Topic Modeling with NMF and LDA

Articles and Headlines

Manjari Senthilkumar

24 July 2021

Abstract

Reading, understanding and processing text is an enormous aspect in both our daily life and in any project. In today's age of technological advancements, the natural next step is to let machines take over this tedious, yet critical task.

When parsing through a paper, you are likely to skim the text, looking for key ideas that stand out, trying to get a general understanding of what the text is about, and condensing it to information you can retain. Natural Language Processing does the same in theory: It parses through text to add useful numerical structure, resolve ambiguity and draw out key ideas.

Topic Modeling is useful In the specific case where you want to process a large quantity of documents, and draw out information both specific to the document and overarching concepts. It relies on weighting words as a representation of importance and then finding shared instances between the documents.

Topic modeling and machine learning in general are key to eliminating human bias and introducing more objective diversity into our results.

There are several algorithms that do this effectively. A notable few are Non-negative Matrix Factoring (NMF), Latent Dirichlet Allocation (LDA), Principal Component Analysis (PCA), and Latent Semantic Analysis (LSA). In this paper, I will go through the

algorithms, NMF and LDA, and highlight some interesting differences we see in data that is processed through each as a reflection of the algorithm itself.

Data

Two datasets were used in this paper. The first being a collection of 337 Medium articles. Medium is an American open online publishing platform that resides in the heart of the technology, and specifically data science, community. As a result the content mostly revolves around computer science, programming languages, and machine learning.

	author	claps	reading_time	link	title	text
0	Justin Lee	8.3K	11	https://medium.com/swlh/chatbots-were-the-next...	Chatbots were the next big thing: what happene...	Oh, how the headlines blared:\nChatbots were T...
1	Conor Dewey	1.4K	7	https://towardsdatascience.com/python-for-data...	Python for Data Science: 8 Concepts You May Ha...	If you've ever found yourself looking up the s...
2	William Koehrsen	2.8K	11	https://towardsdatascience.com/automated-featu...	Automated Feature Engineering in Python – Towa...	Machine learning is increasingly moving from h...
3	Gant Laborde	1.3K	7	https://medium.freecodecamp.org/machine-learni...	Machine Learning: how to go from Zero to Hero ...	If your understanding of A.I. and Machine Lear...
4	Emmanuel Ameisen	935	11	https://blog.insightdatascience.com/reinforcem...	Reinforcement Learning from scratch – Insight ...	Want to learn about applied Artificial Intelli...

The second dataset used is a collection of 1226258 news headlines from ABC (Australian Broadcasting Corporation) published over the last 18 years. ABC is a popular reputable australian news source. Naturally, we would assume that the news headlines would converge to major headline news in australia as topics

	publish_date	headline_text
0	20030219	aba decides against community broadcasting lic...
1	20030219	act fire witnesses must be aware of defamation
2	20030219	a g calls for infrastructure protection summit
3	20030219	air nz staff in aust strike for pay rise
4	20030219	air nz strike to affect australian travellers

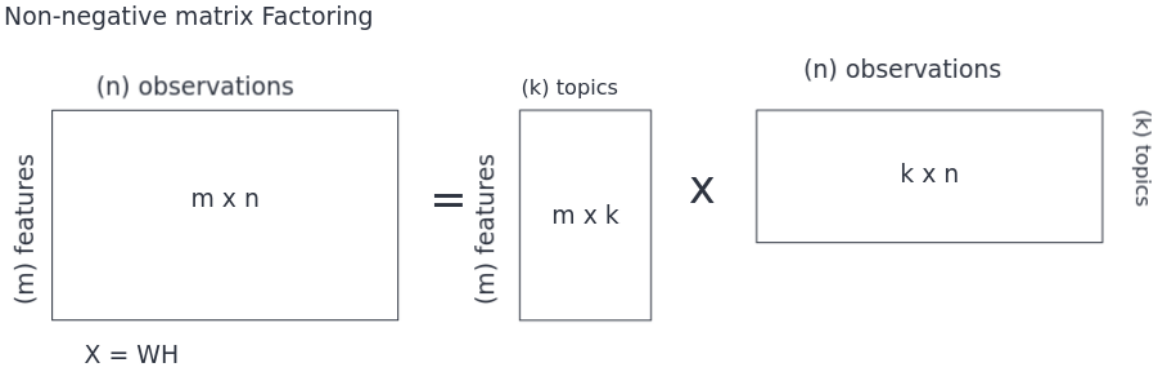
Algorithms

1. NMF

NMF uses the logic of the “distributional hypothesis”, which is the idea that “words that are similar in meaning occur in similar contexts”. This enables us to ignore both semantic and syntactic features of text because we assume word order and word meaning do not matter.

Given a corpus of data, the algorithm will first create a matrix with each document as a column vector and every word in a document will be represented by the row entries. This matrix is then converted into a non-negative matrix, because negativity is insignificant in representing words.

Then we can decompose the matrix such that the 2 matrices are also non-negative as shown below.



In the above diagram X is our non-negative matrix that represents the corpus and all its terms. W and H are it’s factored form.

The number of topics, k, is a hyperparameter, and thus it is up to the programmer to choose a value k that potentially represents the number of topics you expect to see in the corpus. By the nature of matrix multiplication each entry in matrix X should represent the respective dot product of the row vector of W and the column vector of H. Thus each word is composed of a linear combination of all the possible topics for that word and all the possible topics in that document.

This method of factoring accomplishes several things.

(1) Reduce the dimensionality of the system which makes it easier to deal with large datasets
(2) Draw out key topics

In addition, to ensure that the data is trained efficiently and does not include common words. The data is removed of stop words and a weighting scheme is applied to reflect how important each word is to the document.

A popular weighting scheme is tf-idf (term frequency-inverse document frequency), which uses the proportion at which each word appears to add relative importance to each word. The formula is as given below:

$$TF(t_i, d) = \left\{ \frac{f(t_i, d)}{|N_d|} \right\}$$

Ultimately, the goal of NMF is to optimize the cost function ||X- HF|| since it's unlikely that X will factor completely, so we must find the closest approximation.

There are 2 common ways to properly factor the equation: Gradient descent on coordinate matrix or applying the Multiplicative step method, as given below.

$$\mathbf{H}^{n+1}_{[i,j]} \leftarrow \mathbf{H}^{n}_{[i,j]} \frac{((\mathbf{W}^n)^T \mathbf{V})_{[i,j]}}{((\mathbf{W}^n)^T \mathbf{W}^n \mathbf{H}^n)_{[i,j]}}$$

and

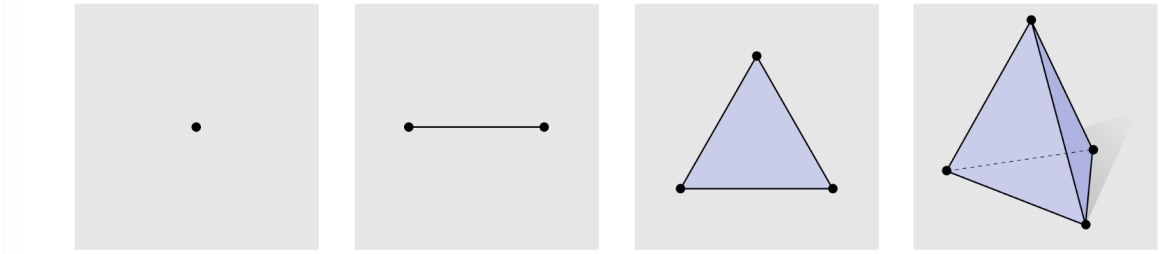
$$\mathbf{W}^{n+1}_{[i,j]} \leftarrow \mathbf{W}^{n}_{[i,j]} \frac{(\mathbf{V}(\mathbf{H}^{n+1})^T)_{[i,j]}}{(\mathbf{W}^n \mathbf{H}^{n+1}(\mathbf{H}^{n+1})^T)_{[i,j]}}$$

2. LDA

LDA functions on a lot of the same assumptions of NMF, like the lack of semantic and syntactic meaning. In addition, it greatly relies on relative knowledge. In essence, LDA assumes in performing it's algorithm that all other word to topic associations are true, and sorts the current word based on its relation to the other words in the corpus.

2.1. Dirichlet Distribution

LDA takes a geometric and probabilistic approach to sorting the corpus of data. It creates a k-th dimensional simplex where each corner represents 1 of the k topics.



It then distributes each of the documents on the spectrum of topics. For instance if a corpus had the topics sports, science and politics. The documents would be distributed on a 2-dimensional simplex. If a document was observed to be 100% sports it would be a point on the corner assigned to sports. If there were a document that was 50% sports and 50% politics it would be assigned midway through the segment of science and politics.

Dirichlet distributions have two key constants, α and β. α represents the distribution of data:

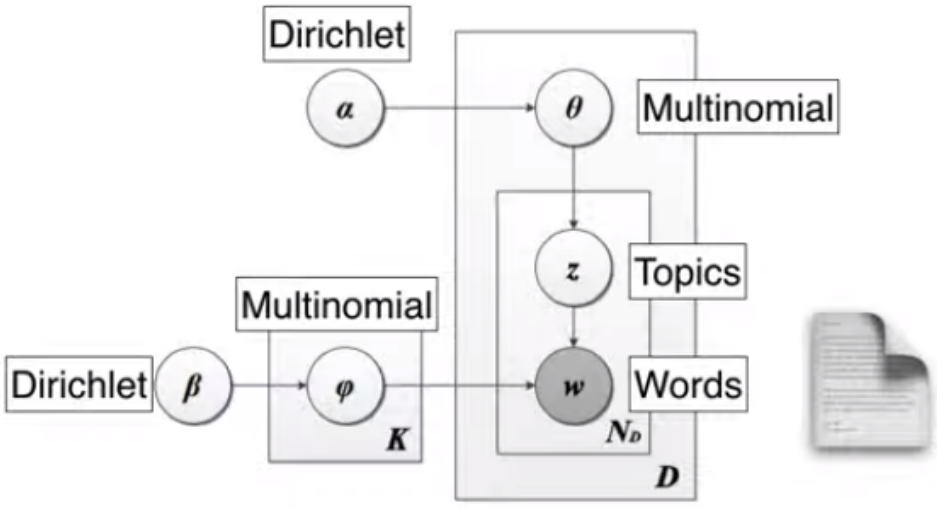
α or $\beta = 1$: evenly distributed

α or $\beta > 1$: higher density of data in the center

α or $\beta < 1$: higher density of data in the corners

2.2. Method

The goal of LDA is to essentially recreate the exact document given the following distributions of data. Obviously, the probability that the *exact* document is reproduced is extremely low, but in effect of pursuing that goal, the topic sorting accuracy will also be optimized.



LDA has two dirichlet distributions that are fed into the algorithm. The first being the documents distributed over the topics, and the second being the topics distributed over the words in each document. They are each represented by dirichlet distributions with the parameters $(\theta_i;\alpha)$ and $(\varphi_i;\beta)$ respectively. And the following equation, that we attempt to optimize, to give us the probability that we can reproduce the exact document

$$P(\mathbf{W}, \mathbf{Z}, \theta, \phi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} | \theta_j) \; P(W_{j,t} | \phi_{Z_{j,t}})$$

$\prod_{j=1}^{M} P(\theta_j; \alpha)$ represents the Dirichlet distribution of the documents on the topics

$\prod_{i=1}^{K} P(\phi_i; \beta)$ represents the Dirichlet distribution of the topics distributed over each word in the document.

$\prod_{t=1}^{N} P(Z_{j,t} | \theta_j)$ represents the multinomial distribution of the topics of the words in a document

$P(W_{j,t} | \phi_{Z_{j,t}})$ represents the probability a word is a particular topic.

One effective method of optimizing this probability is by optimizing the probability that a word is a particular topic.

We can use the probability rule below to update the probability after each step. The words are initialized randomly with topics, and then updated until the probability begins to stabilize.

```
p(word w with topic t) = p(topic t | document d) * p(word w | topic t)
```

Results

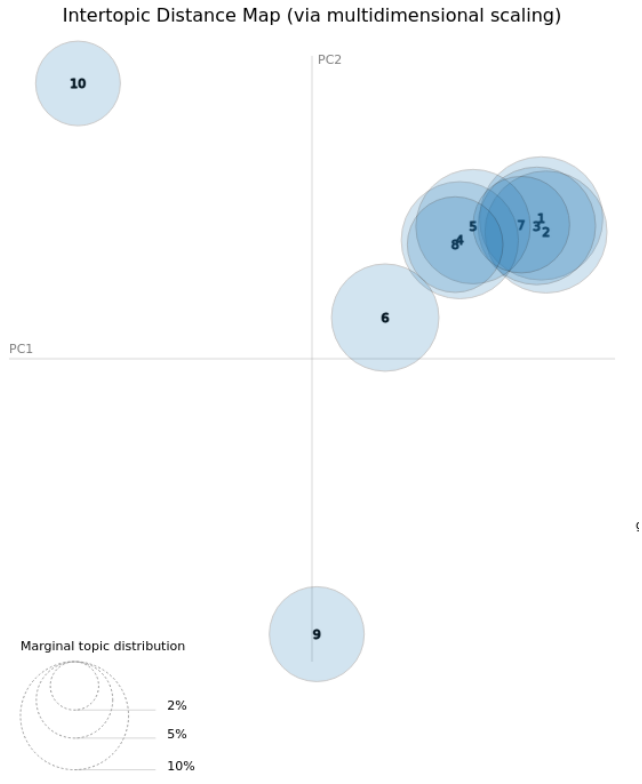
1. **Headlines**

   NMF

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	interview	to	in	for	the	on	of	over	with	at
1	michael	be	killed	calls	and	police	out	man	charged	and
2	extended	urged	man	new	drum	and	accused	charged	man	police
3	nrl	new	found	call	from	new	police	police	and	after
4	john	set	crash	up	is	down	ahead	court	markets	by
5	david	from	after	search	says	us	new	death	extended	fire
6	james	govt	new	police	world	after	and	murder	police	dies
7	smith	council	australia	jailed	rise	as	guilty	jailed	new	new
8	andrew	up	police	more	about	govt	warns	arrested	speaks	out
9	afl	get	sydney	no	up	track	man	after	up	man
10	scott	help	dies	set	speaks	council	warned	concerns	after	dead
11	mark	return	us	out	australian	focus	by	win	win	home
12	peter	police	court	support	year	back	death	fined	wednesday	takes
13	will	back	dead	missing	business	rise	us	charges	tuesday	risk
14	and	australia	two	and	back	report	confident	fire	deal	up
15	george	face	injured	push	thursday	attack	court	council	murder	as
16	white	speaks	as	man	monday	puts	murder	woman	signs	us
17	cricket	us	fire	fined	wednesday	work	year	govt	monday	from
18	north	plan	up	council	friday	coast	australian	attack	friday	back
19	price	boost	melbourne	water	tuesday	fire	says	by	thursday	sydney

LDA

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	restriction	new	melbourne	australian	coronavirus	victoria	australia	u	police	nsw
1	china	man	death	fire	case	border	coronavirus	donald	woman	queensland
2	covid	court	state	crash	trump	vaccine	election	news	record	wa
3	found	coronavirus	lockdown	final	say	canberra	government	south	family	year
4	house	live	coronavirus	island	sydney	hit	health	two	coronavirus	coronavirus
5	coronavirus	first	adelaide	dy	covid	report	home	biden	child	covid
6	beach	quarantine	andrew	car	nt	community	day	world	market	sa
7	abc	school	business	time	call	joe	covid	coronavirus	open	coast
8	north	murder	inquiry	million	pandemic	farmer	change	nrl	charged	tasmania
9	finance	life	drum	high	win	set	afl	job	national	bushfire
10	water	test	darwin	dead	one	rain	qld	season	minister	hotel
11	west	worker	fear	killed	could	kohler	victorian	royal	amid	back
12	war	trial	labor	india	premier	hong	morrison	commission	help	brisbane
13	northern	people	end	coronavirus	christmas	top	scott	four	show	regional
14	street	face	second	find	budget	kong	update	industry	student	plan
15	resident	protest	make	pay	return	economy	care	program	take	gold
16	testing	get	asx	road	president	johnson	hospital	trade	missing	council
17	weather	attack	service	aboriginal	outbreak	still	federal	tourism	covid	ban
18	bushfires	uk	speaks	cricket	cut	story	indigenous	announces	may	rule
19	western	former	friday	white	act	future	travel	cup	victim	storm



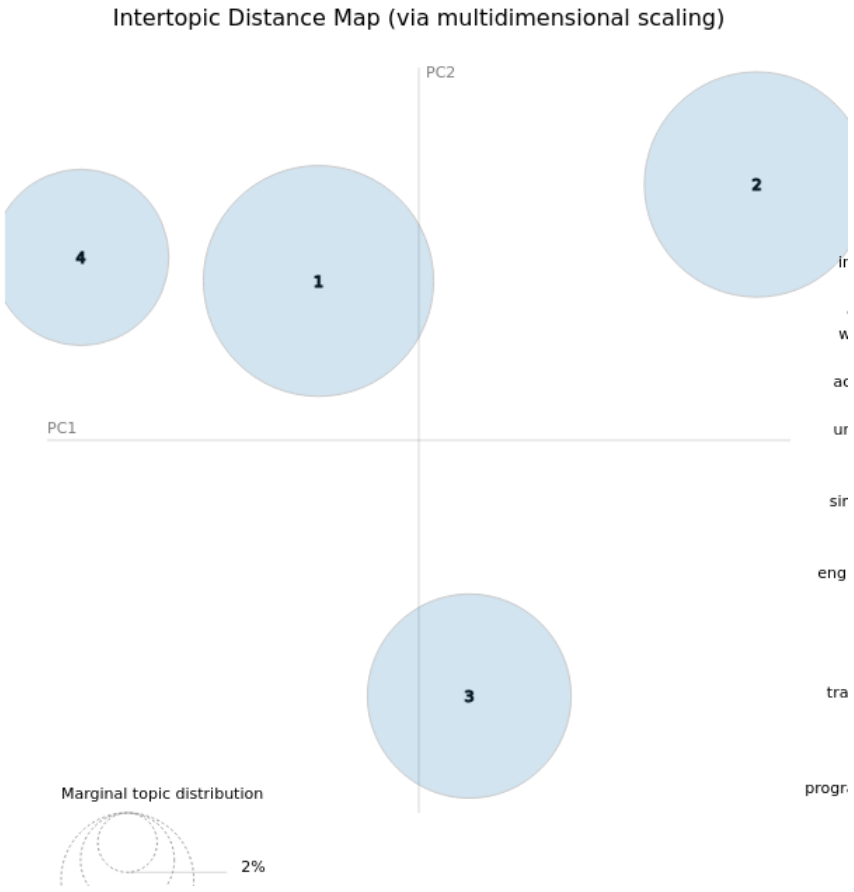
2. Articles

NMF

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	the	de	cheat	learning	the	stars	cnn	you	de	we
1	and	que	sheet	machine	of	github	image	to	la	to
2	of	deep	https	data	is	on	region	the	que	of
3	to	learning	datacamp	models	to	python	the	it	en	our
4	in	do	numpy	facebook	network	of	bounding	and	et	and
5	that	fast	www	supervised	in	and	to	your	re	that
6	is	ai	en	to	neural	for	boxes	this	intelligence	in
7	it	com	com	big	this	source	and	that	vector	is
8	with	to	the	google	and	library	of	for	machines	the
9	as	machine	scikit	clustering	for	projects	we	if	no	for
10	on	natural	python	feature	layer	open	object	is	pre	this
11	are	https	and	ai	output	tensorflow	in	can	me	can
12	ai	50	community	artificial	training	at	faster	in	natural	it
13	for	as	to	regression	are	the	pixel	of	words	data
14	be	en	blog	story	function	framework	cnns	on	part	table
15	was	you	keras	classification	model	with	this	with	local	an
16	they	this	is	from	neuron	to	objects	up	intelligent	be
17	their	show	of	intelligence	on	research	pixels	get	datasets	as
18	human	standing	http	clap	with	quality	fast	be	car	are
19	more	from	machine	cheer	input	brain	google	how	non	have

LDA

	Topic # 01	Topic # 02	Topic # 03
0	neuron	star	cnn
1	activation	review	box
2	policy	rating	pixel
3	player	average	table
4	woman	interview	translation
5	men	university	cpu
6	man	weighted	sequence
7	simulation	engineering	sheet
8	zero	music	region
9	sigmoid	programming	batch
10	supervised	analysis	sentence
11	derivative	skill	house
12	alphago	option	gpu
13	distribution	github	gtx
14	backpropagation	page	rnn
15	facial	participant	price
16	behavior	matrix	card
17	reinforcement	cover	letter
18	relu	assignment	market
19	tree	estimated	lstm



Key Observations

NMF included a lot of stop words, likely due to poor implementation.

PCA on LDA topics was a very useful method of analyzing the effectiveness of the algorithm itself. It allowed me to tune the hyperparameter, k, that represents the number of topics.

NMF was especially good at drawing out topics that were used in similar contexts. For instance, NMF was able to group a lot of the python ML libraries together in the same group.

LDA was especially good at drawing out topics that were related to each other. For instance, all the days of the week were together even though the rest of the topic words didn't really associate with the days of the week as much. Similarly LDA associated star to review to grades, which may speak to the common ground of weighted averages.