概率主题模型及其应用

易磊,李晓东

大连理工大学软件学院

May 24, 2012

Information overload

- As more information becomes available, it becomes more difficult to find and discover what we need.

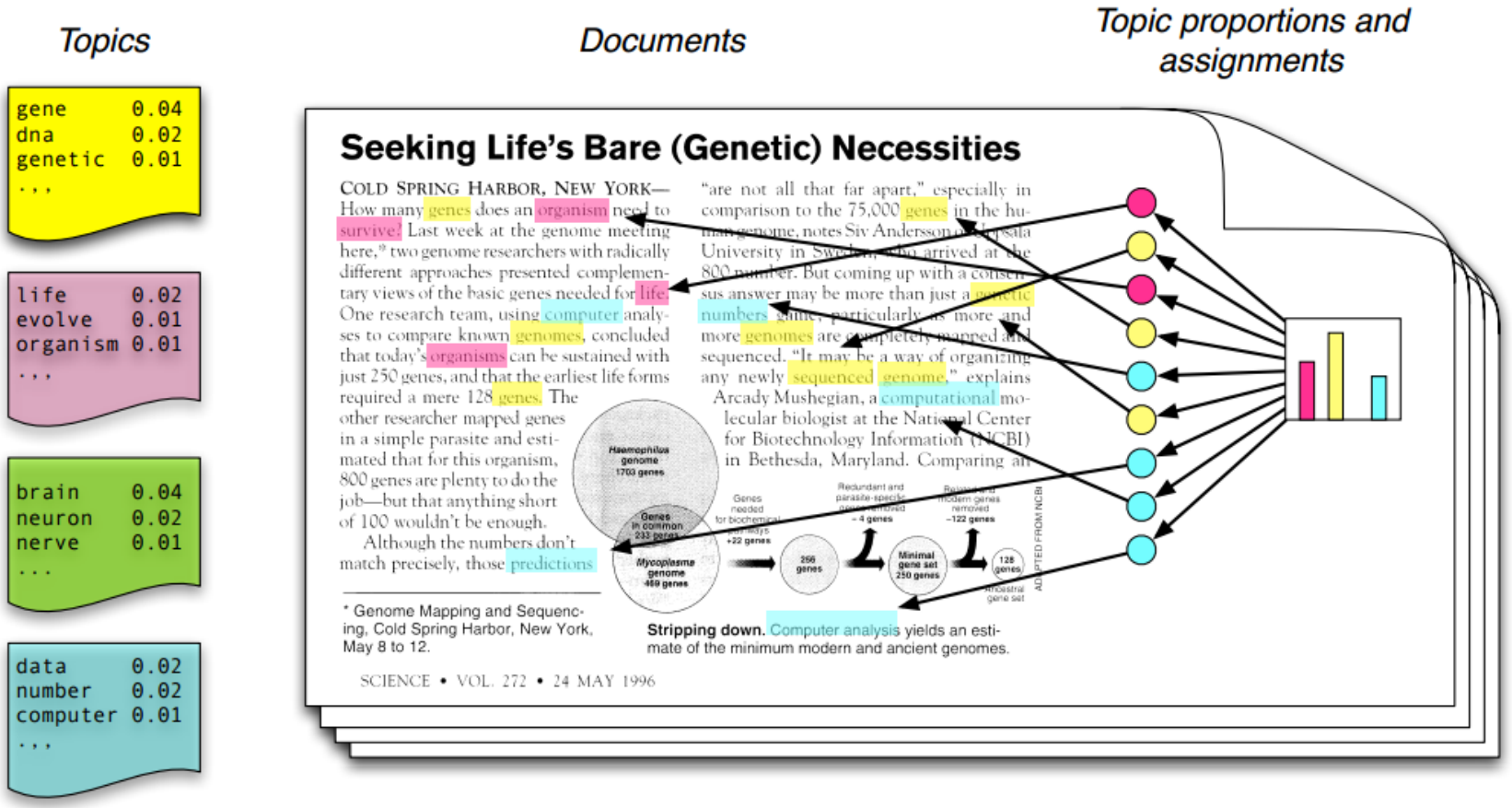
Information overload

- As more information becomes available, it becomes more difficult to find and discover what we need.
- We need new tools to help us organize, search, and understand these vast amounts of information.

Information overload

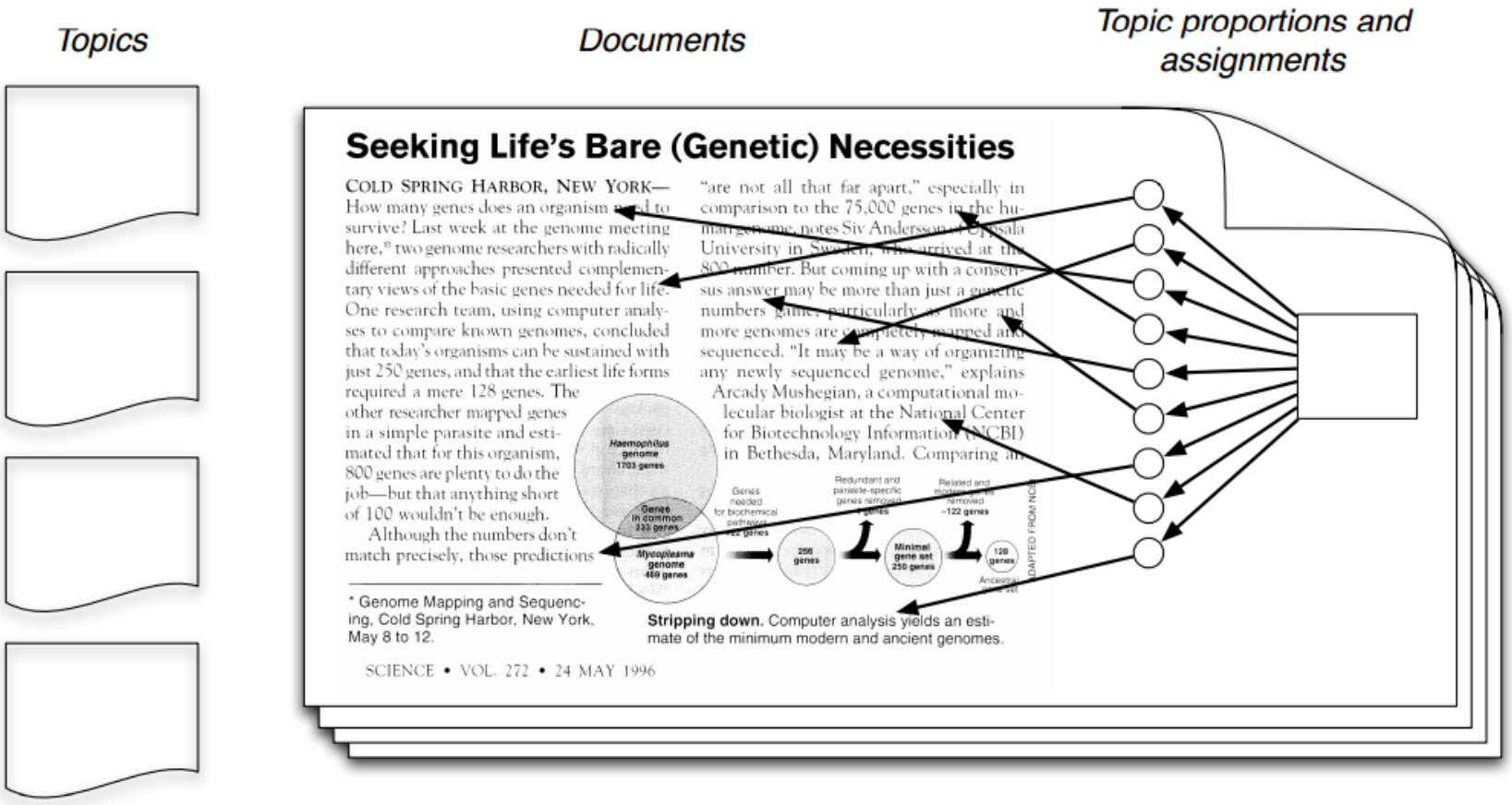
- As more information becomes available, it becomes more difficult to find and discover what we need.
- We need new tools to help us organize, search, and understand these vast amounts of information.
- Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

Latent Dirichilet allocation



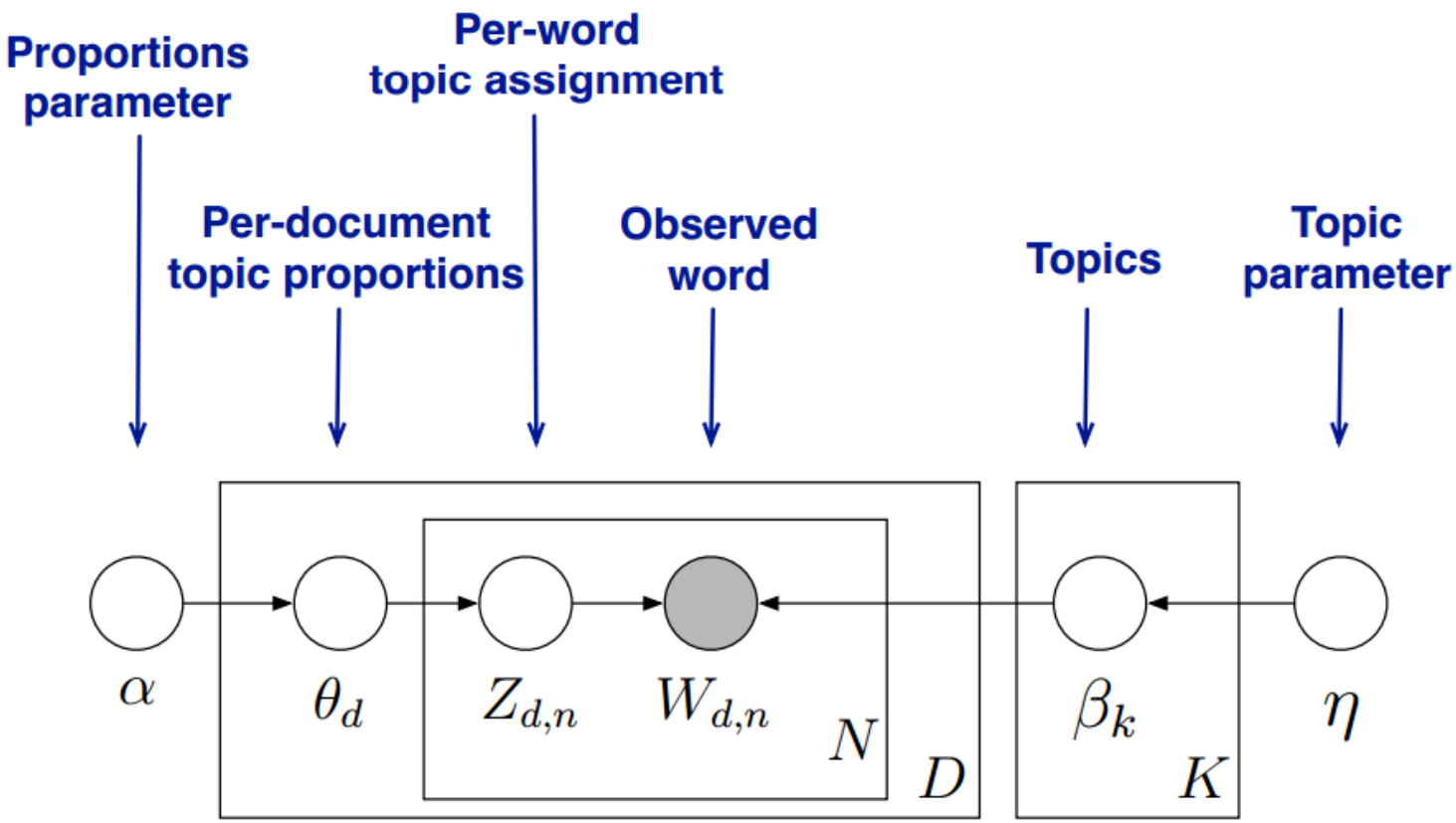
- Each topic is a distribution over words.
- Each document is a mixture of topics.
- Each word is drawn from one of those topics.

The posterior distribution



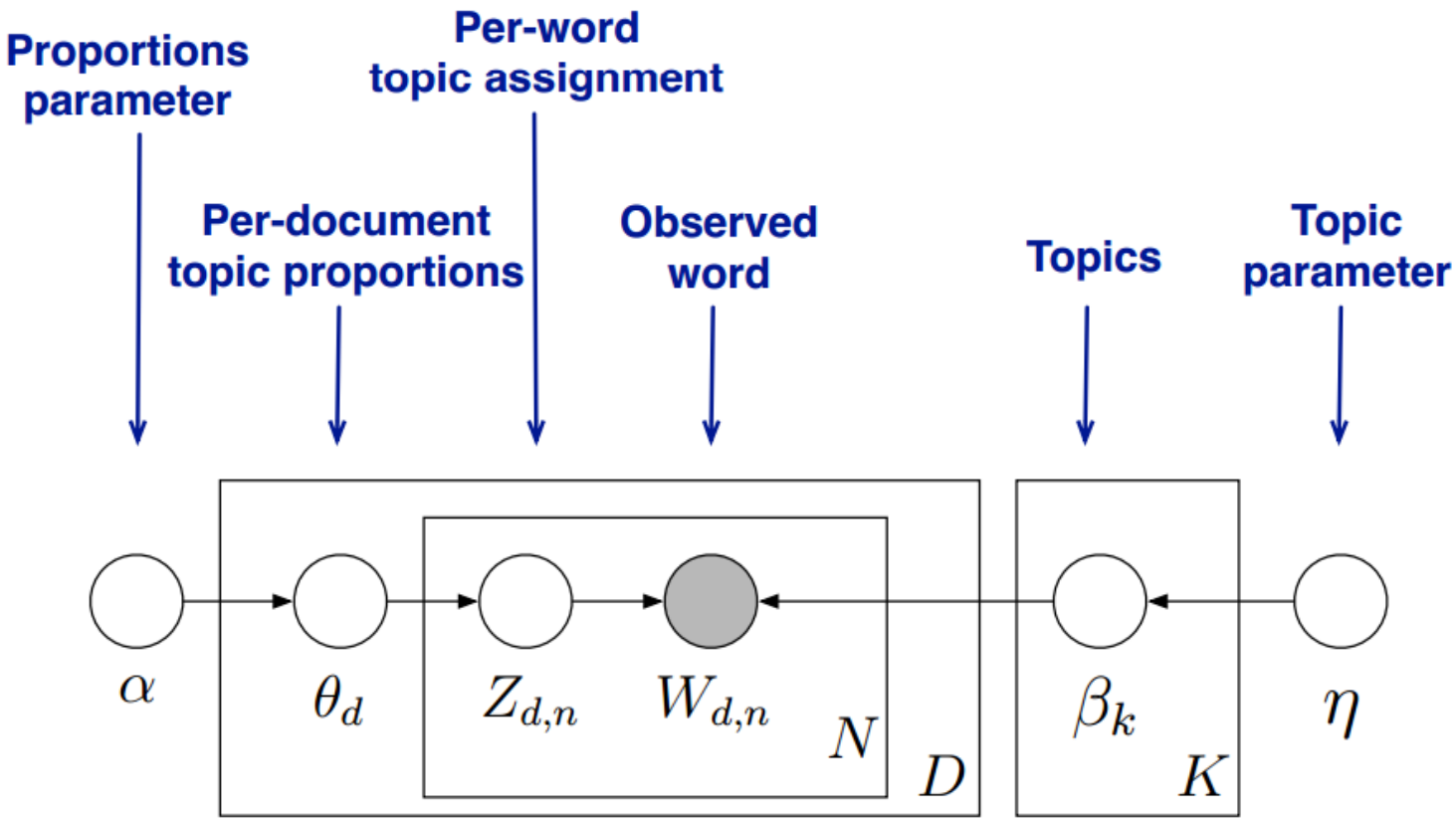
- In reality, we only observe the documents.
- The other structure are hidden variables
- Our goal is to infer the hidden variables.

LDA as a graphicial model



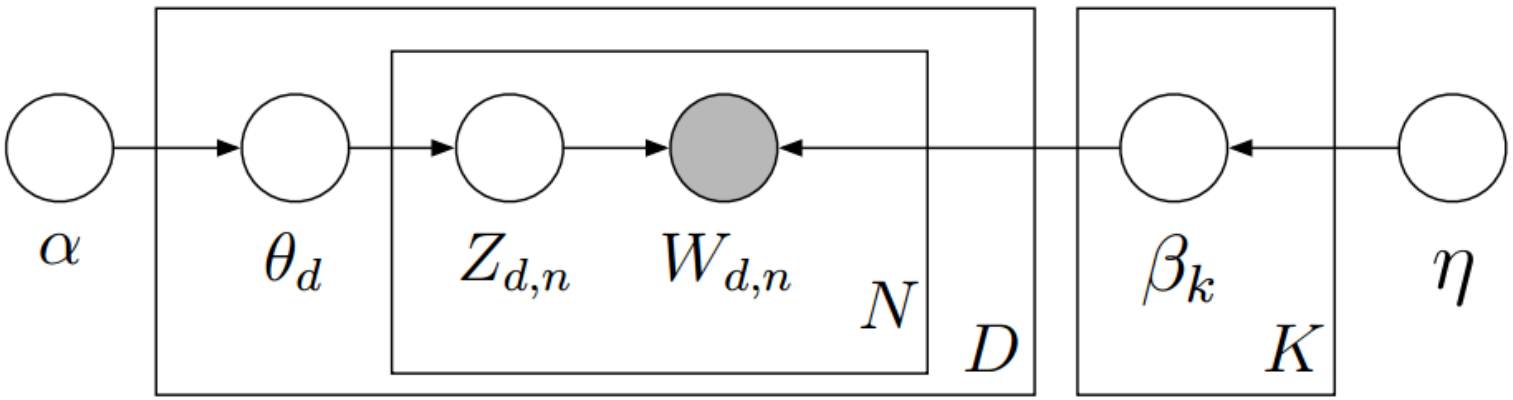
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

LDA as a graphicial model



$$\prod_{i=1}^{K} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \Big(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, Z_{d,n}) \Big)$$

LDA



This joint defines a posterior.approximate posterior inference algorithms:

- Gibbs sampling (Griffiths and Stryvers, 2002)
- variational inference (Teh et at., 2006)

分析新浪微博知名微博主的微博主题(just a toy)

- 实验数据
  - 李开复(创新工场董事长兼首席执行官) 4000条最新微博
  - 杨幂(中国大陆知名女演员) 2000条最新微博
  - 36氪(关注互联网创业的科技博客) 7700条最新微博
- 预处理
  - 分词
  - 去除停用词
- LDA建模，使用LDA的开源实现gensim
- 实验环境及参数
  - 机器配置: Intel(R) Core(TM)2 Duo CPU E7400 @ 2.80GHz，内存2G，操作系统Ubuntu 11.04
  - 主题个数设为10

实验结果

- 李开复的微博主题

```
2012-05-22 21:42:11,659 : INFO : topic #0: 0.004*公司 + 0.004*中国 + 0.004*创新 + 0.003*创业 + 0.003*说 + 0.003*美国 + 0.003*工场 + 0.003*
3*做 + 0.003*开
2012-05-22 21:42:11,663 : INFO : topic #1: 0.004*创新 + 0.004*中国 + 0.004*斯 + 0.004*公司 + 0.003*工场 + 0.003*创业 + 0.003*说 + 0.003*用
3*投资 + 0.003*美国
2012-05-22 21:42:11,667 : INFO : topic #2: 0.004*中国 + 0.003*美国 + 0.003*创新 + 0.003*做 + 0.003*用户 + 0.003*工场 + 0.003*乔布 + 0.003*
003*斯 + 0.002*李
2012-05-22 21:42:11,671 : INFO : topic #3: 0.004*公司 + 0.004*工场 + 0.004*创新 + 0.003*创业 + 0.003*中国 + 0.003*斯 + 0.003*投资 + 0.003*
3*美国 + 0.003*谷
2012-05-22 21:42:11,676 : INFO : topic #4: 0.005*创新 + 0.004*中国 + 0.004*公司 + 0.004*美国 + 0.004*工场 + 0.003*斯 + 0.003*做 + 0.003*说
谷 + 0.003*创业
2012-05-22 21:42:11,680 : INFO : topic #5: 0.004*斯 + 0.004*中国 + 0.003*创业 + 0.003*公司 + 0.003*创新 + 0.003*工场 + 0.003*说 + 0.003*谷
复 + 0.003*做
2012-05-22 21:42:11,684 : INFO : topic #6: 0.004*中国 + 0.004*创新 + 0.004*公司 + 0.004*工场 + 0.004*斯 + 0.003*李 + 0.003*做 + 0.003*开 +
 + 0.003*用户
2012-05-22 21:42:11,689 : INFO : topic #7: 0.005*中国 + 0.004*工场 + 0.004*斯 + 0.003*公司 + 0.003*谷 + 0.003*创新 + 0.003*创业 + 0.003*说
美国 + 0.003*乔布
2012-05-22 21:42:11,692 : INFO : topic #8: 0.005*创新 + 0.004*公司 + 0.004*工场 + 0.004*斯 + 0.003*中国 + 0.003*做 + 0.003*说 + 0.003*开 +
国 + 0.003*用户
2012-05-22 21:42:11,697 : INFO : topic #9: 0.004*中国 + 0.004*创新 + 0.003*公司 + 0.003*创业 + 0.003*美国 + 0.003*用户 + 0.003*斯 + 0.003*
003*谷 + 0.003*苹果
```

- 杨幂的微博主题

```
2012-05-22 21:54:07,284 : INFO : topic #0: 0.011*爱 + 0.007*心 + 0.006*泪 + 0.005*◐ + 0.004*我的 + 0.004*谢谢 + 0.004*今天 + 0.004*耶 + 0.
0.003*幸福
2012-05-22 21:54:07,285 : INFO : topic #1: 0.010*心 + 0.009*爱 + 0.006*◐ + 0.005*泪 + 0.005*好 + 0.005*谢谢 + 0.004*幸福 + 0.004*" + 0.00
0.004*耶
2012-05-22 21:54:07,294 : INFO : topic #2: 0.008*爱 + 0.007*◐ + 0.007*心 + 0.007*泪 + 0.005*幸福 + 0.005*我的 + 0.004*谢谢 + 0.004*好 + 0.
0.003*快乐
2012-05-22 21:54:07,296 : INFO : topic #3: 0.009*爱 + 0.007*心 + 0.006*◐ + 0.005*好 + 0.005*泪 + 0.004*我的 + 0.004*耶 + 0.004*谢谢 + 0.00
0.003*今天
2012-05-22 21:54:07,298 : INFO : topic #4: 0.009*爱 + 0.006*◐ + 0.005*心 + 0.005*泪 + 0.005*我的 + 0.005*好 + 0.004*今天 + 0.004*谢谢 + 0.
+ 0.003*你的
2012-05-22 21:54:07,308 : INFO : topic #5: 0.009*爱 + 0.007*心 + 0.005*泪 + 0.005*我的 + 0.004*◐ + 0.004*谢谢 + 0.004*好 + 0.004*今天 + 0.
+ 0.003*快乐
2012-05-22 21:54:07,309 : INFO : topic #6: 0.007*心 + 0.006*爱 + 0.005*◐ + 0.005*泪 + 0.005*幸福 + 0.004*我的 + 0.004*好 + 0.003*谢谢 + 0.
+ 0.003*耶
2012-05-22 21:54:07,311 : INFO : topic #7: 0.009*爱 + 0.008*心 + 0.005*◐ + 0.005*泪 + 0.004*幸福 + 0.004*谢谢 + 0.004*耶 + 0.004*我的 + 0.
0.003*快乐
2012-05-22 21:54:07,312 : INFO : topic #8: 0.008*爱 + 0.007*心 + 0.006*好 + 0.005*◐ + 0.005*泪 + 0.005*谢谢 + 0.004*快乐 + 0.004*我的 + 0.
0.004*"
2012-05-22 21:54:07,314 : INFO : topic #9: 0.008*心 + 0.006*◐ + 0.005*泪 + 0.005*爱 + 0.005*幸福 + 0.004*今天 + 0.004*耶 + 0.004*好 + 0.00
```

实验结果(续)

- 36氪的微博主题

```
2012-05-22 22:28:55,374 : INFO : topic #0: 0.010*公司 + 0.009*应用 + 0.007*美元 + 0.007*用户 + 0.006*万 + 0.006*创业 +
0.005*发布 + 0.005*网 + 0.004*苹果 + 0.004*手机
2012-05-22 22:28:55,377 : INFO : topic #1: 0.008*公司 + 0.008*应用 + 0.007*用户 + 0.006*美元 + 0.006*万 + 0.005*发布 +
0.005*苹果 + 0.005*移动 + 0.005*网 + 0.004*手机
2012-05-22 22:28:55,380 : INFO : topic #2: 0.009*应用 + 0.009*用户 + 0.007*公司 + 0.006*美元 + 0.006*苹果 + 0.006*创业
+ 0.006*手机 + 0.005*移动 + 0.005*万 + 0.004*发布
2012-05-22 22:28:55,383 : INFO : topic #3: 0.010*公司 + 0.010*应用 + 0.008*美元 + 0.007*苹果 + 0.007*万 + 0.007*用户 +
0.006*发布 + 0.005*手机 + 0.005*创业 + 0.005*移动
2012-05-22 22:28:55,386 : INFO : topic #4: 0.011*应用 + 0.010*公司 + 0.009*用户 + 0.008*美元 + 0.006*苹果 + 0.006*万 +
0.005*创业 + 0.005*推出 + 0.005*手机 + 0.004*平台
2012-05-22 22:28:55,389 : INFO : topic #5: 0.010*应用 + 0.010*用户 + 0.008*公司 + 0.006*美元 + 0.006*手机 + 0.005*网 +
0.005*万 + 0.005*创业 + 0.005*苹果 + 0.004*移动
2012-05-22 22:28:55,393 : INFO : topic #6: 0.009*应用 + 0.009*用户 + 0.008*美元 + 0.008*公司 + 0.006*万 + 0.006*苹果 +
0.005*手机 + 0.005*创业 + 0.005*移动 + 0.005*发布
2012-05-22 22:28:55,395 : INFO : topic #7: 0.010*公司 + 0.008*美元 + 0.008*应用 + 0.007*用户 + 0.007*苹果 + 0.007*创业
+ 0.006*手机 + 0.005*发布 + 0.005*万 + 0.005*移动
2012-05-22 22:28:55,399 : INFO : topic #8: 0.008*应用 + 0.007*公司 + 0.007*美元 + 0.006*用户 + 0.006*手机 + 0.006*万 +
0.006*创业 + 0.005*苹果 + 0.005*网 + 0.005*移动
2012-05-22 22:28:55,401 : INFO : topic #9: 0.009*应用 + 0.007*用户 + 0.006*公司 + 0.006*美元 + 0.006*创业 + 0.005*苹果
+ 0.005*网 + 0.005*推出 + 0.005*手机 + 0.004*万
```

主题归纳

-  创新，中国，创业，投资，用户，硅谷，苹果
-  爱，心，幸福，谢谢，快乐，耶
-  公司，应用，美元，用户，创业，发布，移动，苹果，手机

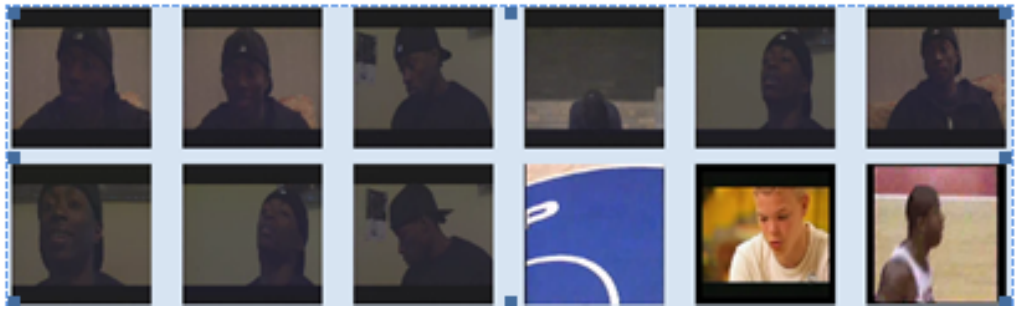
Topic model in image processing

- 实验数据: youtube22数据集，包含22个概念的数据，实验只使用其中的一个概念，篮球。一共使用了50个与篮球相关的视频，提取的关键帧包括2632张图片。
- 特征提取：我了提取SIFT特征，SIFT算法是一种提取局部特征的算法，在尺度空间寻找极值点，提取位置，尺度，旋转不变量。
- 关键帧的向量表示,使用BOF模型。
- 实验参数及环境设置：SIFT的采样数目设为1万个，聚类为500类，每张图片用一个500维的特征向量表示。然后通过LDA进行主题分析，主题个数我们设置为10，

实验结果

- Topic 2:

- Topic 3:

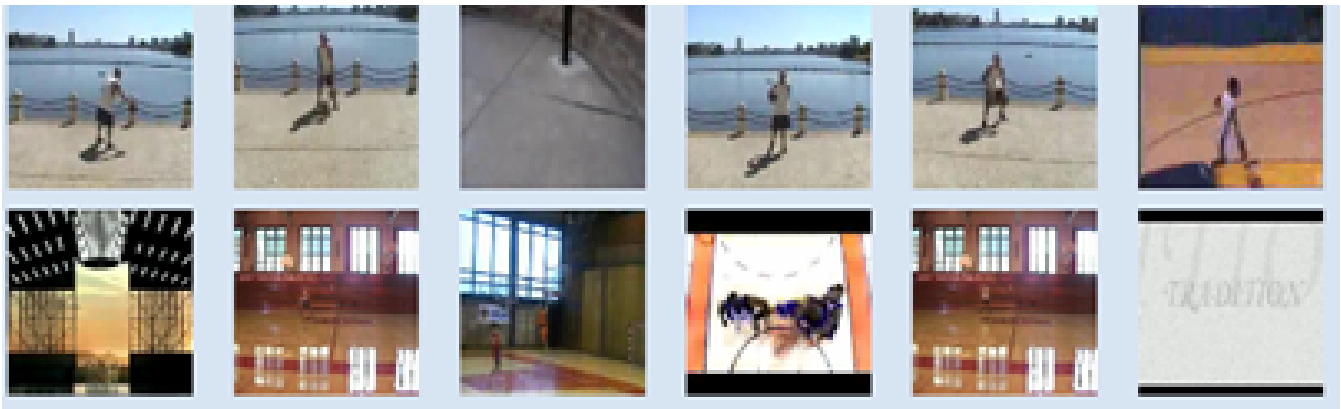


- Topic 5:

实验结果(续)

- Topic 1:

- Topic 4:

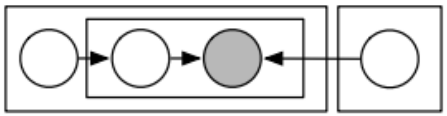


Beyond latent Dirichlet allocation

- Modeling richer assumptions
  - Dynamic topic models
  - Correlated topic models
- Supervised topic models
  - Supervised LDA
  - Relational topic models
- Bayesian nonparametric topic models

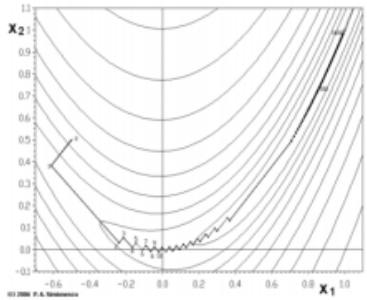
Summary

Assumptions

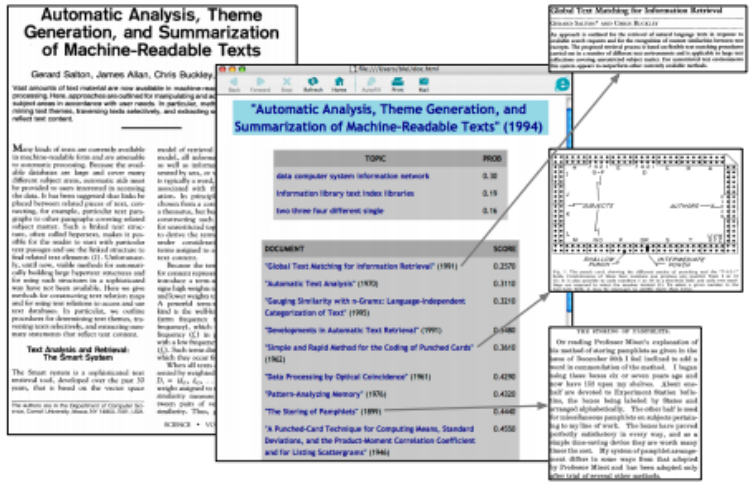


Data

Inference algorithm



Discovered structure



Reference

- David M. Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. JMLR (3) 2003 pp. 993-1022.
- David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM.2011
- David M. Blei, Jon D. McAuliffe. Supervised Topic Models. NIPS (2007)
- David M. Blei, John D. Lafferty. Dynamic Topic Models. ICML (2006)
- Jonathan Chang, David Blei. Relational Topic Models for Document Networks. AIStats (2009)

Q&A

Thanks!