Assignment 3: Topic Models for Healthcare

Issued: 11/12/2018
Due: 11/20/2018

Total points: 100

The goal of this assignment is to give you hands on experience with topic models for a real application. Thus, you will implement and experiment with some topic models, and write a short report about your experiences and findings. As you know, topic models capture the most important topics in a text corpus.

Scenario:

You are a newly employed text analyst at a data analytics company. Congratulations!
Your first project is for a new client, a (new) healthcare company. The company is in the process of collecting a very large dataset of patient comments from all their clinical locations throughout the country and is interested in getting insights on patient experience from this collection (i.e., insights into patient comments about doctors, nurses, clinics, healthcare services, etc.). They will deliver the data in 6 months.

However, since this is the first project on healthcare for your company, your boss would like you to start working on it asap and is asking you to perform some explorative analysis of a similar dataset (of your choice) and prepare a preliminary report on what kinds of insights can be generated from some a kind of data.
For this assignment, you have to sit down and decide on the design decisions you need to make to solve the task, then test your models on a relevant healthcare dataset, and write a report on the results.

To help you in this process, you are asked to work on a series of tasks which describe the exploratory analytics process for this application.

1) **<u>Task#1: Corpus collection and Corpus Descriptive analysis</u>** [20 points]

First, you have to find and collect a dataset that is similar to the one the healthcare client will provide later. The task is rather challenging since such secondary data are difficult to find due to compliance issues. However, after considerable research, you manage to find a freely-available patient review dataset from RateMD (http://ratemds.com), one of the most popular platforms for physician reviews in the United States.

RateMD Data Description:
Founded in 2004, RateMD has the largest number of user-submitted reviews with narratives by a large margin. In RateMD every doctor is given an ID which uniquely specifies a doctor's profile information: name, gender, location, specialization. The website also provides the average rating for a doctor (on a a Likert scale of 1 (low) to 5), the review text.

You decide to crawl the website for all the comments for a period of 10 years (2007 - 2018) using a Java WebCrawler script, thus generating 101,902 reviews. The institutional review board approval was obtained for this study.

Here is an example of an entry in this dataset:

Dr. Shirley A. Thomas Female Fishers, IN Gynecologist (OBGYN) Overall rating: 4.75 Best doctor in the world. She not only is beyond knowledgeable from her 40 years of practice but she cares about us.. A lot. She doesn't need the money, she does this because it gives her joy delivering babies. I would fly from Cali to see her, that's how much I trust her.

Each entry in the corpus consists of 6 tab-separated fields:
[Dr's Name; Gender; Location; Specialization; Overall rating; Review]

You have to work on the following problems:

<u>Problem#1:</u>
Do a descriptive analysis of your corpus and provide (in the table below): the distribution of reviews per gender and sentiment (show both count and percent coverage). Here the sentiment can be only positive or negative -- determined by mapping the overall ratings at most 3 into negative (i.e., [1,3]) and those at least 4 into positive (i.e., [4,5]). E.g., the overall rating of the example above maps into positive sentiment.

Counts

Gender	**Sentiment**		**Total**
	Positive	**Negative**	20,421
Female	2,686	2,120	4,806
Male	9,877	5,738	15,615
Total	12,563	7,858	20,421

Percentages: What % of [gender] reviews are [sentiment]?

Gender	Sentiment		Total
	Positive	Negative	
Female	56%	44%	24%
Male	63%	37%	76%
Total	62%	38%	100%

Also provide and comment on the size of the reviews in the corpus: i.e., length of the smallest review and of the largest review, as well as the average length of the reviews in the corpus. Here we consider a coarse definition of review length as the number of raw tokens (i.e., any sequence of characters separated by space and/or beginning/end of review).

Problem#2:
Why is this dataset from RateMD a valid, relevant corpus for your project?
For this, you are referred to the corpus design principles discussed in class (Lecture 5). In particular, consider the following helping questions (your reference corpus is the corpus to be provided by the healthcare company) and fill in the entries:

No.	Questions	RateMD corpus	Healthcare company's corpus
1	What is the corpus language variety (i.e., genre)?	Reviews written by patients of doctors who are not necessarily part of the company's clinic.	Reviews written by patients of the company's clinics
2	What is the size of the corpus?	20,421 reviews	500,000 reviews
3	What meta-data is provided with the reviews?	Doctor's name, doctor's gender, doctor specialty, clinic location, review sentiment (0-5), qualitative rating (text)	Doctor's name, gender, clinic location; review sentiment
4	What socio-demographic information is provided about the patients who wrote the reviews?	None systematically exists, however, some may exist within the reviews.	Gender, age, economic and educational status
5	Is the corpus balanced along the meta-data dimensions considered? (look only at sentiment and gender)	The corpus is not balanced, we have more men, and the men tend to be rated more positively than women.	No (but the distribution of meta-data dimensions exhibits the natural distribution)

Compare the answers to the questions in table above. Identify and comment on one important disadvantage of using this corpus as a good, relevant corpus for this project (i.e., 'good, relevant' here means how similar is it to the corpus the healthcare company will provide in the future).
Hint: Think of who is writing the reviews in RateMD? How does this compare with the healthcare company's data?

- RateMD contains no (clear) patient information, as the ratings from this company would
- RateMD is more focused on reviewing doctors, rather than on reviewing the practice itself (which includes doctor, staff, nurses, etc.)
- RateMD is will include a lot of selection bias, because reviewers feel motivated enough to go on the web and write a review. This company's reviews, however, will likely be solicited and contain more unbiased reviews.

2) **Task#2: Exploratory Analysis of Corpus with LDA** [40 points]

You have to write a python program (lda_run.py) that takes as input the corpus, a given number of topics k, and generates these topics. For this task you will experiment with LDA (Latent Dirichlet Allocation).

Specifically, as explained in class, you have to consider a number of steps:

Step 1: Clean the corpus
Your text corpus has to be cleaned before you use it as input to the topic model.
Thus, you have to convert the text reviews to lowercase, tokenize them, and then remove punctuations. Of course, you also have to remove stop words. You also want to experiment with lemmatization as well, so you have to test your LDA model *with* and *without* lemmatization.

Step 2: Create the dictionary
Here you will create the term dictionary from your corpus. Recall that in this process every unique term is assigned an index.

Step 3: Do more preprocessing
You also decide to filter the terms which occurred less than 10 times. How large is your vocabulary?

- I filtered out English stop words
- The most frequent 200 words in the reviews
- Words that occurred fewer than 10 times in the reviews
- The doctors last names
- All numbers
- After this manipulation, my vocabulary is 2,893 words

Step 4: Convert list of documents (i.e., reviews) into Document Term Matrix using dictionary prepared at Step 3.

Step 4: Run the LDA model on the document term matrix

Step 5: For each of the k topics, print the top 10 words

You have to work on the following problems:

Problem#1:

Here you run the LDA model without lemmatization.

Place the topics in one or two tables (showing the top 10 words per topic as done in class). Then analyze the goodness of your topics – meaning, can you manually label each topic with a topic word or phrase? Could you find a label for all the topics? Which ones were the easy to label and which were more noisy (and thus, not easy to label)?

Family Planning	Cancer/Research	Good Bedside Manner - caring	Long Wait	Good Bedside Manner - smart	Unknown	Bad Reviews	Surgery/ Unknown	Serious Illness	Unknown
son	husband	amazing	waiting	thorough	months	worst	procedure	medication	listen
child	cancer	truly	phone	concerns	saw	money	skin	diagnosed	else
baby	breast	cares	hours	listens	second	horrible	look	blood	side
daughter	hospital	compassionate	hour	health	later	wrong	face	months	health
pregnancy	er	thank	apt	easy	weeks	test	removed	hospital	medicine
children	knowledge	awesome	waited	talk	knee	pay	foot	condition	try
delivered	top	everyone	calls	knowledgeable	opinion	bad	body	right	seems
pregnant	free	helped	exam	things	year	unprofessional	hip	tests	think
old	area	comfortable	front	understand	ago	terrible	bad	symptoms	issues
husband	research	friends	late	makes	procedure	nothing	area	home	gives

Problem#2:

Do Problem#1 above, but with lemmatization this time.

Unknown	Eyecare	Breast Augmentation	Dental	Unknown	Unknown	Family/ Unknown	Book Appointment/ Unknown	Unknown	Clinic/ Unknown
doc	test	procedure	daughter	mri	husband	son	mother	medication	appt
cancer	eye	breast	tooth	dad	thyroid	bill	look	try	wait
state	come	implant	exam	vein	prescription	child	clinic	start	week
month	pap	follow	baby	specialist	seem	father	hospital	birth	pay
speak	order	explain	decide	scar	leak	wife	refill	arm	think
symptom	script	able	money	condition	tip	nose	cigna	body	flu
phone	later	late	cavity	kid	ovary	hour	sign	psychiatrist	ask
lab	lasik	code	routine	program	ear	give	schedule	severe	steroid
record	front	want	therapy	comp	sister	hand	colonoscopy	use	walk
receive	side	infection	health	ivf	forget	hip	book	tech	woman

Problem#3:

Compare the LDA model's output with and without lemmatization. Which of these preprocessing settings generates better topics?

I believe that the model without lemmatization generated better, clearer topics. It seems to make more natural groupings of words that have clear themes.