07 – Topic identification, topic modelling

IA161 Advanced Techniques of Natural Language Processing

Jiří Materna

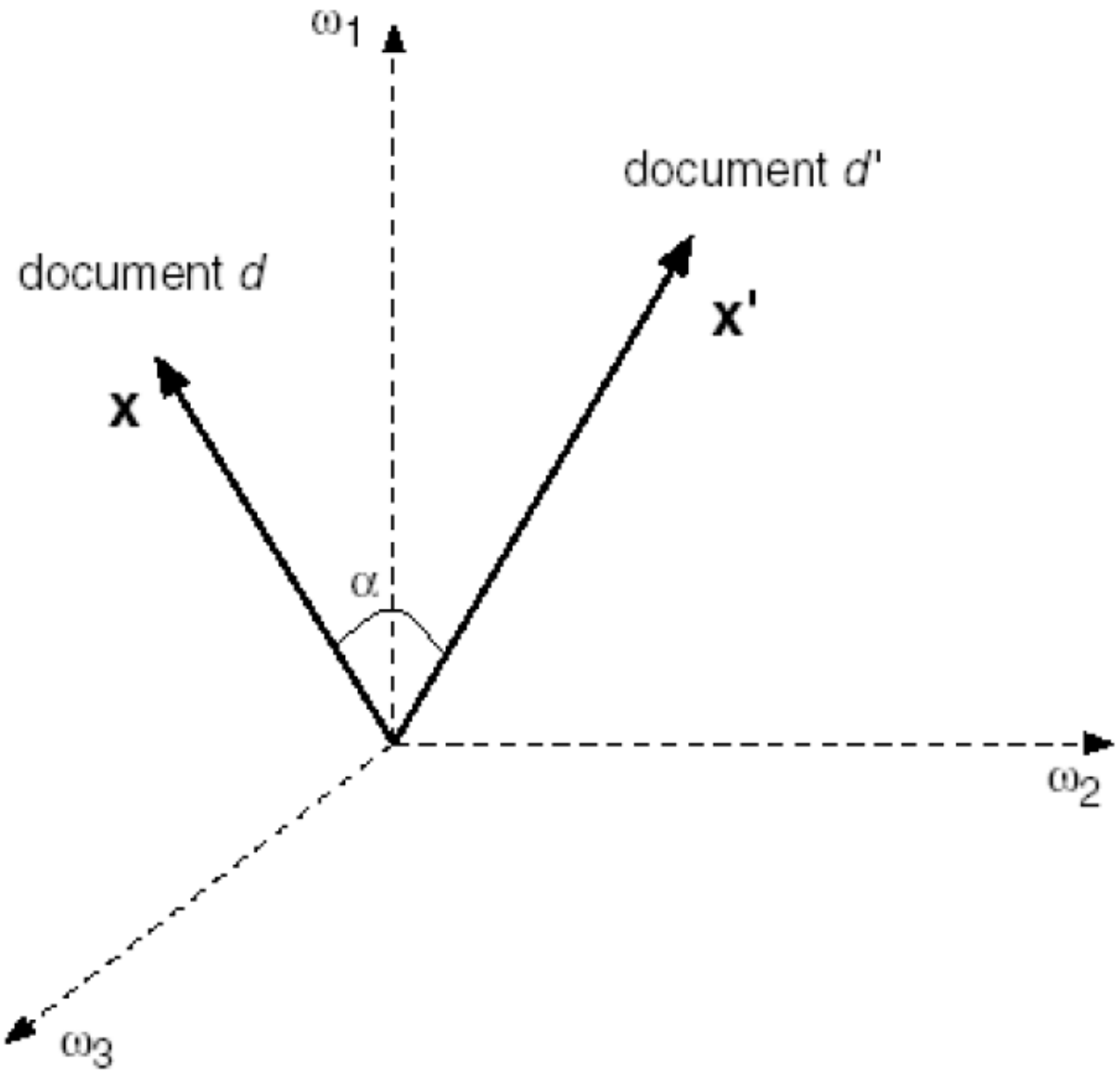
Seznam.cz, a.s.

November 9, 2015

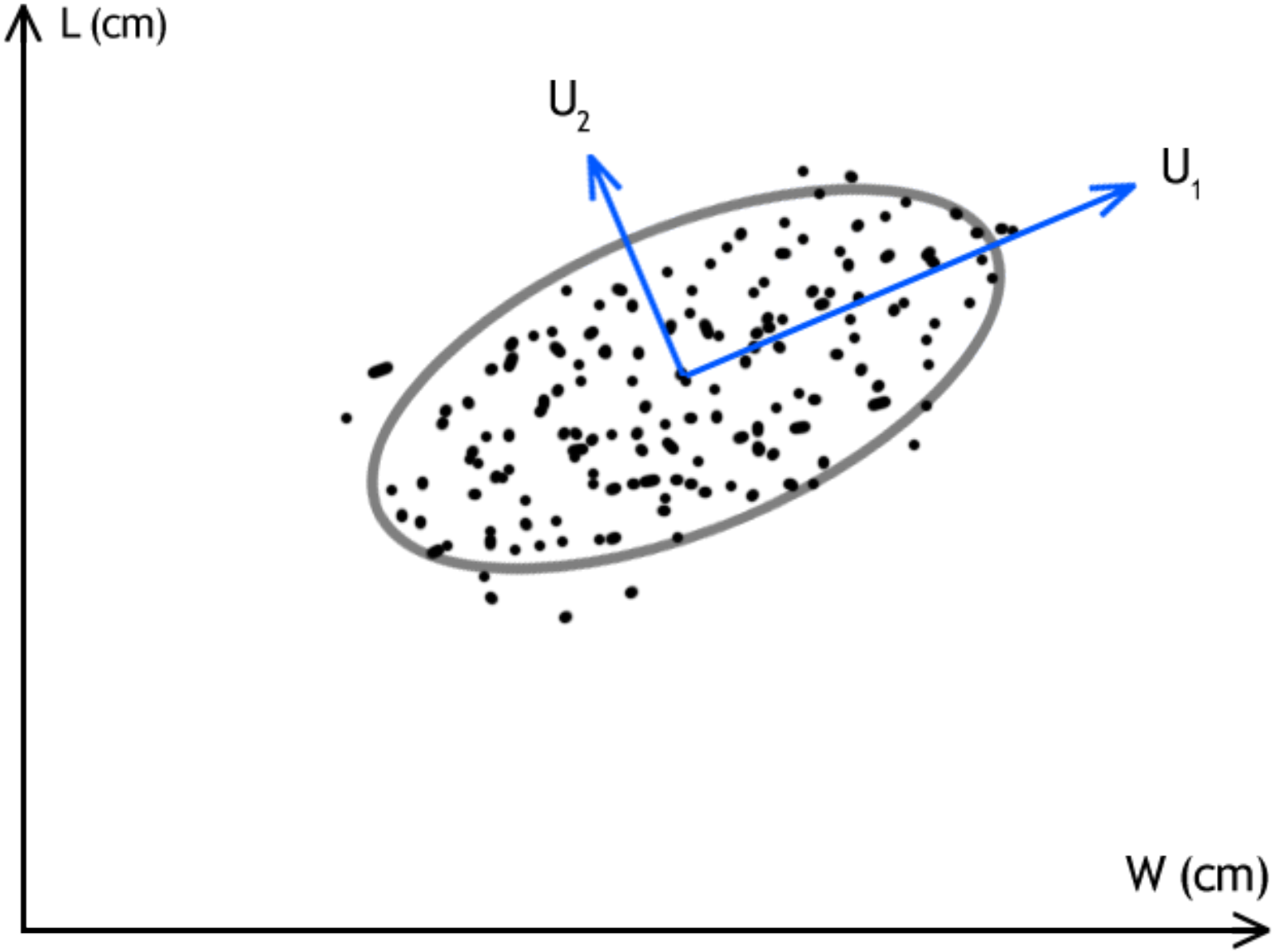
Outline

- Introduction to topic modelling
- Latent Semantic Analysis
- Latent Dirichlet Allocation
- Gensim

Topic modelling

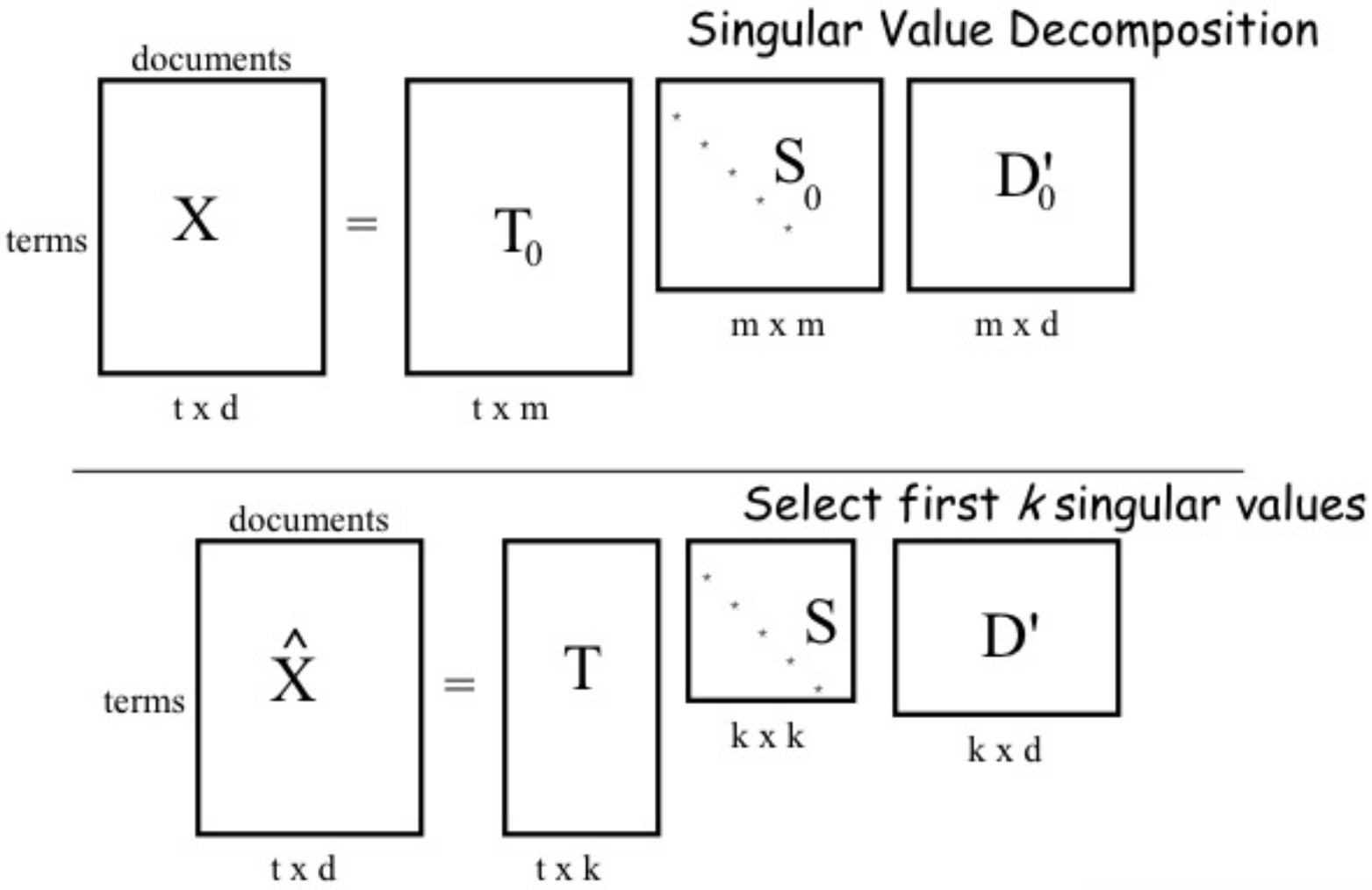


Singular Value Decomposition



source: https://commons.wikimedia.org

Latent Semantic Analysis



source: `http://csee.wvu.edu/~timm/cs591o/old/FSS.html`

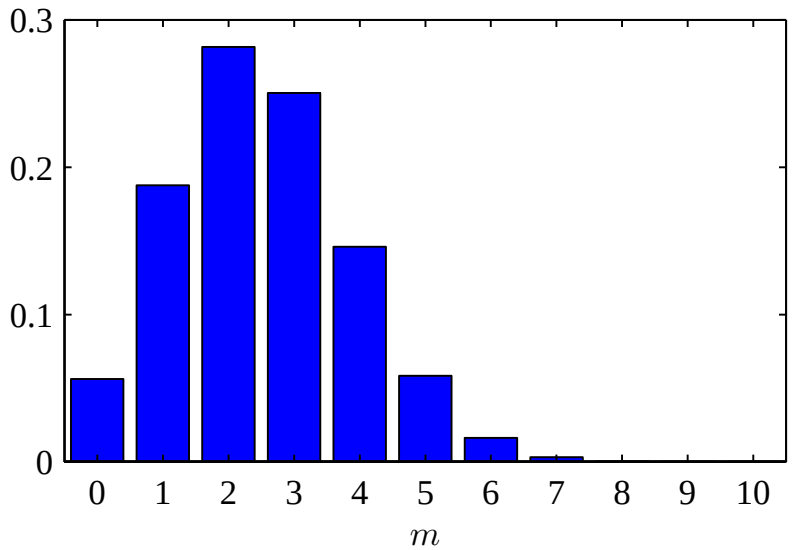
Important distributions

- Binomial

$$\mathrm{Bin}(k|n,p) = \binom{n}{k} p^k.(1-p)^{n-k} =$$

$$\mathrm{Bin}(x_1,x_2|p_1,p_2) = \frac{(x_1+x_2)!}{x_1!x_2!} p_1^{x_1}.p_2^{x_2}$$

$$p_1 + p_2 = 1$$



Example: $n = 10$, $p = 0.25$

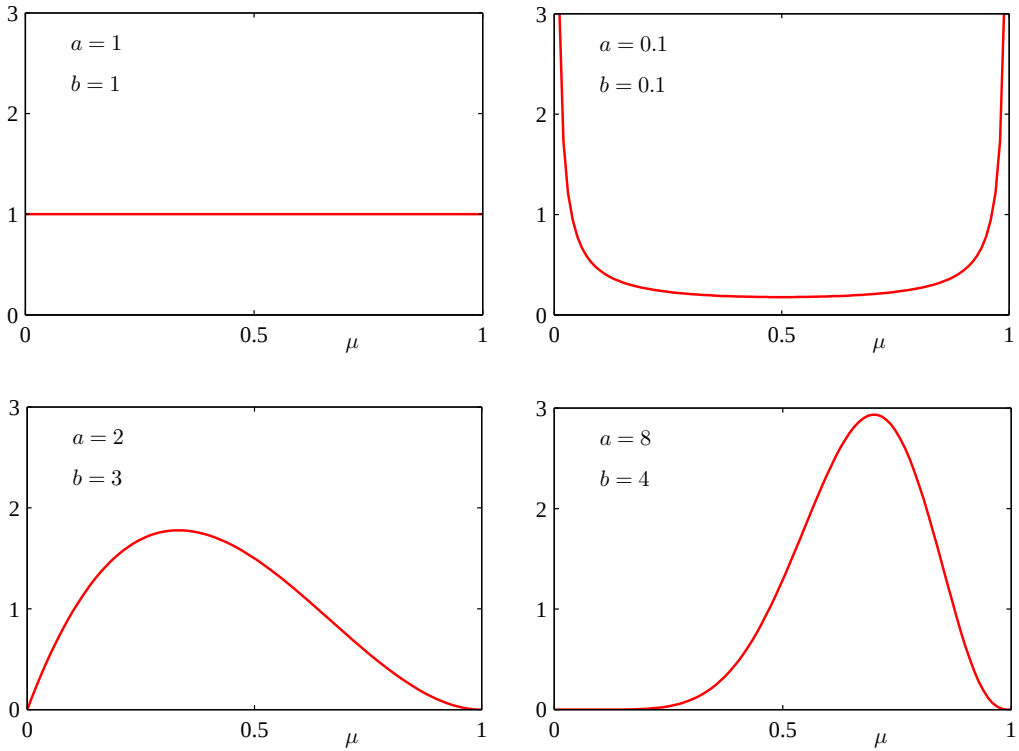
Important distributions

- Beta

$$\text{Beta}(p_1, p_2|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_1^{\alpha-1}.p_2^{\beta-1}$$

$$p_1 + p_2 = 1$$

$$\Gamma(x) = (x-1)!$$



Important distributions

- Multinomial

$$\text{Mult}(x_1 \dots x_n | p_1 \dots p_n) = \frac{(\sum x_i)!}{\prod x_i!} \prod_{i=1}^{n} p_i^{x_i}$$

- Dirichlet

$$\text{Dir}(p_1 \dots p_n | \alpha_1 \dots \alpha_n) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1}$$

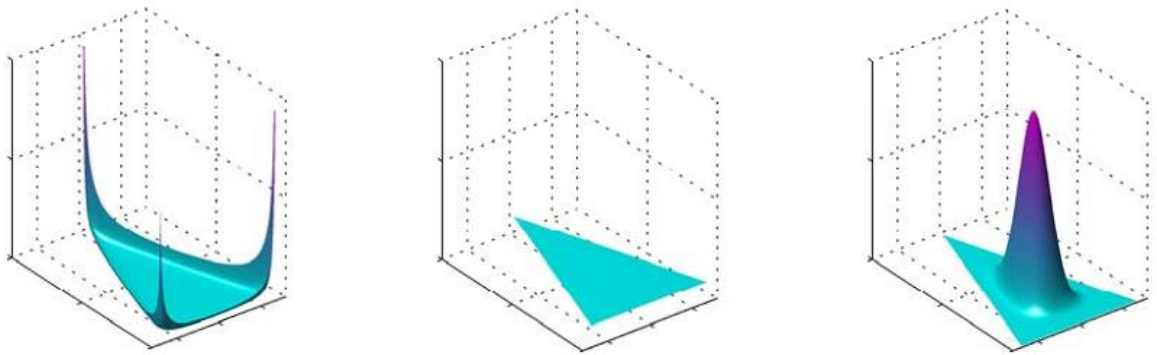
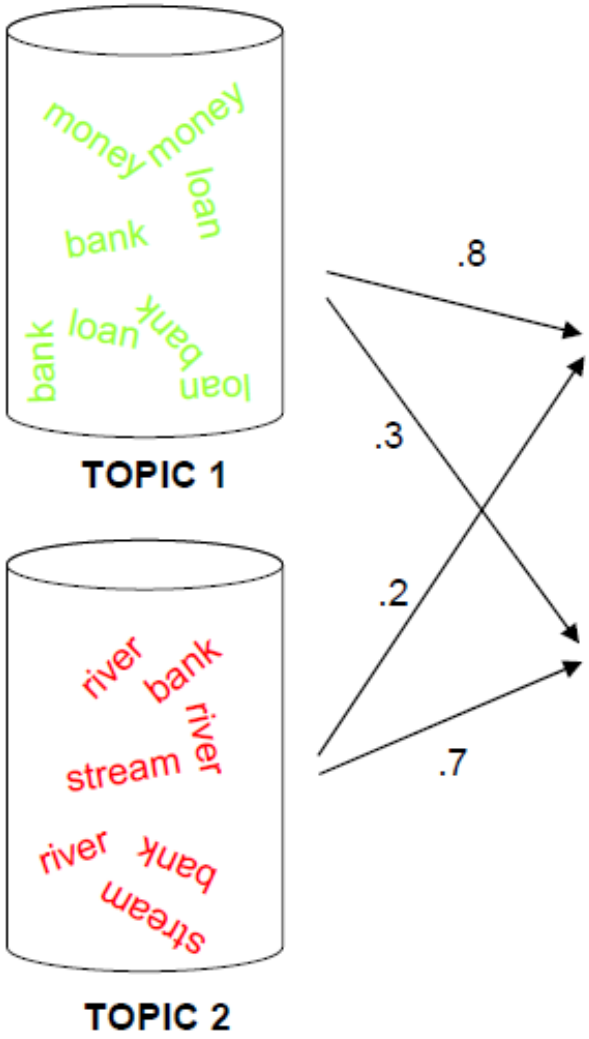


Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\} = 10$ in the right plot.

Latent Dirichlet Allocation



DOCUMENT 1: money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] money[1] stream[2] bank[1] money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] bank[1] money[1] stream[2]

DOCUMENT 2: river[2] stream[2] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] bank[1] stream[2] river[2] loan[1] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] bank[2] stream[2] bank[2] money[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] money[1] bank[1] stream[2] river[2] bank[2] stream[2] bank[2] money[1]

Gensim

```
>>> from gensim import corpora, models, similarities
>>>
>>> # Load corpus iterator from a Matrix Market file on disk.
>>> corpus = corpora.MmCorpus('/path/to/corpus.mm')
>>>
>>> # Initialize Latent Semantic Indexing with 200 dimensions.
>>> lsi = models.LsiModel(corpus, num_topics=200)
>>>
>>> # Convert another corpus to the latent space and index it.
>>> index = similarities.MatrixSimilarity(lsi[another_corpus])
>>>
>>> # Compute similarity of a query vs. indexed documents
>>> sims = index[query]
```

References I


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3:993 – 1022.


Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988).
Using Latent Semantic Analysis to Improve Access to Textual Information.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, pages 281–285, New York, NY, USA. ACM.


Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet processes .
Journal of the American Statistical Association, 101:1566 – 1581.