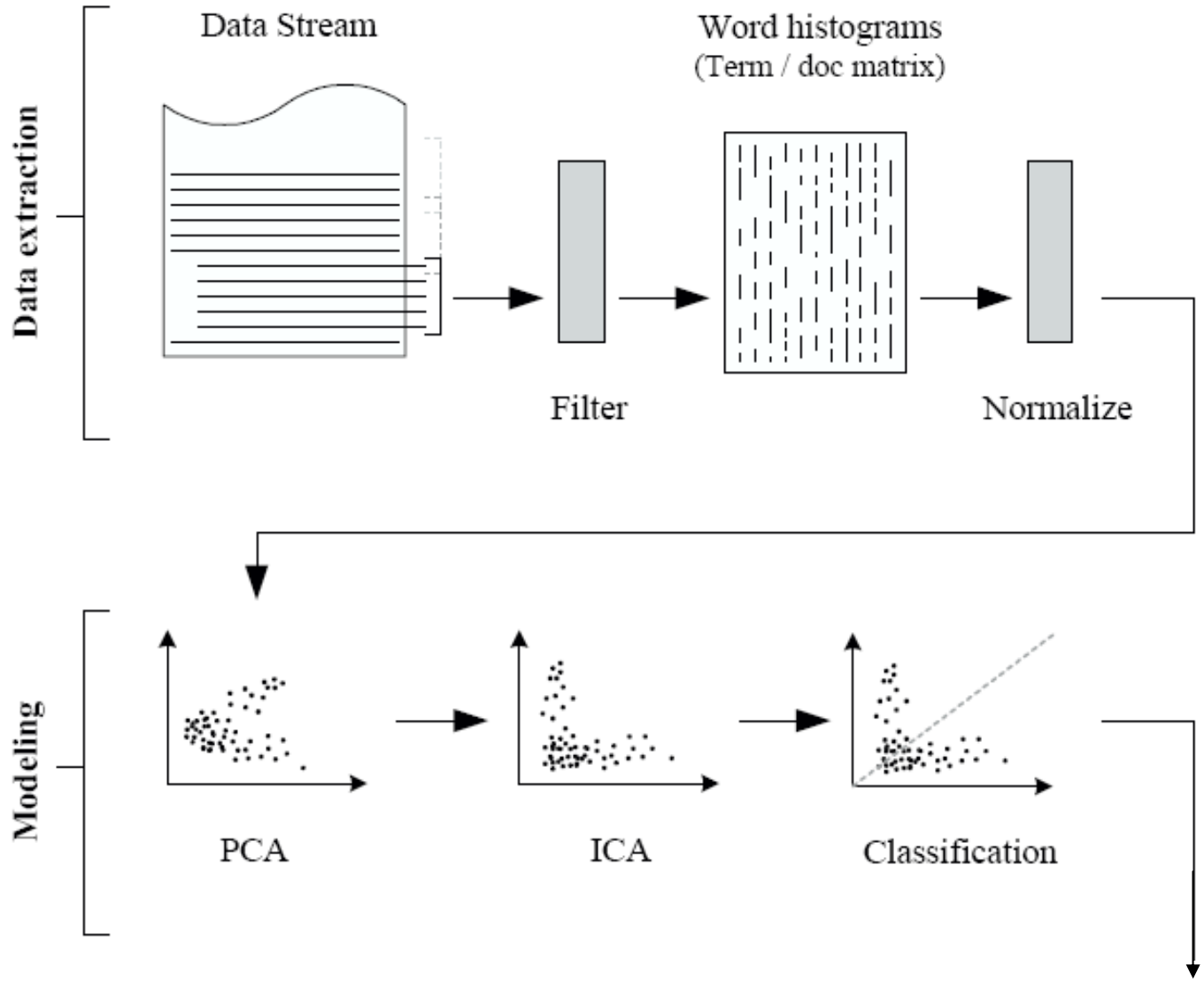
LSA – Algorithms

Ph.D.-student

Lasse Lohilahti Mølgaard

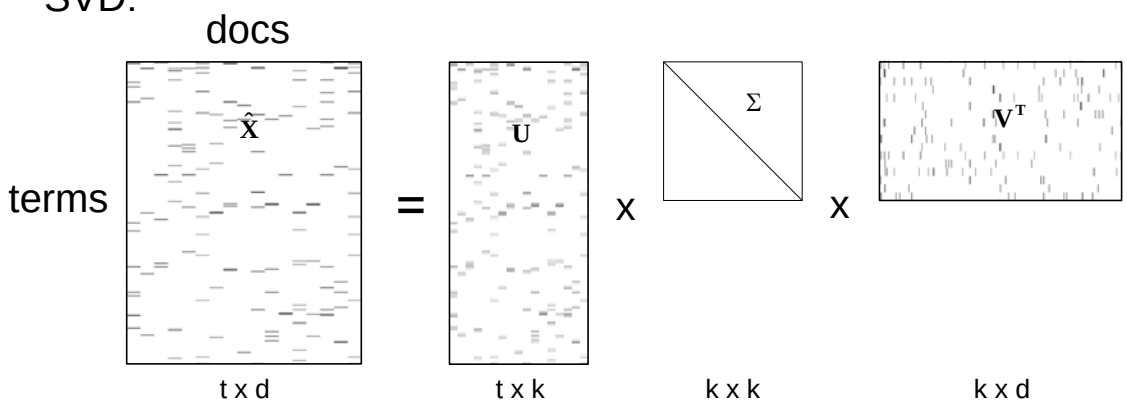
Agenda

- Basics of LSA
  - Mathematical model
- PLSI
  - Probabilistic interpretation
  - Results in retrieval
- Demos
  - Castsearch
  - Wikipedia


Data extraction
Data Stream
Word histograms
(Term / doc matrix)
Filter
Normalize
Modeling
PCA
ICA
Classification

LSA - algorithm

SVD:



Reduced singular value decomposition of the term x document matrix, X. Where:

T has orthogonal, unit-length columns (T' T = I)
D has orthogonal, unit-length columns (D' D = I)
S is the diagonal matrix of singular values

t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq \min(t,d)$)
k is the chosen number of dimensions in the reduced model ($k \leq m$)

SVD properties

- SVD solves the eigenvalue-problem for the matrices,

$$\mathbf{U} : \mathbf{X}\mathbf{X}^T \text{ and } \mathbf{V} : \mathbf{X}^T\mathbf{X}$$

- Eigenvectors for these matrices are orthogonal, i.e.

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n] \text{ is orthogonal if:}$$

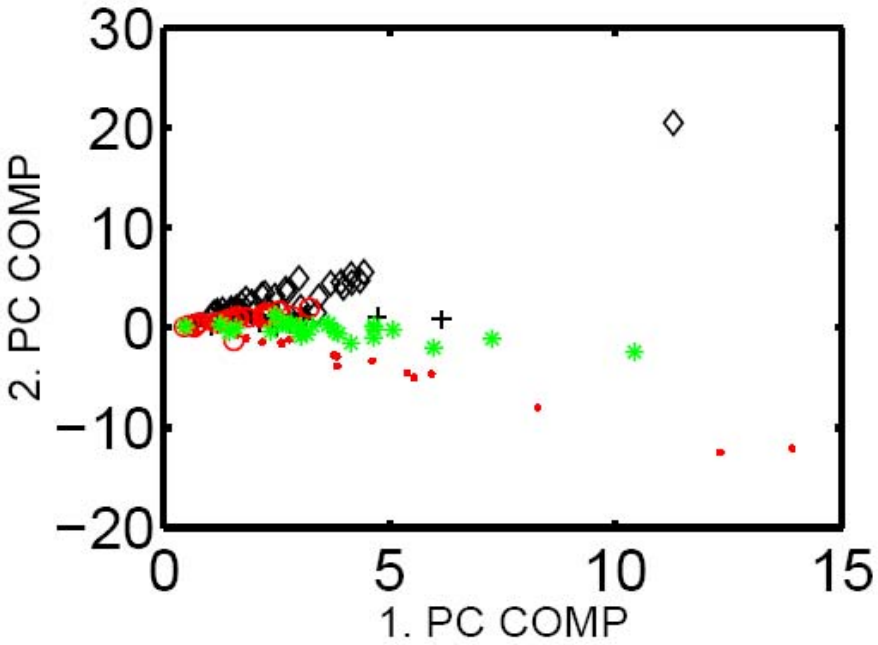
$$\mathbf{q}_i^T \mathbf{q}_j = 0, i \neq j \text{ and } \mathbf{q}_i^T \mathbf{q}_j \neq 0, i = j$$

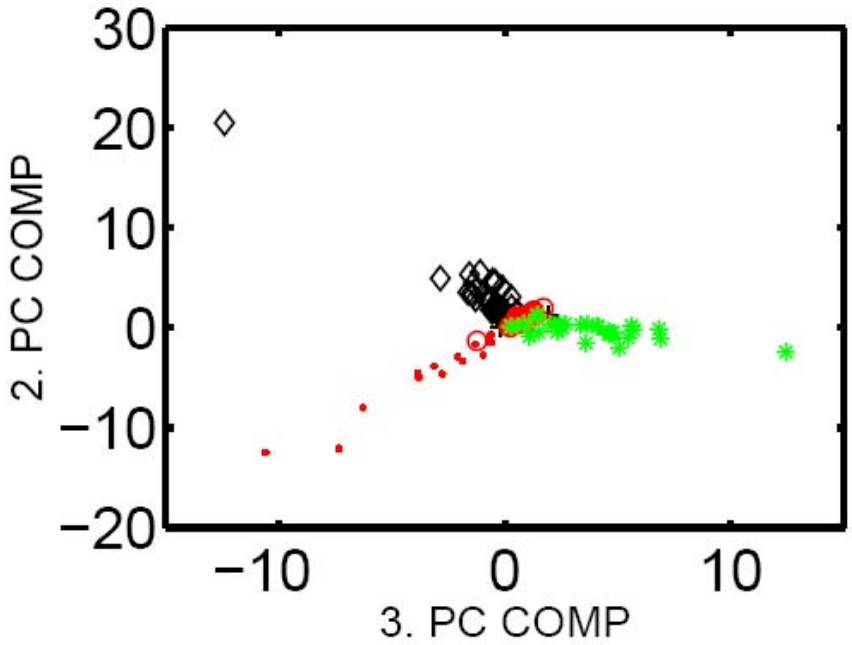
- LSA uses the vectors corresponding to the largest eigenvalues.

SVD properties

- The extracted vectors correspond to the directions containing most of the variance.
- SVD can be solved very efficiently because of these properties and the sparsity of the termdoc matrix

Text corpus example



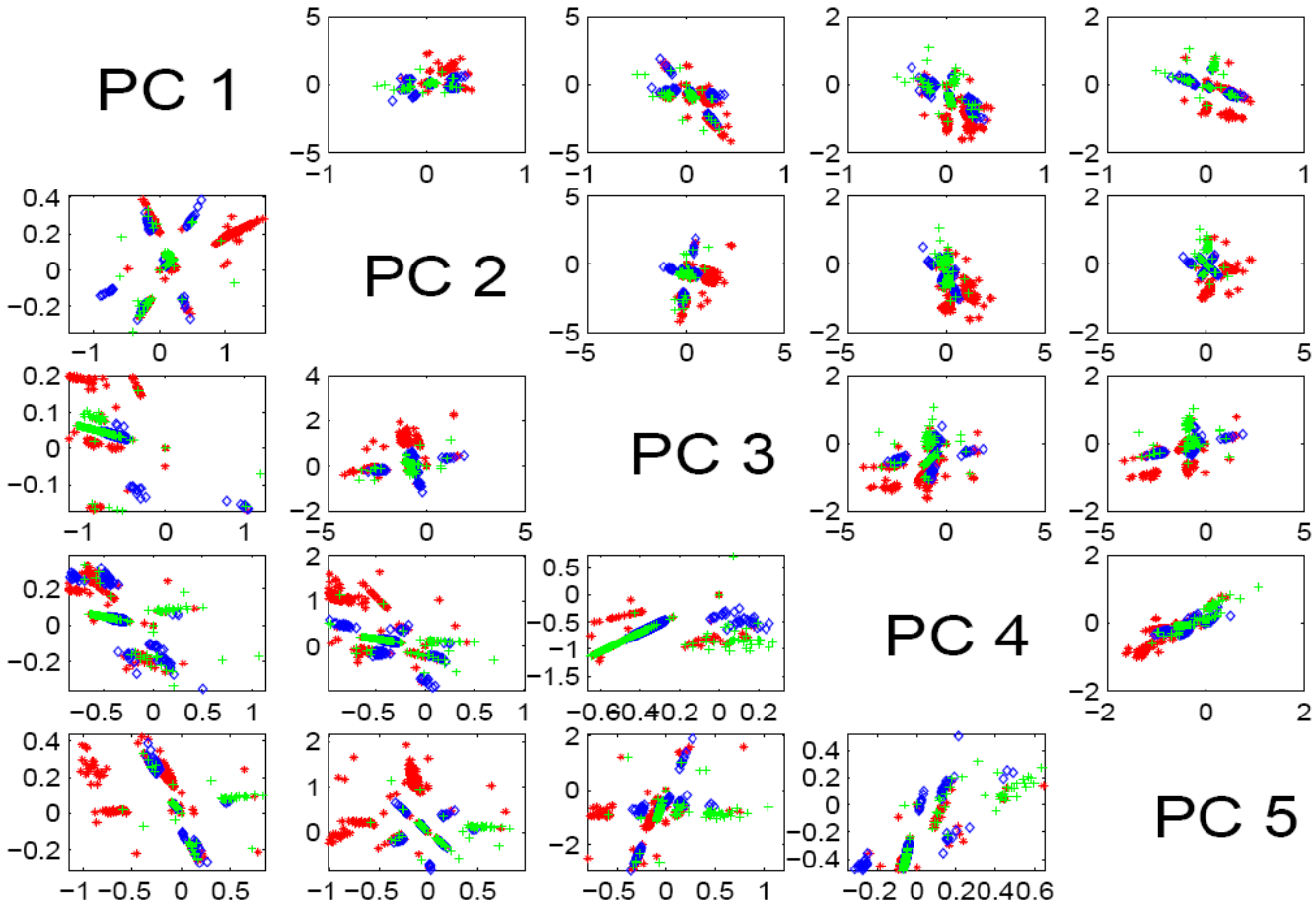


LSA - shortcomings

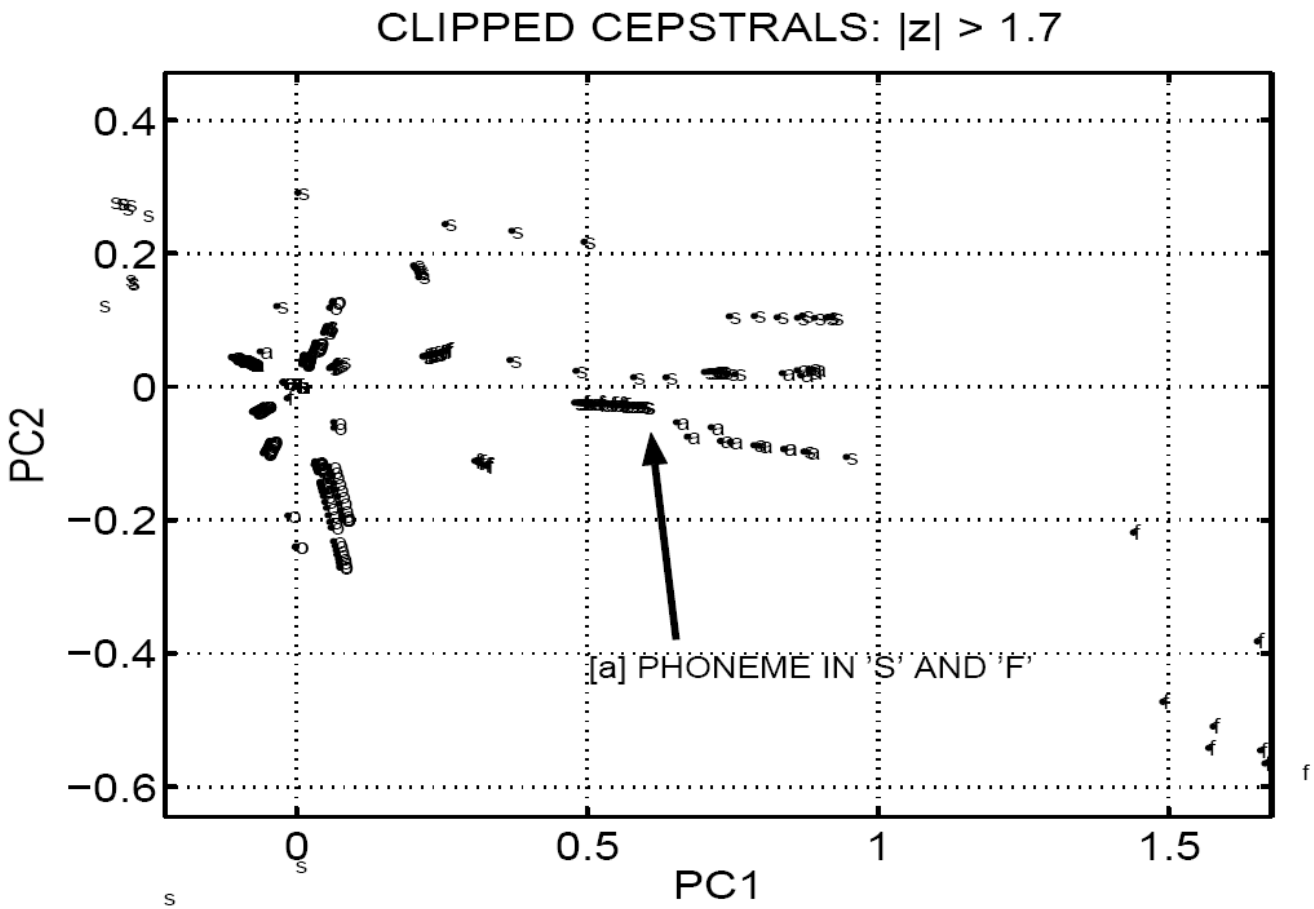
- LSA components are orthogonal
- Components are not directly interpretable – subsequent clustering necessary
- Missing a way to find number of components
- May return negative entries in X

- LSA-components represent space containing most of the variance. Is this the most useful model?

Cepstral features in music [Hansen, Feng]



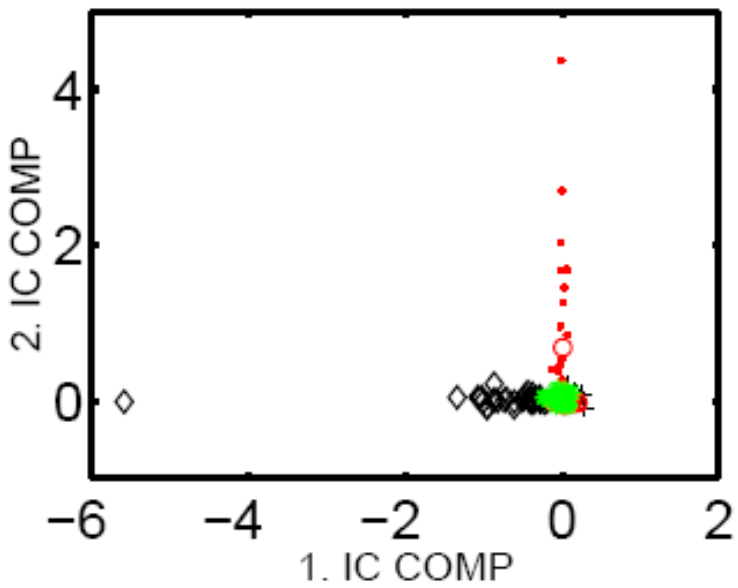
Cepstral Features of Phonemes [Hansen, Feng]

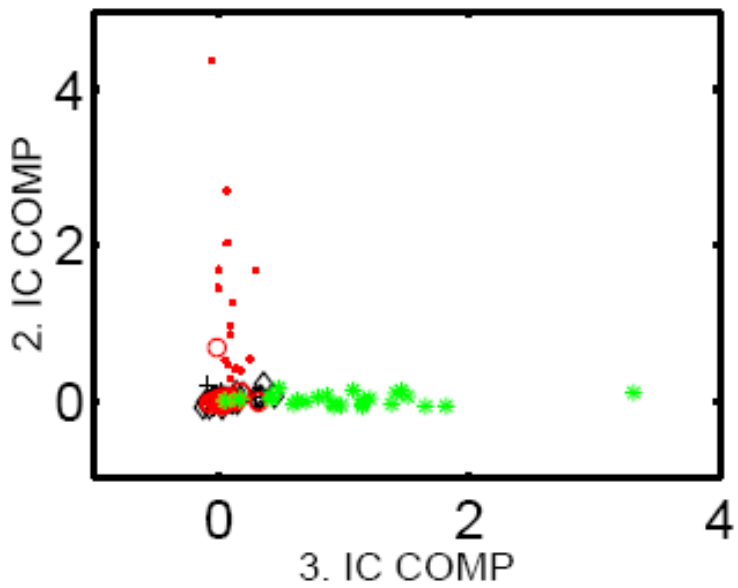


Related algorithms

- ICA – Independent Component Analysis [Kolenda et al.]
  - Components independent
- PLSA - Probabilistic LSA [Hoffman]
  - Similar to ICA
  - Feasible algorithms for optimisation.
- LDA – Latent Dirichlet Association [Blei,Ng,Jordan]
  - An extension of the PLSA model

ICA example





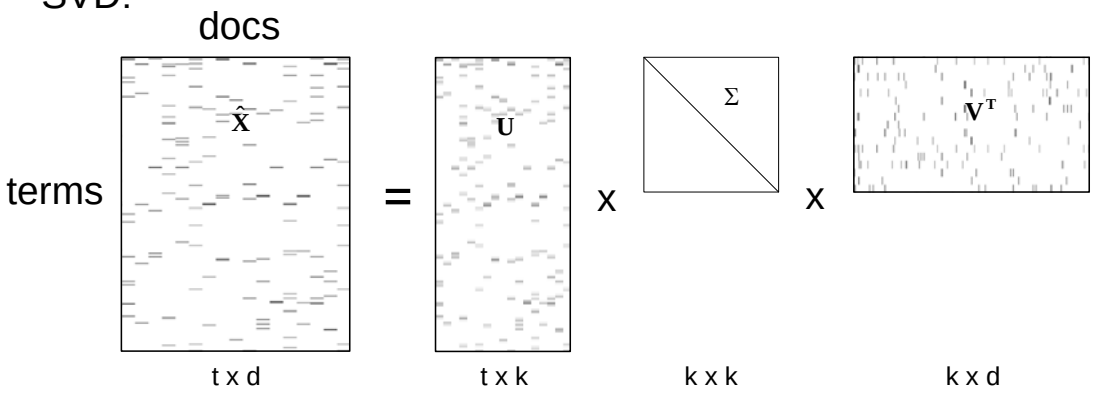
PLSI

- Probabilistic latent variable model
  - Generative model with an unobserved class variable k
  - Joint probability model over terms and documents

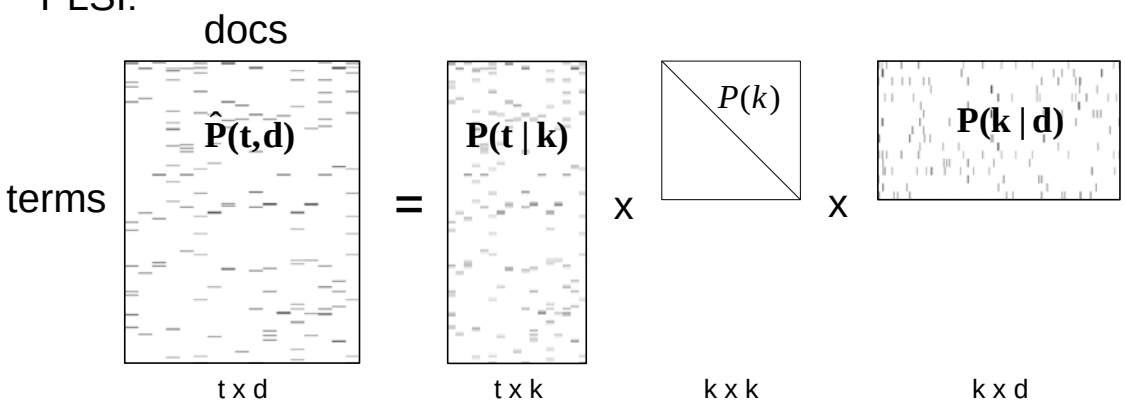
$$p(t,d) = \sum_{k=1}^{K} p(t|k)p(d|k)p(k)$$

  - Components are found using an iterative EM-algorithm
  - Analogous structure, but with the benefits of a probabilistic framework – similarities as probabilities vs. cosines.

SVD:



PLSI:

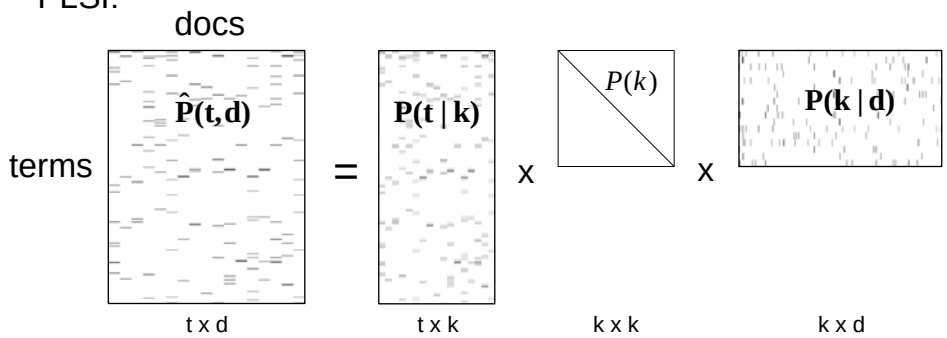


Same structure but different interpretation.

PLSI shortcomings

- EM algorithm requires full matrix reconstruction of $\hat{P}(t,d)$

PLSI:



- E-step

$$P(k|d,t) = \frac{P(k)P(d|k)P(t|k)}{\sum_{k' \in K} P(k')P(d|k')P(w|k')}$$

- M-step

$$\begin{aligned} P(t|k) &= \frac{\sum_d n(d,t)P(k|d,t)}{\sum_{d,t'} n(d,t')P(k|d,t')} \\ P(d|k) &= \frac{\sum_w n(d,t)P(k|d,t)}{\sum_{d',w} n(d',t)P(k|d',t)} \\ P(k) &= \frac{1}{R}\sum_{d,t} n(d,t)P(k|d,t); R \equiv \sum_{d,w} n(d,t) \end{aligned}$$

- Infeasible for large termdoc-matrices
  - 100.000 documents x 20.000 terms -> 15 GB td-matrix
- Model can be approximated using Nonnegative Matrix Factorisation

Query relevance

- Query: a new document d* with equal probability for terms

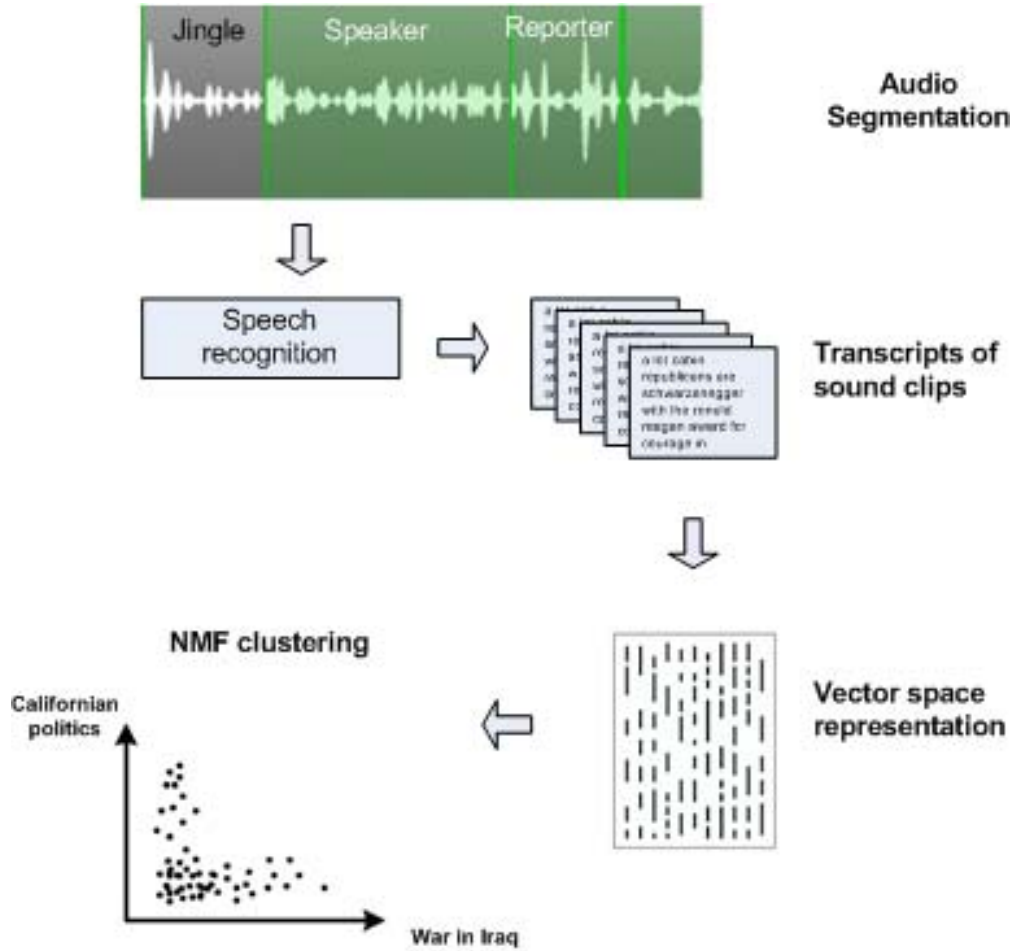
$$p(d|d^*) = \sum_{k=1}^{K} p(d|k)p(k|d^*)$$

$$p(d|k) = \mathbf{H}_{\mathbf{k,d}}$$

$$p(k|d*) = \sum_{t} p(k|t)p(t|d^*)$$

Castsearch

Castsearch.imm.dtu.dk



Context evaluations

- Data:
  - CNN hourly podcasts continuously retrieved.
    In the period: April 4th 2006 – September 9th 2006
  - 2099 CNN-News podcasts processed
- Vector space representation
  - 30977 speaker documents
  - Vocabulary of 37791 words (excluding stop words)

Text segmentation

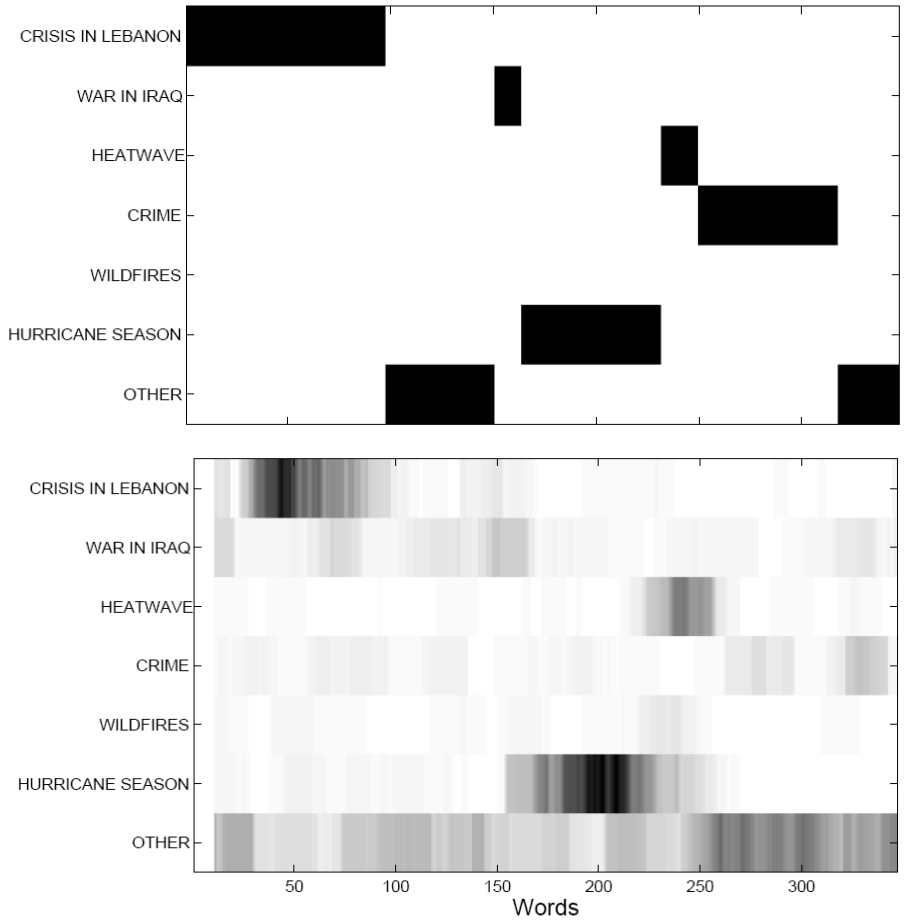
- Resegment text according to topics
- NMF-model:
  - Assign topic given pseudo-documents as query

d*

last month viacom demanded you to remove more than a hundred thousand unauthorized clips || attorney general **alberto gonzalez**

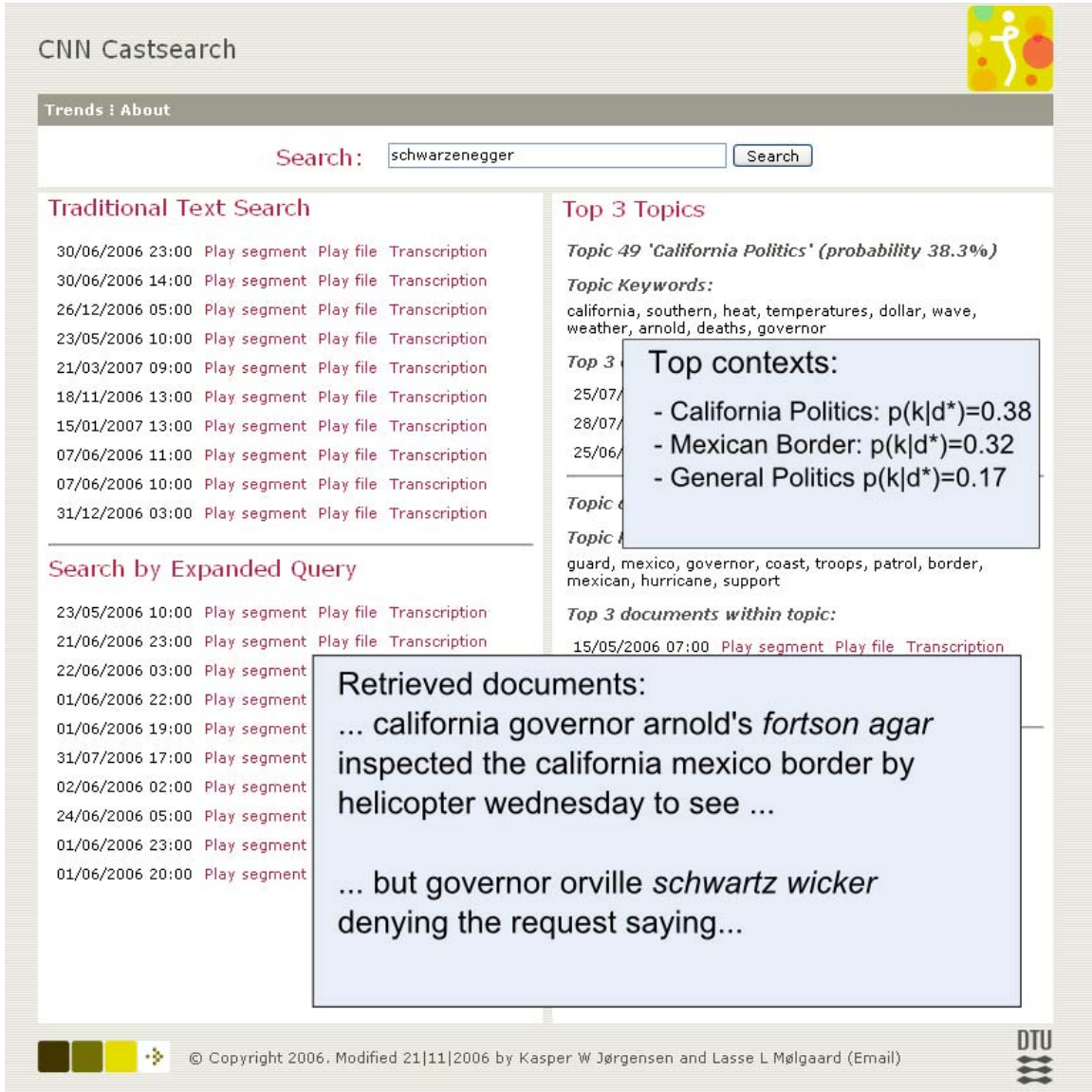
- Tests performed on a manually segmented subset of podcasts

Text segmentation example



Castsearch Demo

- Webdemo
  - Automatic retrieval of CNN podcasts and preprocessing
  - 70 contexts annotated manually
  - castsearch.imm.dtu.dk


CNN Castsearch
Trends : About
Search: schwarzenegger Search
Traditional Text Search
30/06/2006 23:00 Play segment Play file Transcription
30/06/2006 14:00 Play segment Play file Transcription
26/12/2006 05:00 Play segment Play file Transcription
23/05/2006 10:00 Play segment Play file Transcription
21/03/2007 09:00 Play segment Play file Transcription
18/11/2006 13:00 Play segment Play file Transcription
15/01/2007 13:00 Play segment Play file Transcription
07/06/2006 11:00 Play segment Play file Transcription
07/06/2006 10:00 Play segment Play file Transcription
31/12/2006 03:00 Play segment Play file Transcription
Search by Expanded Query
23/05/2006 10:00 Play segment Play file Transcription
21/06/2006 23:00 Play segment Play file Transcription
22/06/2006 03:00 Play segment
01/06/2006 22:00 Play segment
01/06/2006 19:00 Play segment
31/07/2006 17:00 Play segment
02/06/2006 02:00 Play segment
24/06/2006 05:00 Play segment
01/06/2006 23:00 Play segment
01/06/2006 20:00 Play segment
Top 3 Topics
Topic 49 'California Politics' (probability 38.3%)
Topic Keywords:
california, southern, heat, temperatures, dollar, wave, weather, arnold, deaths, governor
Top 3
25/07/
28/07/
25/06/
Topic
Topic
guard, mexico, governor, coast, troops, patrol, border, mexican, hurricane, support
Top 3 documents within topic:
15/05/2006 07:00 Play segment Play file Transcription
Top contexts:
- California Politics: p(k|d*)=0.38
- Mexican Border: p(k|d*)=0.32
- General Politics p(k|d*)=0.17
Retrieved documents:
... california governor arnold's fortson agar inspected the california mexico border by helicopter wednesday to see ...
... but governor orville schwartz wicker denying the request saying...
© Copyright 2006. Modified 21|11|2006 by Kasper W Jørgensen and Lasse L Mølgaard (Email)
DTU

References

- [Hansen,Feng]: Hansen, L. K., Feng, L., *Cogito Componentiter Ergo Sum*, ICA2006 - 6th International Conference on Independent Component Analysis and Blind Source Separation, pp. 446 - 453, 2006
- [Kolenda et al.] : Kolenda, T., Hansen, L. K., Sigurdsson, S., *Indepedent Components in Text*, Advances in Independent Component Analysis, pp. 229-250, 2000
- [Hoffman] : Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval,* SIGIR '99.
- [Blei,Ng,Jordan] : Blei, D., Ng, A., Jordan, M., Latent Dirichlet Allocation, *Journal of Machine Learning Research 3* (2003) 993-1022
- Castsearch: Mølgaard, L. L., Jørgensen, K.W., Hansen, L.K., Castsearch - Context Based Spoken Document Retrieval ICASSP, 2007