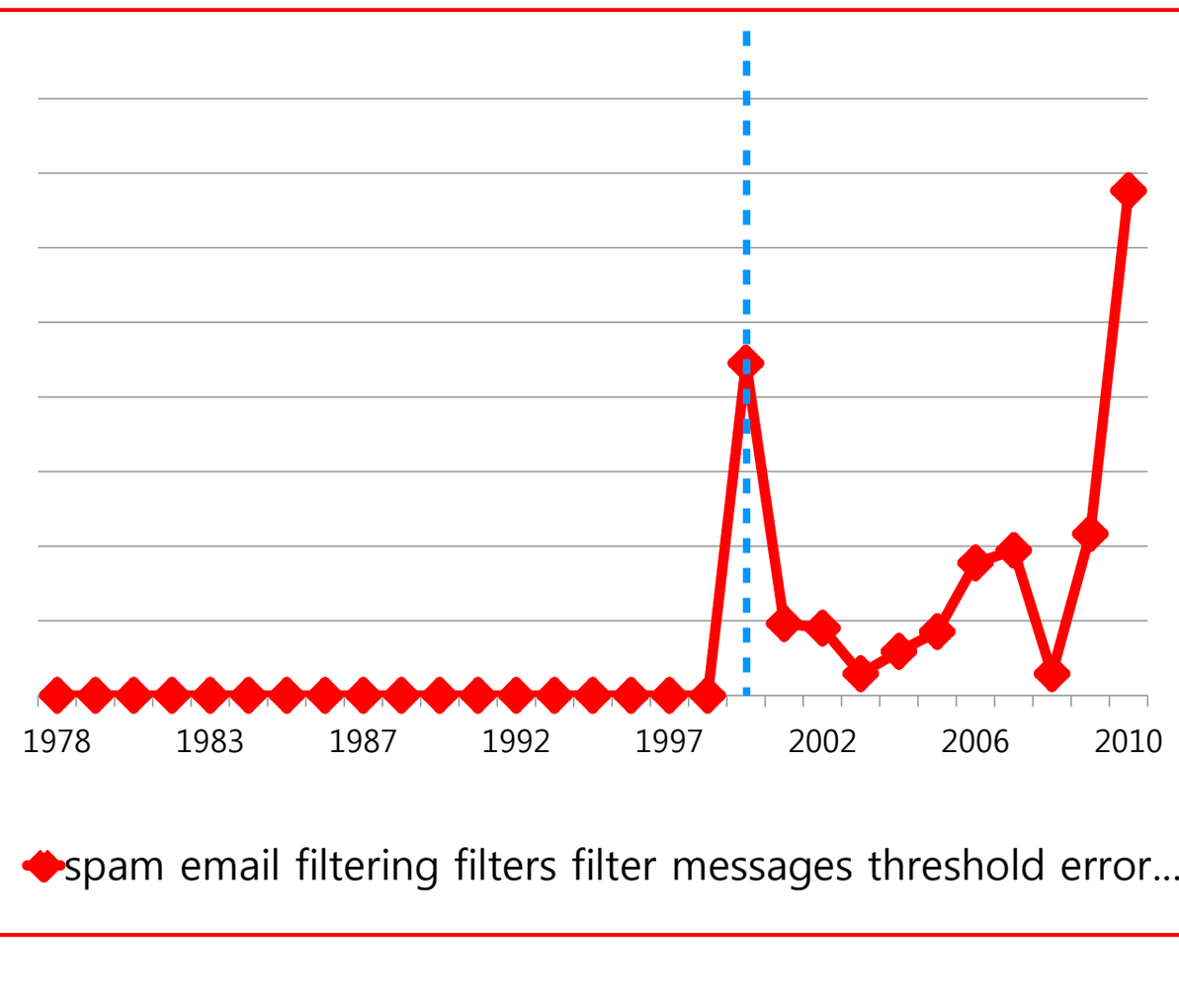
- Four conference abstract datasets
  - SIGIR, SIGMOD, SIGGRAPH, NIPS
- Using two distance functions for measuring time distances between documents
- Posterior sampling by Gibbs sampler

Results

CRF



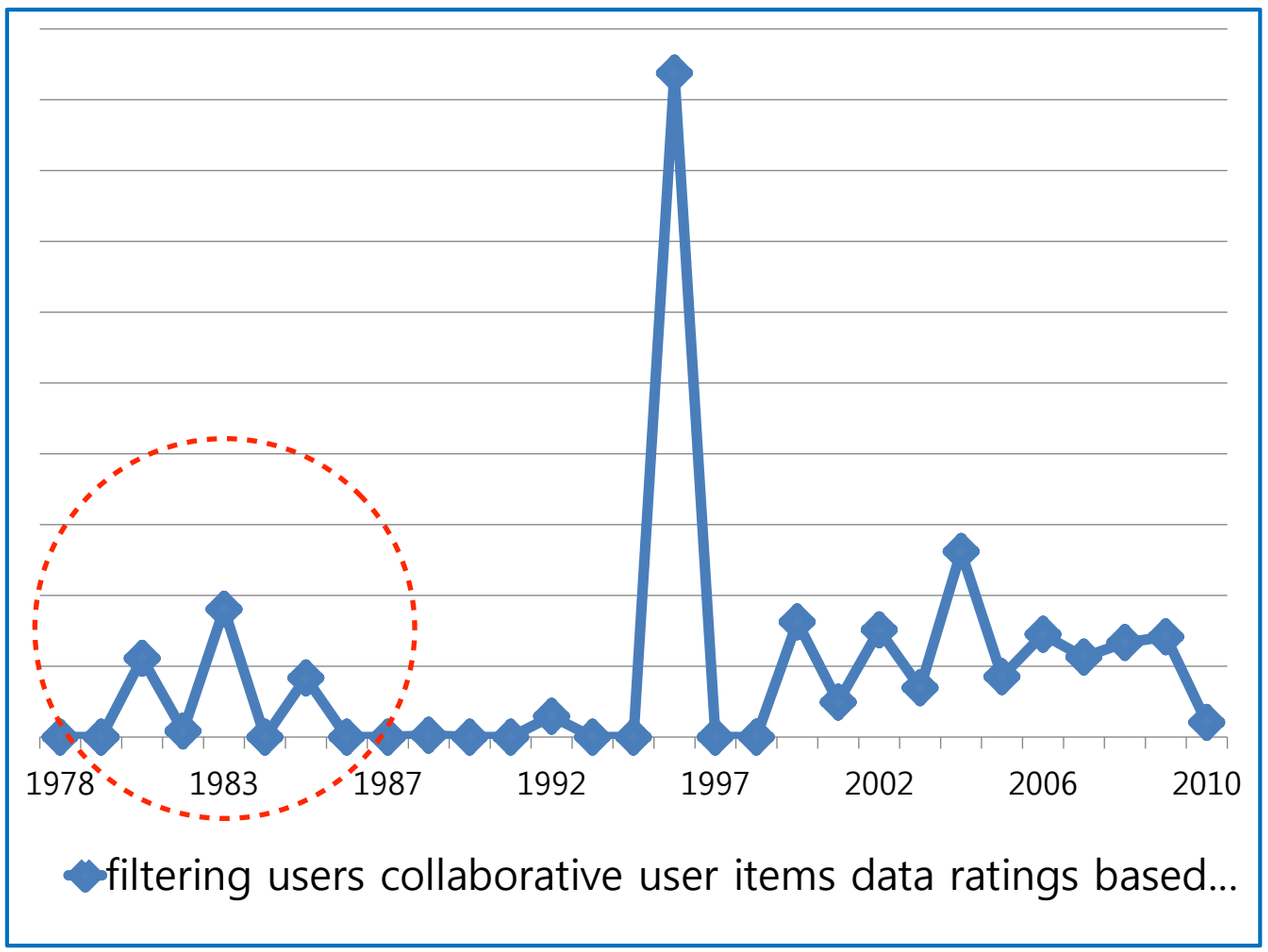
ddCRF



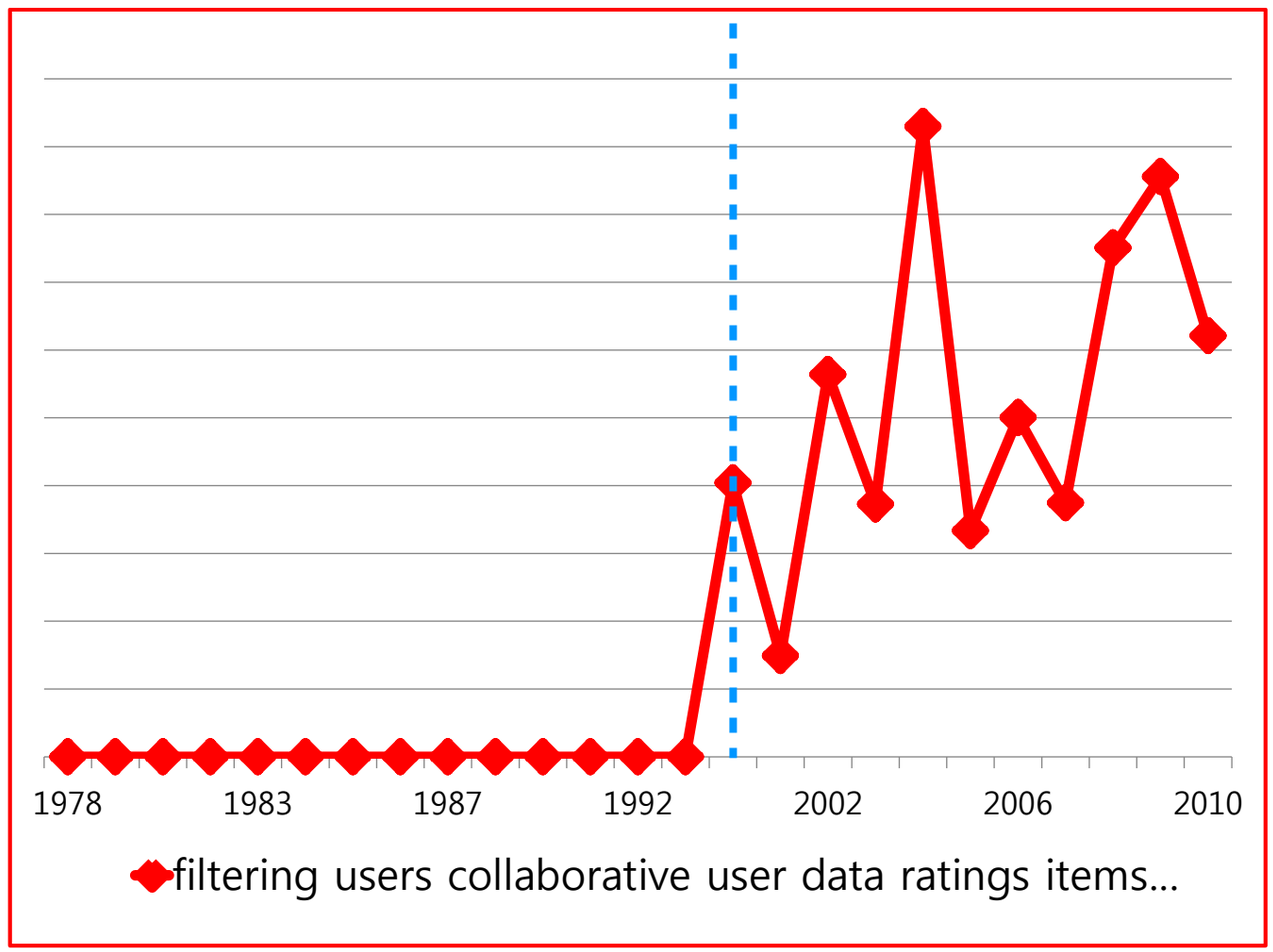
Spam filtering topic emerge at 2000 (from SigIR)

Results

CRF

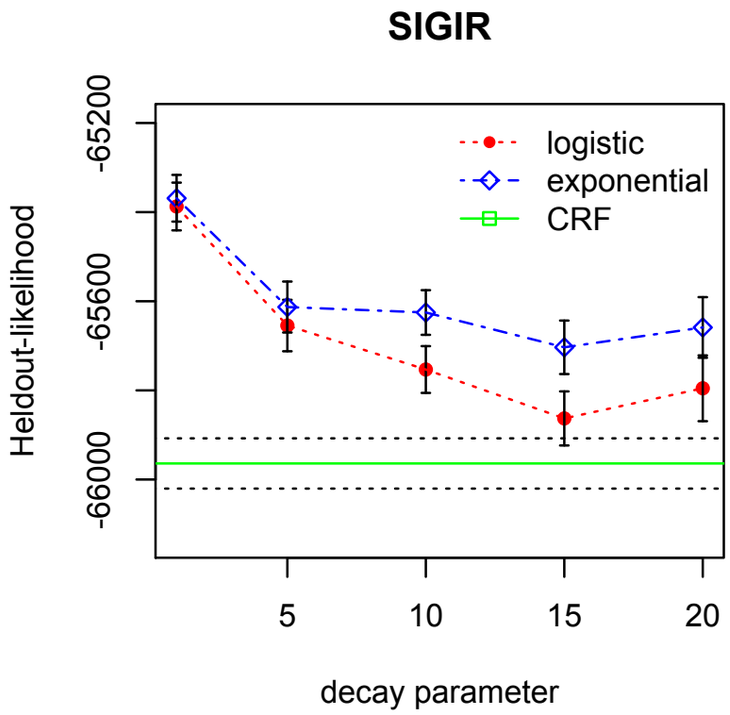


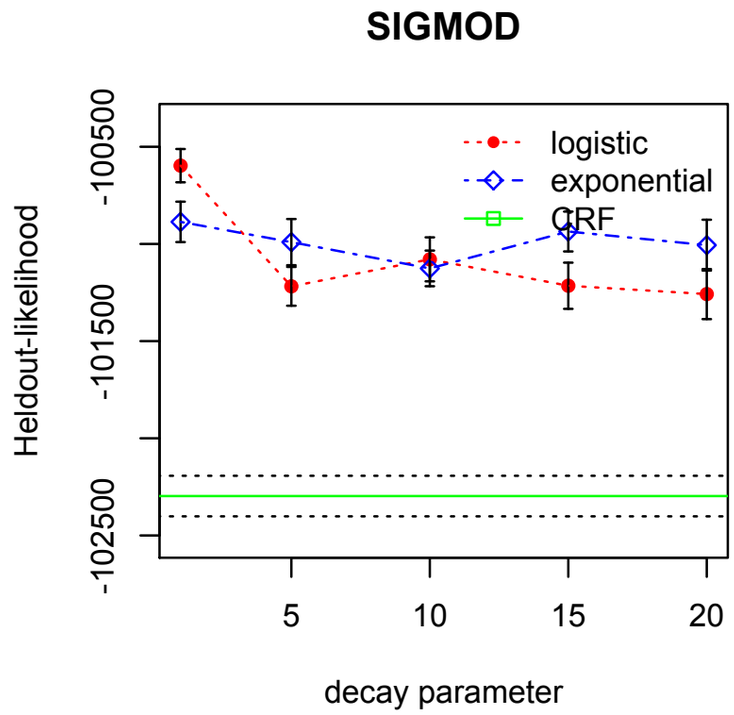
ddCRF

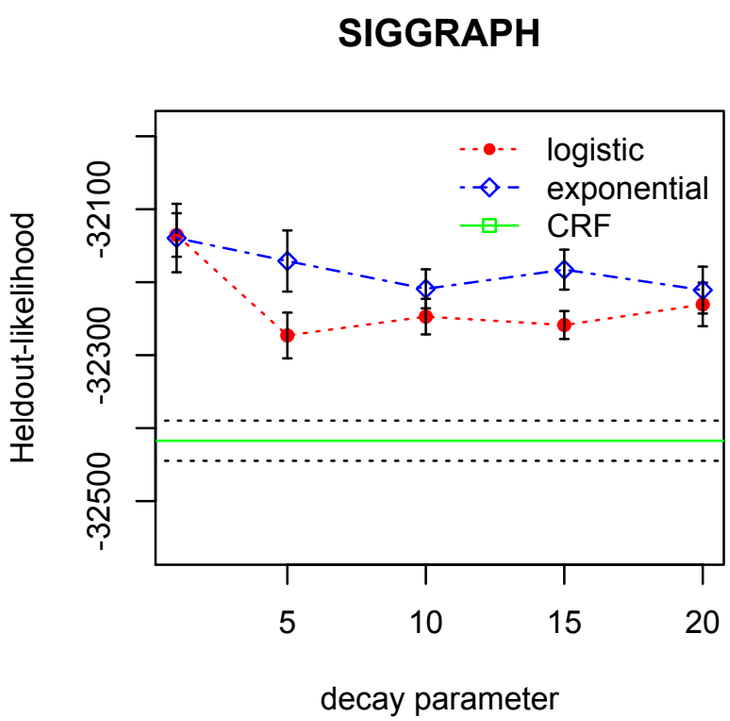


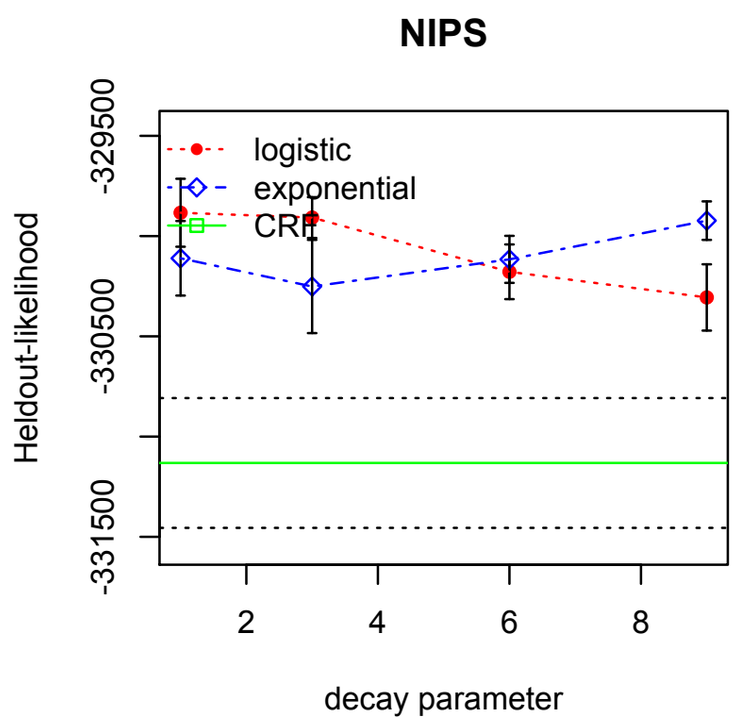
Collaborative filtering topic emerge at 1999 (from SigIR)

Held-out Likelihood

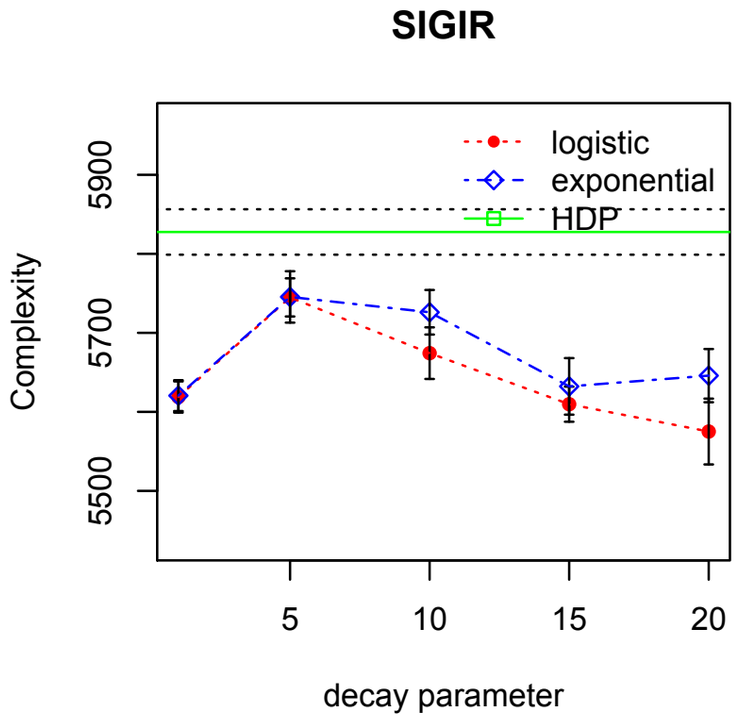


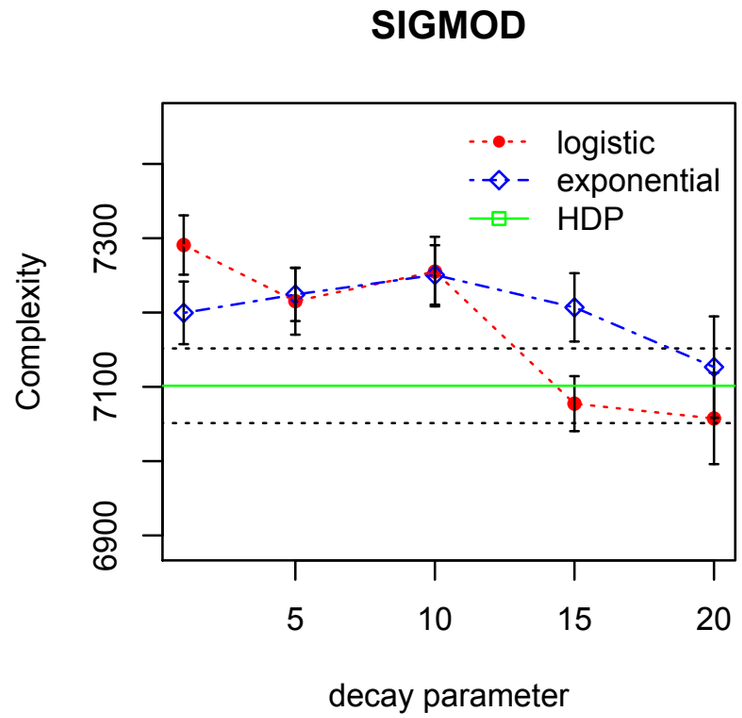


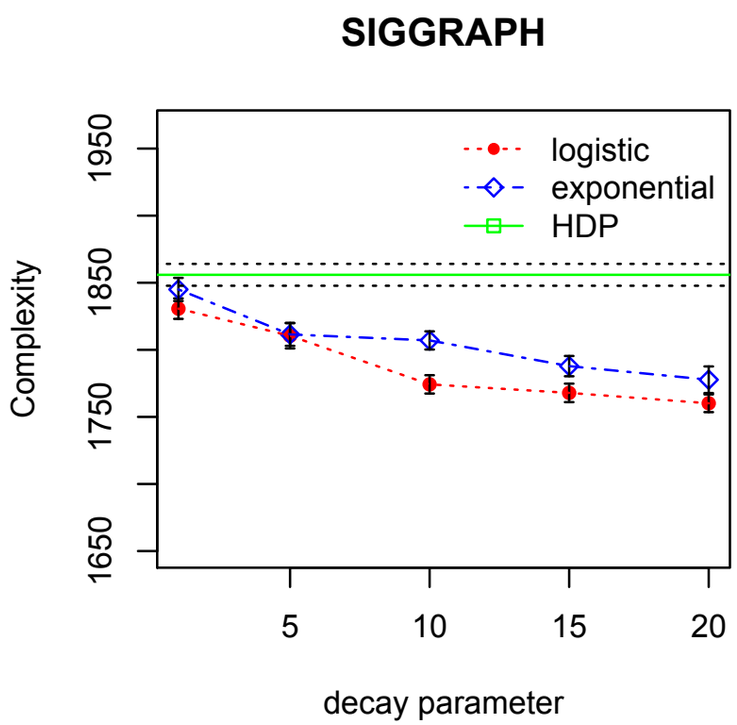


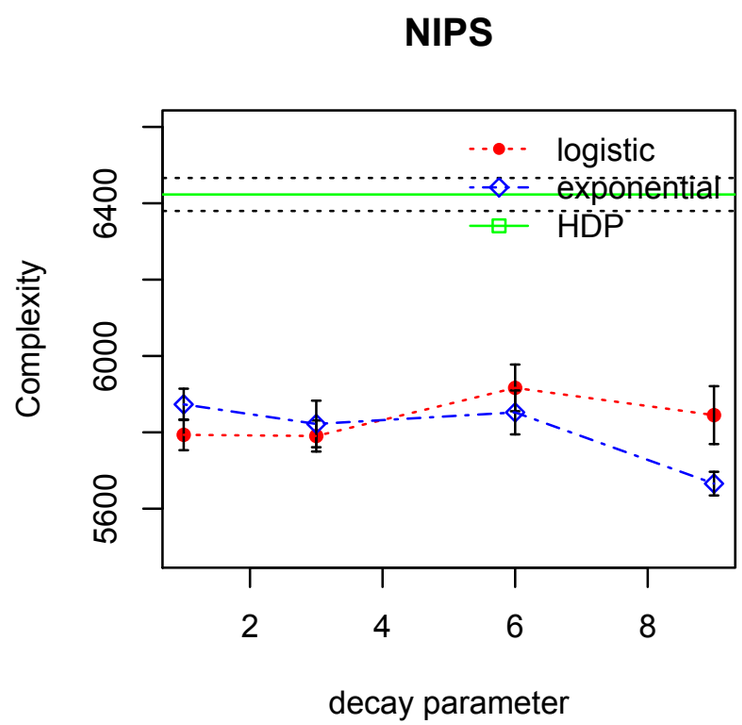


Complexity


SIGIR
Complexity
5500
5700
5900
logistic
exponential
HDP
5
10
15
20
decay parameter


SIGMOD
Complexity
6900
7100
7300
logistic
exponential
HDP
5
10
15
20
decay parameter


SIGGRAPH
Complexity
1650
1750
1850
1950
logistic
exponential
HDP
5
10
15
20
decay parameter


NIPS
Complexity
5600
6000
6400
logistic
exponential
HDP
2
4
6
8
decay parameter

Conclusion and Further work

- Modeled topics from four different corpora to capture temporal patterns of topics
  - Quantitative evaluation shows improved performance over LDA(parametric topic model) and HDP(non-parametric topic model)
  - Qualitative evaluation shows interesting temporal patterns of topic emergence
- Future work will explore various definitions of distance: time dimension, spatial dimension, or some other dimension
  - Also it can be interesting to combine two more dimensions into one distance function

Thank you!