北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Topic Detection - LDA

Name: Chen Muqiao
Name: Fu Lemeng meimei
Name: Liu Jintian didi
Name: Lu Xisuo
Name: Wang Jingqi

德以明理 学以精工

Background

目录 | CONTENTS

1 What is TDT

2 why

3 Framework

1 What is TDT

TDT research begins with a constantly arriving stream of text from newswire and from automatic speech-to-text systems that are monitoring selected television, radio, and Web broadcast news shows. Roughly speaking, the goal of TDT is to break the text down into individual news stories, to monitor the stories for events that have not been seen before, and to gather the stories into groups that each discuss a single news topic.

The initial motivation for research in TDT was to provide a core technology for an envisioned system that would monitor broadcast news and alert an analyst to new and interesting events happening in the world. Analysts are very

1 Grasp the main idea

Suppose you have a huge number of documents(corpus) New York times

Topic detection provide a corpus-level intuition of major themes

1 Track Topics on Social Medias

Getting raw

Before we can complete our review of your developer account application, we need some more details about your use case, and who will be using your product.

Please reply to this email with the following information:

- A list of the government or public sector entities that will have access to Twitter content, or information derived from Twitter content, under this use case.
- The specific use cases of your product or service by government or public sector entities.

Framework

Getting the Data

Text Modeling

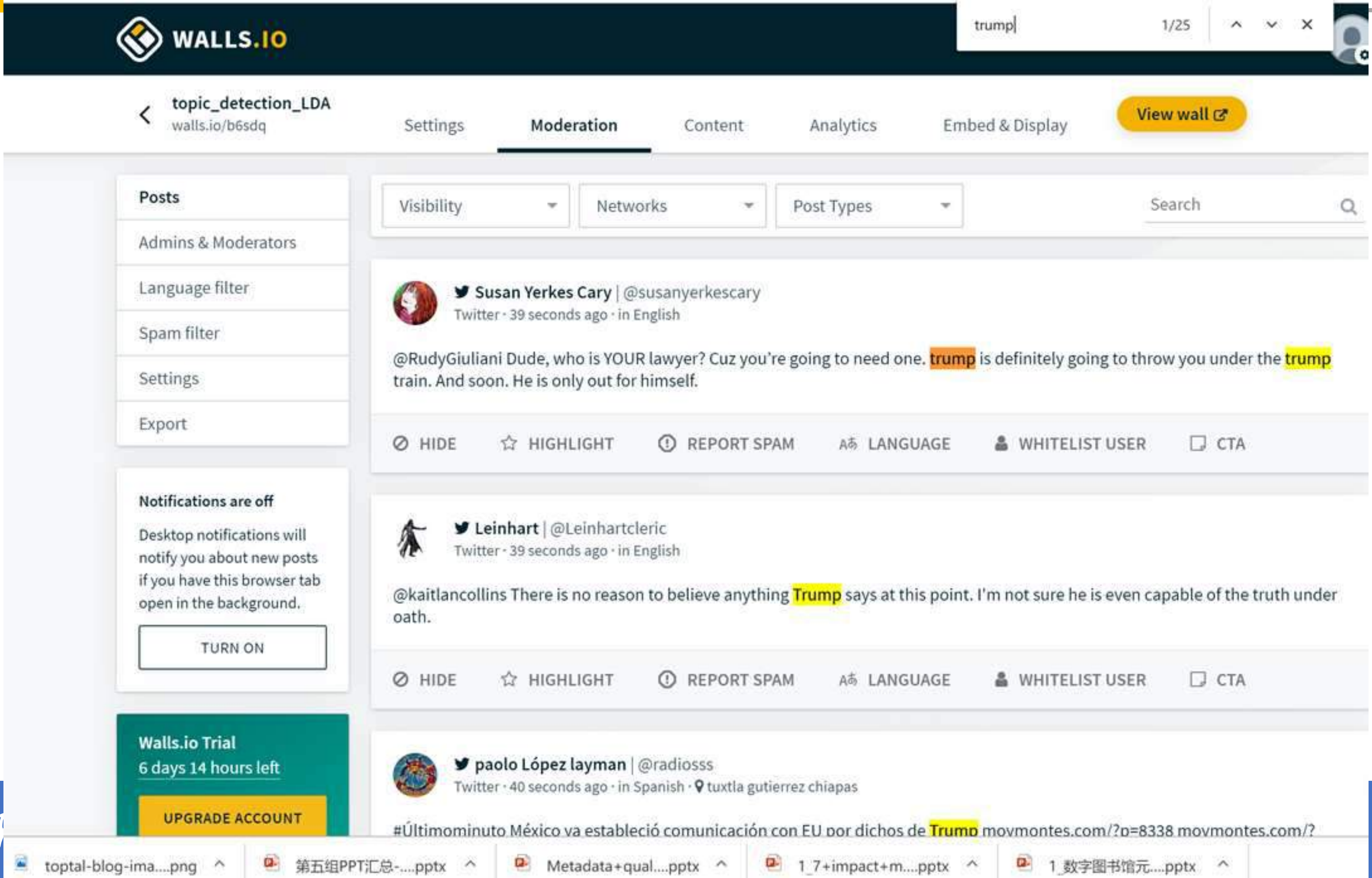
Cleaning the Data

Clustering

1

Getting the Data

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Comment	Type	External name	External fulln	External image	External created	Status	Is highligh	Location	Longitude	Latitude	Language	
2	Imagine the Trump/Rocky picture in a history book some day.	twitter	LeahKraus	Leah Kraus	https://pbs.twim	2019/11/27 20:18	1	0	Memphis, TN	-88.4831	35.76146	en	
3	Personally, I don't put much belief in 'polls', too easy to slant the opinions, bias. If	twitter	pink2yoo	j k miller	https://pbs.twim	2019/11/27 20:18	1	0	Wi USA	-88.7879	43.78444	en	
4	@KlassLib @Atomreisfleisch Der weitgehend erfolglose Obama hat sicherlich das	twitter	DamasiusPugna	Luke Grumpl	https://pbs.twim	2019/11/27 20:18	1	0	Grumplor			de	
5	@Osiris_Rep_UPR @GuiguiVincenti @CH_Gallois @idrissaberkane c'est tant mieux ! Et ce n'est pas pour rien que les USA de Trump ont promis d'aider l'Angleterre lors du Brexit. Mais le Brexit n'existe toujours pas... donc s'appuyer sur le fait "qu'il n'y a pas de pb avec le Brexit" reste une escroquerie intellectuelle de certains #UPR.	twitter	MarcChinal	Marc Chinal	https://pbs.twim	2019/11/27 20:18	1	0	Lyon	4.828693	45.77007	fr	
6	@LindseyGrahamSC No, you have that wrong @LindseyGrahamSC Trump is invit	twitter	Trumpisaphony	Duncan Scott	https://abs.twim	2019/11/27 20:18	1	0	USA	-95.7129	37.09024	en	
7	鈥樉merica First鈥?is another Trump fraud: Watchdog reveals how foreign compa	twitter	BLKROCKET	ROBERT	https://pbs.twim	2019/11/27 20:18	1	0	Naked Blue Planet			en	
8	17 States Sue Trump Administration to Keep Endangered Species Act Intact https:	twitter	susanLa60	susan Lane	https://pbs.twim	2019/11/27 20:18	1	0	Australia	-94.5258	39.00882	en	
9	The fact that these idiots can vote is why trump was able to steal an election. @cc	twitter	Arigon52	Greg Butler	https://pbs.twim	2019/11/27 20:18	1	0	BKNY US	-73.9442	40.67818	en	
10	There is a real danger of us not getting the no deal we require to sell the NHS to tRump I fixed it for you	twitter	skybluebint	kathy	https://pbs.twim	2019/11/27 20:18	1	0	salford.	-2.29013	53.48752	en	
11	Trump 2020. Read it and weap "Democrats "	twitter	ChachiButt	Bob Martin	https://abs.twim	2019/11/27 20:18	1	0				en	
12	#Trump's #Bribery Apprentice? How Failed Bribery Plot Put Governor In #Prison For 14 Years | #MSNBC https://t.co/pMy60q5a9V via @YouTube	twitter	Gjallarhornet	Oden	https://pbs.twim	2019/11/27 20:18	1	0	Helsingborg, SI	12.69451	56.04647	en	
13	Thank #God we have #Leaders like @SpeakerPelosi Protecting #America #Americ	twitter	Ramon36069	Ramon 360	https://pbs.twim	2019/11/27 20:18	1	0	Metro NY Area	-74.0005	40.71864	en	
14	Boo. Just got a 12 hour ban. My first. I'm so proud. I wonder which trump cried. M	twitter	Al_K_Hall64	Huw Jones馃	https://pbs.twim	2019/11/27 20:18	1	0	In Your Dream	-97.4721	35.21885	en	
15	@ccollao @Silc_nj @Travon There isnt a fucking wall. All a lie like everything else	twitter	reznov194513	why so seriou	https://pbs.twim	2019/11/27 20:18	1	0				en	
16	@Emolclause @WFrance26 @Paula_White Another irrational idiot who Trump sur	twitter	JayKevinJohnso	Jay Kevin Joh	https://pbs.twim	2019/11/27 20:18	1	0	Palm Springs, C	-116.524	33.82508	en	
17	Reminds me of when I tweeted that I was wondering what Canadians were thinkir	twitter	hsiyinl	Evangeline	https://pbs.twim	2019/11/27 20:18	1	0	Earth	4.392025	52.14478	en	
18	@thehill Yeah trump has his self surrounded by some real cut throat politicians. t	twitter	c_blabeblabe	james c blade	https://pbs.twim	2019/11/27 20:17	1	0	Pineville, La	-92.3944	31.34387	en	
19	Trump hates himself. And he hates those that kiss up to him, because he hates hir	twitter	Mindy94694520	Sillybird 1313	https://pbs.twim	2019/11/27 20:17	1	0				en	
20	I'm sure I'm missing a ton of good Twitter content today because I have to speed-scroll past all of those godawful photoshopped trump/Rocky images. I bet the WH staff will soon be instructed to make that image their icon on all social media platforms.	twitter	Resist_Agitate	Really Mom	https://pbs.twim	2019/11/27 20:17	1	0	Tennessee, US	-86.3172	35.83052	en	
21	Trump Security Adviser Kash Patel sues NY Times over Ukraine story https://t.co/zl	twitter	raybae689	RAY BAEZ	https://pbs.twim	2019/11/27 20:17	1	0	Queens, NY	-73.7976	40.75009	en	
	@JamesEl09687611 @piersmorgan Lighten up. Why can't the left meme, or												

2019-11-27-20-18-32-b6sdq.csv = (~\Downloads) - GVIM

文件(F) 编辑(E) 工具(T) 语法(S) 缓冲区(B) 窗口(W) 帮助(H)


```
"Comment","Type","Post link","Post image","Post video","Post id","External name","External fullname","External image","External created","Status","Is highlighted","Location","Longitude","Latitude","Language"
"Imagine the Trump/Rocky picture in a history book some day.","twitter","https://twitter.com/LeahKraus/status/1199784512382799872","","","1199784512382799872","LeahKraus","Leah Kraus","https://pbs.twimg.com/profile_images/1008518609784705024/O0Jjp3VU.jpg","2019-11-27 20:18:13","1","0","Memphis, TN","-88.4831490789100","35.7614567537170","en"
"Personally, I don't put much belief in 'polls', too easy to slant the opinions, bias. If Americans can't see the parallel of Trump's actions & those crime families that eventually get taken out by the Ricco laws, then many folks are blind of afraid o truth https://t.co/OTQPEdo0fO","twitter","https://twitter.com/pink2yoo/status/1199784510080061440","","","1199784510080061440","pink2yoo","j k miller","https://pbs.twimg.com/profile_images/1181211330600148992/g1oVTw3v.jpg","2019-11-27 20:18:12","1","0","Wi USA","-88.78786780000000","43.78443970000000","en"
```

Common data cleaning steps on all text:

1. Split (, . ‘ ‘)
2. Remove stopwords (a, the, or)
3. stemming(ed, ing)

BIT

Text Modeling

德以明理 学以精工

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录 | CONTENTS

1 LDA & variations

2 LDA

3 Gibbs Sampling

Latent Dirichlet Allocation

- Input: Document-Term Matrix, number of topics, number of iterations
- Gensim will go through the process of finding the best word distribution for each topic and best topic distribution for each document.
- Output: The top words in each topic. It is your job as a human to interpret this and see if the results makes sense. If not, try altering the parameters - terms in the document-term matrix, number of topics, number of iterations, etc. Stop when the topics make sense.

LDA Variations

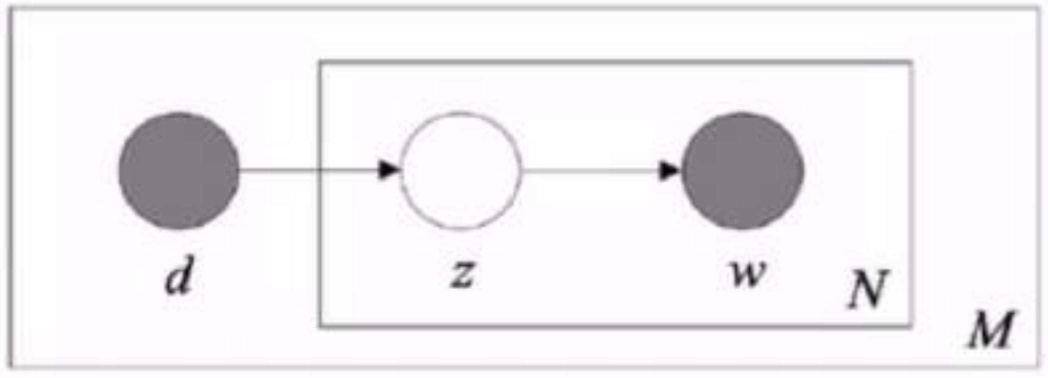
	Supervised	Un_Supervised
Hierachical	PLSA, LDA	HDP
Non-Hierachical	Labeled LDA	HSLDA

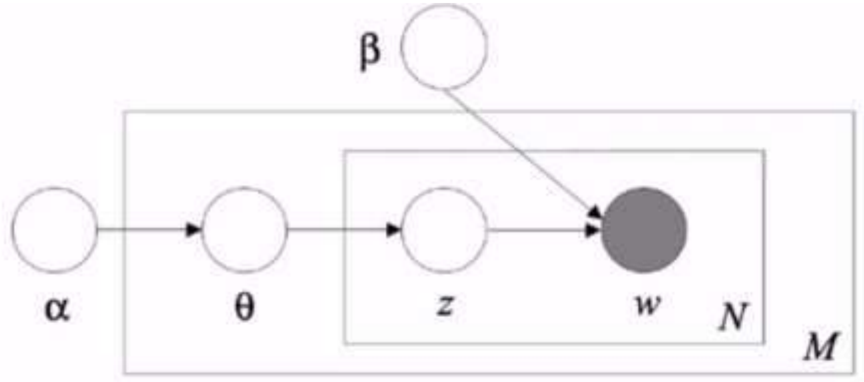
Assumptions of LDA

All topic models are based on the same basic assumption:

- each **document** consists of a mixture of ***topics***, and
- each ***topic*** consists of a collection of **words**.

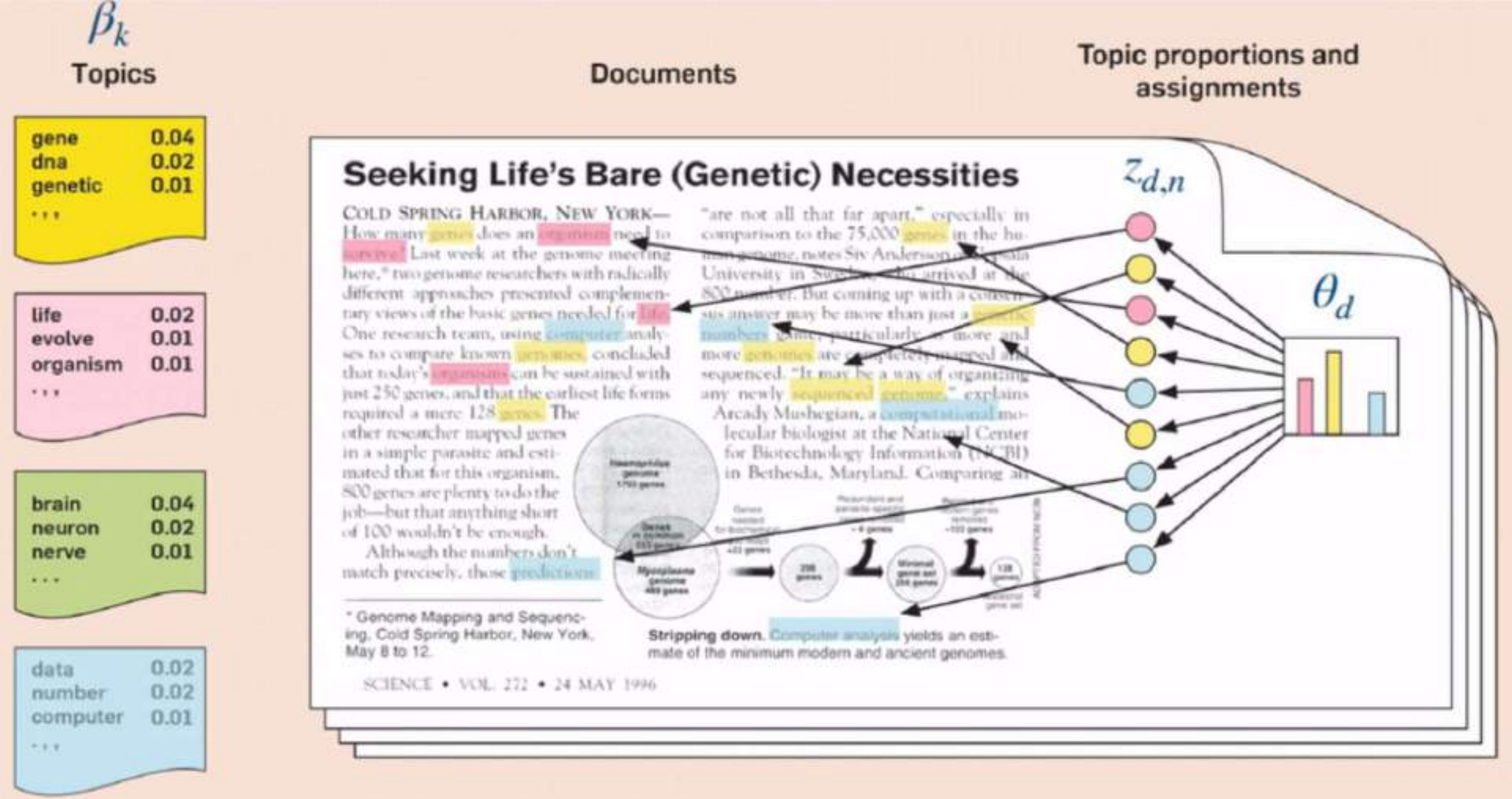
1 LDA



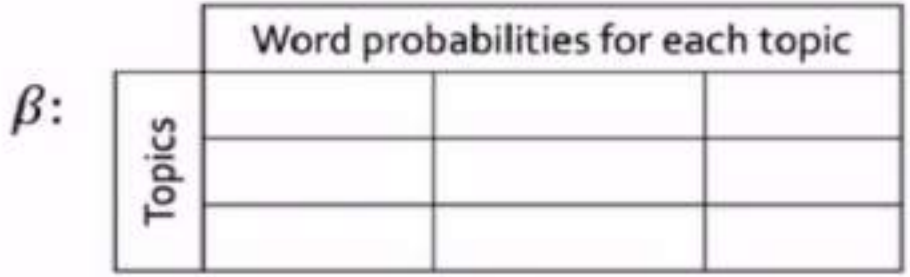


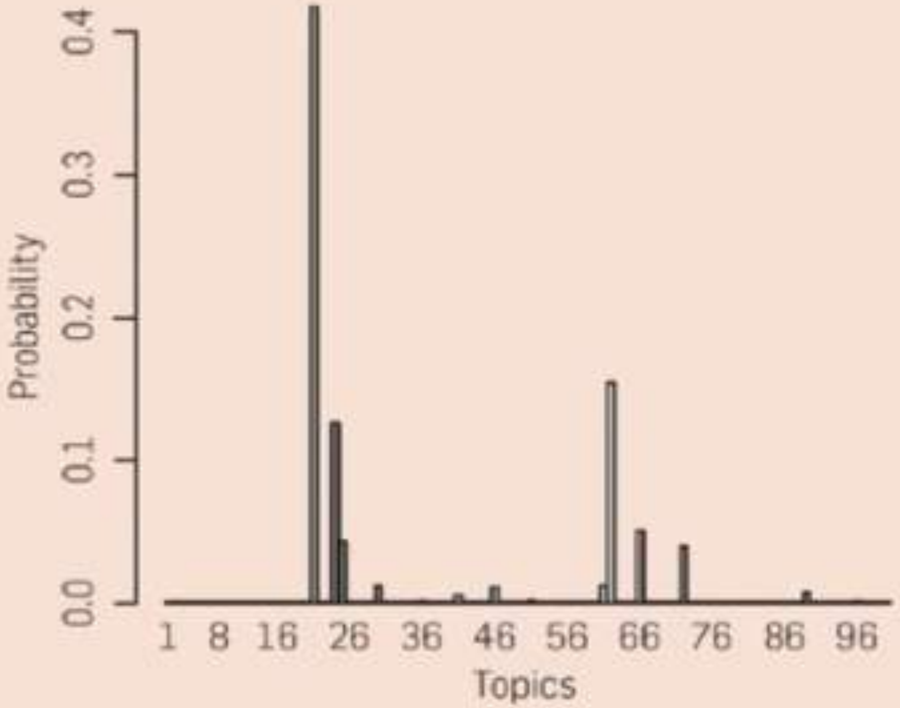
$$p(\mathbf{w}) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta)p(z_n|\theta) \right) p(\theta;\alpha)d\theta,$$

LDA improves PLSA by imposing Dirichlet priors on the model parameters



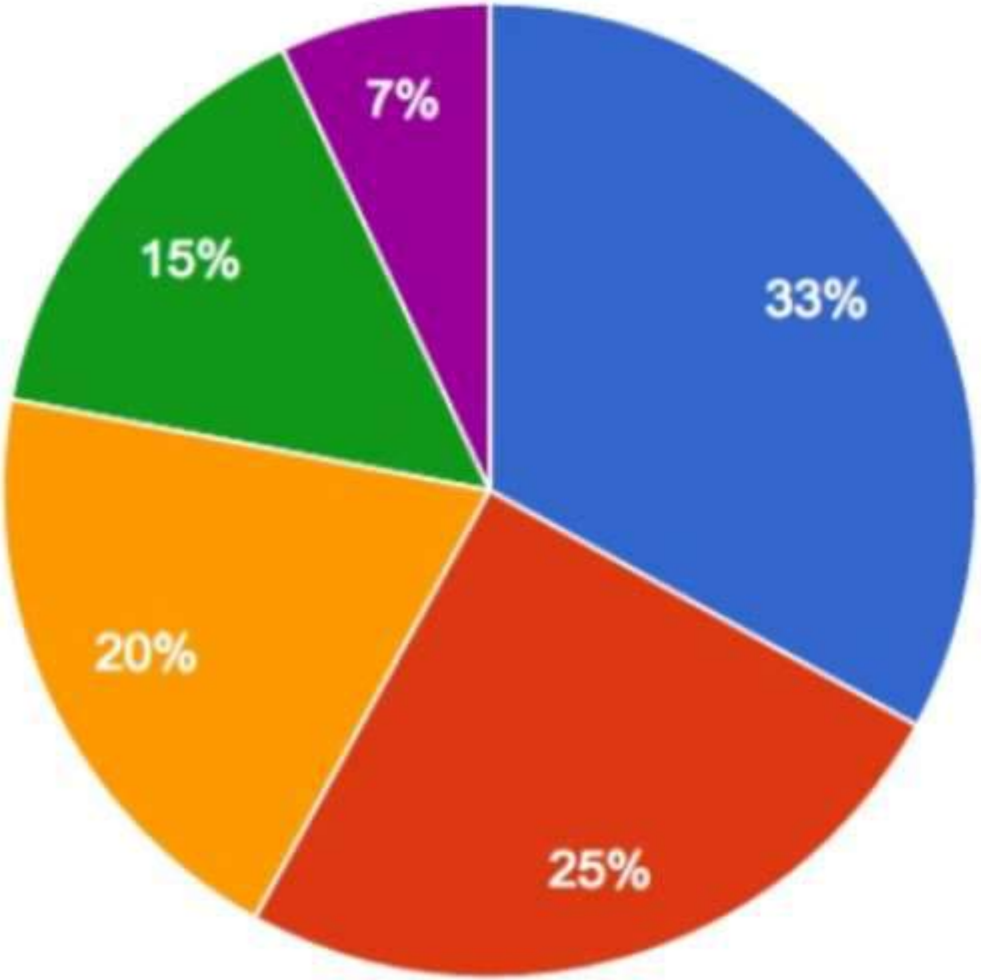
β_k
Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
$z_{d,n}$
θ_d
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996





"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Multinominal distribution



A Pie Chart!

The multinomial distribution is when there are **multiple** identical independent trials where each trial has k possible outcomes.

The categorical distribution is when there is **only one** such trial.

Development of LDA

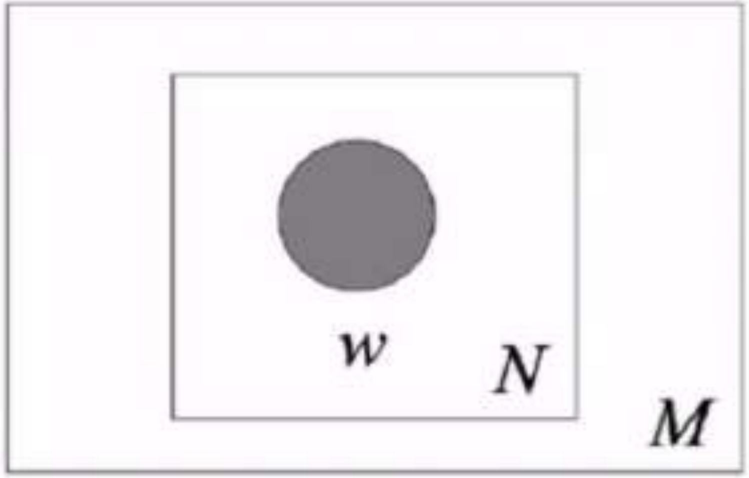
Uni-Gram

PLSA

Mixture unigram

LDA

Unigram Model



- the words of every document are drawn independently from a single multinomial distribution

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

Grasp the main idea

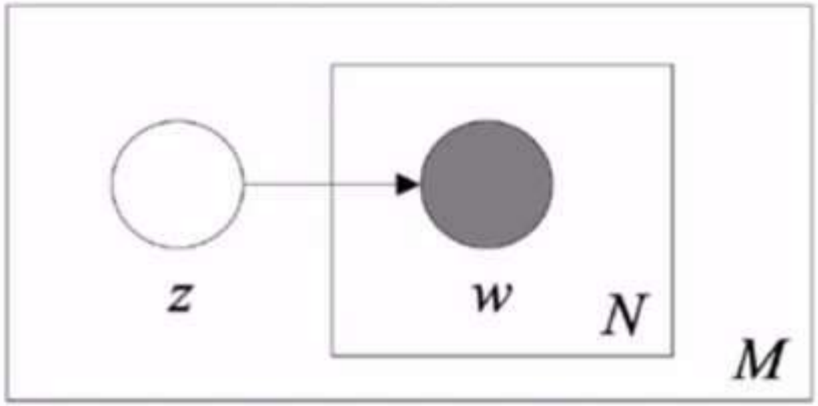
Chimpanzee seated at a typewriter

Infinite monkey theorem

From Wikipedia, the free encyclopedia

The **infinite monkey theorem** states that a monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type any given text, such as the complete works of William Shakespeare. In fact, the monkey would almost

Mixture of Unigrams Model



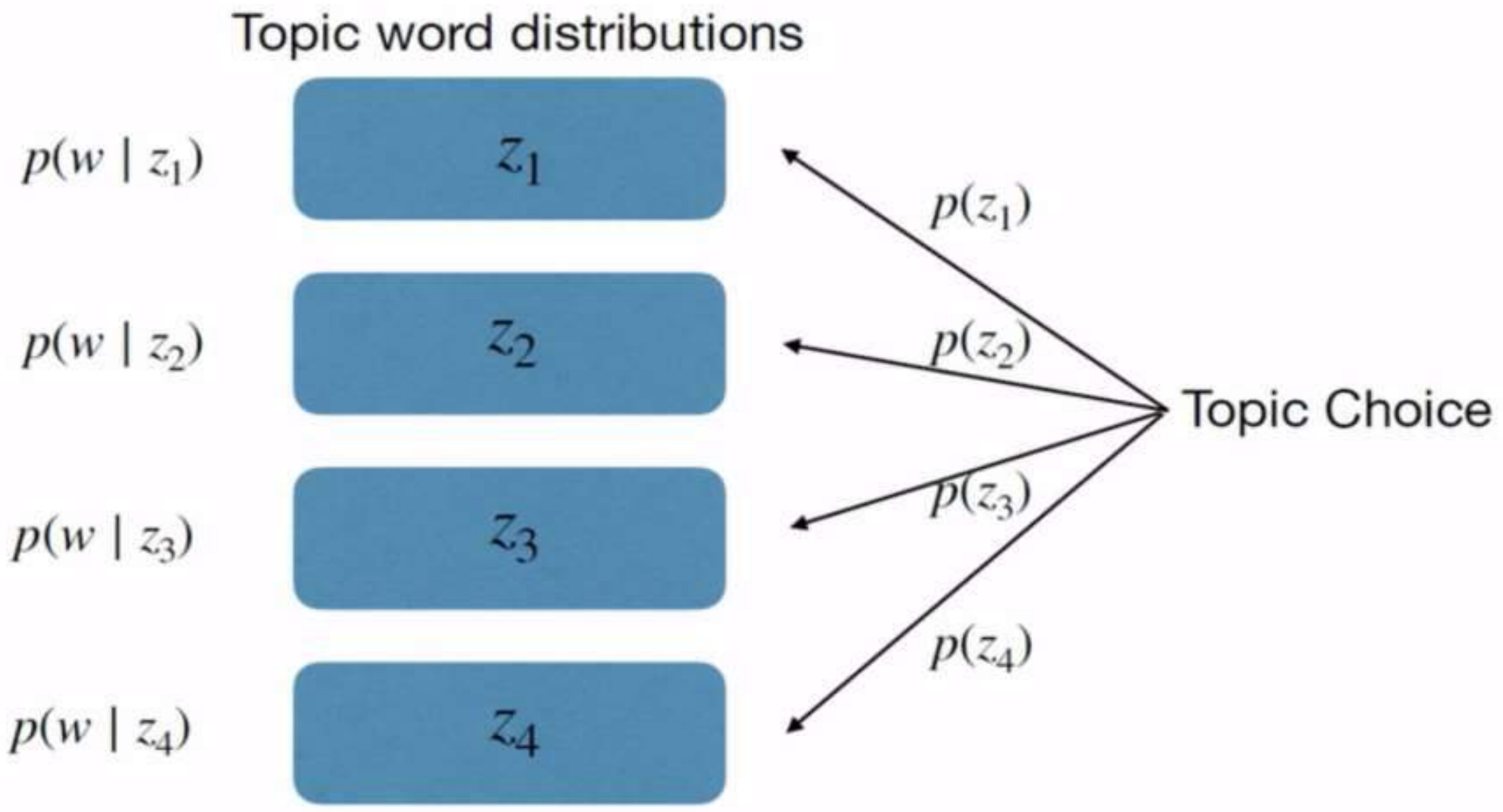
- Argument the unigram model with a discrete random topic variables z and obtain a mixture of unigrams model.

- Each document is generated by first choosing a topic z and then generating words independently from the conditional multinomial

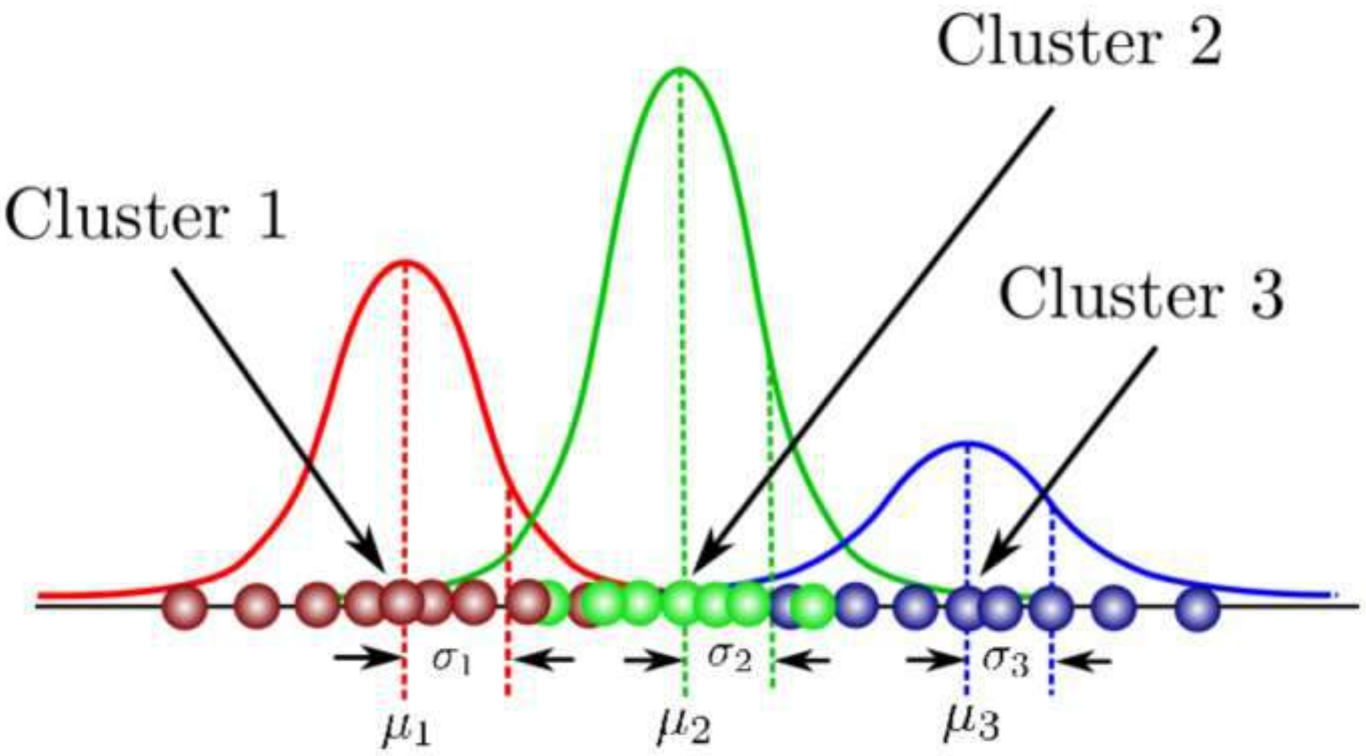
$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

Mixture of Unigram

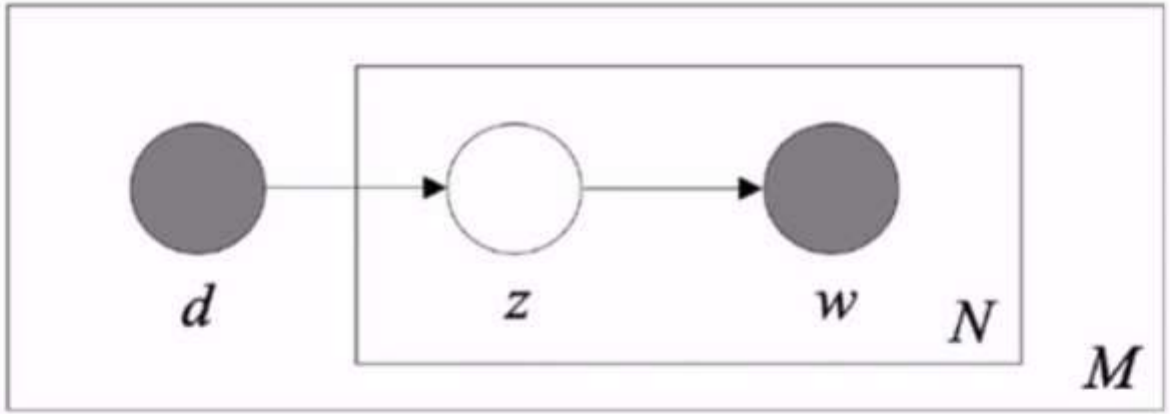
$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^{N} p(w_n \mid z).$$



Mixture of Gaussian process



pLSI



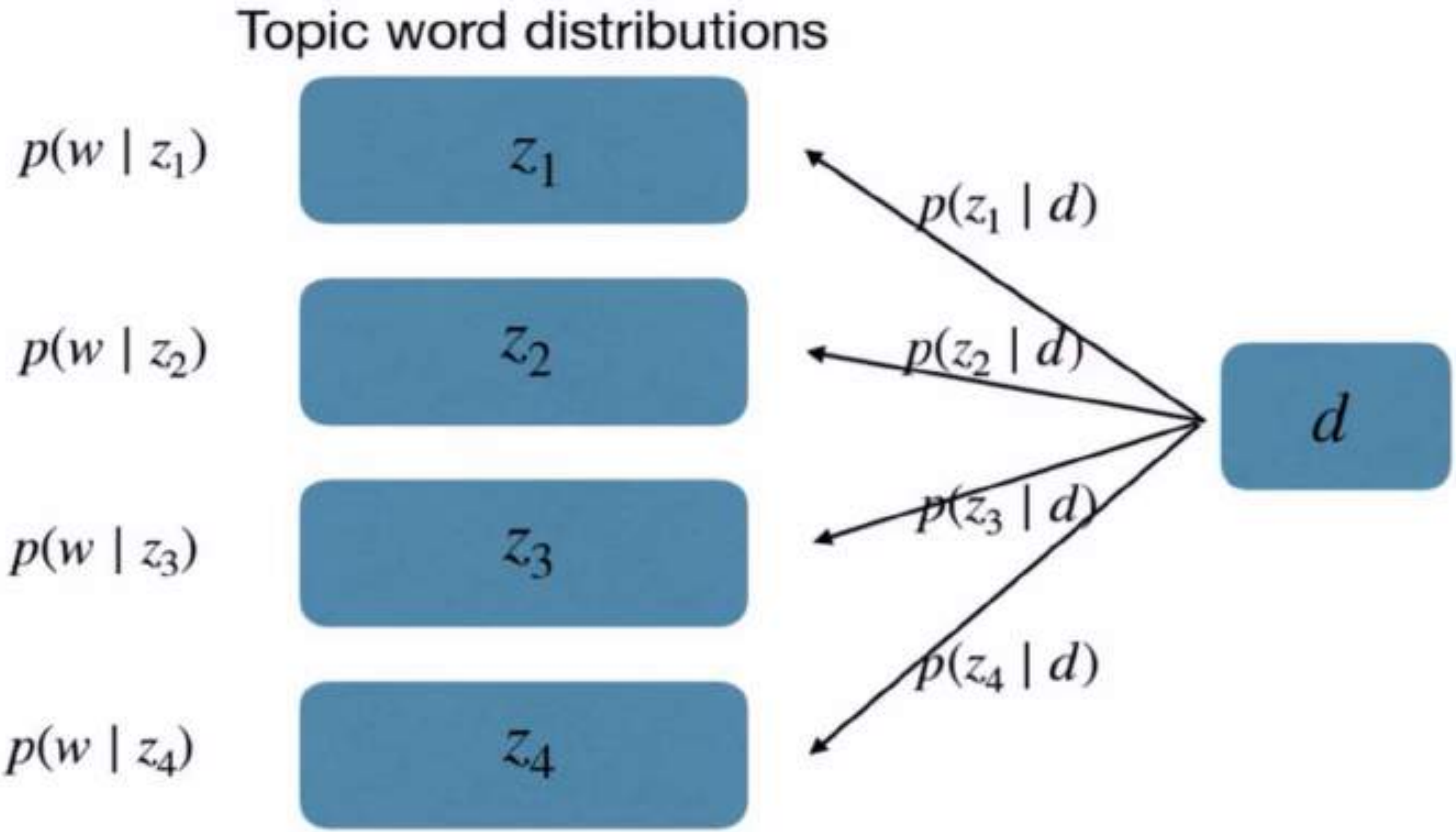
- The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic.

- Given all parameters, we want to infer the distribution z a word is from

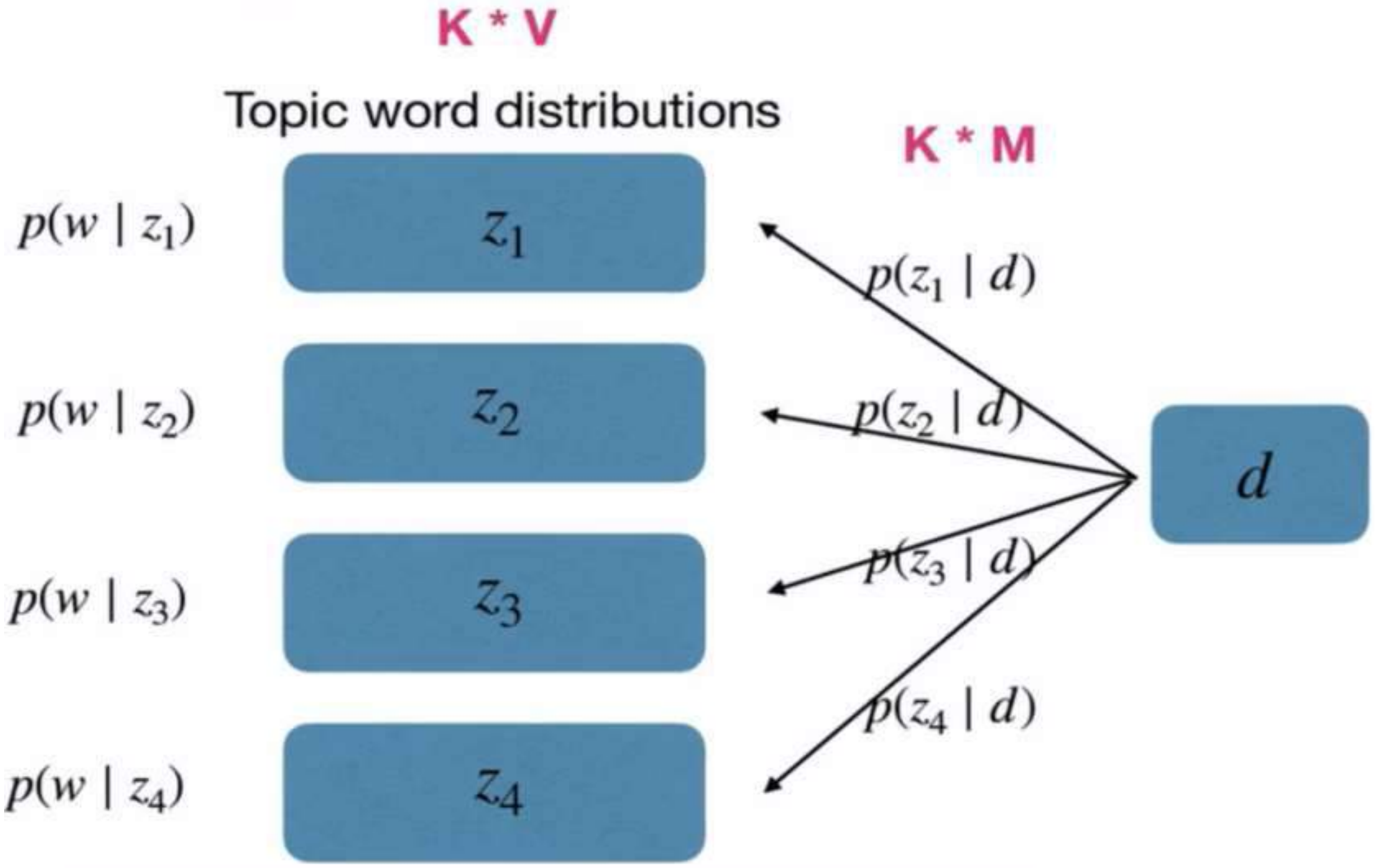
$$p(d, w_n) = p(d) \sum_{z} p(w_n \mid z) p(z \mid d).$$

PLSA

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d).$$



Restrains of PLSA



$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d).$$

1 Restrains of PLSA

for a particular document d. However, it is important to note that d is a dummy index into the list of documents in the *training set*. Thus, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of

documents on which it is trained. For this reason, pLSI is not a well-defined generative model of
documents; there is no natural way to use it to assign probability to a previously unseen document.

number of training documents. The parameters for a k-topic pLSI model are k multinomial distributions of size V and M mixtures over the k hidden topics. This gives $kV + kM$ parameters and

1 Dirichlet

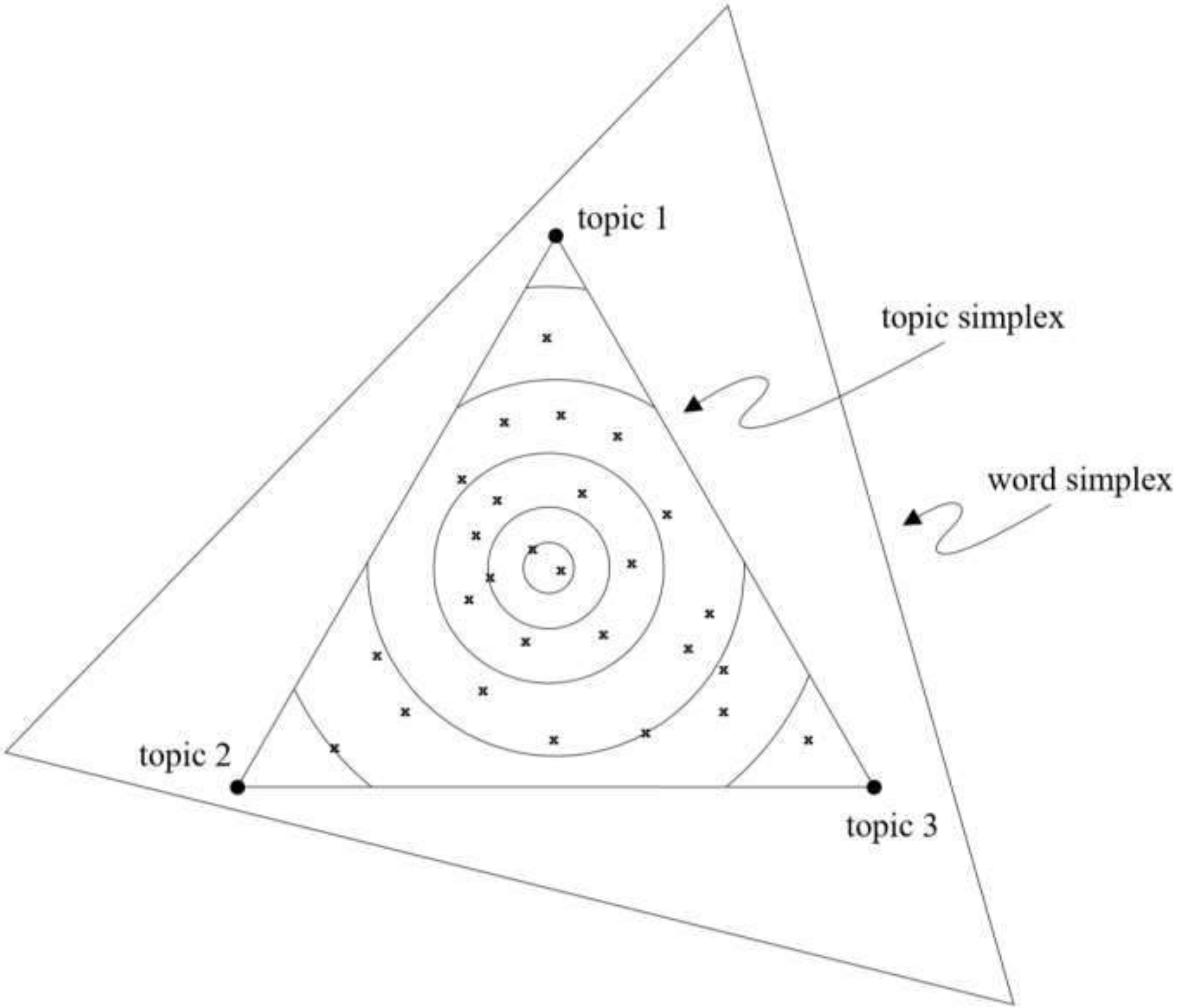
$$f(\mathbf{x}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_i\right)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

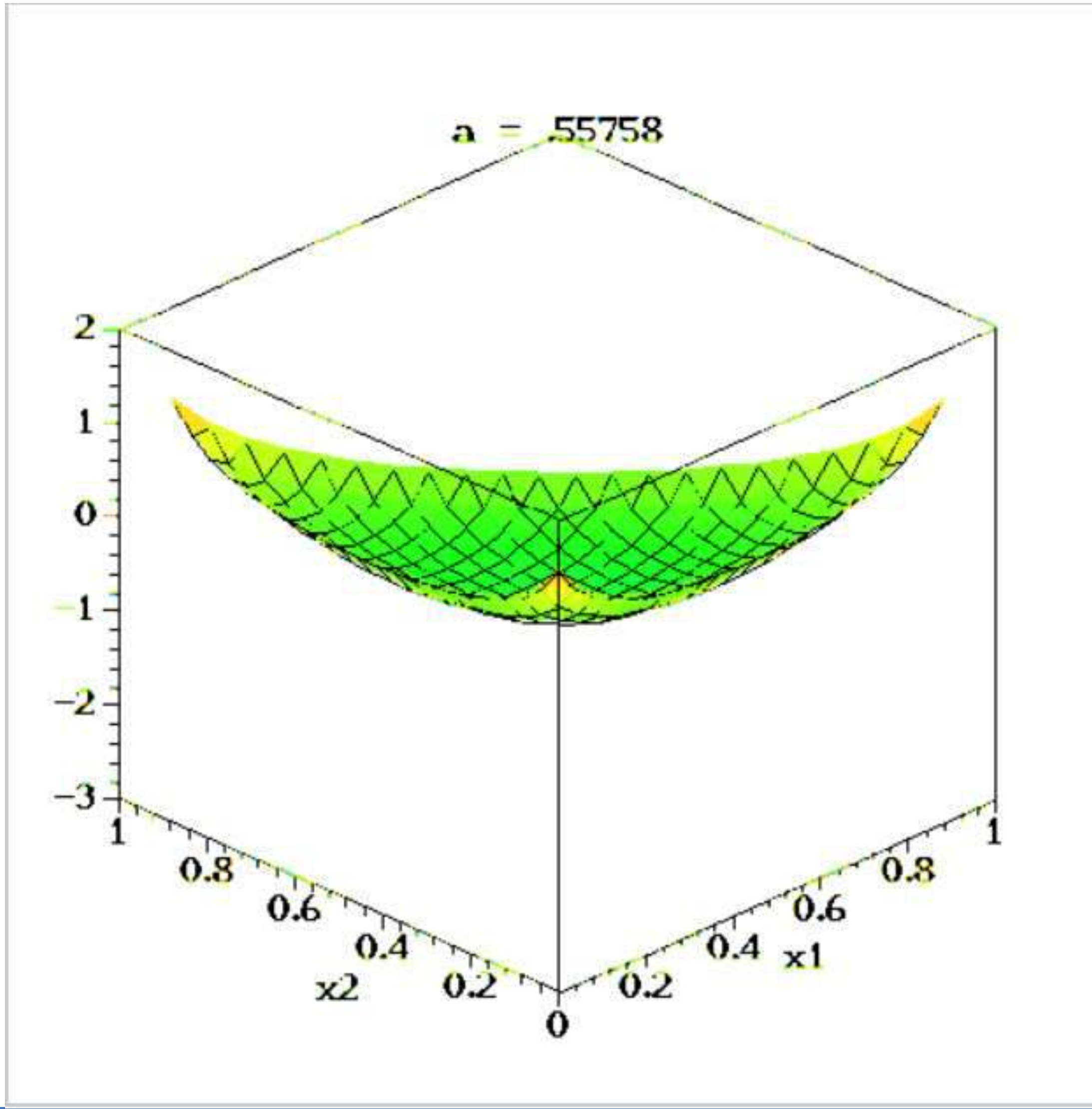
- The Dirichlet distribution is a generalization of the Beta distribution for multiple random variables
- The Dirichlet distribution is over vectors whose values are all in the interval $[0, 1]$ and the sum of values in the vector is 1.

1 Insight

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

Dirichlet distributions are just a convenient family for representing distributions over the simplex (the set of N-vectors whose components sum to 1), so it's a useful prior distribution on discrete probability distributions over categorical variables -- in fact it is the conjugate prior to the categorical and multinomial distributions (meaning that multiplying a Dirichlet prior by a multinomial or categorical likelihood will yield another Dirichlet distribution of a certain form). Whether the concentration parameter is below or above 1 controls whether sparse categorical distributions are preferred. The assumptions made are fairly weak, and similar to those of any continuous density; it mostly depends on the shape and concentration parameters selected.


topic 1
topic simplex
word simplex
topic 2
topic 3



a = .55758
2
1
0
−1
−2
−3
1
0.8
0.6
0.4
0.2
x2
0
0.2
0.4
0.6
0.8
1
x1

Insight

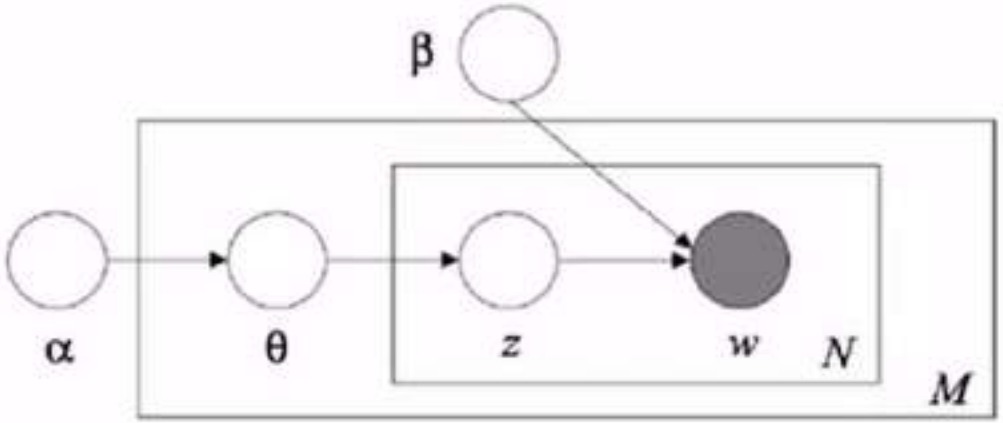
Beta distribution gives us a single probability,

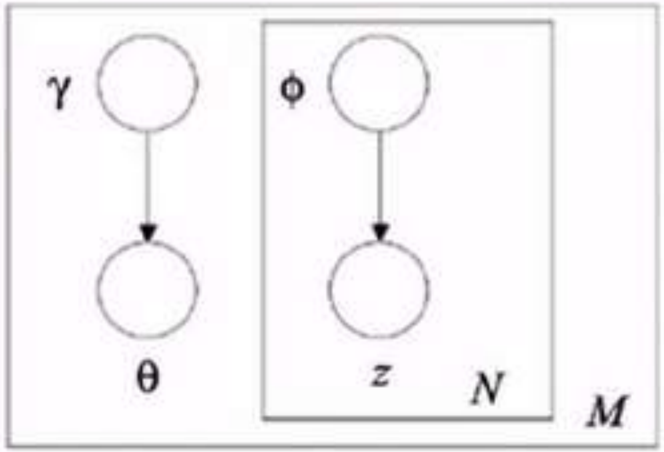
The Dirichlet distribution gives us K probabilities that
Define a probability distribution over K-d vector that sum
To one

Gibbs Sampling & Variational inference

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document

Unfortunately, this distribution is intractable to compute in general





北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Algorithm

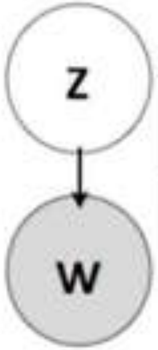
1. For each iteration i:
   1.1 For each document d and word n currently assigned to z_{old}:
      1.1.1 Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
      1.1.2 Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
      1.1.3 Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

$$p(z_{d,n}=k|\vec{z}_{-d,n},\vec{w},\alpha,\lambda)=\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i}\frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$$

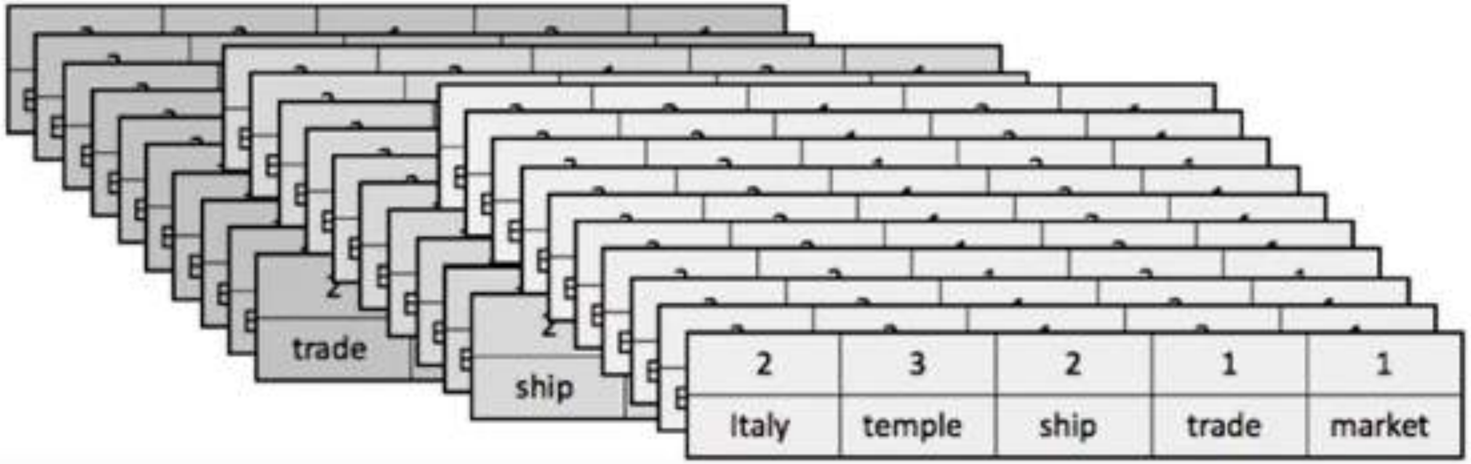
- Number of times document *d* uses topic *k*
- Number of times topic *k* uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic *k*
- How much this topic likes word $w_{d,n}$

Random Assignment



3	2	1	3	1
Etruscan	trade	price	temple	market

m



Total topic counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Sample this word

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**8**	1
...			

Decrement its count

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			

How much does this document like each topic

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

$$\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i}\frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$$

Total topic counts

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

How much each topic like the word

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

	1	2	3
trade	10	7	1

$$\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i}\frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$$

Geometric interpretation

Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
...			

Algorithm

1. For each iteration i:

   1.1 For each document d and word n currently assigned to z_{old}:

      1.1.1 Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$

      1.1.2 Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$

      1.1.3 Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

topic modeling


```
# Create a new document-term matrix using only nouns and adjectives, also remove common words with max_df
cvna = CountVectorizer(stop_words=stop_words, max_df=.8)
data_cvna = cvna.fit_transform(data_nouns_adj.transcript)
data_dtmna = pd.DataFrame(data_cvna.toarray(), columns=cvna.get_feature_names())
data_dtmna.index = data_nouns_adj.index
data_dtmna
```

```
# Create the gensim corpus
corpusna = matutils.Sparse2Corpus(scipy.sparse.csr_matrix(data_dtmna.transpose()))

# Create the vocabulary dictionary
id2wordna = dict((v, k) for k, v in cvna.vocabulary_.items())
```

```
# Let's start with 2 topics
ldana = models.LdaModel(corpus=corpusna, num_topics=2, id2word=id2wordna, passes=10)
ldana.print_topics()
```

Output of LDA


```
[(0,
  '0.009*"joke" + 0.005*"mom" + 0.005*"parents" + 0.004*"hasan" + 0.004*"jokes" + 0.004*"anthon
y" + 0.003*"nuts" + 0.003*"dead" + 0.003*"tit" + 0.003*"twitter"'),
 (1,
  '0.005*"mom" + 0.005*"jenny" + 0.005*"clinton" + 0.004*"friend" + 0.004*"parents" + 0.003*"hu
sband" + 0.003*"cow" + 0.003*"ok" + 0.003*"wife" + 0.003*"john"'),
 (2,
  '0.005*"bo" + 0.005*"gun" + 0.005*"guns" + 0.005*"repeat" + 0.004*"um" + 0.004*"ass" + 0.004
*"eye" + 0.004*"contact" + 0.003*"son" + 0.003*"class"'),
 (3,
  '0.006*"ahah" + 0.004*"nigga" + 0.004*"gay" + 0.003*"dick" + 0.003*"door" + 0.003*"young" +
0.003*"motherfucker" + 0.003*"stupid" + 0.003*"bitch" + 0.003*"mad"')]
```

These four topics look pretty decent. Let's settle on these for now.

- Topic 0: mom, parents
- Topic 1: husband, wife
- Topic 2: guns
- Topic 3: profanity

Clustering

K-means

Decision Tree

Naïve Bayes

SVM

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

THANKS

德以明理 学以精工