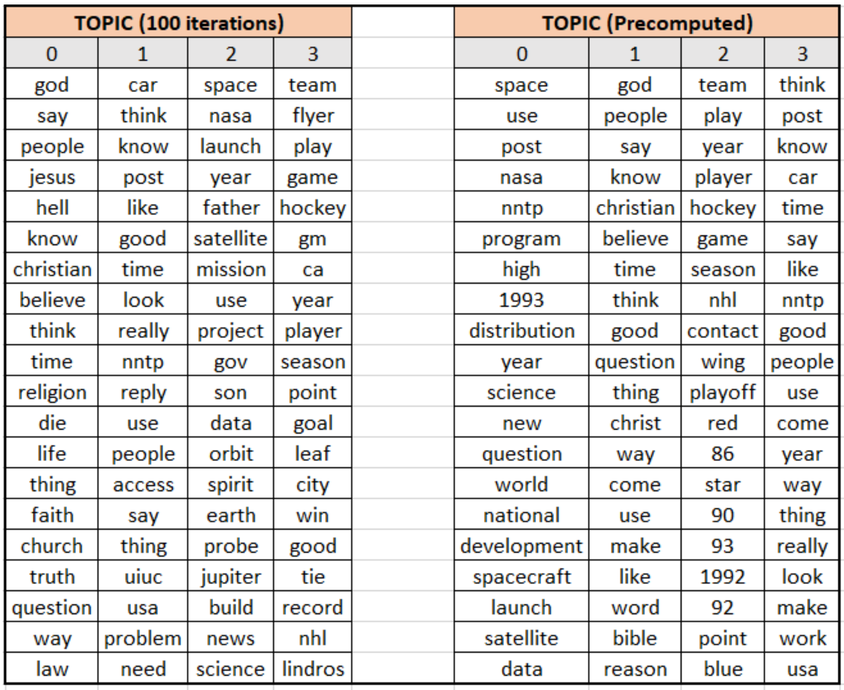
ANIRUDH SETH
aniset@kth.se

Tutorial 9
Latent Dirichlet Allocation for topic clustering using python

2020-02-25

Newsgroup Dataset

The algorithm was implemented for the newsgroup dataset. The figure below shows the top 20 words for all the topics(the best on the top).In the figure below we compare the results after running the algorithm for 100 iterations with the provided results(200 iterations).



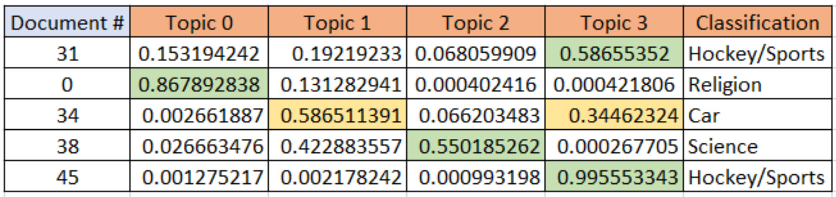
TOPIC (100 iterations)					TOPIC (Precomputed)			
0	1	2	3		0	1	2	3
god	car	space	team		space	god	team	think
say	think	nasa	flyer		use	people	play	post
people	know	launch	play		post	say	year	know
jesus	post	year	game		nasa	know	player	car
hell	like	father	hockey		nntp	christian	hockey	time
know	good	satellite	gm		program	believe	game	say
christian	time	mission	ca		high	time	season	like
believe	look	use	year		1993	think	nhl	nntp
think	really	project	player		distribution	good	contact	good
time	nntp	gov	season		year	question	wing	people
religion	reply	son	point		science	thing	playoff	use
die	use	data	goal		new	christ	red	come
life	people	orbit	leaf		question	way	86	year
thing	access	spirit	city		world	come	star	way
faith	say	earth	win		national	use	90	thing
church	thing	probe	good		development	make	93	really
truth	uiuc	jupiter	tie		spacecraft	like	1992	look
question	usa	build	record		launch	word	92	make
way	problem	news	nhl		satellite	bible	point	work
law	need	science	lindros		data	reason	blue	usa

Figure 1: Top 20 words for the newsgroup dataset

SOME RESULTS

- The order of topics is not the same. This was expected since the algorithm is unsupervised.
- Topic 0 has several words similar to the Topic 1 from the provided samples. The words god, jesus, christian, religion, faith, church etc suggest that the most likely topic is ”Christian Religion”.
- Topic 2 has several words similar to the Topic 0 from the provided samples. The words space, nasa, science, launch, satellite etc suggest that the most likely topic is ”Space Missions/Science”.
- Topic 3 has several words similar to the Topic 2 from the provided samples. The words team, play, game, hockey, season etc suggest that the most likely topic is ”Sports/Hockey”.
- Topic 1 has several words similar to the Topic 4 from the provided samples.The likely topic is not clearly evident.This can be attributed to the small vocabulary size.It could be related to cars as it the top word.

INFERENCE ON TEST SET



Document #	Topic 0	Topic 1	Topic 2	Topic 3	Classification
31	0.153194242	0.19219233	0.068059909	0.58655352	Hockey/Sports
0	0.867892838	0.131282941	0.000402416	0.000421806	Religion
34	0.002661887	0.586511391	0.066203483	0.34462324	Car
38	0.026663476	0.422883557	0.550185262	0.000267705	Science
45	0.001275217	0.002178242	0.000993198	0.995553343	Hockey/Sports

Figure 2: Inference on 5 random documents from the test set

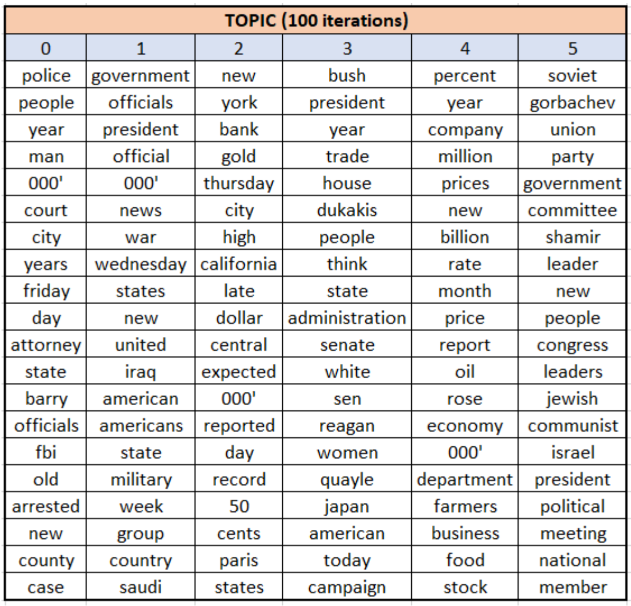
- After training the model on 200 documents for 100 iterations, the estimated parameters α and β can be used to calculate ϕ_i and γ_i for each of the new document.
- The γ_i gives the the mixture components for each topic for a document i. There are 50 test documents. For the purpose of analysis , 5 documents were randomly selected to inspect the topic mixtures.Figure 2 shows the results of topic assignments for each of them.
- Document 31 is an email with the team scores and updates about votes for NHL. This was correctly classified to Topic 3 i.e Sports/Hockey.
- Document 0 is an email talking about religion , Christianity and some excerpts from the bible. It was correctly classified to Topic 0.
- Document 34 is a small email with 2 sentences about a person complaining about predictions about a team winning the cup. The model assigns higher weight to Topic 1 and Topic 3. This was incorrectly classified to Topic 1. This can be attributed to size of the document being too small.It can also be observed that the several top words from this topic like ’think’ ,’know’ etc have been used.
- Document 38 talks about science fiction in movies and answers some questions about NASA and space. It was correctly classified to Topic 2.
- Document 45 talk about President’s cup and was also correctly classified to Topic 3.

OPTIONAL DATASETS

Associated Press docs dataset

SET UP

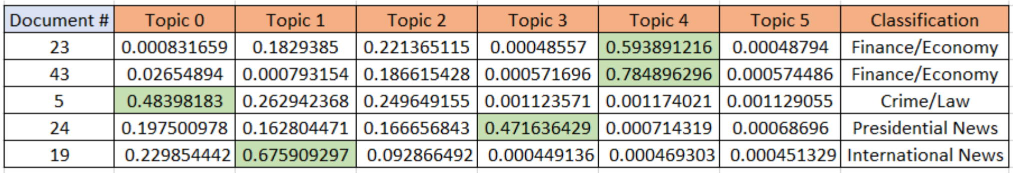
- Algorithm was run for 100 iterations with number of topics ,k=6.



TOPIC (100 iterations)					
0	1	2	3	4	5
police	government	new	bush	percent	soviet
people	officials	york	president	year	gorbachev
year	president	bank	year	company	union
man	official	gold	trade	million	party
000'	000'	thursday	house	prices	government
court	news	city	dukakis	new	committee
city	war	high	people	billion	shamir
years	wednesday	california	think	rate	leader
friday	states	late	state	month	new
day	new	dollar	administration	price	people
attorney	united	central	senate	report	congress
state	iraq	expected	white	oil	leaders
barry	american	000'	sen	rose	jewish
officials	americans	reported	reagan	economy	communist
fbi	state	day	women	000'	israel
old	military	record	quayle	department	president
arrested	week	50	japan	farmers	political
new	group	cents	american	business	meeting
county	country	paris	today	food	national
case	saudi	states	campaign	stock	member

Figure 3: Top 20 words for the AP dataset

SOME RESULTS



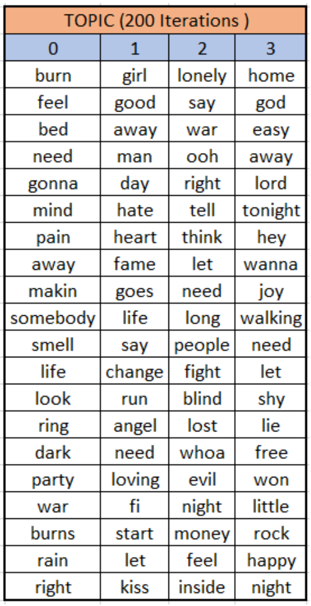
Document #	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Classification
23	0.000831659	0.1829385	0.221365115	0.00048557	0.593891216	0.00048794	Finance/Economy
43	0.02654894	0.000793154	0.186615428	0.000571696	0.784896296	0.000574486	Finance/Economy
5	0.48398183	0.262942368	0.249649155	0.001123571	0.001174021	0.001129055	Crime/Law
24	0.197500978	0.162804471	0.166656843	0.471636429	0.000714319	0.00068696	Presidential News
19	0.229854442	0.675909297	0.092866492	0.000449136	0.000469303	0.000451329	International News

Figure 4: Inference on 5 random documents from the test set

- Document 23 talks about shares and investments about Air Wis Services. Topic 4 is the correct assignment.
- Document 43 talks about stock prices and stock market in general.Topic 4 is the correct assignment.
- Document 5 talks about train derailment and phosphorus fire incident and filing a lawsuit against a company.Topic 0 is the correct assignment.
- Document 24 talks about a deal signed by President Reagan. Topic 3 is the correct assignment.It consists of many words related to the US president like Bush,Reagan,president,campaign etc.
- Document 19 talks about the support of Roman Catholic calls for Christians to boycott a new American film about Jesus Christ in Britain. It was assigned to Topic 1. Out of the available topics , this topic is the closest match, It consists of a lot countries like Iraq,Saudi,America. Likely topic could be International news.

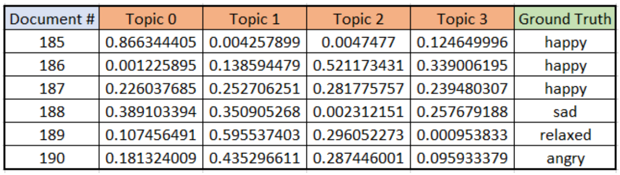
Moody Lyrics dataset

- The results on the lyrics dataset with the default settings were not good. An improvement was seen by increasing the iterations to 200. However, the results were still not acceptable . This could be attributed to very small vocabulary and number of documents. It should be noted that happy words like love , forever can also occur in sad ,angry songs but with negation prefix.To alleviate this problem, the vocabulary was increased from one word to 2-gram. This increased the size of vocabulary from 500 to over 18000 resulting in extremely slow performance on my personal computer. Below results are from one word vocabulary and 200 iterations.



TOPIC (200 Iterations)			
0	1	2	3
burn	girl	lonely	home
feel	good	say	god
bed	away	war	easy
need	man	ooh	away
gonna	day	right	lord
mind	hate	tell	tonight
pain	heart	think	hey
away	fame	let	wanna
makin	goes	need	joy
somebody	life	long	walking
smell	say	people	need
life	change	fight	let
look	run	blind	shy
ring	angel	lost	lie
dark	need	whoa	free
party	loving	evil	won
war	fi	night	little
burns	start	money	rock
rain	let	feel	happy
right	kiss	inside	night

Figure 5: Top 20 words for the Moody Lyrics dataset



Document #	Topic 0	Topic 1	Topic 2	Topic 3	Ground Truth
185	0.866344405	0.004257899	0.0047477	0.124649996	happy
186	0.001225895	0.138594479	0.521173431	0.339006195	happy
187	0.226037685	0.252706251	0.281775757	0.239480307	happy
188	0.389103394	0.350905268	0.002312151	0.257679188	sad
189	0.107456491	0.595537403	0.296052273	0.000953833	relaxed
190	0.181324009	0.435296611	0.287446001	0.095933379	angry

Figure 6: Comparison with the ground truth

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (March 2003), 993–1022.

Unused Plots

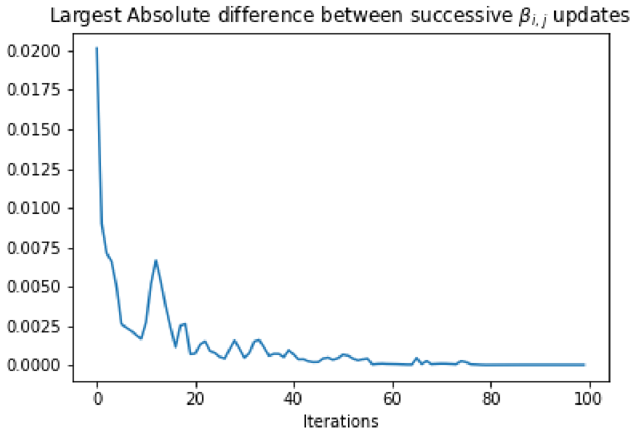


Figure 7: Newsgroup

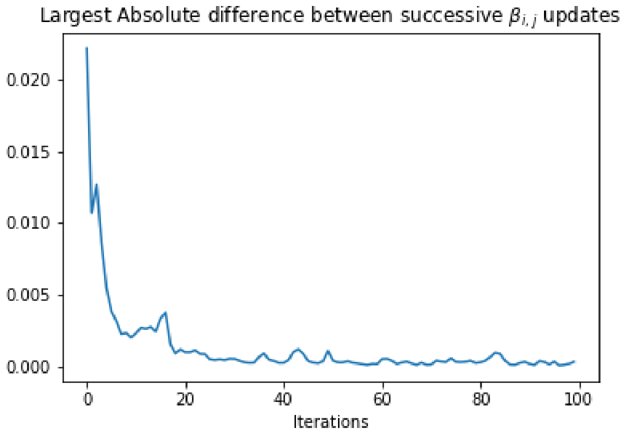


Figure 8: AP Dataset

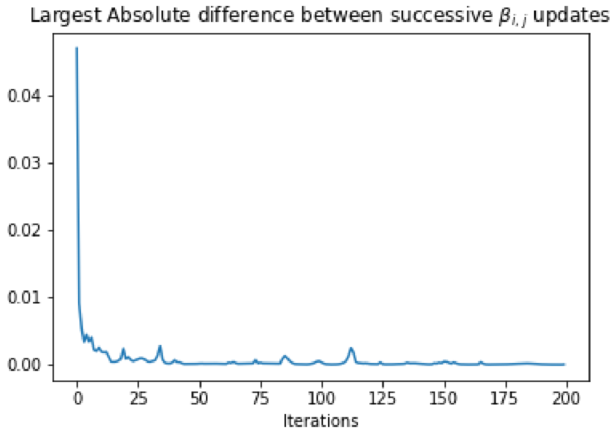


Figure 9: Moody Lyrics