Bornholm Text analysis

Finn Årup Nielsen

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Email: fn@imm.dtu.dk
WWW: http://www.imm.dtu.dk/~fn

OVERVIEW

- Analysis types
- Vector space model
- Tokenization
- Removal of stop words
- Word stemming
- Term weighting
- "Analysis" types
  - Latent semantic indexing
  - Independent component analysis
  - Non-negative matrix factorization
  - Mixture models
- Statistical language processing
- Examples
  - Text classication
  - Authorship with N-grams
  - Topic spotting in chat room.
- Usage analysis

ANALYSIS TYPES

- Endogeneous
  - Term analysis, words and phrases: The "vector space model".
  - Formatting: Title, emphasized text.
- Exogeneous
  - Link analysis, hyperlinks, citations. Posible in scientific text, patents, laws, Internet.
  - Usage analysis.
    - amazon.com: "Customers who bought this book also bought"
    - ResearchIndex: "Users who viewed this document also viewed"

VECTOR SPACE MODEL

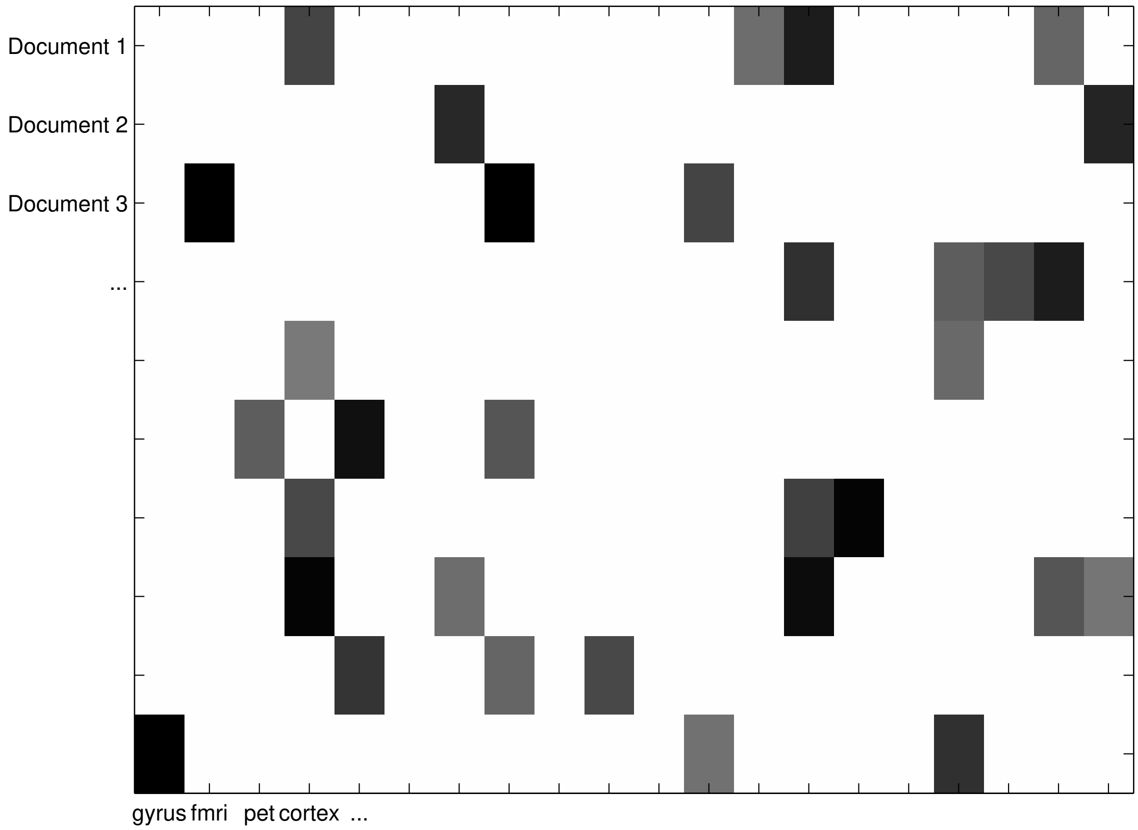

Figure 1: Bag-of-words matrix.

- *Vector space model* or "bag-of-words". $\mathbf{X}(N \times Q)$ with N documents and Q words/terms. Represented in hash array.
- A vector for each document containing the presence or frequency of words in the document.
- Historical (Luhn, 1958), "SMART" system (Salton, 1971).
- The ordering of words is not relevant.

TOKENIZATION

- Tokenization: Identification of terms/tokens.
- Corresponding to “lexical analysis” in compiler construction (Aho et al., 1986).
- Special tokens: Names, titles (journal article), telephone numbers. These tokens contain white spaces!
- Web search engines
  - Letters (case-insensitive) and digits as characters
  - Everything else: whitespace. Several whitespace is one whitespace.
  - Example with same search on Google

    ```
    "ISSN: 1053-8119"
    "ISSN       1053      8119"
    ```

    Zero is returned for

    ```
    "ISSN1053-8119"
    ```

WORDS AND PHRASES

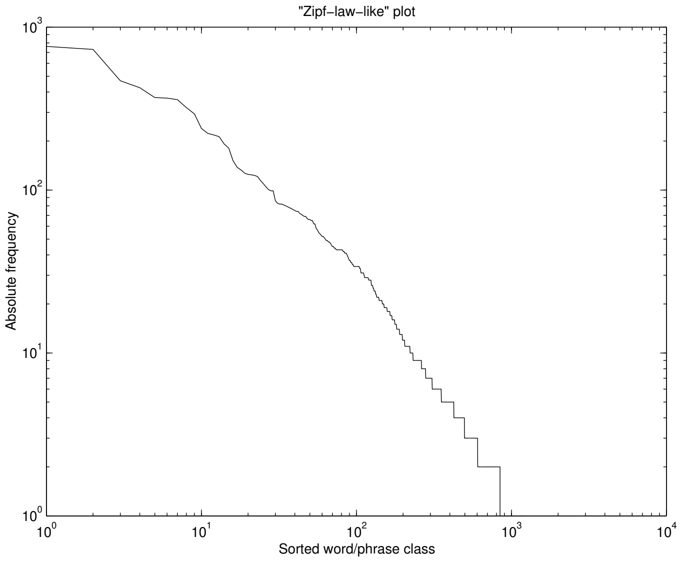

Figure 2: Example on distribution of words.

- N-grams, bigram, trigram: sequence of tokens.
- Character N-grams for automatic generation of indexing terms (Cohen, 1995).
- N-grams on characters for authorship (Kjell and Frieder, 1992).
- Multiword phrases ($N > 2$) can be regarded as a sequence of bigrams.

- Eliminate single word instances when comparing documents.

REMOVAL OF STOP WORDS

- Remove frequent words ("stop words"): the, and, of, ...
- For storage reasons
- Approximately 500 words in stop list.
- Should be domain specific, language, topic.
- Stop word list for Pubmed contains, e.g., "significantly" as a stop word.

  http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords.

- Implemented in web search engine such as Google? "Google ignores common words and characters such as 'where' and 'how' " (http://www.google.com/help/basics.html). However,

  ```
  "to be or not to be"
  ```

  does give Shakesperean links.

WORD STEMMING

- Removes non-important pre- and postfixes,
- Grammatical variations: talks, talked, talk $\rightarrow$ talk.
  - May depend on the language, e.g., little benefit in English (Cohen, 1995, p. 163).
- Names, reduction to surname + 2 initials, e.g., implemented in ISI and Pubmed.
- Journal titles vary a lot, e.g., "Human Brain Mapping"

```
204  Hum. Brain Mapp.
164  Hum. Brain Map.
101  Hum. Brain Mapping
 33  Human Brain Mapping
 19  Human Brain Map.
  7  Human Brain Mapp.
  3  Hum Brain Map.
  2  J. Hum. Brain Map.
  2  Hum. Brain. Mapp.
  2  Hum. Brain. Map.
  1  Hum. Brain. Mapping
  1  Hum. Brain Mapping (Suppl.)
  1  Hum. Brain Mapp
  1  Hum. Brain Map
```

  3-stemming: Convert the journal name to "hum bra map".

- Approximate string matching. Edit distance (Levenshtein, 1965). Google example: `mircosoft` $\rightarrow$ "Did you mean: microsoft".

TERM WEIGHTING

Functions	Description	Reference
Term frequency		
$a_{qn} = 1$	Binary, indicating if the term is present	
$a_{qn} = f_{qn}$	Raw frequency — term frequency (tf)	Luhn, 1957)
$a_{qn} = 0.5 + 0.5 f_{qn} / \max_q (f_{qn})$		
$a_{qn} = \log(f_{qn} + 1)$		Robertson and Sparck Jones, 1976)
$a_{qn} = \sqrt{f_{qn}}$		Cutting et al., 1992)
$a_{qn} = f_{qn}/(f_{qn} + \text{constant})$		Robertson and Walker, 1994)
Collection frequency component		
$b_q = 1$		
$b_q = \log(N/N_q)$	Inverse document frequency (idf)	Sparck Jones, 1972)
$b_q = \log[(N - N_q)/N_q]$	"Probabilistic inverse collection frequency"	
Normalization		
$c_n = 1$	No normalization	
$c_n = 1/\sqrt{\mathbf{x}_n^\mathsf{T}\mathbf{x}_n}$	Cosine normalization	

Table 1: Weighting functions for frequency-based vectors. f_{qn} is the (absolute) frequency of q'th term in the n'th document, N is the total number of documents and N_q is the number of document where the q'th term appear. Entries taken from (Sahami, 1999, section 2.1.2) and (Salton and Buckley, 1988, table 1).

- Combination of term frequency and inverse document frequency *tfidf*.

$$x_{nq} = f_{nq} \cdot \log(N/N_q) \tag{1}$$

Cosine normalization

$$\tilde{\mathbf{x}}_n = \mathbf{x}_n / \sqrt{\mathbf{x}_n^\mathsf{T}\mathbf{x}_n}. \tag{2}$$

ANALYSIS TYPES

- Unsupervized
  - Principal component analysis (PCA) / Singular value decomposition (SVD): "Latent semantic analysis".
  - Independent component analysis
  - Non-negative matrix factorization
  - Gaussian mixture model
- Supervized
  - Artificial neural network
  - Generalizable Gaussian mixture (GGM) models
  - ...
- Time consuming to label data.
  - Combining unsupervized and supervized algorithms.

LATENT SEMANTIC INDEXING

- Latent sematic indexing (Deerwester et al., 1989; Deerwester et al., 1990).
- Singular value decomposition of the document $\times$ word matrix
$$\mathbf{ULV}^{\mathsf{T}} = \mathsf{svd}\left[\mathbf{X}\right] \tag{3}$$
- Sparse sigular value decomposition.
- If one of the dimensions is small compute the eigen-vectors from the inner/outer product matrix.
- Models the covariance between the words
  - Handles synonyms and variations.

ICA

- Independent component analysis in text (Isbell Jr. and Viola, 1999; Kolenda and Hansen, 1999).
- Molgedey-Schuster ICA time-delay decorrelation (Molgedey and Schuster, 1994).

  Consider the autoregression from 0 to τ

  $$\mathbf{X}_\tau = \mathbf{X}\mathbf{W} + \mathbf{U}. \tag{4}$$

  $$\hat{\mathbf{W}} = \left[\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1} \mathbf{X}^\mathsf{T}\mathbf{X}_\tau. \tag{5}$$

  With $\mathbf{X} = \mathbf{Z}\mathbf{A}$ and independent sources assumption $\mathbf{Z}^\mathsf{T}\mathbf{Z} = \mathbf{\Lambda}$ the mixing matrix $\mathbf{A}$ is found as

  $$\mathsf{leig}[\hat{\mathbf{W}}] = \mathbf{A}^{-1}\mathbf{L}\mathbf{A}. \tag{6}$$

- Model for singular data (Hansen and Larsen, 1998)

NON-NEGATIVE MATRIX FACTORIZATION

- Factorization of a positive matrix $\mathbf{X}(N \times P)$ into two positive matrices where $\mathbf{W}(N \times K)$ and $\mathbf{H}(K \times P)$ into two matrices

$$\mathbf{WH} = \mathsf{nmf}(\mathbf{X}), \tag{7}$$

where K is the size of the hidden space (Lee and Seung, 1999; Lee and Seung, 2001).

- Sparse coding
- Update for for euclidean cost function

$$E_{\text{``eucl''}} = ||\mathbf{X} - \mathbf{Y}||_F^2 = \sum_{np} (x_{np} - y_{np})^2 \tag{8}$$

where the matrix $\mathbf{Y} = \mathbf{WH}$.

$$\mathbf{H}_{kp} \leftarrow \mathbf{H}_{kp} \frac{\left(\mathbf{W}^\mathsf{T}\mathbf{X}\right)_{kp}}{\left(\mathbf{W}^\mathsf{T}\mathbf{WH}\right)_{kp}} \tag{9}$$

$$\mathbf{W}_{nk} \leftarrow \mathbf{W}_{nk} \frac{\left(\mathbf{XH}^\mathsf{T}\right)_{nk}}{\left(\mathbf{WHH}^\mathsf{T}\right)_{nk}}. \tag{10}$$

MIXTURES MODELS

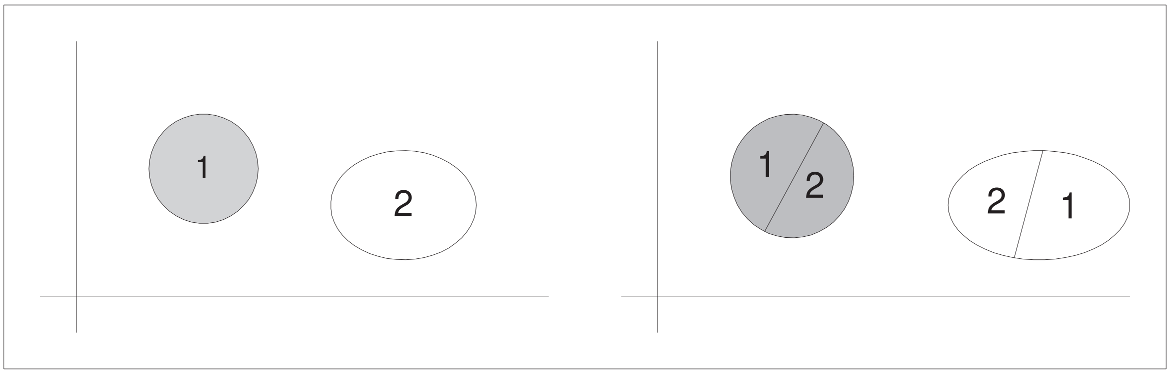

Figure 3: Unsupervized and supervized. From (Larsen, 2001, p. 26).

- Mixture model with k mixtures

$$p(\mathbf{x}) \approx \sum_k^K p(\mathbf{x}|k)P(k) \tag{11}$$

where $P(k)$ is a prior for the k'th component.

- Gaussian kernel

$$p(\mathbf{x}|k) = |2\pi\mathbf{\Sigma}_k|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\right]. \tag{12}$$

- Supervized, with classes y

$$p(\mathbf{x}|y) = \sum_k^K p(\mathbf{x}|k)\ P(k|y) \tag{13}$$

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_y p(\mathbf{x}|y)\ P(y)} \tag{14}$$

"GENERALIZABLE GAUSSIAN MIXTURES"

- Generalizable Gaussian Mixtures (GGM) (Hansen et al., 2000).

$$\boldsymbol{\mu}_0 \leftarrow N^{-1}\mathbf{X}^\mathsf{T}\mathbf{1}_N \tag{15}$$
$$\boldsymbol{\Sigma}_0 \leftarrow N^{-1}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}_0{}^\mathsf{T})^\mathsf{T}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}_0{}^\mathsf{T}) \tag{16}$$
$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{17}$$
$$\boldsymbol{\Sigma}_k \leftarrow \boldsymbol{\Sigma}_0 \tag{18}$$
$$P(k) \leftarrow 1/K. \tag{19}$$

The following steps are repeated until convergence

$$\mathcal{S}_\mu, \mathcal{S}_\Sigma \leftarrow \text{Split } n = 1..N \text{ randomly into two sets} \tag{20}$$
$$\mathcal{K}_k \leftarrow \{n : \max_k \, p(k|\mathbf{x}_n)\} \tag{21}$$
$$\boldsymbol{\mu}_k \leftarrow N_{k,\mu}^{-1} \sum_{n \in \mathcal{K}_k \wedge n \in \mathcal{S}_\mu} \mathbf{x}_n \tag{22}$$
$$\boldsymbol{\Sigma}_k \leftarrow \begin{cases} \sum_{n \in \mathcal{K}_k \wedge n \in \mathcal{S}_\Sigma} \frac{(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}}{N_{k,\Sigma}} & \text{if } N_k > P+1 \\ \boldsymbol{\Sigma}_0 & \text{if } N_k \leq P+1 \end{cases} \tag{23}$$
$$P(k) \leftarrow N_k/N \tag{24}$$

where $\mathcal{K}_k$ is the N_k sized set of assigned objects to the k'th component.

- Iteration over different Ks, e.g., $K = 1..N/2$ and the best model is selected via the AIC criteria.

STATISTICAL LANGUAGE PROCESSING

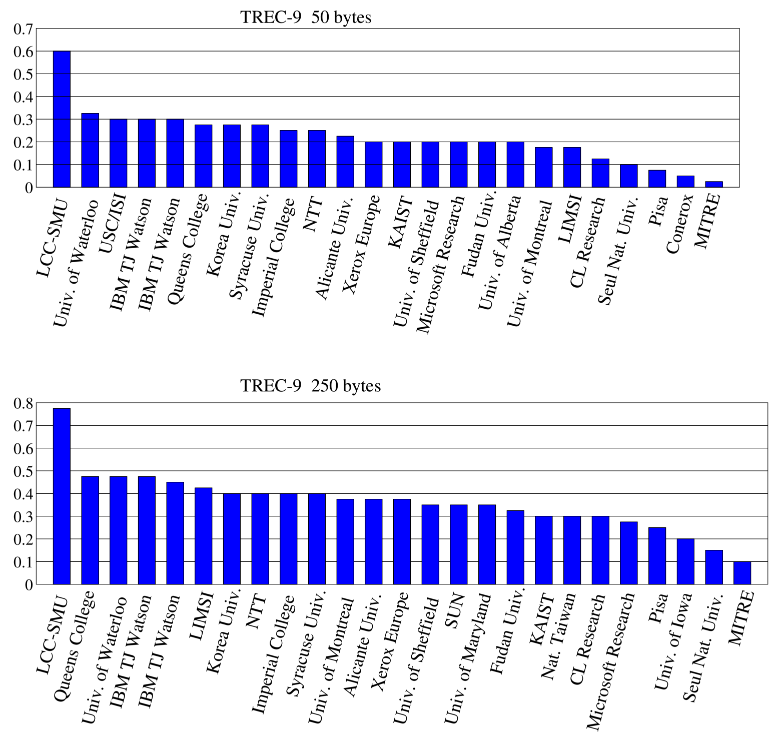

Figure 4: TREC-9 Q/A results. From (Pasca and Harabagiu, 2001, fig. 6).

- NIPS tutorial 2001 (Manning, 2001).
- Sophisticated natural language processing system (Pasca and Harabagiu, 2001) well ahead winner of TREC-9 question-answer comparison (Voorhees, 2000).
  - Statistical parser (Collins, 1996).
  - WordNet, http://www.cogsci.princeton.edu/~wn/, electronic lexical database with thesaurus and explanations (Miller et al., 1990). .

EXAMPLE ON TEXT CLASSIFICATION

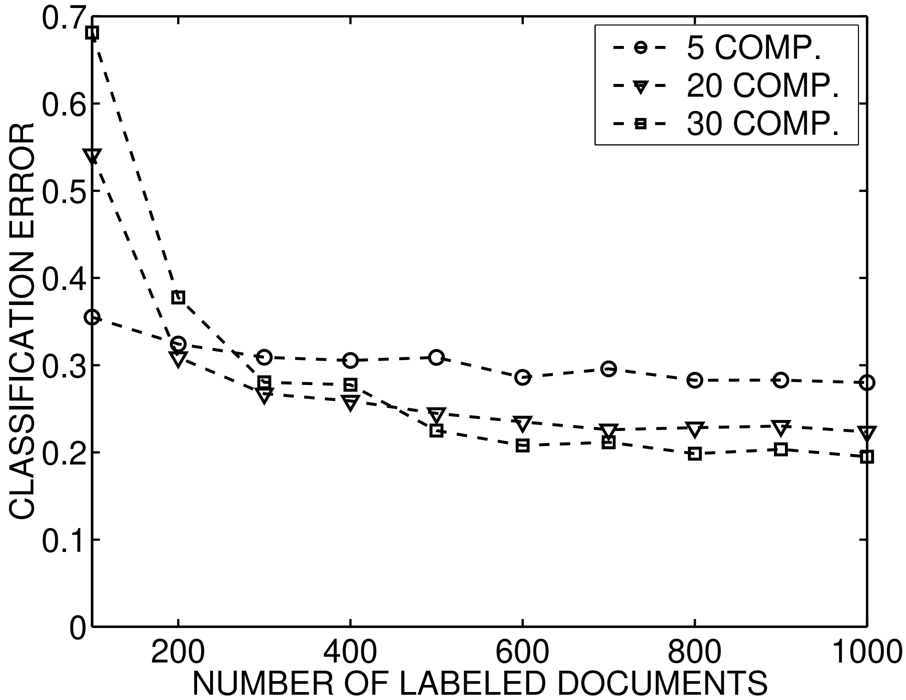

Figure 5: Learning curve for text classification. From (Hansen et al., 2000)

- Supervized classification of text (Hansen et al., 2000).
- Data from CMU WebKM repository, Manually labeled university data set: Course/Faculty/Project/Student. http://www.cs.cmu.edu/~textlearning/
- PCA to a subspace of dimension 5, 20, 30
- Generalizable Gaussian Mixture (GGM) model.
- Learning curve
- Novelty detection

EMAIL SPAM IDENTIFICATION

```
From nod93@post.com Sat May  6 00:35:00 MET 2000
Received: from serv1.imm.dtu.dk (root@serv1.imm.dtu.dk [130.225.85.9])
        by eivind.imm.dtu.dk (8.9.3 (PHNE_18979)/8.8.6) with ESMTP id AAA21493
        for <fnielsen@eivind.imm.dtu.dk>; Sat, 6 May 2000 00:34:59 +0200 (METDST)
From: nod93@post.com
Received: from ns.finfort.com (ns.finfort.com [195.184.199.129])
        by serv1.imm.dtu.dk (8.9.3 (PHNE_18979)/8.9.3) with ESMTP id AAA23245
        for <fn@imm.dtu.dk>; Sat, 6 May 2000 00:34:44 +0200 (METDST)
Received: from 172.137.178.75 (AC89B24B.ipt.aol.com [172.137.178.75])
        by ns.finfort.com (8.8.7/8.8.7) with SMTP id RAA24730;
        Sat, 6 May 2000 01:26:17 +0300
Message-ID: <000026006417$00002a92$00006cf8@>
To: <Undisclosed.Recipients@ns.finfort.com>
Subject: Financial Independence ..
Date: Fri, 05 May 2000 12:24:00 -0500
X-Priority: 3
X-MSMail-Priority: Normal
Reply-To: nod93@post.com
Status: RO
X-Status:
```

Figure 6: Spam email: Difference in return address and originating address.

- Term analysis in the email body. (Larsen et al., 2002; Szymkowiak et al., 2001): Keywords found "remove, call, address, free, day, business". Gaussian Mixture model used.
- Other information: Email header: Return address is often different than the the computer it has been sent from, investigate spamming domains.

AUTHORSHIP OF BIBLICAL TEXTS

- Determine authorship of chapters in *book of Daniel*, *Song of Songs* and *Ecclesiates* (Soboroff et al., 1997).
- Character N-grams (3-grams).
- LSI (SVD) of the chapter by N-grams matrix.

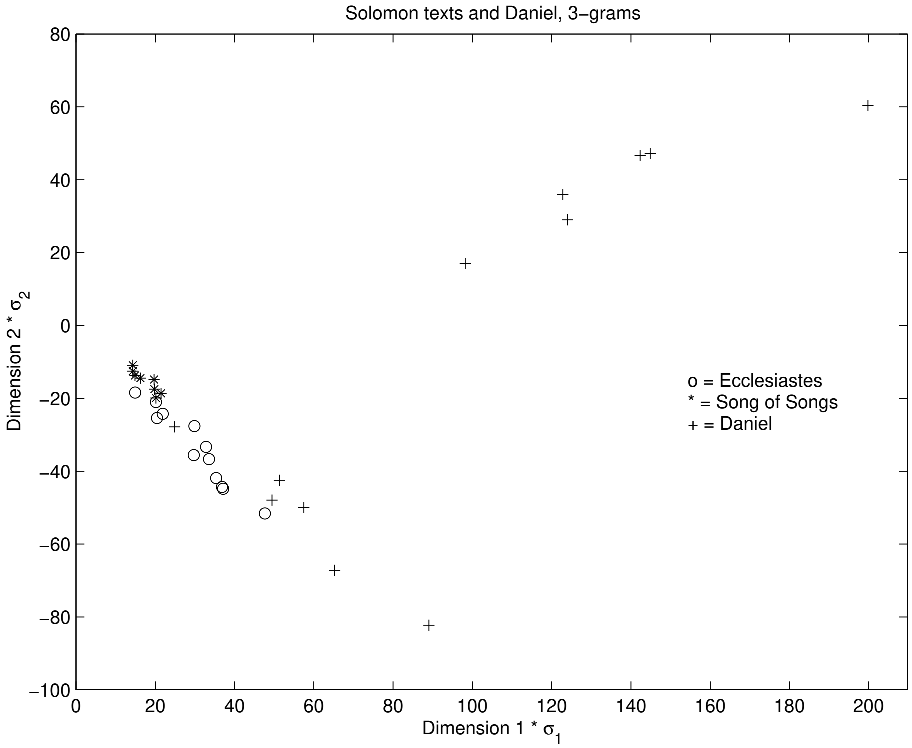

Figure 7: SVD plot of the two first dimensions. Extracted from (Soboroff et al., 1997, figure 1).

- Second dimension distinguishes between Hebrew and Aramaic chapters of *Daniel*.
- Other early determination of authorship (Cox and Brandwood, 1959): Sentence endings.

CHAT ROOM ANALYSIS

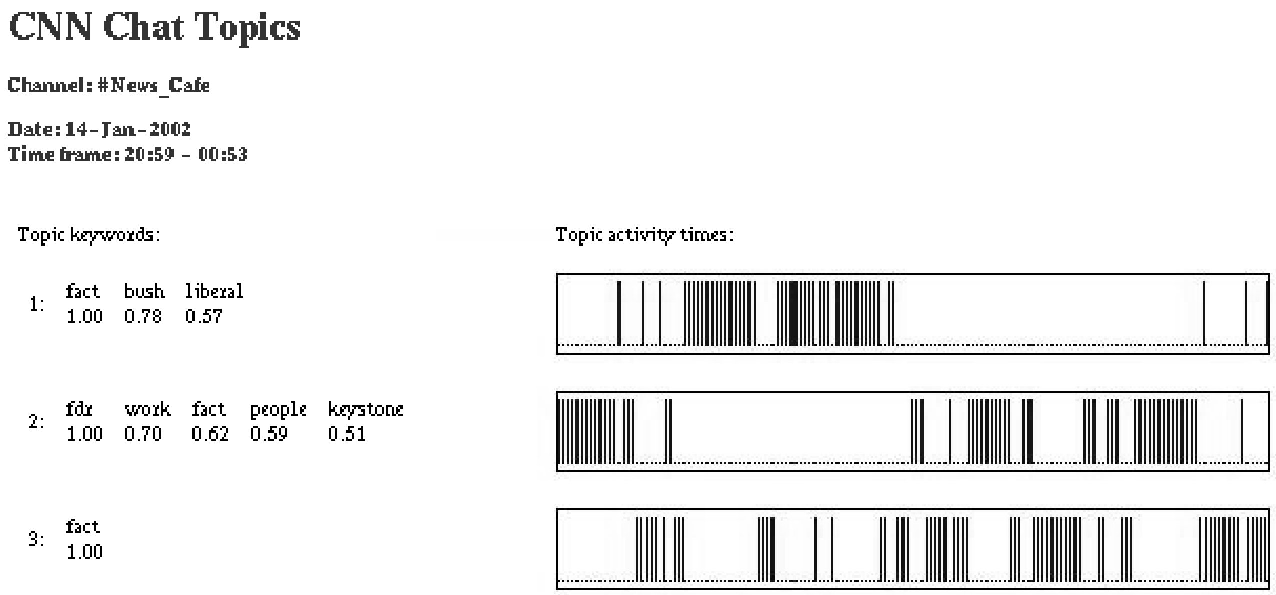

Figure 8: Chat room analysis from http://base.imm.dtu.dk/webchat/.

- "Topic spotting" in data from chat room on CNN.com (Kolenda et al., 2001).
- Sliding window over chat text converted to bag-of-words, e.g, 1114 pseudo-documents and 2498 terms.
- Singular independent component analysis with the Molgedey-Schuster dynamic decorrelation (Molgedey and Schuster, 1994).
- Model selection with probabilistic assumptions: Bayesian information criterion, isotropic gaussian noise, time correlated Gaussian distribution for source distribution.

USAGE ANALYSIS

- Server logs, buying behavior.
- On the web the user can be identified from: Cookie-mekanism, login, or CGI scripts (hidden fields).
- Commercial companies for online marketing, e.g., DoubleClick Inc., http://www.doubleclick.com/.
- Classifying e-customers based on web-server logs (Hørlyck et al., 2001).
- Comparing publisher expectation (through path length analysis) user behavior (page access co-occurrence) (Kato et al., 2000).
- (Page, 1997): A hyperlink analysis is better than usage analysis since it requires more effort to construct a hyperlink (than just clicking).

References

Aho, A. V., Sethi, R., and Ullman, J. D. (1986). *Compilers. Priciples, Techniques, and Tools*. Addison-Wesley series in Computer Science. Addison-Wesley Publishing Company, world student series edition edition. ISBN 0–201–10194–7, [defkat.dk — bibliotek.dk — amazon.com — bn.com]. "The Dragon Book".

Cohen, J. D. (1995). Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 46(3):162–174.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*.

Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *Journal of the Royal Statistical Society, Series B*, 21:195–200.

Cutting, D. R., Karger, D. R., Penderson, J. O., and Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *The 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. http://www.si.umich.edu/~furnas/POSTSCRIPTS/LSI.JASIS.paper.ps. ResearchIndex: http://citeseer.nj.nec.com/deerwester90indexing.html.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E., and Streeter, L. Q. (1989). Computer information retrieval using latent semantic structure. United States Patent No. 4,839,853.

Hansen, L. K. and Larsen, J. (1998). Source separation in short image sequences using delayed correlation. In Dalsgaard, P. and Jensen, S. H., editors, *Proceedings of NORSIG'98, Vigsø, Denmark*, pages 253–256.

Hansen, L. K., Sigurðsson, S., Kolenda, T., Nielsen, F. Å., Kjems, U., and Larsen, J. (2000). Modeling text with generalizable Gaussian mixtures. In *Proceedings of ICASSP'2000*. Turkey.

Hørlyck, J., Larsen, J., Hansen, L. K., and Christiansen, T. (2001). Are all e-comstumers alike? In *1st Nordic Workshop on Electronic Commerce*.

Isbell Jr., C. L. and Viola, P. (1999). Restructuring sparse high dimensional data for effective retrieval. In Kearns, M. S., Solla, S. A., and Cohen, D., editors, *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, pages 480–486, Cambridge, Massachusetts. MIT Press. NIPS-11.

Kato, H., Nakayama, T., and Yamane, Y. (2000). Navigation analysis tool based on the correlation between contets distribution and access patterns. In *Workshop on Web Mining for E-Commerce — Challenges and Opportunities (WEBKDD'2000)*. ACM.

Kjell, B. and Frieder, O. (1992). Visualization of literary style. In *IEEE International Conference on Systems, Man and Cybernetics*.

Kolenda, T. and Hansen, L. K. (1999). Independent components in text. Submitted to NIPS'99, but not accepted.

Kolenda, T., Hansen, L. K., and Larsen, J. (2001). Signal detection using ICA: Application to chat room topic spotting. In *ICA'2001*.

Larsen, J. (2001). Mining the web, a learning based approach. Tutorial at the course Statistical Methods for Datamining, Società Italiana di Statistica, Rome, Italy.

Larsen, J., Szymkowiak, A., and Hansen, L. K. (2002). Probabilistic hierarchical clustering with labeled and unlabeled data. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 6(1):56–62. http://eivind.imm.dtu.dk/publications/2001/larsen.kes.pdf.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization.

Levenshtein, V. I. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Russian Problemy Peredachi Informatsii*.

Luhn, H. P. (1957). A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Luhn, H. P. (1958). Automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165.

Manning, C. (2001). Probabilistic linguistics and probabilistic models of natural language processing. NIPS 2001 Tutorial.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235–213.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637. PMID: 10056251. http://www.theo-physik.uni-kiel.de/thesis/molgedey94.ps.gz. ResearchIndex: http://citeseer.nj.nec.com/molgedey94separation.html.

Page, L. (1997). PageRank: Bringing order to the web. Stanford Digital Libraries Working Paper 1997-0072, Stanford University.

Pasca, M. and Harabagiu, S. (2001). High-performance question/answering. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*.

Robertson, S. E. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science*, 27:129–146.

Robertson, S. E. and Walker, S. (1994). Some simple effective approximation to the 2-Poisson model for probabilistic weighted retrieval. In Croft, W. B. and van Rijsbergen, C. J., editors, *SIGIR 94, Proceedings of the Seventh International Conference on Research and Development in Information Retrieval*, pages 232–241. Springer–Verlag.

Sahami, M. (1999). *Using Machine Learning to Improve Information Access*. Ph.d. thesis, Computer Science Department, Stanford University, California. http://www.sims.berkeley.edu/courses/is296a-3/f98/readings/sonia.ps.

Salton, G. (1971). *The SMART retrieval system*. Prentice Hall sereis in Automatic Computation. Prentice-Hall, Englewood Cliffs, New Jersey.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Soboroff, I. M., Nicholas, C., and Kukla, J. M. (1997). Visualizing document authorship using N-grams and latent semantic indexing. In Ebert, D. S. and Nicholas, C. K., editors, *Workshop on New Paradigms in Information Visualization and Manipulation*, pages 43–48, New York, New York. The Association of Computing Machinery, ACM Press.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Szymkowiak, A., Larsen, J., and Hansen, L. K. (2001). Hierarchical clustering for datamining. In Babs, N., Jain, L. C., and Howlett, R. J., editors, *Proceedings of KES-2001 Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies*, pages 261–265.

Voorhees, E. M. (2000). Overview of the TREC-9 question answering track. In *he Ninth Text REtrieval Conference (TREC 9)*, NIST Special Publication 500-249, pages 71+. NIST.