JOHANNES KEPLER UNIVERSITY LINZ | JKU

Mixture Models in Text Mining

Bettina Grün

Universität Innsbruck, March 20 2012

This research is supported by the Austrian Science Fund (FWF): V170-N18.

Text Mining

- Text is the most common vehicle for the formal exchange of information.
- Learn meaningful information from natural language text.
- Availability of large electronic document collections requires automatic statistical tools for analyzing text.
  E.g., abstracts of scientific journals, mailing group archives, etc.
- However, text is highly unstructured.

Bag-of-Words Models

- In general a corpus containing different text documents is the input data.
- For bag-of-words models the text is represented as an unordered collection of words, disregarding grammar and word order.
  $\Rightarrow$ Exchangeability assumption.
- Each of the input documents is mapped to a frequency vector how often each of the words occured.
- Practical problems for data pre-processing:
  - How to define a “word”.
  - What to do with numbers.
  - . . .

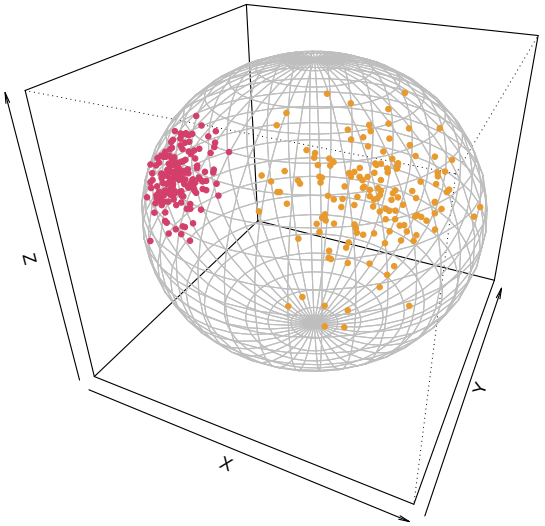
Document-Term Matrix (DTM)

- Contains the frequency of terms that occur in a collection of documents.
- The rows correspond to documents in the collection. The columns correspond to terms. All terms constitute the vocabulary.
- In general this will be a sparse matrix.
- Pre-Processing:
  - Convert characters to lower case
  - Remove punctuation
  - Tokenizing
  - Remove numbers
  - Stemming
  - Remove stopwords
  - Set minimum word length
  - Set minimum frequency in corpus / in document

Mixture Models for DTMs

- Finite mixtures of von Mises-Fisher distributions:
  - Each document is assumed to belong to one cluster.
  - Cluster membership does only depend on the relative frequencies of word occurences and not on the length of the document.
    $\Rightarrow$ Directional data.
- Topic models:
  - Co-occurences of words are explained using latent topics.
  - Mixed-membership model:
    - Each document is assumed to be a mixture of several topics.
    - These compositions of topics differ over documents.

Model-Based Clustering



von Mises-Fisher Distribution

- The von Mises-Fisher distribution has as support the $(p-1)$-dimensional sphere on $\mathbb{R}^p$.
- The density is given by

$$f(x;\theta) = C_p(\|\theta\|)\exp(\theta^\top x),$$

where the normalizing constant is equal to

$$C_p(\|\theta\|) = \frac{\|\theta\|^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\|\theta\|)}$$

and $I_\nu(\cdot)$ is the modified Bessel function of the first kind and order ν.

von Mises-Fisher Distribution / 2

- Another possible parameterization is

$$\mu = \frac{\theta}{\|\theta\|}, \qquad \kappa = \|\theta\|.$$

μ is denoted as the expected direction and κ as concentration.

- The von Mises-Fisher distribution can be obtained by conditioning a suitable p-dimensional normal distribution:

$$x \sim N(\mu, \kappa^{-1} I_p) \quad \text{with } \|\mu\| = 1$$

then

$$x | \|x\| = 1 \sim \text{vMF}_p(\mu, \kappa).$$

von Mises-Fisher Distribution / 3

- The maximum likelihood estimates of the parameters are given by

$$\hat{\mu} = \frac{r}{\|r\|} = \frac{\sum_{i=1}^{n} x_i}{\|\sum_{i=1}^{n} x_i\|},$$
$$\frac{I_{p/2}(\hat{\kappa})}{I_{p/2-1}(\hat{\kappa})} = \frac{\|r\|}{n} = \bar{r}.$$

- The estimation of κ is analytically not possible and needs to be determined using numeric or asymptotic methods.
- Different variants are for example proposed in Banerjee et al. (2005), Tanabe et al. (2007) and Sra (2011).

Extension to Mixture Models

$$h(x|\vartheta) = \sum_{k=1}^{K} \pi_k f_{\text{vMF}}(x|\theta_k),$$

with $\pi_k > 0$ for all k and $\sum_{k=1}^{K} \pi_k = 1$.

- For each of the components von Mises-Fisher distributions are assumed.
- Estimation using the EM algorithm: The parameters in the M-step are determined using weighted ML methods.
- Equivalent to spherical k-means, where the distance measure is 1 minus the cosine between the vectors, if
  - the component sizes are restricted to be equal and
  - the concentration parameters are all fixed at infinity.

Topic Models

- Generative model which provides a probabilistic framework for the term frequency occurrences in documents in a given corpus.
- The observed interdependences between occurences of words in documents are explained using latent topics.
- Each document consists of several topics:
  - Document-specific topic distribution
  - Mixed-membership model
- Each topic contains the terms with certain frequencies:
  - Topic-specific term distribution

Latent Dirichlet Allocation (LDA)

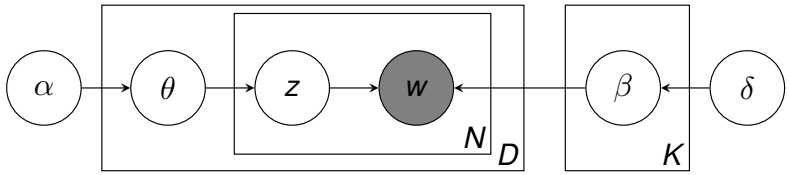
1. Draw the term distribution for the K topics

$$\beta_k \sim \text{Dirichlet}(\delta).$$

2. For each document w with N words:
   1. Draw the proportions of the topic distribution

$$\theta \sim \text{Dirichlet}(\alpha).$$

   2. For each of the N words:
      1. Choose a topic $z_i \sim \text{Multinomial}(\theta)$.
      2. Choose a word $w_i \sim \text{Multinomial}(\beta_{z_i})$.



Dirichlet Distribution

- The support of the K-dimensional Dirichlet distribution is the open $(K-1)$-dimensional simplex.
- The density is given by

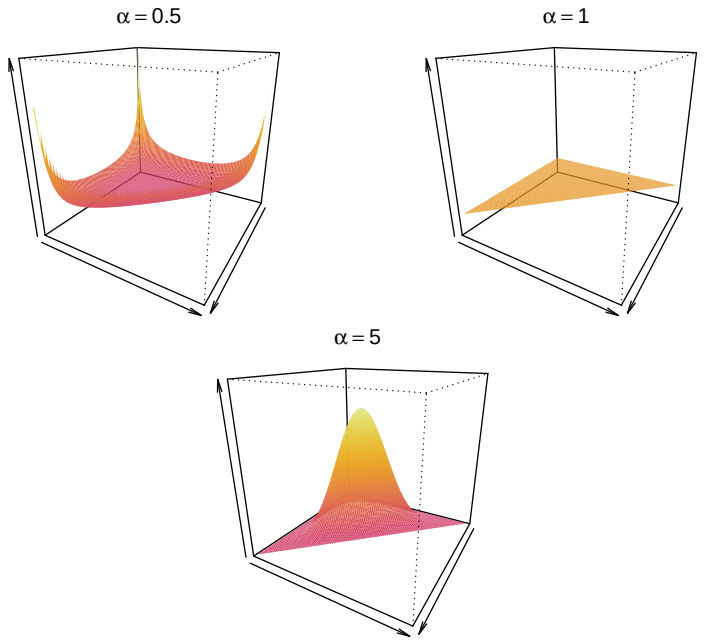
$$f(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $x_i > 0$ and $\alpha_i > 0$ for all i and $\sum_{i=1}^{K} x_i = 1$ and

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}.$$

- The Dirichlet distribution is the conjugate prior of the multinomial distribution.
- For LDA symmetric distributions are used, i.e., $\alpha_i = \alpha$ for all i.

Dirichlet Distribution / 2



Estimation

- From a collection of documents infer
  - topic assignments z for each word,
  - topic distributions θ for each document,
  - term distributions β for each topic.
- Standard estimation methods:
  - Variational Expectation-Maximization (VEM; Blei, Ng, and Jordan, 2003)
  - Gibbs sampling (Griffiths and Steyvers, 2004)

Estimation: Variational Expectation-Maximization

- The log-likelihood is maximized with respect to the parameters α and β of the Dirichlet priors for the topic distributions of the documents and the term distributions for the topics.
- For one document the log-likelihood is given by

$$\begin{aligned}\ell(\alpha, \beta) &= \log\left(p(w|\alpha, \beta)\right) \\ &= \log \int \left\{ \sum_z \left[\prod_{i=1}^{N} p(w_i|z_i, \beta) p(z_i|\theta) \right] \right\} p(\theta|\alpha) d\theta.\end{aligned}$$

- For application of the EM algorithm:
  Missing data are the topic distributions θ for each document and the topic assignments z of each word.

Estimation: Variational Expectation-Maximization / 2

- The posterior $p(\theta, z|w, \alpha, \beta)$ in the E-step is computationally intractable.
  $\Rightarrow$ The posterior is replaced by a variational distribution $q(\theta, z|\gamma, \phi)$ with document specific parameters γ and ϕ.
- The variational distribution is set equal to

$$q(\theta, z|\gamma, \phi) = q_1(\theta|\gamma) \prod_{i=1}^{N} q_2(z_i|\phi_i),$$

  where
  - $q_1()$ is a Dirichlet distribution with parameters γ.
  - $q_2()$ is a multinomial distribution with parameters ϕ_i.
- The parameters are determined by minimizing the Kullback-Leibler divergence of the variational posterior and the true posterior.

$$\mathrm{D_{KL}}(q(\theta, z|\gamma, \phi)\|p(\theta, z|w, \alpha, \beta)) = \mathrm{E}_q \left[\log \frac{q(\theta, z|\gamma, \phi)}{p(\theta, z|w, \alpha, \beta)}\right]$$

Estimation: Variational Expectation-Maximization / 3

- Using Jensen's inequality implies

$$\begin{aligned}
\log p(w|\alpha, \beta) &= \log \int \sum_z p(\theta, z, w|\alpha, \beta) d\theta \\
&= \log \int \sum_z \frac{p(\theta, z, w|\alpha, \beta) q(\theta, z)}{q(\theta, z)} d\theta \\
&\geq \int \sum_z q(\theta, z) \log p(\theta, z, w|\alpha, \beta) - q(\theta, z) \log q(\theta, z) d\theta \\
&= \mathrm{E}_q[\log p(\theta, z, w|\alpha, \beta)] - \mathrm{E}_q[\log q(\theta, z)].
\end{aligned}$$

- The difference between the left- and the right-hand side is equal to the Kullback-Leibler divergence of the posteriors.

Estimation: Variational Expectation-Maximization / 4

- The VEM algorithm iterates between

  E-step: For each document find the optimal values of the variational parameters $\{\gamma, \phi\}$ for the LDA model.

  M-step: Maximize the resulting lower bound on the log-likelihood with respect to the model parameters α and β for the LDA model.
- Each iteration of the VEM algorithm is guaranteed to increase the lower bound on the log-likelihood of the observed data.
- The iterations are terminated when the lower bound converges to a local maximum.
- Inference of θ and z uses the variational parameters based on the assumption that the variational posterior probability is a good approximation of the true posterior probabilities.

Estimation: Gibbs Sampling

- A Dirichlet prior with parameter δ is assumed for the term distributions of the topics

$$\beta \sim \text{Dirichlet}(\delta).$$

- The hyperparameters for the Dirichlet priors α and δ are fixed a-priori.
- Griffiths und Steyvers (2004) recommend to use

$$\alpha = \frac{50}{K}, \qquad \delta = 0.1.$$

Estimation: Gibbs Sampling / 2

- Draws from the posterior distribution $p(z|w)$ are obtained by sampling from

$$p(z_i = k|w, z_{-i}) \propto \frac{n^{(j)}_{-i,k} + \delta}{n^{(.)}_{-i,k} + V\delta} \frac{n^{(d_i)}_{-i,k} + \alpha}{n^{(d_i)}_{-i,.} + K\alpha},$$

  where
  - V denotes the vocabulary size.
  - j indicates that w_i is equal to the jth word in the vocabular.
  - $n^{(j)}_{-i,k}$ gives how often the jth word of the vocabulary is currently assigned to topic k without the ith word.
  - The . implies that summation over this index is performed.
  - d_i indicates the document in the corpus to which word w_i belongs.

Estimation: Gibbs Sampling / 3

- The predictive distributions of β and θ given w and z are given by

$$\hat{\beta}^{(j)}_k = \frac{n^{(j)}_k + \delta}{n^{(.)}_k + V\delta},$$
$$\hat{\theta}^{(d)}_k = \frac{n^{(d)}_k + \alpha}{n^{(.)}_k + K\alpha},$$

  for $j = 1, \ldots, V$, $d = 1, \ldots, D$ and $k = 1, \ldots, K$.
- For Gibbs sampling the log-likelihood is given by

$$\log(p(w|z)) = K \log\left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V}\right) + \sum_{k=1}^{K}\left\{\left[\sum_{j=1}^{V}\log(\Gamma(n^{(j)}_k + \delta))\right] - \log(\Gamma(n^{(.)}_k + V\delta))\right\}.$$

Model Selection

- Perplexity:
  - Measure to evaluate the model on a test dataset.
  - Equivalent to the inverse of the geometric mean of the per-word likelihood.
  - If the model is estimated using Gibbs sampling, the topic distributions for new documents are either
    - estimated using Gibbs sampling where the term distributions of the topics are fixed or
    - the a-priori distribution is used.
- Marginal likelihoods if the model is estimated using Gibbs sampling.

Model Extensions and Variants

- Modify model assumptions:
  - Correlated Topic Model (CTM):
    - Relax the assumption that topics are uncorrelated.
    - Replace the Dirichlet prior for the topic distributions with a transformation of a multivariate normal distribution.
- Include additional information:
  - Relational Topic Model (RTM; Chang and Blei, 2009)
    - Links between documents such as for example citations are available.
  - Author-Topic Models (Rosen-Zvi et al., 2004):
    - For each document author information is available.

Tools in R: Preprocessing

- Package **tm** for data pre-processing (Feinerer, Hornik and Meyer, 2008; Feinerer, 2010):
  - Reads in text from different sources and constructs a corpus.
  - Transforms the corpus to a document-term matrix with data-preprocessing.
  - The document-term matrix is stored in a sparse format (simple triplet representation from package **slam**; Hornik, Meyer and Buchta, 2010).
- Package **OAIHarvester** (Hornik, 2010) allows to harvest metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 2.0.

Tools in R: Modelling

- Package **movMF** (Hornik and Grün, 2011) for drawing from and fitting finite mixtures of von Mises-Fisher distributions.
- Package **lda** (Chang, 2010) fits the LDA model and related models using collapsed Gibbs sampling.
- Package **topicmodels** (Grün and Hornik, 2011):
  - Allows to use different estimation methods through interfaces to
    - C code by David M. Blei et al., the original authors of the LDA model and the CTM.
    - C++ code for fitting the LDA model using Gibbs sampling by Xuan-Hieu Phan.
  - Builds on and extends functionality from package **tm** (Feinerer et al., 2008; Feinerer, 2011) for text mining.

Application: JSS abstracts

- Abstracts of papers published in *Journal of Statistical Software*
- Data downloaded using package **OAIHarvester**
- Already available as package **corpus.JSS.papers** at the repository `http://datacube.wu.ac.at/`

Application: JSS abstracts / 2

```
> install.packages("corpus.JSS.papers",
+                  repos = "http://datacube.wu.ac.at/",
+                  type = "source")

> data("JSS_papers", package = "corpus.JSS.papers")
> JSS_papers <-
+   JSS_papers[JSS_papers[,"date"] < "2010-08-05",]
> JSS_papers <-
+   JSS_papers[sapply(JSS_papers[, "description"],
+                     Encoding) == "unknown",]

> JSS_abstracts <- JSS_papers[, "description"]
```

Application: Build DTM

```
> library("tm")
> corpus <- Corpus(VectorSource(JSS_abstracts))
> JSS_dtm <- DocumentTermMatrix(corpus,
+    control = list(stemming = TRUE, stopwords = TRUE,
+      minWordLength = 3, removeNumbers = TRUE,
+      removePunctuation = TRUE))
> dim(JSS_dtm)
```

```
[1]  348 4217
```

Application: Build DTM / 2

- Determine the importance of a word in the corpus using **tf-idf** which is the product of
  - **tf**, the term frequency, and
  - **idf**, the inverse document frequency.

```
> library("slam")
> term_tfidf <-
+   tapply(JSS_dtm$v/row_sums(JSS_dtm)[JSS_dtm$i],
+          JSS_dtm$j, mean) *
+     log2(nDocs(JSS_dtm)/col_sums(JSS_dtm > 0))
> summary(term_tfidf)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02323 0.08443 0.11110 0.13660 0.15350 1.53700
```

```
> JSS_dtm <- JSS_dtm[,term_tfidf >= 0.1]
> JSS_dtm <- JSS_dtm[row_sums(JSS_dtm) > 0,]
> dim(JSS_dtm)
```

```
[1]  348 2465
```

Application: movMF

```
> set.seed(201107)
> library("movMF")
> jss_mix <- movMF(JSS_dtm, k = 30,
+                  control = list(nruns = 20))
```

Application: movMF / 2

- Which papers are in the cluster with the highest concentration?

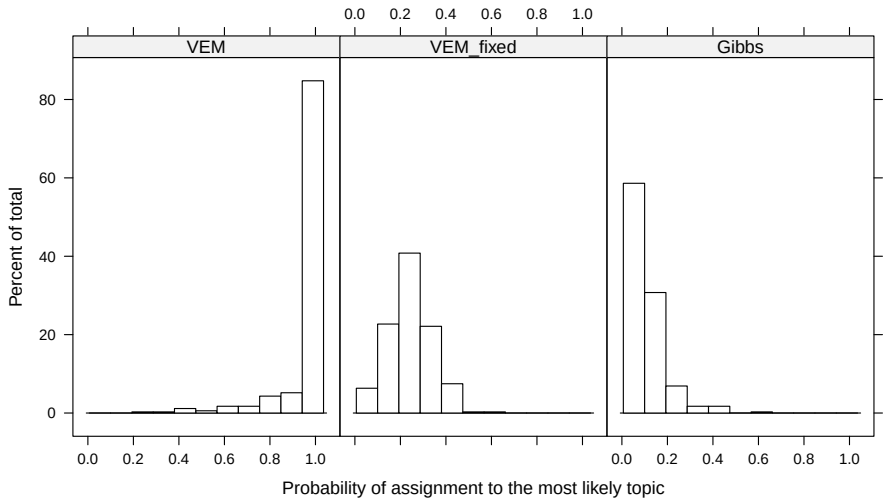
```
> kappa <- skmeans:::row_norms(jss_mix$theta)
> p <- which(max.col(jss_mix$P) %in% which(kappa >= max(kappa)))
> unlist(JSS_papers[p, "title"])
```

```
XLISP-Stat Tools for Building Generalised Estimating Equation
Models
Clustering in an Object-Oriented Environment
MIXNO: a computer program for mixed-effects nominal logistic
regression
A CLUE for CLUster Ensembles
The R Package geepack for Generalized Estimating Equations
Model-based Methods of Classification: Using the mclust
Software in Chemometrics
clValid: An R Package for Cluster Validation
Fitting Latent Cluster Models for Networks with latentnet
rEMM: Extensible Markov Model for Data Stream Clustering in R
```

Application: Topic Models

```
> k <- 30
> SEED <- 2010
> jss_TM <-
+   list(VEM = LDA(JSS_dtm, k = k,
+          control = list(seed = SEED)),
+        VEM_fixed = LDA(JSS_dtm, k = k,
+          control = list(estimate.alpha = FALSE,
+            seed = SEED)),
+        Gibbs = LDA(JSS_dtm, k = k, method = "Gibbs",
+          control = list(seed = SEED, burnin = 1000,
+            thin = 100, iter = 1000)))
```

Application: Topic Models / 2



Application: Topic Models / 3

- The estimated topics for a document and estimated terms for a topic can be obtained using the convenience functions `topics()` and `terms()`. The most likely topic for each document is obtained by

```
> Topic <- topics(jss_TM[["VEM"]], 1)
```

- The five most frequent terms for each topic are obtained by

```
> Terms <- terms(jss_TM[["VEM"]], 5)
> Terms[,1:4]
```

```
     Topic 1  Topic 2      Topic 3       Topic 4
[1,] "popul"  "correl"     "multivari"   "vista"
[2,] "captur" "gee"        "subset"      "loglinear"
[3,] "cell"   "qls"        "aspect"      "condit"
[4,] "anim"   "multilevel" "autoregress" "lispstat"
[5,] "interv" "bias"       "criterion"   "project"
```

Application: Topic Models / 4

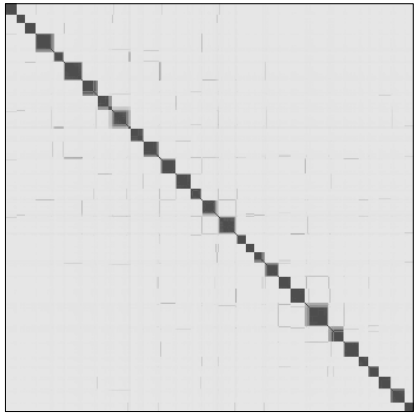
- Have a look at the most likely topics for papers published in Volume 24, which is a Special Issue on “Statistical Modeling of Social Networks with ‘statnet’”.

```
> p_v24 <- grep("/v24/", JSS_papers[, "identifier"])
> avg <- colMeans(posterior(jss_TM[["VEM"]])$topics[p_v24,])
> topics_v24 <- which.max(avg)
> most_frequent_v24 <- which.max(tabulate(topics_v24))
> terms(jss_TM[["VEM"]], 10)[, most_frequent_v24]
```

```
 [1] "network"  "mathemat" "text"     "ratio"    "learn"
 [6] "statnet"  "condit"   "social"   "bar"      "creat"
```

Application: Topic Models / 5

- Hellinger distance matrix of the abstracts using their topic distributions.



Summary & Outlook

- Basic infrastructure for text mining is provided in R by package **tm**.
- Packages providing specific methods to analyze this kind of data are for example **movMF** and **topicmodels**.
- Future work:
  - Apply methods to different datasets.
  - Extend methods to allow for certain data situations:
    - Is a general, extensible framework possible?
  - Modify the models to reduce the number of parameters, e.g., by imposing sparsity assumptions.
  - Develop methods for analyzing and visualizing the results.

References

Banerjee A., Dhillon I.S., Ghosh J., Sra S. (2005). Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6, 1345−1382.

Blei D.M., Lafferty J.D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1(1):17−35.

Blei D.M., Lafferty J.D. (2009). Topic Models. In A. Srivastava, M. Sahami (eds.), *Text Mining: Classification, Clustering, and Applications*, Chapman & Hall/CRC Press.

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

Chang J. (2010). **lda**: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.2.3.

Chang J., Blei D.M. (2009). Relational Topic Models for Document Networks. In AISTATS '09: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, pp. 81−88.

Feinerer I. (2011). **tm**: Text Mining Package. R package version 0.5-6.

References / 2

Feinerer I., Hornik K., Meyer D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5):1−54.

Griffiths T.L., Steyvers M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228−5235.

Grün B., Hornik K. (2011). **topicmodels**: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1−30.

Grün B., Hornik K. (2011). **topicmodels**: Topic Models. R package version 0.1-0.

Hornik K. (2011). **OAIHarvester**: Harvest Metadata Using OAI-PMH v2.0. R package version 0.1-3.

Hornik K., Grün B. (2011). **movMF**: Mixtures of von Mises-Fisher Distributions. R package version 0.0-0.

Hornik K., Meyer D., Buchta C. (2011). **slam**: Sparse Lightweight Arrays and Matrices. R package version 0.1-22.

References / 3

Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. (2004). The Author-Topic Model for Authors and Documents. In UAI'04 Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 487−494.

Steyvers M., Griffiths T. (2007). Probabilistic Topic Models. In T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (eds.), Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates.

Sra S. (2011). A Short Note on Parameter Approximation for von Mises-Fisher Distributions: and a Fast Implementation of $I_s(x)$. *Computational Statistics*. Accepted for publication.

Tanabe A., Fukumizu K., Oba S., Takenouchi T., Ishii S. (2007). Parameter Estimation for von Mises-Fisher Distributions. *Computational Statistics*, 22, 145−157.

Witten I.A. (2005). Text Mining. In M.P. Singh (ed.), The Practical Handbook of Internet Computing, pp. 14-1−14-22. Chapman & Hall/CRC Press.