W4995 Applied Machine Learning

LSA & Topic Models

04/05/17

Andreas Müller

Beyond Bags of Words

Limitations of bag of words:

- Semantics of words not captured
- Synonymous words not represented
- Very distributed representation of documents

Latent Semantic Analysis (LSA)

- Reduce dimensionality of data.
- Can't use PCA: can't subtract the mean (sparse data)
- Instead of PCA: Just do SVD, truncate.
- "Semantic" features, dense representation.
- Easy to compute – convex optimization

LSA with TruncatedSVD

```
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer(stop_words="english", min_df=4)
X_train = vect.fit_transform(text_train)
```

```
X_train.shape
```

```
(25000, 30462)
```

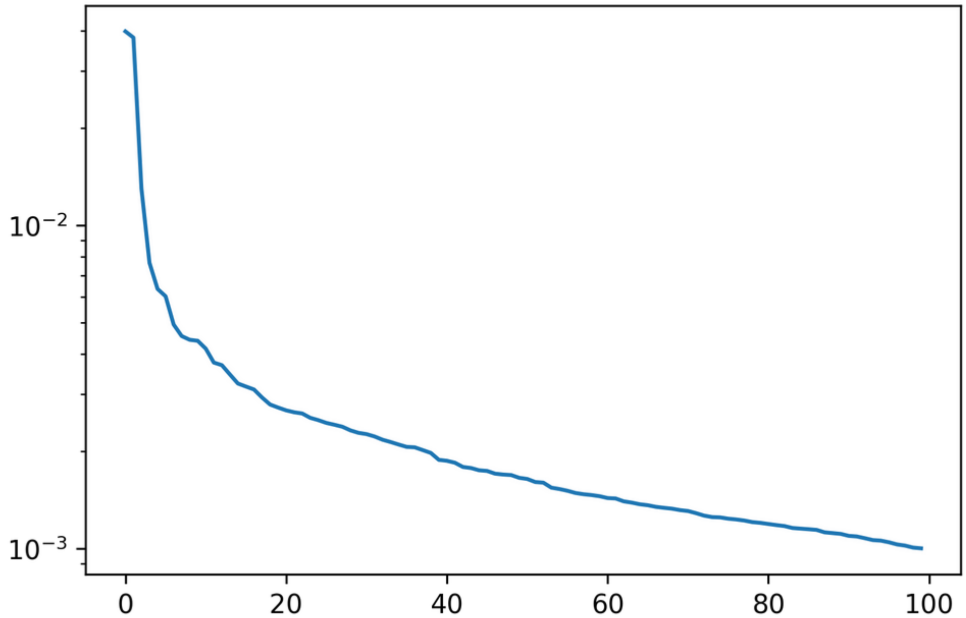
```
from sklearn.decomposition import TruncatedSVD
lsa = TruncatedSVD(n_components=100)
X_lsa = lsa.fit_transform(X_train)
```

```
lsa.components_.shape
```

```
(100, 30462)
```

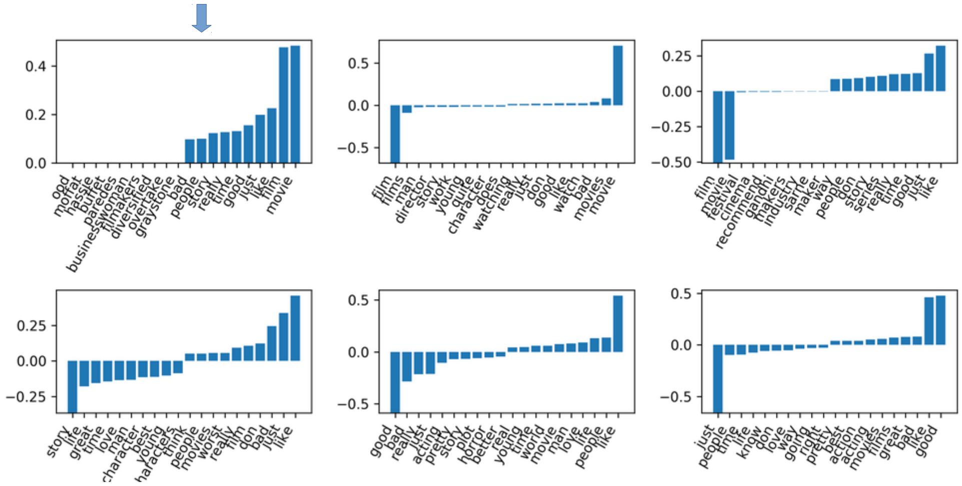
```
plt.semilogy(lsa.explained_variance_ratio_)
```

```
[<matplotlib.lines.Line2D at 0x7f55d1d4dd68>]
```



First Six eigenvectors

Points in direction of mean



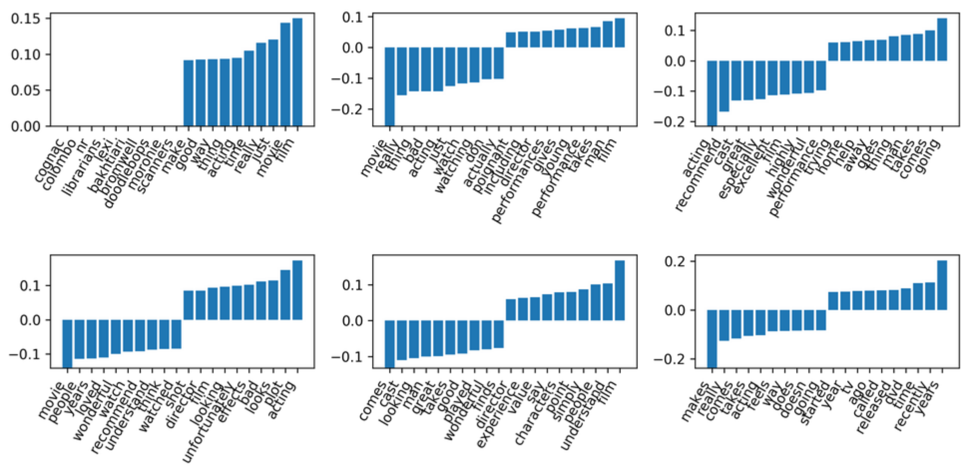
Scale before LSA

```
from sklearn.preprocessing import MaxAbsScaler
scaler = MaxAbsScaler()
X_scaled = scaler.fit_transform(X_train)

lsa_scaled = TruncatedSVD(n_components=100)
X_lsa_scaled = lsa_scaled.fit_transform(X_scaled)
```

"Movie" and "Film" was dominating first couple of components.
Try to get rid of that effect.

Eigenvectors after scaling

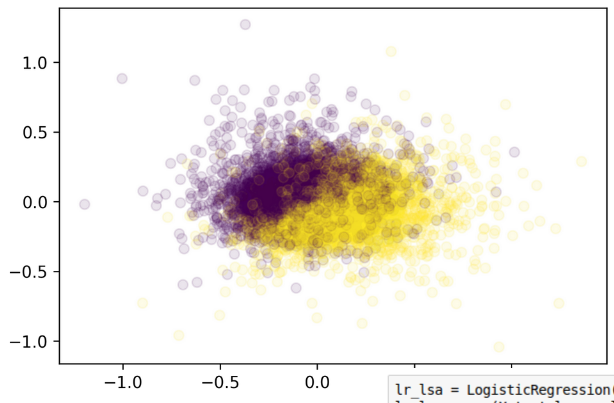


Movie and film still important, but not that dominant any more.

Some Components Capture Sentiment

```
plt.scatter(X_lsa_scaled[:, 1], X_lsa_scaled[:, 3], alpha=.1, c=y_train)
```

```
<matplotlib.collections.PathCollection at 0x7f55ca2bbb00>
```




```
lr_lsa = LogisticRegression(C=100).fit(X_lsa_scaled[:, :10], y_train)
lr_lsa.score(X_test_lsa_scaled[:, :10], y_test)
```

```
0.82711999999999997
```

```
lr_lsa.score(X_lsa_scaled[:, :10], y_train)
```

```
0.82808000000000004
```

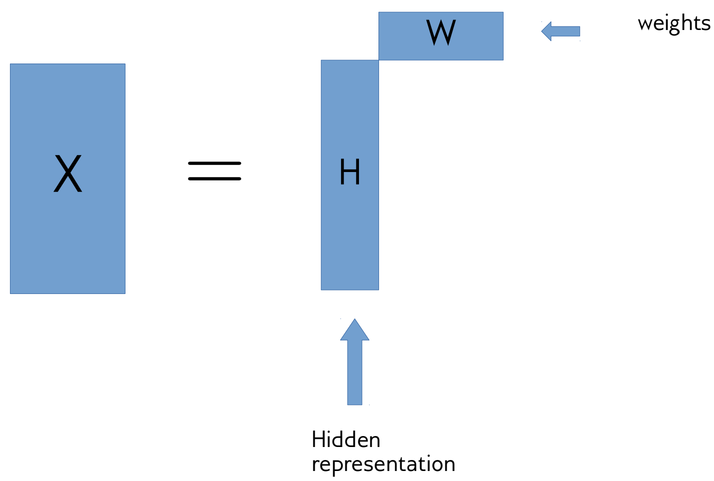
Not competitive but reasonable with just 10 components!

Topic Models

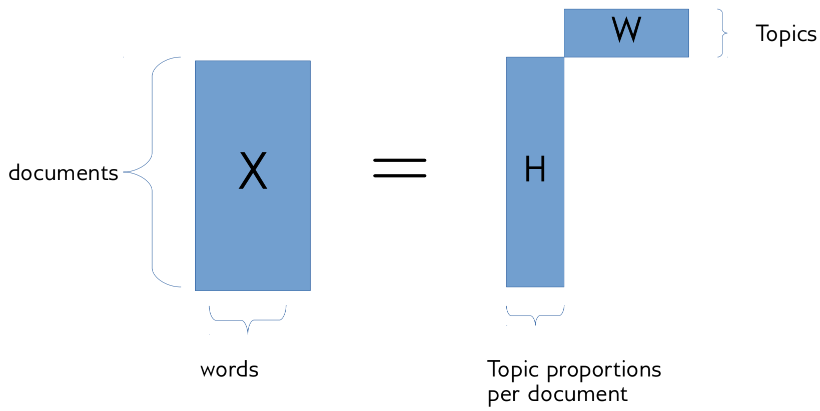
Motivation

- Each document is created as a mixture of topics
- Topics are distributions over words
- Learn topics and composition of documents simultaneously
- Unsupervised (and possibly ill-defined)

NMF for topic models



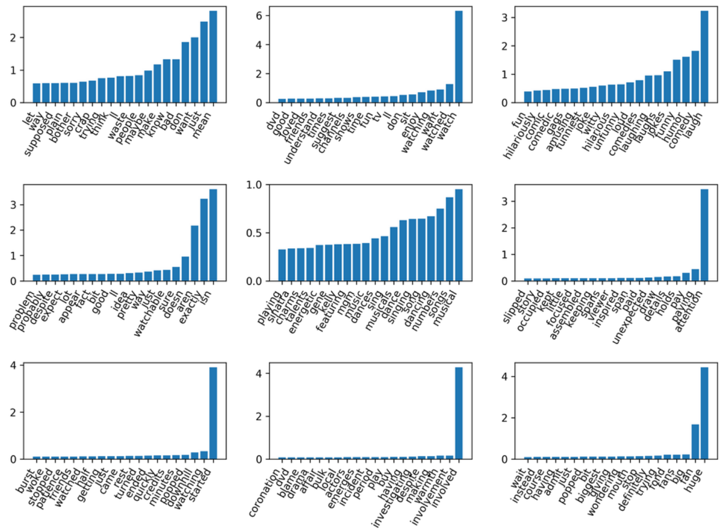
NMF for topic models



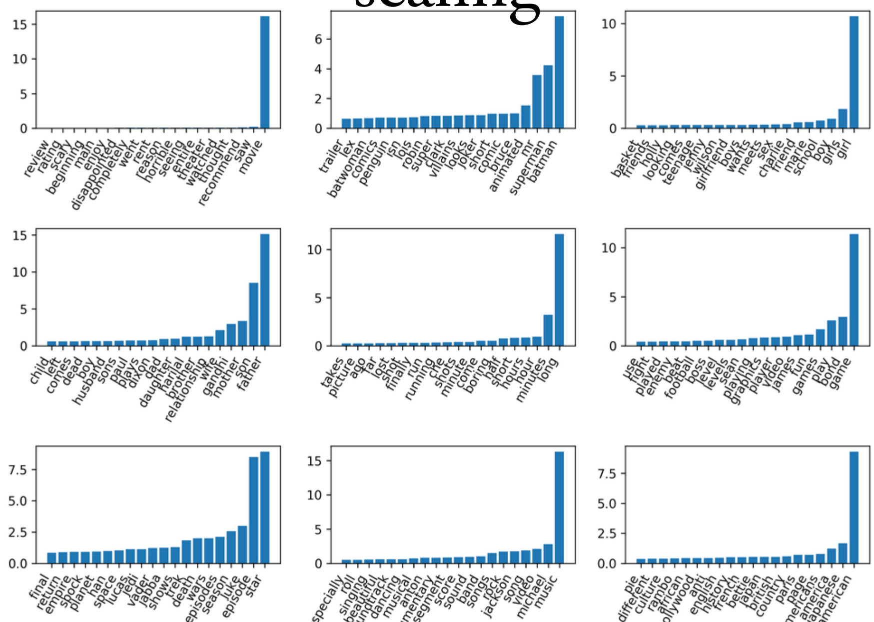
Each row of W corresponds to one “topic”

NMF on Scaled Data

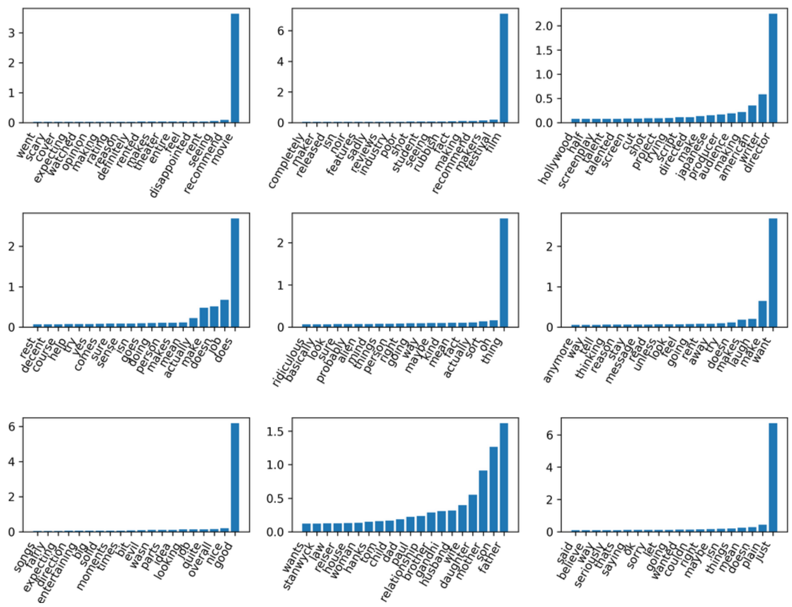
```
from sklearn.decomposition import NMF
nmf = NMF(n_components=100, verbose=10, tol=0.01)
nmf.fit(X_scaled)
```



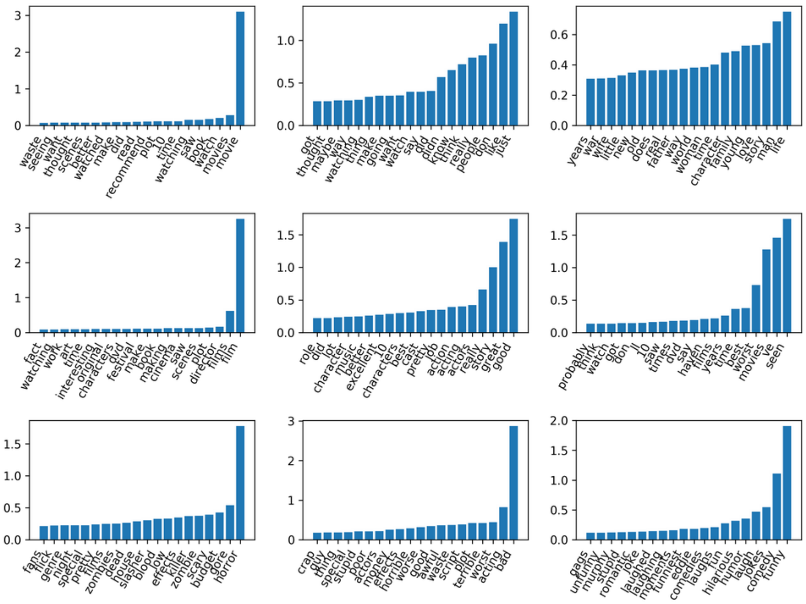
NMF components without scaling



NMF with tfidf



NMF with tfidf and 10 components

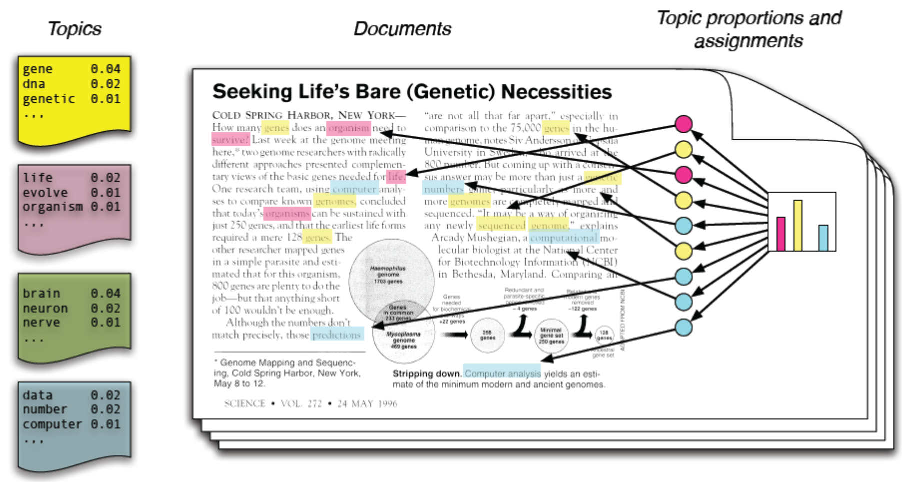


Latent Dirichlet Allocation (the other LDA)

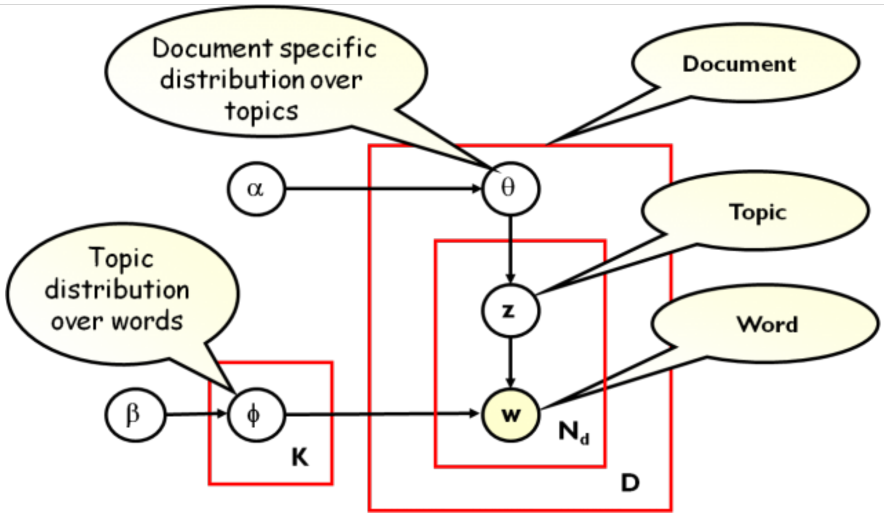
LDA motivation

- Generative probabilistic model (similar to mixture model)
- Bayesian graphical model
- Learning is probabilistic inference
- Non-convex optimization (even harder than mixture models)

The LDA Model



(Stolen from Dave and John)



1. For each topic k, draw $\beta_k \sim Dirichlet(\eta),\ k = 1...K$
2. For each document d, draw $\theta_d \sim Dirichlet(\alpha),\ d = 1...D$
3. For each word i in document d:

    a. Draw a topic index $z_{di} \sim Multinomial(\theta_d)$
    b. Draw the observed word $w_{ij} \sim Multinomial(beta_{z_{di}}.)$

(taken from Yang Ruan, Changsi An http://salsahpc.indiana.edu/b649proj/proj3.html)

Estimated Parameters

- K topics = multinomial distributions over words
- “mixture weights” for each document:
  - How important is each topic for this document
  - Each document contains multiple topics!

Two Schools (of solvers)

Gibbs sampling

- Implements MCMC
- Standard procedure for any probabilistic model.
- Very accurate
- Very slow

Variational Inference

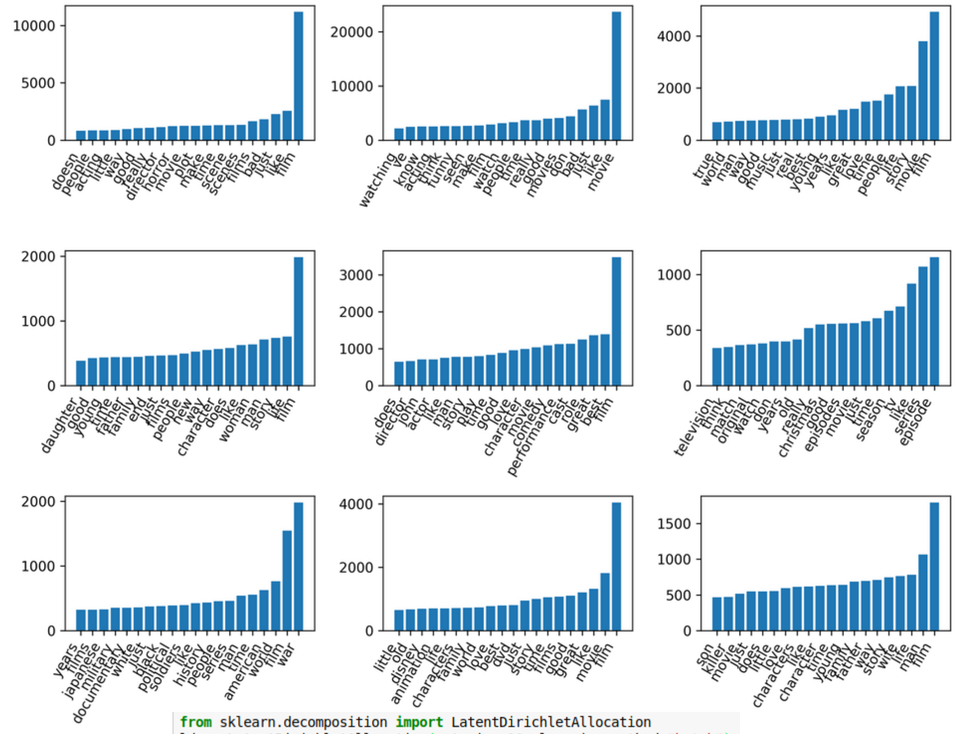
- Extension of expectation-maximization algorithm
- Deterministic
- fast(er)
- Less accurate solutions
- Championed by Dave Blei

Pick a solver

- “Small data” (<= 10k? Documents):
  - Gibbs sampling (lda package, MALLET in Java!)
- “Medium data” (<= 1M? Documents):
  - Variational Inference (scikit-learn current default)
- “Large Data” (>1M? Documents):
  - Stochastic Variational Inference (scikit-learn future default)
  - SVI allows online learning (partial_fit)

- Remember SGD Lecture (and Leon Bottou):
  More data beats better inference (often)

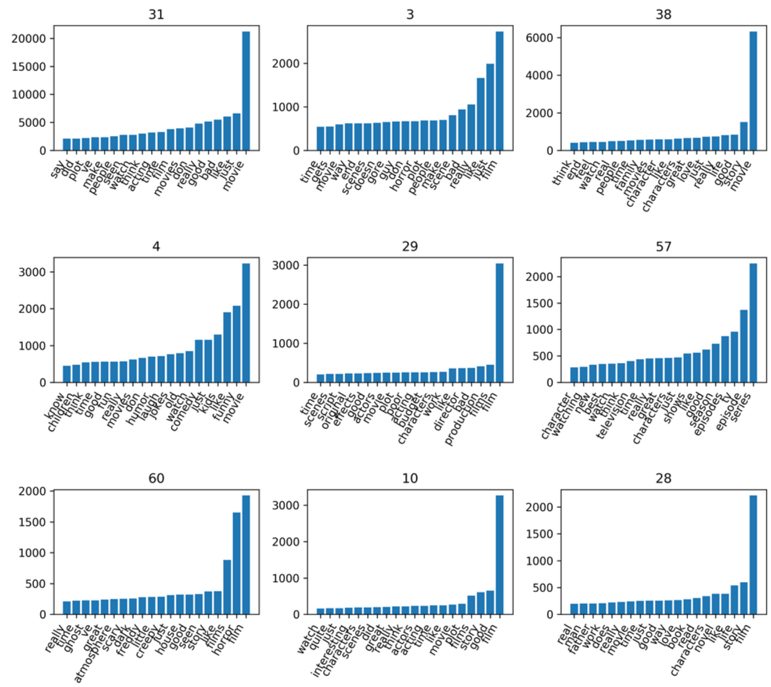
- Edward by Dustin Tran: http://edwardlib.org/
  Tensor-flow based framework for stochastic variational inference.




```
from sklearn.decomposition import LatentDirichletAllocation
lda = LatentDirichletAllocation(n_topics=10, learning_method="batch")
X_lda = lda.fit_transform(X_train)
```

Very generic, similar to NMF(n_components=10).
TV Series, “family drama”, and “history / war” topics

```
lda100 = LatentDirichletAllocation(n_topics=100, learning_method="batch")
X_lda100 = lda100.fit_transform(X_train)
```

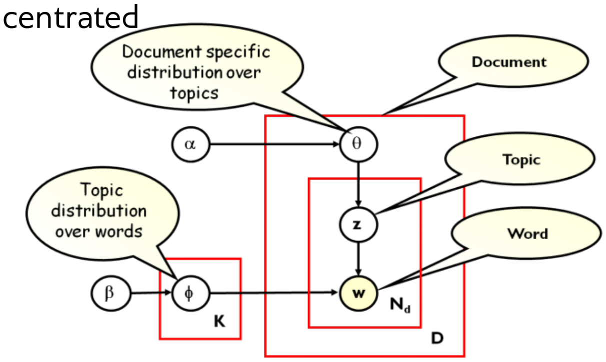


topic 31	topic 3	topic 38	topic 4	topic 29	topic 57	topic 60	topic 10
movie	film	movie	movie	film	series	film	film
just	just	story	funny	films	episode	horror	good
like	like	good	like	production	tv	films	story
bad	really	life	kids	bad	episodes	like	films
good	bad	really	just	director	season	story	plot
really	scene	just	comedy	like	good	seen	movie
don	make	love	watch	work	like	good	like
movies	people	great	old	characters	shows	house	time
film	plot	characters	jokes	budget	just	just	acting
time	horror	like	laugh	acting	characters	creepy	actors
acting	don	character	humor	poor	great	little	think
think	guy	movies	don	plot	really	freddy	really
watch	gore	family	movies	movie	time	dark	great
seen	doesn	time	really	actors	television	scary	did
people	scenes	people	fun	good	think	atmosphere	scenes
make	end	real	good	effects	watch	great	characters
ve	way	watch	time	original	best	ve	interesting
plot	movie	feel	think	script	new	ghost	just
did	gets	end	children	scenes	watching	time	quite
say	time	think	know	time	character	really	watch

topic 28	topic 48	topic 2	topic 69	topic 99	topic 62	topic 22	topic 12
film	film	movie	music	film	movie	action	film
story	love	good	film	story	god	film	killer
life	people	action	musical	time	guy	fight	movie
like	story	bad	songs	films	bad	martial	horror
novel	movie	like	song	life	like	arts	like
characters	life	watch	dance	man	good	fu	halloween
read	just	just	rock	like	just	scenes	good
book	characters	time	singing	young	funny	movie	slasher
love	like	film	band	love	know	kong	just
way	time	guy	dancing	work	does	kung	night
good	character	10	best	just	bruce	fighting	story
just	way	really	great	way	did	jackie	man
time	young	fun	numbers	woman	time	chan	scene
movie	great	want	time	characters	little	like	know
really	little	movies	number	beautiful	make	movies	people
does	world	make	kelly	husband	gets	hong	carpenter
work	beautiful	don	story	scene	scene	films	michael
father	real	plot	sing	new	really	lee	didn
man	feel	acting	musicals	director	rap	best	john
real	kelly	way	stage	character	half	good	time

Hyper-Parameters

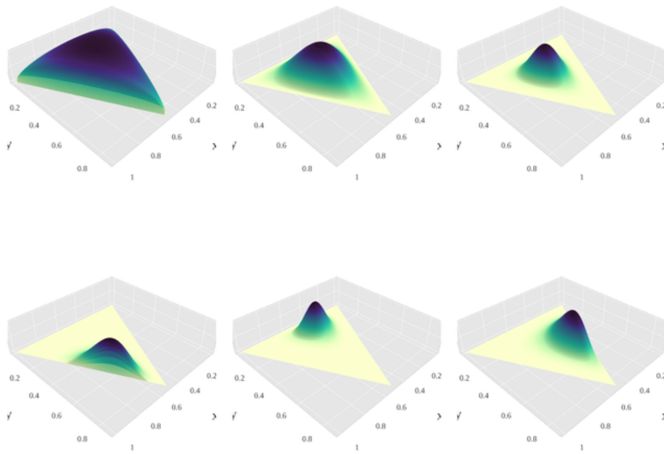
- α (or θ)= doc_topic_prior
- β (or η) = topic_word_prior
- Both dirichlet distributions
- Large value → more dispersed
- Small value → more concentrated



Dirichlet Distribution

PDF	$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

Mean	$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$



Conjugate Prior

- Prior is called "conjugate" if the posterior has the same form as prior.

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{\int p(x \mid \theta')\, p(\theta')\, d\theta'}.$$

- If $p(x|\theta)$ is multinomial (discrete distribution), then $p(\theta)$ = Dirichlet(...) is a conjugate prior.

Multinomial:

pmf	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

Dirichlet

PDF	$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

Further Reading

- Rethinking LDA: Why Priors Matter - Hanna Wallach
- LDA Revisited: Entropy, Prior and Convergence – Zhang et. al.

Homework IV

The task is to do text classification on a dataset of complaints about traffic conditions to the city of Boston. You can find the data here: https://data.boston.gov/dataset/vision-zero-entry

There are two goals:

- First, try to predict the type of complaint ("REQUESTTYPE") from the complaint text.
- Second, try to come up with a better categorization of the data into semantic categories.

_id	X	Y	OBJECT...	GLOBA...	REQUE...	REQUE...	REQUE...	STATUS	*STREET...*	COMMENTS	USERTY...
2	-71.0722...	42.3326...	13608		13608	it's too fa...	2016-01...	Unassig...	0		walks
3	-71.0930...	42.3498...	13609		13609	bike facil...	2016-01...	Unassig...	0	I feel scared biking ...	bikes
4	-71.0915...	42.3491...	13610		13610	bike facil...	2016-01...	Unassig...	0	While I love that the...	bikes
5	-71.0674...	42.3523...	13611		13611	bike facil...	2016-01...	Unassig...	0	Need a bike facility t...	bikes
6	-71.0692...	42.3450...	13612		13612	people s...	2016-01...	Unassig...	0	3 lane, no parking e...	walks
7	-71.0773...	42.3500...	13613		13613	people r...	2016-01...	Unassig...	0	People who are wal...	bikes
8	-71.0953...	42.3315...	14007		14007	people c...	2016-01...	Unassig...	0		travels (...
9	-71.0721...	42.3326...	14008		14008	people r...	2016-01...	Unassig...	0		drives
10	-71.0709...	42.3316...	14009		14009	bike facil...	2016-01...	Unassig...	0		bikes
11	-71.0766...	42.3488...	14010		14010	people d...	2016-01...	Unassig...	0		bikes
12	-71.1041...	42.3169...	14011		14011	people c...	2016-01...	Unassig...	0	The SWC path has ...	walks
13	-71.1098...	42.3220...	14012		14012	people d...	2016-01...	Unassig...	0	People driving out o...	walks
14	-71.1115...	42.3209...	14013		14013	bike facil...	2016-01...	Unassig...	0	An "except bikes" u...	bikes
15	-71.0881...	42.3361...	14014		14014	bike facil...	2016-01...	Unassig...	0	Where is the south...	bikes
16	-71.0895...	42.3450...	14015		14015	bike facil...	2016-01...	Unassig...	0	Hemenway from Bo...	bikes
17	-71.0933...	42.3498...	14016		14016	it's too fa...	2016-01...	Unassig...	0	Huge wide open int...	walks
18	-71.0725...	42.3554...	14017		14017	people d...	2016-01...	Unassig...	0	Cars from Storrow ...	walks
19	-71.0647...	42.3436...	14018		14018	of somet...	2016-01...	Unassig...	0	Always traffic here. ...	drives
20	-71.1025...	42.3433...	14019		14019	it's too fa...	2016-01...	Unassig...	0	It feels like it will tak...	walks
21	-71.0758...	42.3439...	14020		14020	people r...	2016-01...	Unassig...	0	You think they're goi...	walks
102	-71.0638...	42.3204...	17265		17265	people s...	2016-01...	Unassig...	0		walks
22	-71.0943...	42.3471...	14021		14021	it's too fa...	2016-01...	Unassig...	0	The street is one la...	walks

of something that is not listed here	1418
bike facilities don't exist or need improvement	782
people speed	737
people run red lights / stop signs	660
people don't yield while turning	461
people double park their vehicles	426
it’s hard to see / low visibility	384
sidewalks/ramps don't exist or need improvement	301
people don't yield while going straight	263
people cross away from the crosswalks	254
the roadway surface needs improvement	221
the wait for the "Walk" signal is too long	204
there are no bike facilities or they need maintenance	128
there's not enough time to cross the street	121
it's too far / too many lanes to cross	83
there are no sidewalks or they need maintenance	40
the roadway surface needs maintenance	34
people have to wait too long for the "Walk" signal	30
it’s hard for people to see each other	28
people have to cross too many lanes / too far	27
people are not given enough time to cross the street	9