Topic Modeling

Comparison of methods for choosing an optimum value for the number of topics in an LDA model

Patricia Jean Goedecke

Supervised by Dr. Dale Bowman
Mathematical Sciences, Statistics
University of Memphis

June 29, 2017

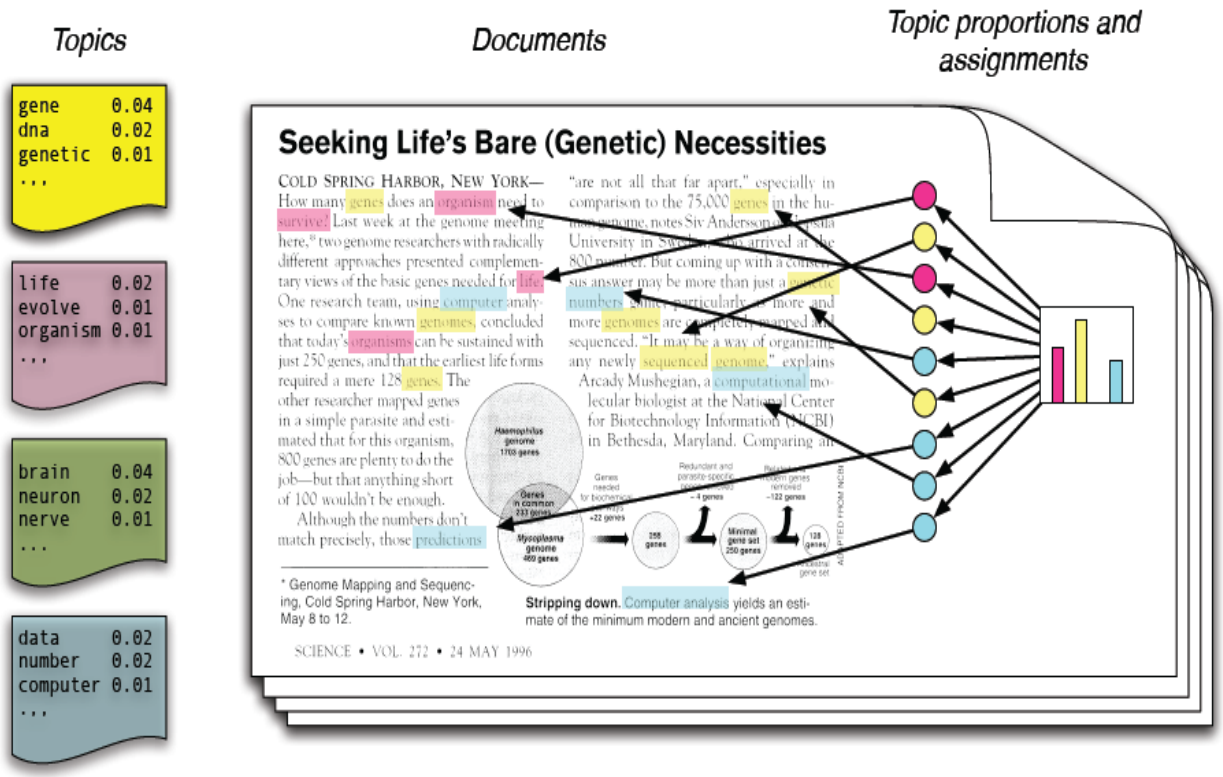
Outline

Introduction

Introduction: What is topic modeling?

- Unsupervised dimension reduction or clustering technique
- Assumes latent topics joining documents to words orthogonally
- Originally used with large corpuses of text data
- Recently applied to other big data including biomedical data sets

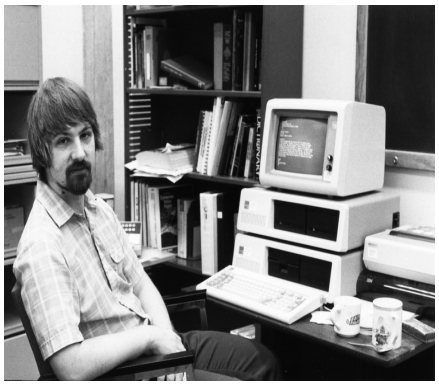
Introduction: What is topic modeling?



[0] Image credit Blei 2011

Methods leading to Latent Dirichlet Allocation

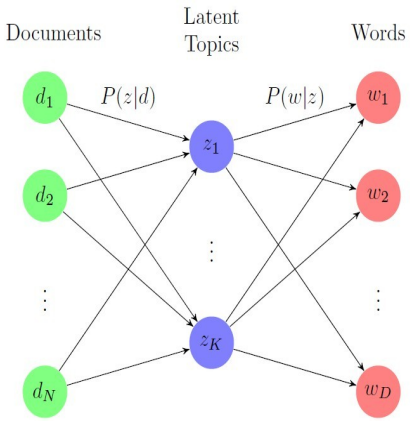
Latent Semantic Analysis (LSA)



- Deerwester et al. (1990)
- unsupervised text modeling technique
- term by document matrix
- orthogonal factors
- New indexing method found "promising"

[1] Photographic Archive, [apf12345], Special Collections Research Center, University of Chicago Library.

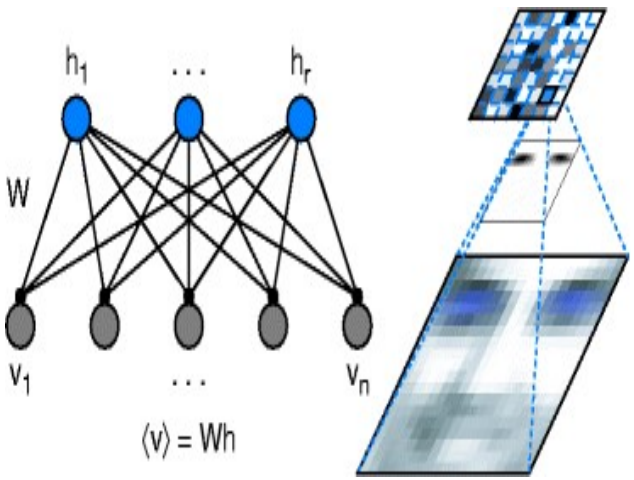
Probabilistic latent semantic indexing (pLSI, pLSA)



- Hofmann (1999)
- uses a mixture decomposition
- maximum likelihood model
- tempered expectation–maximization (EM) iterative algorithm

[2] Image credit: http://www.cnblogs.com/yuzhung/archive.html

Non-negative matrix factorizaton (NMF)

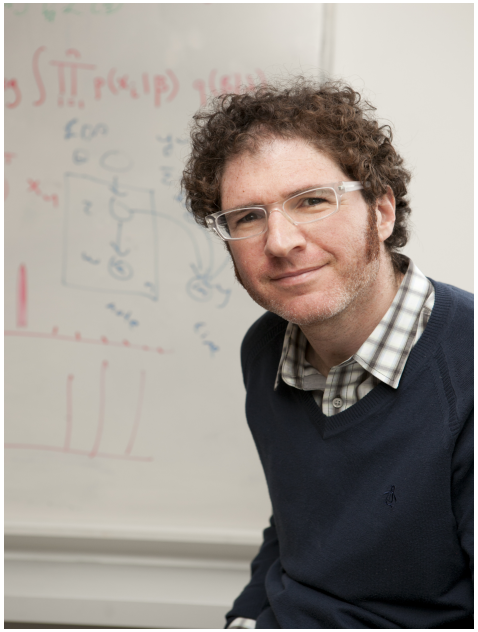


- Lee & Seung (1999)
- parts-based representation
- described as a neural network
- found by Ding (2008) to optimize the same objective function as pLSI

[1] Image credit Nature.com Lee & Seung (1999)

LDA and current trends

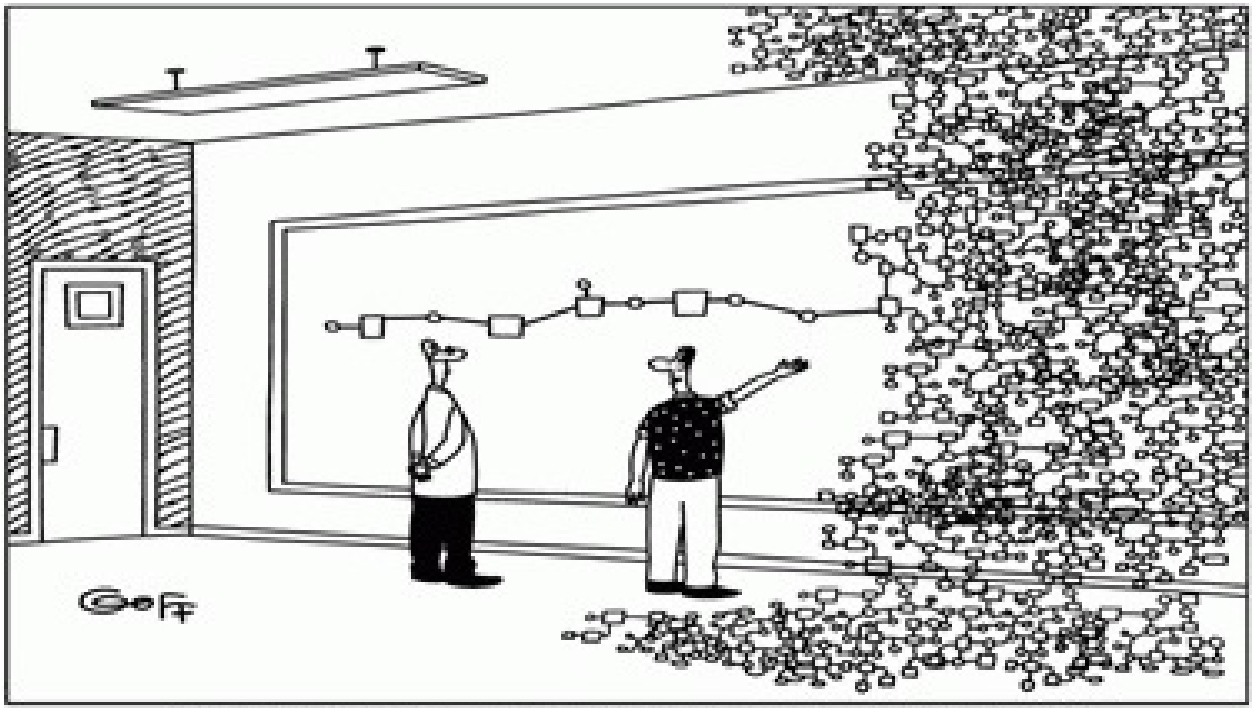
Latent Dirichlet allocation (LDA)



- Blei, Ng and Jordan (2003).
- document is finite mixture over topics
- topic is infinite mixture over probability distribution
- modeled using variational methods, EM algorithm

[2] Photo credit: Denise Applewhite

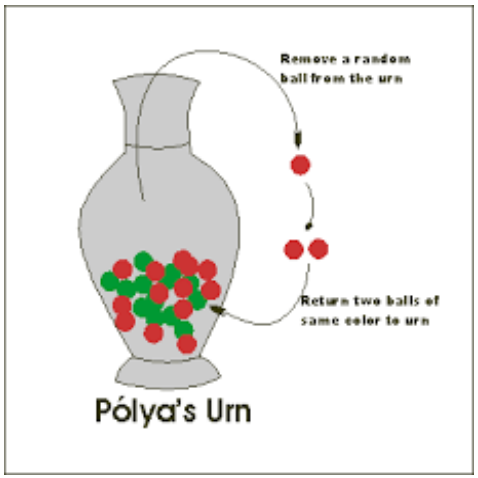
Expanding applications



"This is where the idea for the new EHR starts getting a little complicated."

[1] cldeanrobles.files.wordpress.com

Topic coherence & selecting the number of topics



- Pointwise Mutual Information (PMI), Newman et al. (2010)
- Generalized Pólya Urn model, Mimno et al. (2011)

[1] http://fac.ksu.edu.sa/raguech/blog/173205

Topic modeling using latent Dirichlet allocation

de Finetti's theorem and exchangeability

- Exchangeable random variables are conditionally independent, conditioned on some latent variable
- In topic modeling, words are conditionally independent, conditioned on topic

[1] http://www.novuslight.com.html

LDA hypothetical document generation

Representation $\sim$

1. A corpus consists of D documents
2. A document consists of N words from vocabulary of size V
3. Each word w_n is chosen from one of K topics.

LDA hypothetical document generation

1. Number of words (N) for document selected from Poisson, parameter ξ.
2. Multinomial probability vector θ length K selected from Dirichlet, parameter α
3. The nth word is chosen from the following process:

- Choose a topic, Z_n, from $Mult(1, \theta)$
- Choose a $V \times 1$ vector ϕ from $Dir(\beta_{Z_n})$
- Choose a word, w_n from $Mult(1, \phi)$

LDA hypothetical document generation

In this model

- w_{nj}=1 if the nth word is the jth in the vocabulary
- Z_{ni} = 1 if the nth word is in topic i
- w_{nj} is observed
- Z_{ni} is unobserved
- ϕ_{ij} =$p(w_{nj} = 1|z_{ni} = 1)$

LDA hypothetical document generation

As only one word occurs at each location in a document,

$$\sum_{i=1}^{K} z_{ni} = \sum_{j=1}^{V} w_{nj} = 1.$$

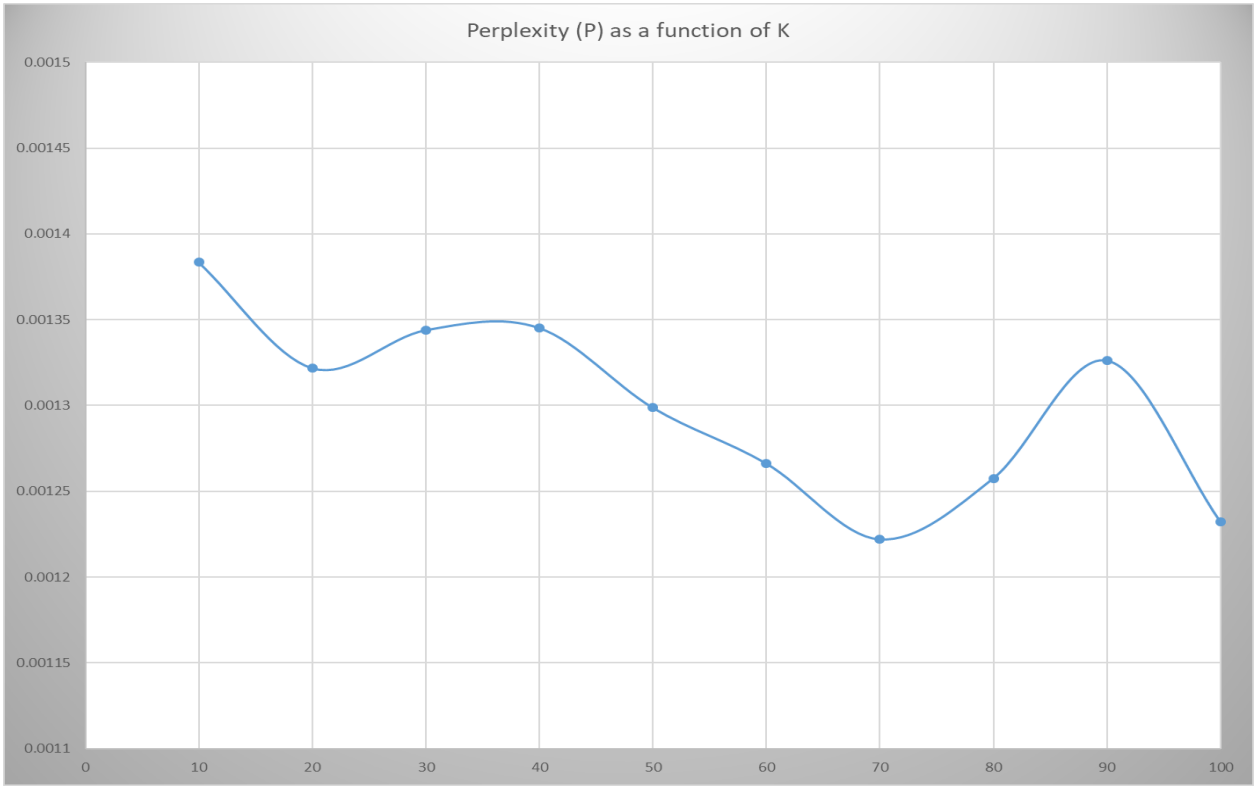
K optimization methods

Perplexity (P)

- Blei et al. (2003)
- Algebraic inverse of geometric mean per-word likelihood
- Lower perplexity scores indicate better performance
- Metric intrinsic to the LDA modeling process

$$P = \exp\left\{ -\frac{\sum_{d=1}^{D_M} \log P(\mathbf{w_d})}{\sum_{d=1}^{D_M} N_d} \right\}$$

Perplexity (P)



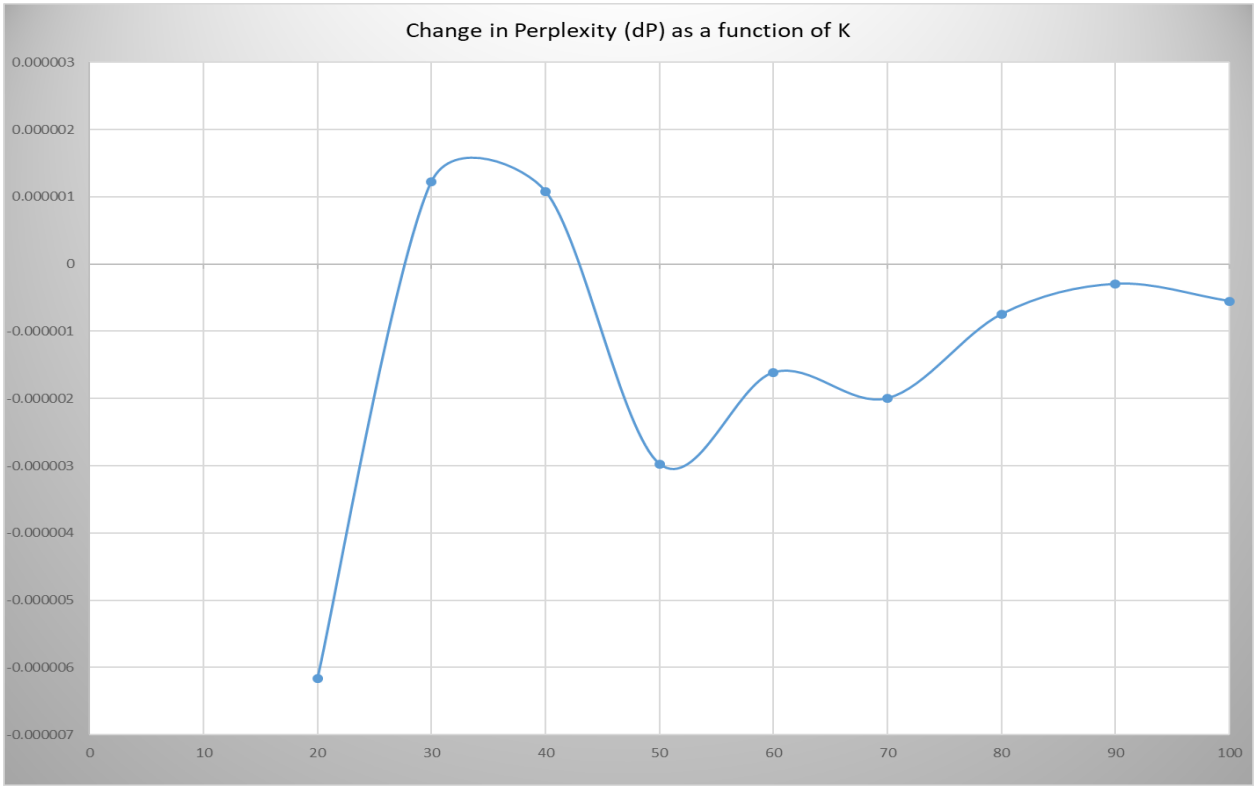
[1] Blei et al., 2003

Change in perplexity (dP)

- Zhao et al. (2015)
- Observed to be more stable and efficient as compared with raw peplexity

$$dP = \frac{P_{i+1} - P_i}{K_{i+1} - K_i}$$

Change in perplexity (dP)

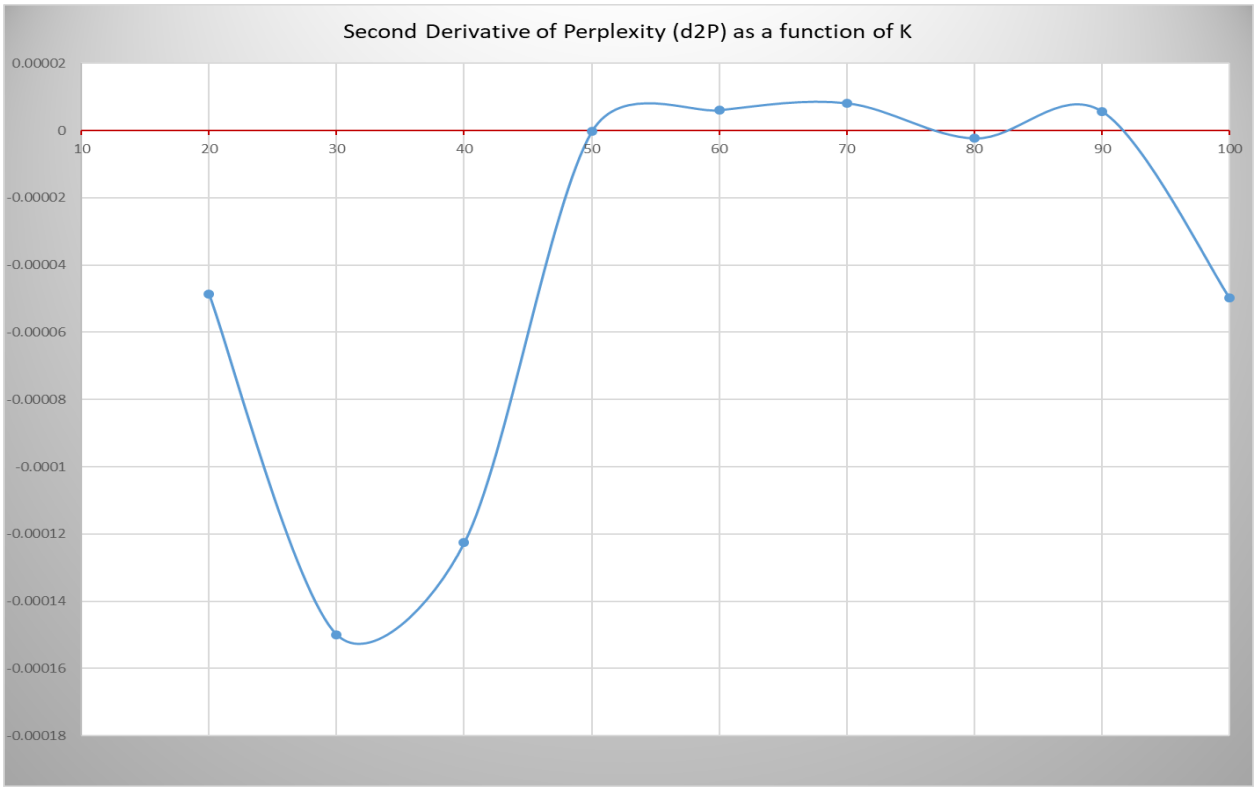


[1] Zhao et al., 2015

2nd derivative of perplexity (d2P)

- The zero line clarifies the change point described by Zhao et al.

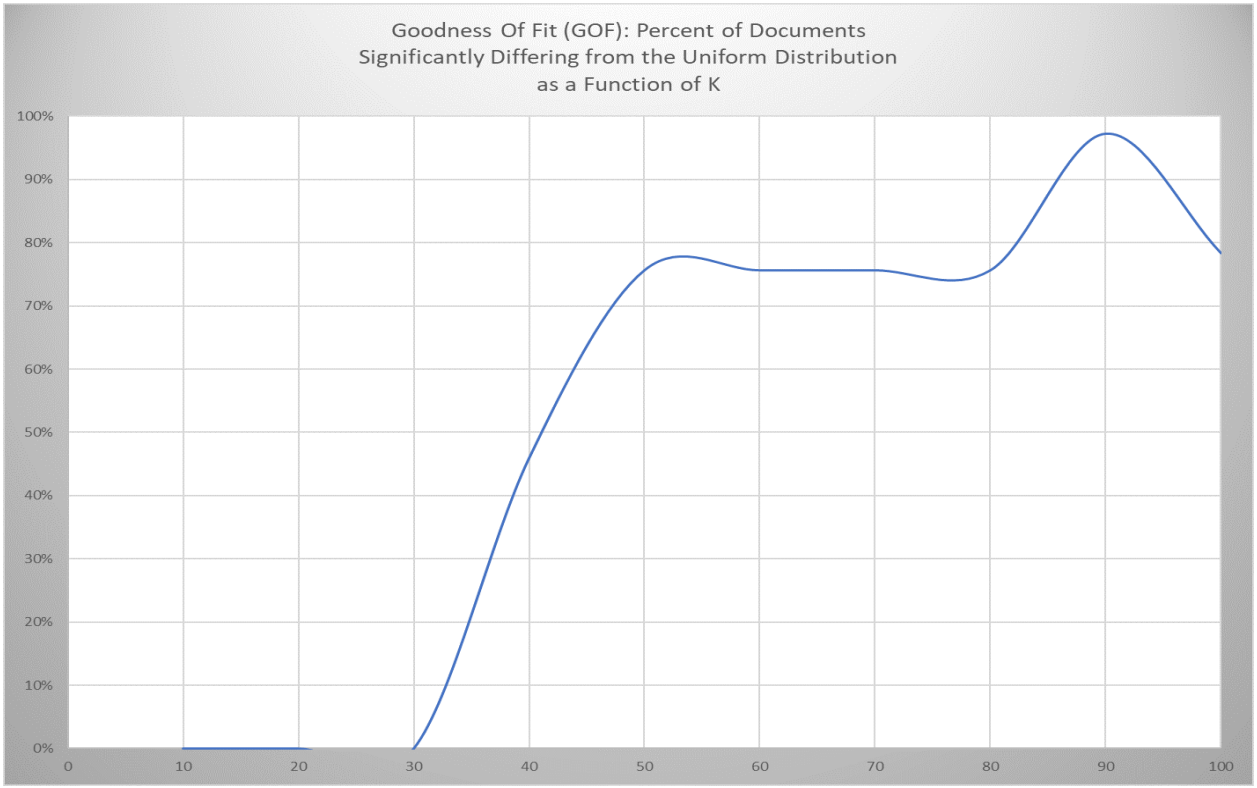
2nd derivative of perplexity (d2P)



Goodness Of Fit (GOF)

- Bowman, Chen and George (in press)
- Statistical analysis of model output
- Proportion of documents significantly different than random as metric to evaluate K
- H_0: Topic assignment not different from random (uniform) assignment

Goodness Of Fit (GOF)



[1] Bowman, Chen and George (in press)

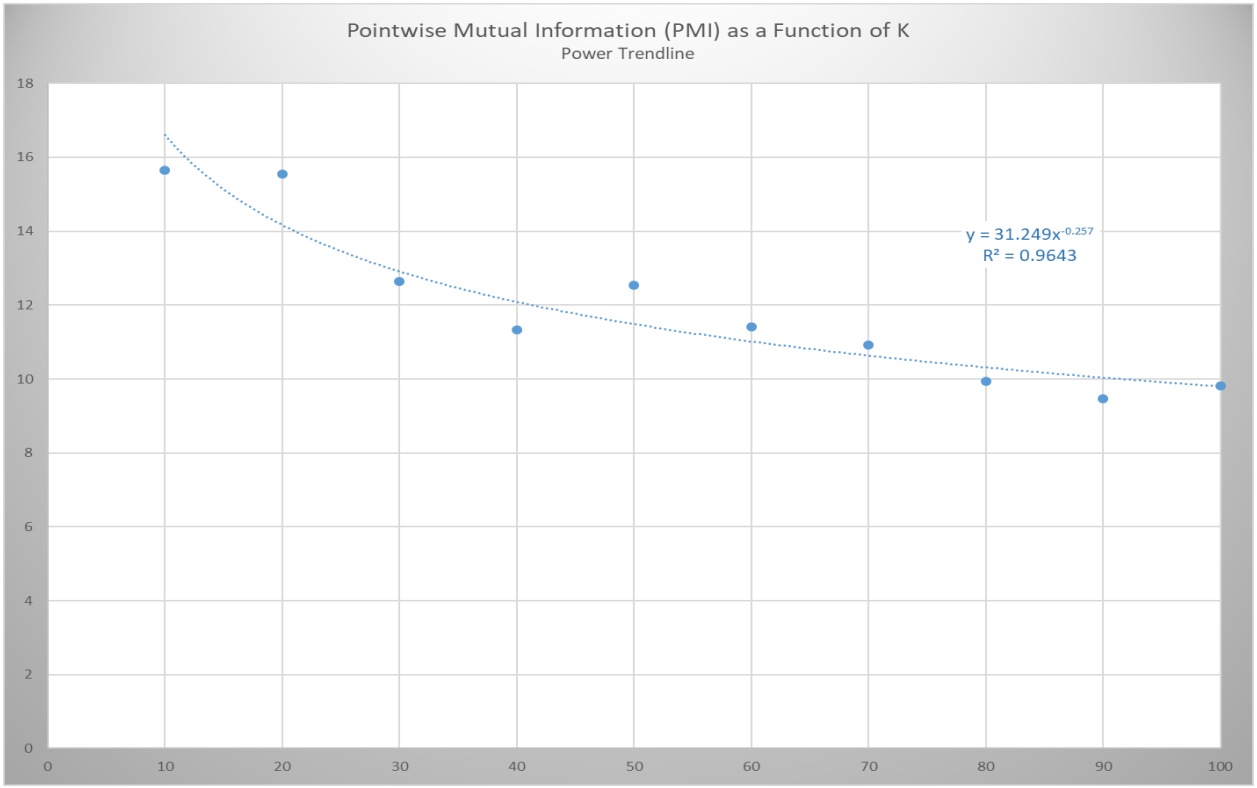
Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

- Newman et al. (2010)
- Extrinsic to modeling process
- Corroborated by human judgment

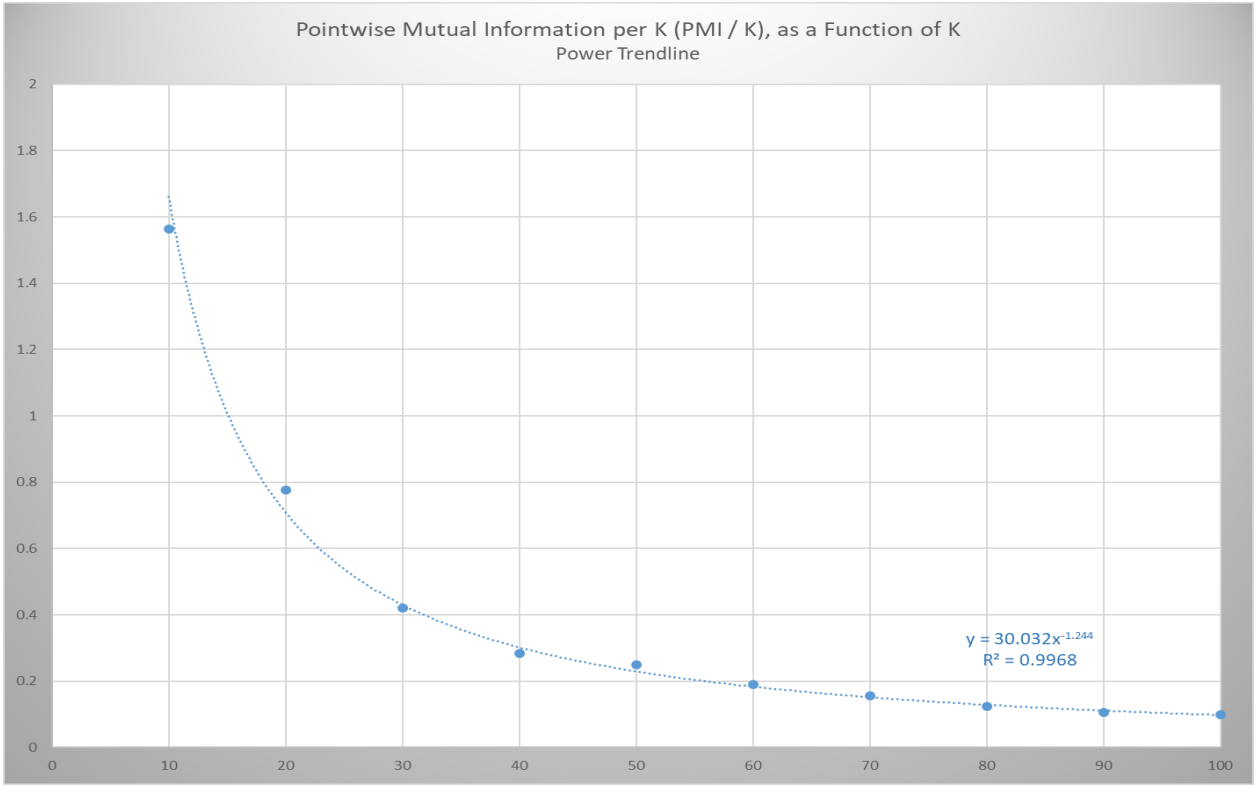
[1] Image credit: http://lintool.github.io/UMD-courses.html

Pointwise Mutual Information (PMI)

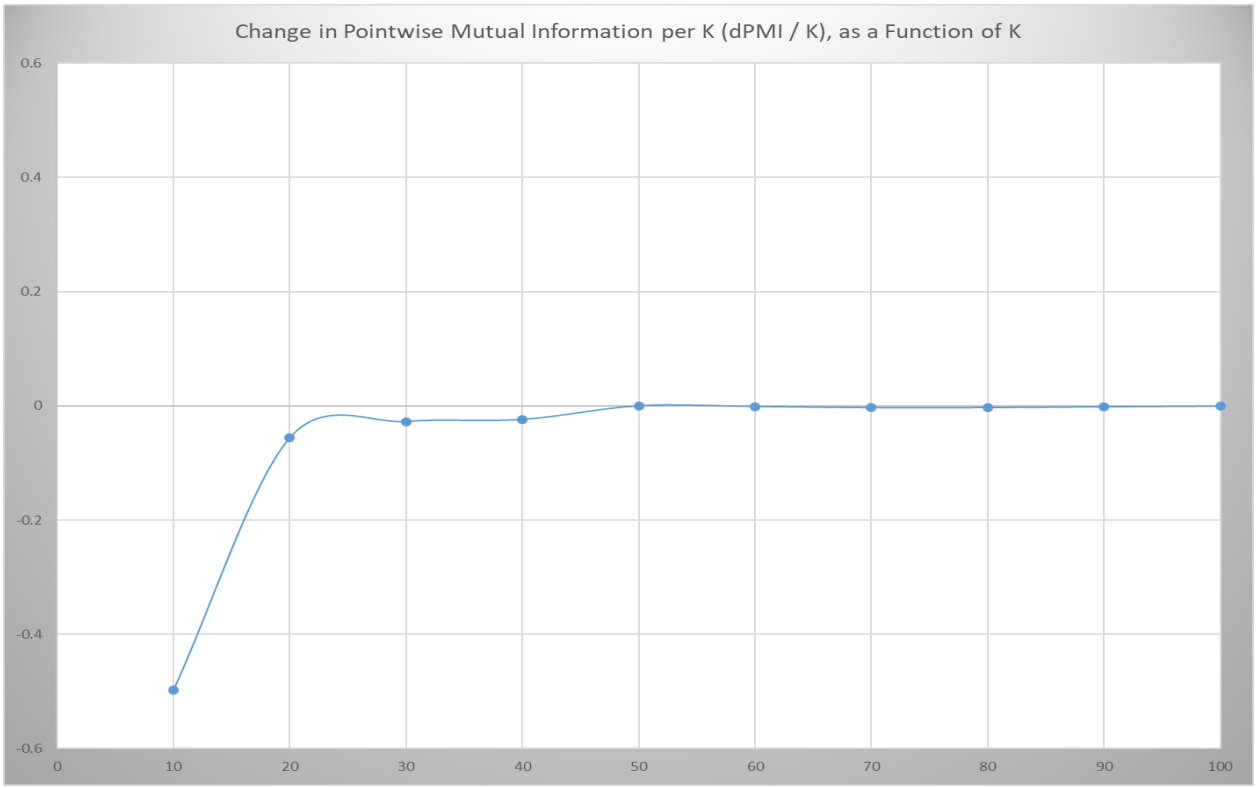


[1] Naraula et al., 2013

Pointwise Mutual Information per K (PMI / K)



Change in Pointwise Mutual Information per K (PMI / K)



Conclusion: Optimization method comparisons

- Methods of optimizing K do not converge
- Different optimization methods will be appropriate for differing purposes
- Judgment is called for

Thank you

$\sim$ *Thank you* $\sim$