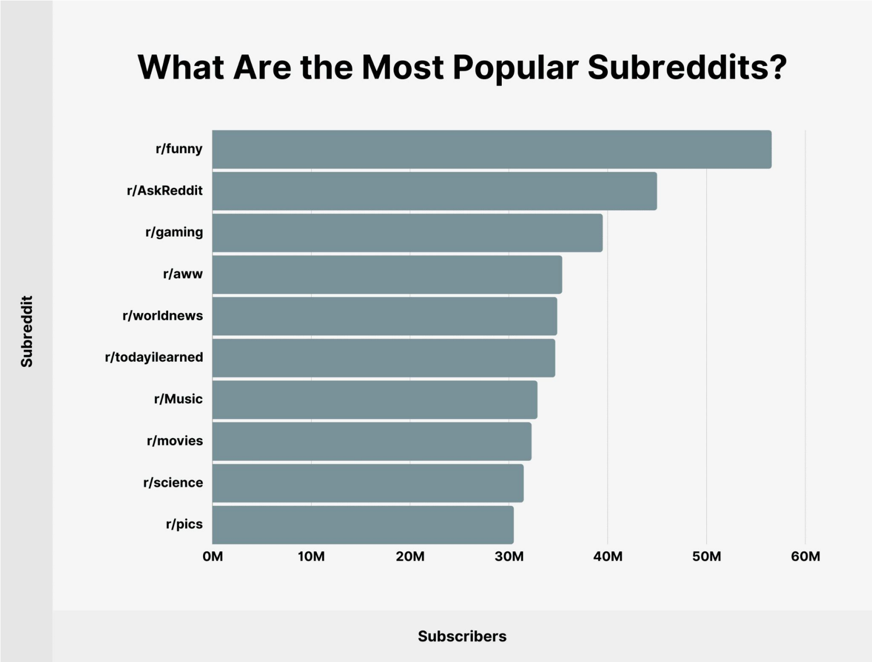
Topic Modeling

What are the major topics being discussed?



https://foundationinc.co/lab/reddit-statistics/

Let's manually assign a few topics...

Egypt's Dahab remains a secret paradise for adventure travelers.

With nowhere to go, she finally took the time to learn a musical instrument.

A visit to Budapest, Hungary offers something for everyone.

No one could recognize what he was playing, but he insisted it was Beethoven.

The Glastonbury music festival attracts visitors from around the world.

What topics are present in this collection and in what ratio?

Let's manually assign a few topics...

Egypt's Dahab remains a secret paradise for adventure travelers.	With nowhere to go, she finally took the time to learn a musical instrument.	A visit to Budapest, Hungary offers something for everyone.	No one could recognize what he was playing, but he insisted it was Beethoven.	The Glastonbury music festival attracts visitors from around the world.

Topic A: Travel

Topic B: Music

Yours may be different and reasonable. There's no single correct answer to this.

Let's manually assign a few topics...

Egypt's Dahab remains a secret paradise for adventure travelers.

Topic A: 100%

With nowhere to go, she finally took the time to learn a musical instrument.

Topic B: 100%

A visit to Budapest, Hungary offers something for everyone.

Topic A: 100%

No one could recognize what he was playing, but he insisted it was Beethoven.

Topic B: 100%

The Glastonbury music festival attracts visitors from around the world.

Topic A: 60%

Topic B: 40%

Topic A: Travel

Topic B: Music

We uncovered the *latent* (hidden) topics within this corpus. This is the goal with topic modelling.

Let’s manually assign a few topics...

Egypt’s Dahab remains a secret paradise for adventure travelers.	With nowhere to go, she finally took the time to learn a musical instrument.	A visit to Budapest, Hungary offers something for everyone.	No one could recognize what he was playing, but he insisted it was Beethoven.	The Glastonbury music festival attracts visitors from around the world.
Topic A: 100%	**Topic B: 100%**	**Topic A: 100%**	**Topic B: 100%**	**Topic A: 60%** **Topic B: 40%**

Topic A: Travel

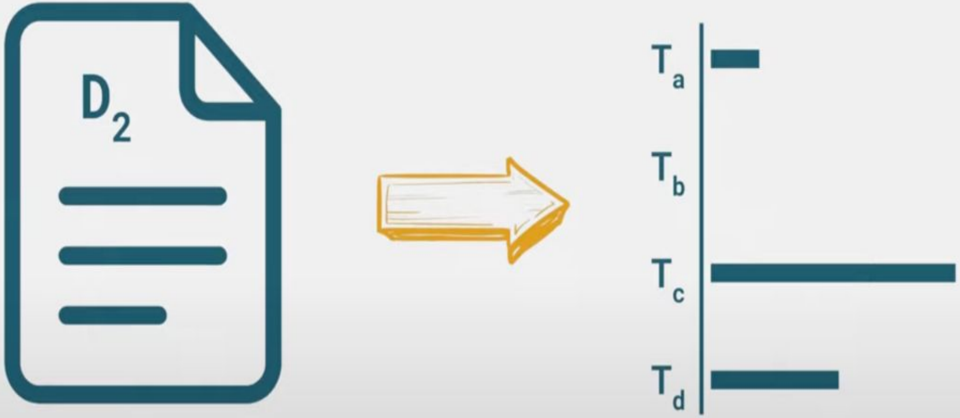
Topic B: Music

We, as humans, are really good at this.
Computers don’t have the same advantages.
But what if there are a million documents?

Latent Dirichlet Allocation (LDA)

Topic models assume two things

1. Every document is a mix of topics.



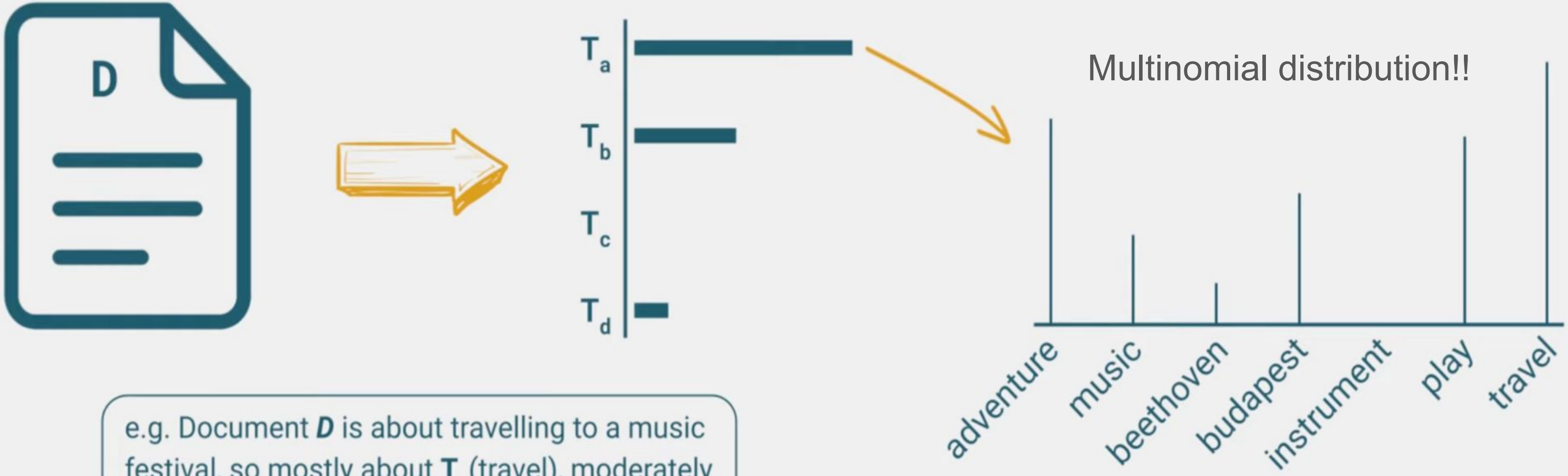
e.g. Document ***D*** is about travelling to a music festival, so mostly about $\mathbf{T_a}$(travel), moderately about $\mathbf{T_b}$(music), and maybe $\mathbf{T_d}$ is about food.

Latent Dirichlet Allocation (LDA)

Topic models assume two things

1 Every document is a mix of topics.

2 Every topic is a mix of words.



e.g. Document ***D*** is about travelling to a music festival, so mostly about $\mathbf{T_a}$(travel), moderately about $\mathbf{T_b}$(music), and maybe $\mathbf{T_d}$ is about food.

Latent Dirichlet Allocation (LDA)

1) Randomly assign a topic to *every* word in *every* document.

D_1	python	general	purpose	dynamic	language
D_2	wish	reciting	monty	python	outlawed
D_3	gather	round	for	bro	tales
. . .					

1) Randomly assign a topic to *every* word in *every* document.

D_1	python	T_1	general	T_3	purpose	T_2	dynamic	T_2	language	T_1
D_2	wish	T_3	reciting	T_2	monty	T_1	python	T_3	outlawed	T_2
D_3	gather	T_3	round	T_1	for	T_2	bro	T_1	tales	T_2
					. . .					

Latent Dirichlet Allocation (LDA)

2) Count the number of times topic *k* occurs in document *d*.

D_1	python T_1	general T_3	purpose T_2	dynamic T_2	language T_1

Topic counts for D_1	
T_1	2
T_2	2
T_3	1

3) Count the number of times every word appears under topic k across corpus.

D_1	python T_1	general T_3	purpose T_2	dynamic T_2	language T_1

Topic counts for D_1	
T_1	2
T_2	2
T_3	1

Word-topic counts for entire corpus

	T_1	T_2	T_3
python	3	0	0
general	22	7	10
purpose	13	12	21
dynamic	0	21	0
language	3	9	12
wish	1	0	30
...			

Latent Dirichlet Allocation (LDA)

4) In current document, unassign a word from its topic

D_1	python	T_1	general	~~T_3~~	purpose	T_2	dynamic	T_2	language	T_1

Topic counts for D_1	
T_1	2
T_2	2
T_3	1

Word-topic counts for entire corpus

	T_1	T_2	T_3
python	3	0	0
general	22	7	10
purpose	13	12	21
dynamic	0	21	0
language	3	9	12
wish	1	0	30
...			

4) In current document, unassign a word from its topic

D_1	python T_1	general ?	purpose T_2	dynamic T_2	language T_1

Topic counts for D_1	
T_1	2
T_2	2
T_3	~~1~~ 0

Word-topic counts for entire corpus

	T_1	T_2	T_3
python	3	0	0
general	22	7	~~10~~ 9
purpose	13	12	21
dynamic	0	21	0
language	3	9	12
wish	1	0	30
...			

Latent Dirichlet Allocation (LDA)

D_1	python T_1	general ?	purpose T_2	dynamic T_2	language T_1

Topic counts for D_1	
T_1	2
T_2	2
T_3	0

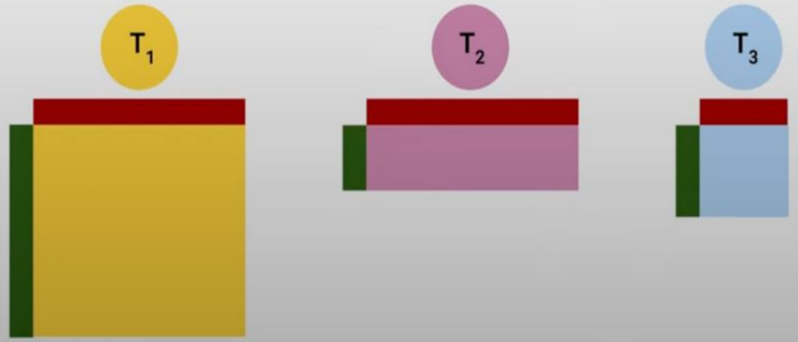
Word-topic counts for *entire* corpus			
	T_1	T_2	T_3
python	3	0	0
general	22	7	9
purpose	13	12	21
dynamic	0	21	0
language	3	9	12
wish	1	0	30
...			

5) Assign $w_{d,n}$ a new topic based on:

a) **The prevalence of each topic in the document.**

b) **The prevalence of the word in each topic.**

$$\frac{n_{d,k} + \alpha}{\sum_i^K n_{d,i} + \alpha} \times \frac{m_{w,k} + \beta}{\sum_i^V m_{i,k} + \beta}$$



Latent Dirichlet Allocation (LDA)

D_1	python T_1	general T_1	purpose T_2	dynamic T_2	language T_1

Topic counts for D_1	
T_1	~~2~~ 3
T_2	2
T_3	0

Word-topic counts for *entire* corpus

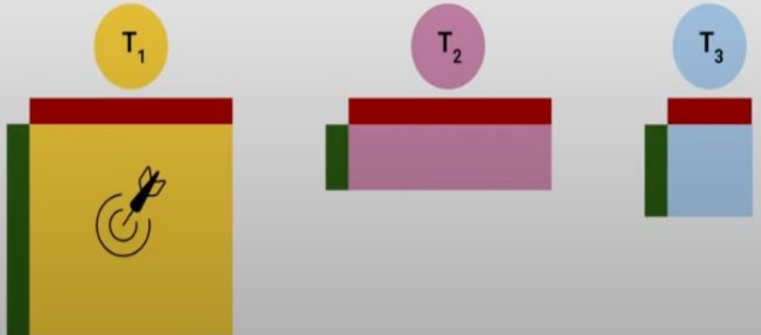
	T_1	T_2	T_3
python	3	0	0
general	~~22~~ 23	7	9
purpose	13	12	21
dynamic	0	21	0
language	3	9	12
wish	1	0	30
...			

5) Assign $w_{d,n}$ a new topic based on:

a) **The prevalence of each topic in the document.**

b) **The prevalence of the word in each topic.**

$$\frac{n_{d,k}+\alpha}{\sum_i^K n_{d,i}+\alpha} \times \frac{m_{w,k}+\beta}{\sum_i^V m_{i,k}+\beta}$$



We can answer these questions from experience as well as on principle. The experiences of camp life show that man does have a choice of action. There were enough examples, often of a heroic nature, which proved that apathy could be overcome, irritability suppressed. Man can preserve a vestige of spiritual freedom, of independence of mind, even in such terrible conditions of psychic and physical stress.

We who lived in concentration camps can remember the men who walked through the huts comforting others, giving away their last piece of bread. They may have been few in number, but they offer sufficient proof that everything can be taken from a man but one thing: the last of the human freedoms – to choose one's attitude in any given set of circumstances, to choose one's way.

And there were always choices to make. Every day, every hour, offered the opportunity to make a decision, a decision which determined whether you would or would not submit to those powers which threatened to rob you of your very self, your inner freedom; which determined whether or not you would become the plaything of circumstance, renouncing freedom and dignity to become molded into the form of the typical inmate.

Topic 1 | Topic 2 | Topic 3

1) Randomly assign a topic to *every* word in *every* document.

2) Count the number of times each topic *k* occurs in document *d*.

3) Count the number of times a word $w_{d,n}$ is assigned a topic *k* across entire corpus.

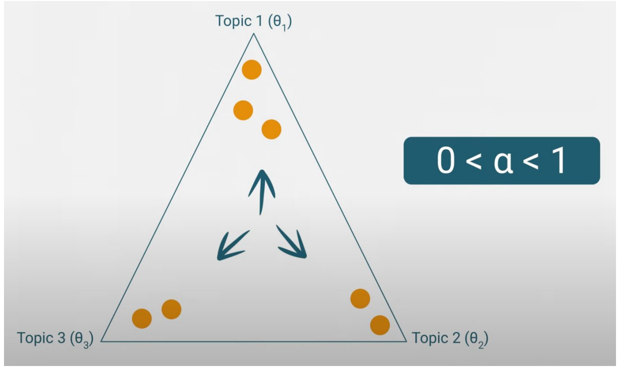
4) In a document d, *unassign* a word $w_{d,n}$ from its topic.

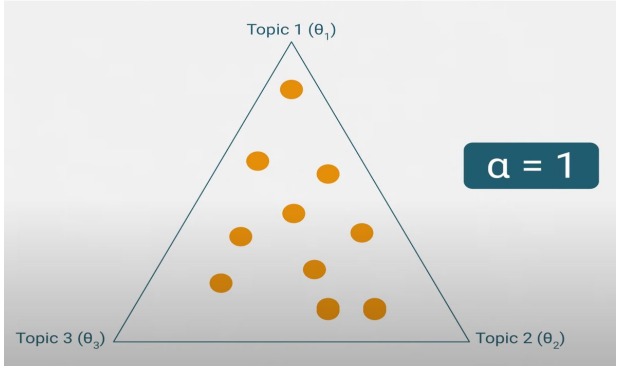
5) Assign $w_{d,n}$ a new topic based on
   a. How much this document *d* likes topic *k*.
   b. How much this topic likes word $w_{d,n}$.

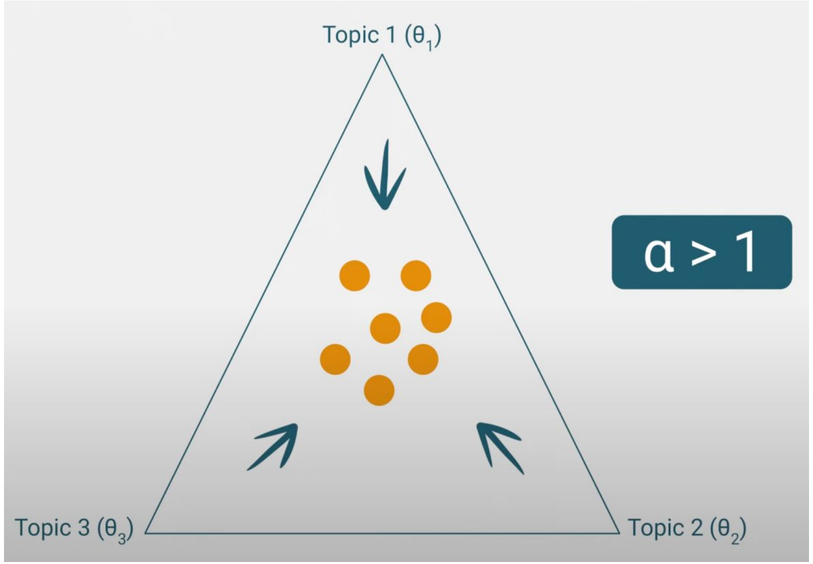
Repeat 2-5 with a different $W_{d,n}$.

What parameters are we exactly learning?

Drawing Topic Distributions

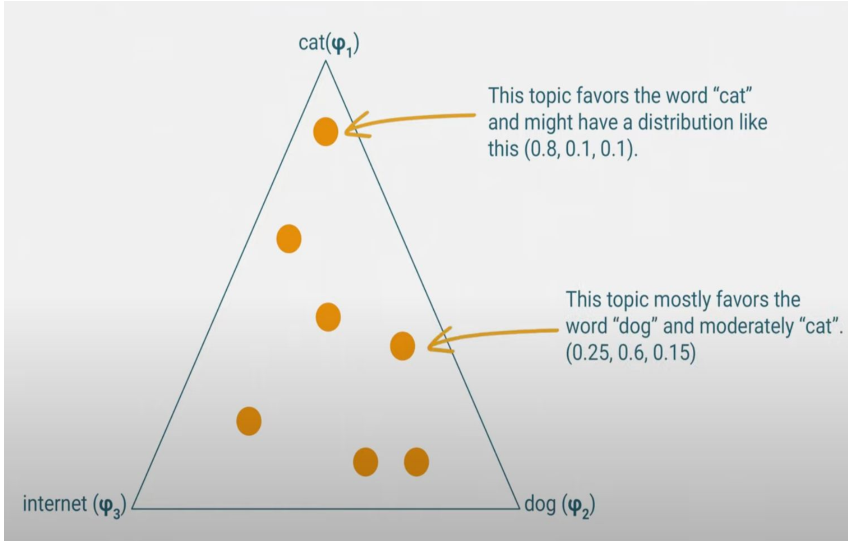


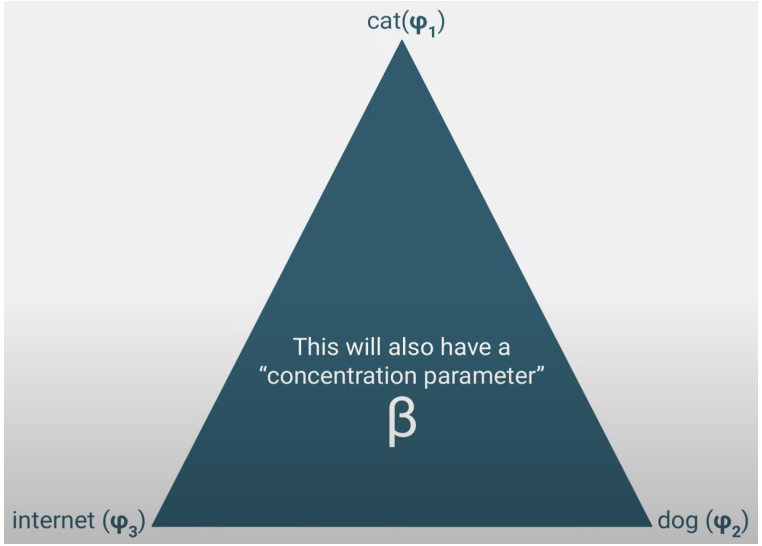




Document= Multinomial distribution of topics with alpha

Drawing Word Distributions





Topic = Multinomial distribution of words with beta

Alpha and Beta are just hyperparameters not something the model learnt

Demo

https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

Any similarity with traditional ML techniques?

SAMPLE TOPICS WITH REPRESENTATIVE WORDS.

Topic Label	Top Weighted Words
Relationship with family (20.8%)	life, relationship, mother, ex, child, father, life, wife, partner, son
Intimate relationship (17.3%)	girlfriend, boyfriend, relationship, dating, upset, feel, pretty, lot, love, guy
Living in shared accommodation (16.5%)	apartment, rent, live, room, living, house, lease, stay, bedroom
Money (7.3%)	pay, rent, saving, buy, job, account, car, loan, afford, cost
Pregnancy concerns in pets (5.5%)	dog, child, husband, child, pregnant, puppy, cat, law, animal, birth
Work (4.4%)	hour, work, boss, company, manager, job, employee, office, shift, week
Appearance judgment (4.2%)	hair, look, wear, white, black, comment, clothes, dress, looked, pretty
Neighborhood (3.3%)	neighbor, phone, email, post, account, people, use, street, yard, facebook

Qualitative analysis and feature engineering

Ruijie Xi and Munindar P. Singh. "The Blame Game: Understanding Blame Assignment in Social Media." *IEEE Transactions on Computational Social Systems* 11, no. 2 (2023): 2267-2276.

Evaluation of Topics

Coherence metric

There is also an extrinsic version

Intrinsic UMass measure

The UMass measure introduced by [Mimno11a] uses as pairwise score function

$$\text{score}_{\text{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

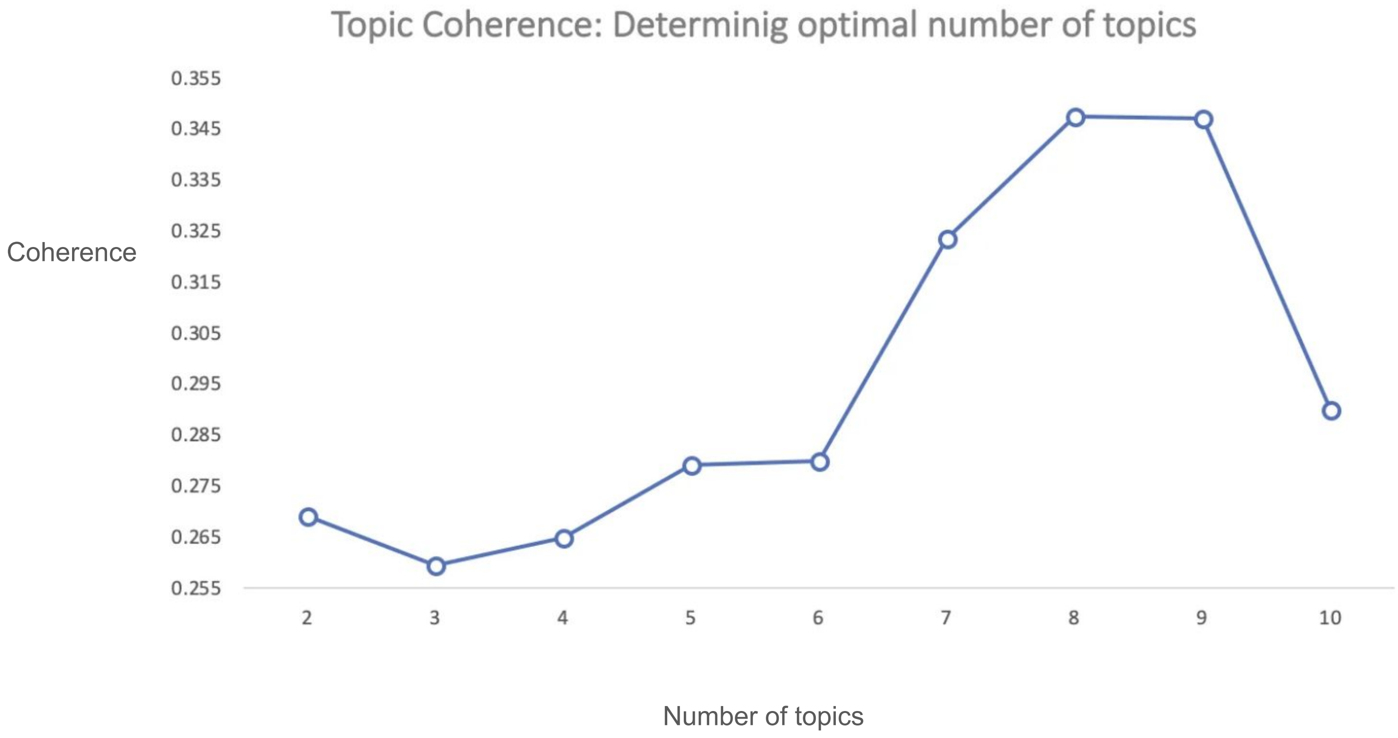
which is the empirical conditional log-probability $\log p(w_j|w_i) = \log \frac{p(w_i,w_j)}{p(w_i)}$ smoothed by adding one to $D(w_i, w_j)$.

The score function is not symmetric as it is an increasing function of the empirical probability $p(w_j|w_i)$, where w_i is more common than w_j, words being ordered by decreasing frequency $p(w|k)$. So this score measures how much, within the words used to describe a topic, a common word is in average a good predictor for a less common word.

As the pairwise score used by the UMass measure is not symmetric, the order of the arguments matters. UMass measure is computing $p(\text{rare word} \mid \text{common word})$, how much a common word triggers a rarer word. However, in human word association, high frequency words are more likely to be used as response words than low frequency words [Steyvers06]. It would be interesting to understand the effect of this choice by doing more experiments and comparing the two options.

https://qpleple.com/topic-coherence-to-evaluate-topic-models/

How to Decide Number of Topics



How to Decide Other Hyperparameters

Once K is final:

Iterate over different values of alpha and beta

Choose the values giving the best coherence

Enjoy and let me enjoy as well!

Acknowledgements

Some Slides from Future Mojo