

NY Times Dataset

I used the scientific python stack for this assignment, particularly SciPy, NumPy, and Scikit-learn. For topic modeling, I used Scikit-learn's (herein referred to as sklearn) implementation of Latent Dirichlet Allocation¹.

The first hurdle for this assignment was reading in the data. Since it wasn't feasible to read all the data in at once, I had to read in the data in chunks and convert those chunks to NumPy arrays. I then concatenated all the NumPy arrays to form a multidimensional array that resembled the original data file. I converted the term frequency counts from the (*docID*, *wordID*, *freq*) format in *docword.nytimes.txt* to a matrix format, which is standard across most topic modeling packages. I did this using the sparse coordinate matrix² functionality from SciPy. This sparse matrix was then converted to a CSR format. I decided to use a sparse matrix from SciPy since creating a term frequency matrix using NumPy arrays would raise memory errors. The resulting matrix had the rows representing documents, the columns representing words, and the data representing word frequency. To make the output from LDA more understandable, I created a look-up dictionary from *vocab.nytime.txt*.

Next was to initialize the LDA objects (this is how sklearn's API operates, initialize models with parameters, then fit) with varying parameters. I set the alpha and beta values to $\frac{1}{\text{number of topics}}$, where the number of topics was set to ten. I then varied the number of iterations for LDA and presented the output in the Appendix of this paper, tables 1 through 5. My output displays the ten highest probability words per topic for ten topics at iterations 2, 5, 10, and 20. I also include hand labeled topics in Table 5. The run time per iteration was ~7 minutes running on three processors.

The results show that as the number of iterations increase, the purity of the topics increases. For example, in Table 1, many of the generated topics are unidentifiable, while in Table 4, it is quite clear what some of the topics are.

MNIST Handwritten Digits Dataset

The processing for this dataset was like what I did for the NY Times dataset. Since the raw data was already in a term frequency matrix format, all I had to do was fit the data to the sklearn LDA object as I did with NY Times. The main difference was examining the topics, where instead of only looking at the top pixels, I used the entire topic for plotting. This can be seen in Table 6. Table 7 and Table 8 display varying the top words in each topic. In Table 8, as the number of iterations increase, we see in some topics a number becoming more clear, while others less clear. I'm not entirely sure if this makes sense. The run time for this dataset was obviously much faster than NY Times.

¹ <http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>

² https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.coo_matrix.html

Conclusion

This assignment was a good hands-on exercise on topic modeling, as I haven't done this before. Had I been approached with the problem of analyzing unlabeled documents with an unsupervised method, I would have simply used clustering. Topic modeling via LDA allows more flexibility and provides results that are relatively easy for humans to interpret. While this was useful for the NY Times dataset and I could clearly see the topics converging as the number of iterations increased; however, the MNIST handwritten digits dataset was less clear.

Appendix

Table 1: 2 Epochs

Words														
zzz_al_gore	president	zzz_bush	zzz_george_bush	campaign	right	zzz_white_house	zzz_clinton	government	republican	plan	official	administration	political	public
palestinian	team	player	season	zzz_israel	play	right	game	zzz_israeli	look	home	family	night	guy	need
school	student	official	government	point	group	right	law	zzz_bush	attack	case	member	zzz_united_states	public	women
team	game	computer	point	program	right	play	show	player	home	student	women	games	percent	million
million	com	company	team	book	season	show	question	art	part	zzz_new_york	business	look	web	information
company	percent	companies	market	stock	million	business	billion	money	bill	industry	analyst	high	election	firm
official	government	zzz_united_states	zzz_bush	zzz_u_s	military	leader	war	group	country	percent	attack	zzz_taliban	president	zzz_afghanistan
season	run	team	percent	home	game	hit	player	right	games	inning	drug	night	play	official
game	cup	yard	minutes	car	add	room	season	home	play	food	run	set	oil	right
million	company	percent	zzz_enron	group	companies	film	market	american	need	system	high	business	money	problem

Table 2: 5 Epochs

Words														
zzz_bush	campaign	zzz_al_gore	president	zzz_george_bush	election	political	republican	zzz_white_house	vote	democratic	voter	zzz_clinton	presidential	zzz_senate
palestinian	team	play	game	player	season	yard	zzz_israel	point	shot	left	right	guy	zzz_israeli	night
school	student	case	law	court	official	police	lawyer	right	government	group	children	death	member	public
game	team	games	computer	point	player	play	season	show	coach	program	won	web	zzz_microsoft	sport
book	film	show	com	movie	music	look	art	character	zzz_new_york	play	director	question	part	american
company	percent	million	companies	market	stock	business	billion	money	cost	industry	plan	firm	analyst	economy
official	government	zzz_united_states	zzz_u_s	military	attack	war	country	leader	zzz_american	zzz_bush	terrorist	zzz_afghanistan	security	zzz_china
team	season	game	run	player	games	hit	inning	right	play	home	left	win	baseball	manager
water	car	room	food	cup	minutes	small	restaurant	add	hour	building	home	look	large	oil
drug	patient	percent	million	doctor	problem	study	research	group	women	cell	disease	health	scientist	human

Table 3: 10 Epochs

Words														
zzz bush	president	campaign	zzz george bush	zzz al gore	election	political	republican	zzz white house	vote	bill	democratic	zzz congress	zzz clinton	zzz senate
team	game	season	play	player	point	yard	shot	coach	left	guy	right	night	zzz israeli	games
school	student	law	case	court	official	police	lawyer	children	family	death	told	right	group	member
game	team	computer	games	player	web	play	point	season	program	site	com	sport	network	won
book	show	film	movie	com	look	music	play	character	friend	art	women	family	zzz new york	love
company	percent	million	companies	market	business	stock	billion	money	industry	cost	firm	plan	customer	sales
official	government	zzz united states	zzz u s	military	attack	war	leader	country	zzz american	terrorist	security	zzz israel	zzz afghanistan	zzz bush
season	team	run	game	hit	player	inning	games	right	baseball	home	race	yankees	manager	play
car	water	room	food	cup	hour	small	minutes	building	home	restaurant	add	house	large	look
drug	patient	doctor	percent	problem	research	cell	health	study	scientist	human	disease	test	million	found

Table 4: 20 Epochs

	Words														
1	zzz bush	president	campaign	zzz george bush	zzz al gore	election	political	republican	zzz white house	vote	bill	zzz congress	democratic	zzz clinton	zzz senate
2	team	game	season	player	play	point	coach	games	shot	win	yard	played	guy	won	left
3	school	student	law	case	court	official	children	lawyer	police	family	told	death	member	group	officer
4	computer	web	com	site	mail	program	network	www	online	system	zzz internet	games	information	zzz microsoft	sites
5	book	show	film	movie	look	music	com	play	women	friend	family	character	love	young	art
6	company	percent	million	companies	market	business	stock	billion	money	industry	cost	firm	plan	sales	customer
7	official	government	zzz united states	zzz u s	military	attack	war	palestinian	leader	zzz american	country	terrorist	zzz israel	security	zzz afghanistan
8	run	season	team	game	hit	inning	player	games	race	baseball	right	home	yankees	manager	won
9	water	car	room	food	cup	hour	building	minutes	small	home	restaurant	add	house	large	town
10	drug	patient	doctor	problem	percent	research	health	scientist	cell	study	human	disease	test	medical	found

Table 5: Topics for Table 4 by index

Topics	1	2	3	4	5	6	7	8	9	10
	Politics	Sports, General	Crime	Computers	Arts	Business	Foreign Policy/International Politics	Sports, Baseball	Classifieds/Adverts	Pharma/Health/Science

Table 6: MNIST Dataset Topic Modeling, all words in topic

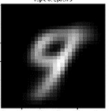
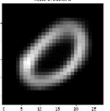
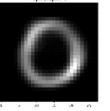
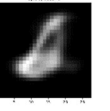
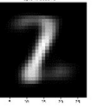
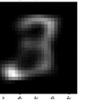
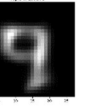
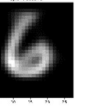
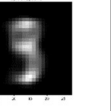
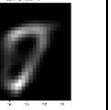
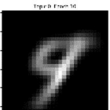
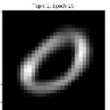
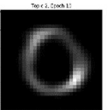
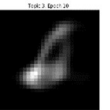
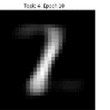
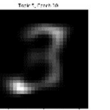
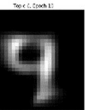
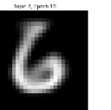
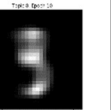
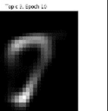
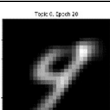
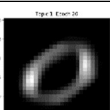
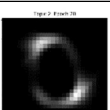
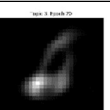
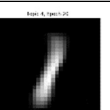
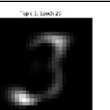
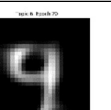
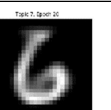
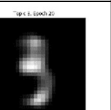
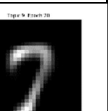
		Topics									
Epochs		1	2	3	4	5	6	7	8	9	10
	5										
	10										
	20										

Table 7: MNIST, top 196 words per topic


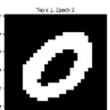
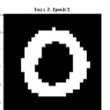
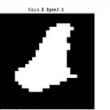




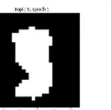

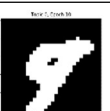
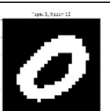
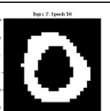

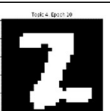


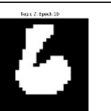
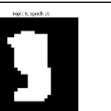
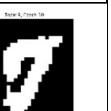
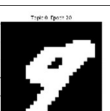
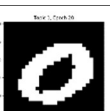
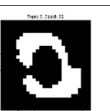
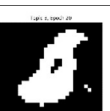
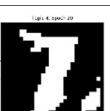







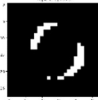

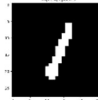



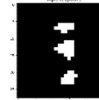
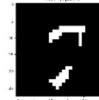


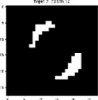

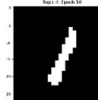
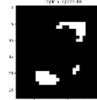
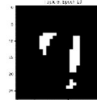



		Topics									
Epochs		1	2	3	4	5	6	7	8	9	10
	5										
	10										
	20										

Table 8: MNIST, top 50 words per topic

Topics

Epochs

	1	2	3	4	5	6	7	8	9	10
5										
10										
20	