Venyatha Manne

Task#2a: Exploratory Analysis of Corpus with LDA

Set 1

Perspective of surgery patients		Perspective of maternity patient	Perspective of patient unhappy with service	Perspective of patient unhappy with service

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
pain	told	first	doctor	office
surgery	doctor	doctor	time	doctor
back	said	baby	office	staff
would	would	would	rude	get
years	went	like	never	call
went	insurance	pregnancy	wait	patient
still	called	see	room	phone
doctor	asked	never	staff	never
said	back	time	see	appointment
told	never	even	like	time

Perspective of satisfied patient	Perspective of very pleased patient	Perspective of very pleased patient	Perspective of surgery patients	Words of thanks after procedure

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
staff	doctor	doctor	surgery	life
recommend	patients	best	surgeon	husband
time	great	years	breast	would
great	time	ever	cancer	god
questions	years	care	great	work
doctor	best	caring	would	dentist
would	always	one	procedure	thank
helpful	patient	good	recommend	years
always	like	patient	results	brown
highly	takes	doctors	performed	many

Set 2

		Perspective of unsatisfied patient	Perspective of patient who had to wait a long time	Perspective of patient who had to wait a long time

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
would	staff	doctor	appointment	doctor
told	time	told	room	get
call	patients	would	doctor	time
get	office	even	time	see
first	care	said	minutes	wait
doctor	always	office	see	hours
back	doctor	like	waiting	patient
said	patient	rude	day	never
went	years	get	office	always
never	get	never	another	back

Venyatha Manne

Perspective of very happy patient impressed with doctor	Perspective of satisfied patient who felt understood			Perspective of very happy patient who is grateful
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
recommend	feel	doctor	back	life
doctor	time	staff	pain	doctor
would	like	rude	would	years
time	questions	office	mri	best
highly	doctor	good	knee	one
always	really	time	said	saved
caring	comfortable	never	never	ever
staff	always	wait	years	man
knowledgeable	made	nice	left	doctors
professional	great	like	could	patient

Payments and insurance	Perspective of very happy patient	Perspective of surgery patient	Perspective of very happy patient impressed with doctor	
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
insurance	staff	surgery	manner	surgery
office	great	pain	doctor	went
medical	office	recommend	best	would
doctor	friendly	would	bedside	told
staff	doctor	years	great	said
pay	helpful	back	ever	never
care	experience	recovery	excellent	still
visit	happy	procedure	caring	done
patient	wonderful	went	knowledgeable	one
service	recommend	surgeon	physician	dentist

Perspective of surgery patient		Perspective of parents whose kids visited the pediatrician	Perspective of patient with a serious prognosis	Perspective of surgery patient
Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
surgery	mother	son	treatment	surgery
surgeon	doctor	children	diagnosis	performed
daughter	schwartz	child	options	removed
would	cancer	love	condition	procedure
plastic	another	kids	concerns	gordon
breast	found	years	medical	two
procedure	nose	else	patient	year
done	problem	pediatrician	shakiba	done
time	went	extra	results	heart
work	patient	parents	different	one

Venyatha Manne

From Set 1, I found Topics 1,2, and 6 hard to label and was ultimately not able to label Topic 2. Also, a few of my labels were repeated over topics. For example Topics 1 and 9 were both “Perspective of surgery patients”. I found Set 2 harder to label than Set 1 as there were more topics that had “filler words”. By this I mean modals like “would” or multiple synonymous verbs like “said” and “told”. These words and verbs are neutral and don't give me a lot of information regarding sentiment or details about the patient.

I think that overall these topics could be a lot more robust. I found that for the most part, the labels that I couldn’t place immediately lacked enough information for me to come up with a label even if I spent more time on it. The topic labels that were intuitive however were indicative of the internal structure/meaning of the ReadMD corpus. It made sense that I was able to separate out the topics mostly based on the type of patient (eg: surgical, pediatric) and their sentiment (eg: happy, unsatisfied, impressed). By having understandable topics that represented the corpus well, this LDA analysis serves the purpose of getting insights on patient experience from this collection. However, work is needed to be done on the corpus or method for the topics to be better, more robust, and more useful

Problem#2:

<u>Set 1</u>

		Perspective of patient who had to wait	Perspective of parents whose kids visited the pediatrician	Perspective of patient unhappy with service
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
call	pain	wait	time	office
get	go	time	child	staff
say	say	appointment	take	rude
would	would	see	like	patient
told	told	hour	feel	call
back	get	get	son	insurance
day	want	minute	care	get
see	back	room	see	bad
go	see	go	make	never
never	even	take	always	like

Perspective of very happy patient		Perspective of very happy patient	Perspective of surgery patients	
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
staff	patient	care	surgery	life
recommend	care	best	surgeon	save
great	time	recommend	procedure	would
time	year	great	pain	breast
question	take	patient	perform	make
office	need	would	year	patel
helpful	always	year	go	year
make	treatment	excellent	result	go
answer	see	highly	remove	one
friendly	health	manner	cancer	family

Venyatha Manne

Set 2

Perspective of patient who had to wait			Perspective of very happy patient	Perspective of patient of family doctor
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
wait	say	go	recommend	life
time	told	say	time	ever
appointment	ask	get	question	treat
see	go	would	staff	save
hour	never	told	answer	child
get	bad	take	take	year
office	see	time	great	know
minute	would	back	would	family
patient	back	day	care	care
staff	like	visit	helpful	patient

Perspective of satisfied patient who	Perspective of very happy patient	Perspective of happy in-ward patient		Perspective of surgery patients
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
feel	care	best	call	surgery
make	excellent	manner	office	perform
like	recommend	bedside	get	problem
comfortable	patient	ever	patient	surgeon
great	year	great	phone	pain
care	family	good	never	back
patient	staff	one	would	remove
felt	highly	care	test	well
seem	would	see	return	result
staff	physician	love	ask	recommend

		Perspective of patient with a serious prognosis	Perspective of maternity patient	Perspective of patient of family doctor
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
patient	staff	patient	baby	son
care	office	care	pregnancy	eye
time	rude	time	first	mother
health	front	problem	deliver	mom
staff	would	treatment	make	old
need	like	take	go	give
well	work	listen	look	child
issue	nurse	diagnosis	breast	see
know	well	see	amaze	year
concern	people	treat	would	use

Perspective of patient unhappy with service			Perspective of happy patient	Perspective of dental patient
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
insurance	pain	year	treatment	surgery
office	year	see	patient	surgeon
pay	go	would	skill	cancer
money	test	know	thank	another
bill	give	old	lack	ray
rude	would	daughter	competent	teeth
company	problem	practice	compassion	go
charge	month	move	god	tooth
staff	knee	man	knowledge	experience
horrible	medication	good	communication	nose

In order to lemmatize the corpus effectively, I tagged the words I was going to lemmatize with their part of speech to ensure more accurate lemmatization. Thus these words have been both noun and verb lemmatized. I found it difficult to label these topics as I felt like they were generally noisy. For example, 8 of the words in the first topic of Set 1 just consist of modals, neutral verbs, and adverbs. This doesn't give me enough information that would allow me to assign a useful label.

Also, I found that some of the labels I did assign weren't very obvious from the top words in that topic. For example, topic 13 in set 2 does have some indication of a patient who has a more serious prognosis but this isn't clear from the majority of the words. There are words that infer longer term issues such as: "time", "treatment", "diagnosis", and "problem" but words like "care", and "see" don't show this as much. I think that I am concerned about how my personal bias of these words affect my labelling and I believe I would greatly benefit from having another student try to label these topics as well.

I think that although the topics I assigned are understandable and do somewhat capture the meaning of the corpus, I don't think they serve the purpose particularly well as there were so many topics I was unable to label. This shows that the goodness of the topics might not be great

Problem#3:

The program outputs with and without lemmatization were pretty similar in that there was a lot of general overlap with the top words under both conditions ("doctor", "recommend", "best", etc). This makes sense because they are using the exact same corpus.

I was able to confidently label 9/10 of the topics for Set 1 without lemmatization but only 6/10 when using the lemmatized corpus. With regards to Set 2, I labeled 15/20 and 14/20 topics when running LDA without and with lemmatization respectively. Surprisingly, I found that running LDA without lemmatization provided me with slightly better results. I'm unsure if this is generally true or if it is due to a poor lemmatization technique or specific to the corpus I am using. I found the “goodness of the topics” to be about the same generally. There was a lot of topic label overlap and I dealt with similar issues when trying to assign labels to the topics.

Because the goodness of the topics is about the same, I can use the runtimes to understand if lemmatizing is worth doing. Without lemmatization, the LDA’s took 130 seconds and 133 seconds respectively. Including lemmatization and the LDA runtime, the runtimes for Part 2b took 267 seconds and 258 seconds for Set 1 and 2 respectively. This shows that lemmatization may not be worth it when conducting LDA analysis with this corpus and this method. I think that in order to improve my topics, I would have to go back and conduct some more data cleaning to make my topics less noisy.

Task#2b: Exploratory Analysis of Corpus with ccLDA

Problem#1:

1) Set1: 10 topics and 2000 iterations

		Logistics of patient experience	Perspective of concerned/curious patient	Perspective of patient who waited a long time
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
go say take told get see would month day could	bigelow see doctor know like year child help go good	call office get insurance phone staff back need return visit	bigelow time question answer make manner doctor patient feel concern	wait time see appointment get hour room minute office doctor

	Perspective of happy patient	Perspective of surgery patient	Perspective of patient with a serious prognosis	Comments on doctor’s specialty
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
make like go first would feel say want come never	bigelow staff recommend would doctor care office anyone great experience	surgery pain procedure bigelow back go would surgeon result look	problem treatment medication condition give medical patient specialist diagnosis bigelow	doctor bigelow patient care ever one treat practice medical physician

2) Set2: 20 topics and 2000 iterations Calculate the runtime of ccLDA in each setting.

	Prescriptions	Perspective of maternity patient	Perspective of dermatology patient	Yearly check-up
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
bigelow know doctor good help life doc think need find	medication take problem get help give med diagnose need try	treatment visit give first experience clinic opinion look one without	say ask told come back want tell even go never	see year doctor time new go since several last another

Perspective of patient who had to wait a long time	Perspective of concerned/curious patient		Perspective of surgery patient	Perspective of very happy patient
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
wait time see appointment hour minute room long get late	question bigelow time answer manner bedside take concern explain listen	get like want work else know bad go think one	pain back go month problem surgery walk could day year	would bigelow recommend doctor anyone care experience need kind great

Payments and Insurance	Perspective on Doctor’s aptitude	Perspective on family doctor	Perspective of surgery patient	Diagnostic Tests
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
insurance office visit pay money service even bill medical make	patient medical bigelow health physician care practice issue knowledge professional	patient care bigelow treat doctor family mother well need life	surgery procedure surgeon perform look remove result breast would scar	test told result blood go day follow month order later

	Appointment logistics	Maternity patient		Perspective of very happy patient
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
make feel like bigelow time take felt really sure talk	call get office appointment day phone back return never week	child bigelow son daughter baby first husband go pregnancy never	doctor bigelow ever care patient one never problem take met	staff bigelow office nurse nice always extremely good experience helpful

Since I found that the quality of my LDA results were similar between texts with and without lemmatization, I decided to run ccLDA with and without lemmatization and take a quick glance between both results and then choose the set of results that I believed would have better topics. By doing this, I realized the quality of the lemmatized results were better because the non-lemmatized version had quite a few instances of the same root word. For example: “doctor” and “doctors” in the same topic. Therefore I continued with ccLDA for the lemmatized corpus.

The LDA runtimes for the lemmatized corpus (excluding time taken to lemmatize) were 142 seconds and 133 seconds for Set 1 and Set 2 respectively. The ccLDA runtimes (excluding time taken to lemmatize) were 1038s and 12m13s respectively. I excluded the lemmatization runtimes as I only needed to do it once to run all 4 commands. As you can see, there is a big difference between these runtimes– ccLDA is much longer.

The results from the ccLDA were noisier than I expected. For example, the name “bigelow” appears numerous times in both set 1 and set 2 which doesn’t provide useful information when labelling topics. I would actually say that the ccLDA produced noisier results than the LDA. For example the eighth topic in set 2 provided almost no topwords of value. However, this isn’t to say that the topics I was able to label were not useful. I actually found it to be easier labelling Set 2 than Set 1 of ccLDA or any other LDA labelling. The topics I was able to gather were more unique than the analyses in the other sections. For example, Set 2 Topic 15 is labelled a diagnostic test and I wasn’t able to find a similar topic in my previous analysis. I believe that the topics produced in the ccLDA analysis represent the RateMD corpus well and provide some insight into patients experience and the topics that arise from their reviews.