**Assignment 3**: Topic Models for Healthcare

Issued: 11/12/2018

Due: 11/23/2018

Total points: 100

1) **Task#1: Corpus collection and Corpus Descriptive analysis** [40 points]

**Problem#1:**

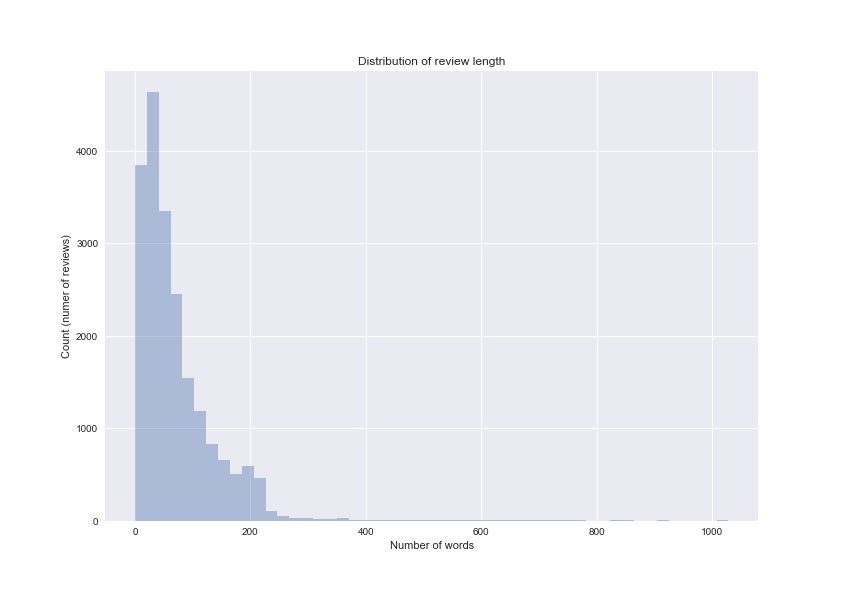
Do a descriptive analysis of your corpus and provide (in the table below): the distribution of reviews per gender and sentiment (show both count and percent coverage). Here the sentiment can be only positive or negative -- determined by mapping the overall ratings of at most 3 into negative (i.e., [1,3])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

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Sentiment** | | **Total** |
| **Positive** | **Negative** |  |
| **Female** | 2,953 (61.4%) | 1,853 (38.6%) | 4,806 (23.5%) |
| **Male** | 10,616 (68.0%) | 4,999 (32.0%) | 15,615 (76.5%) |
| **Total** | 13,569 (66.4%) | 6,852 (33.6%) | 20,421 |

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

Minimum size is 0, i.e. no review. If we discard 0-length review, the smallest review is a single world.

The longest review is 1028 words. The average review is just under 71 words long. The median review is 51 words long. The distribution of review length is considerably right skewed, as shown in the following histogram.



**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 and fill in the entries in the table below.

Note: Your reference corpus is the corpus to be provided by the healthcare company.

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Questions** | **RateMD corpus** | **Healthcare company’s corpus**  **(i.e., reference corpus)** |
| 1 | What is the language variety of the corpus (i.e., genre)? | Doctor’s reviews written by patients. The Doctors in the corpus do not necessarily work for the company’s clinics. The patients writing the reviews might not be patients of the clinics. | 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, gender, clinic location, Doctor’s specialization, numeric rating (0-5), qualitative review | Doctor’s name, gender, clinic location; review sentiment |
| 4 | What socio-demographic information is provided about the patients who wrote the reviews? | None. In some cases, this information could be extracted from the review, but no data about the patient is readily available. | Gender, age, economic and educational status |
| 5 | Is the corpus balanced along the meta-data dimensions considered? (look only at sentiment and gender) | No. As of October 2018, 65% of Doctors in the US are male; however, more than 76% of the reviews in the RateMD corpus are about male Doctors. Moreover, male Doctors are on average rated more positively. | No (the dimensions are not uniformly distributed; they exhibit a natural distribution) |

Compare the answers to the questions in the table above (3rd and 4th columns) and use this comparison to identify and comment on one important disadvantage of using RateMD as a good, relevant corpus for this project (i.e., ‘good, relevant’ here means how similar it is to the corpus the healthcare company will provide in the future).

There are a couple of reasons why RateMD might not be a relevant corpus for this project. First, in the RateMD corpus we don’t have explicit information about the patient writing the review. Second, RateMD only contains review of Doctors, whereas our client is interested in evaluating the overall experience of its patients, which includes interactions with not only Doctors, but also nurses, administrative staff at the clinics, etc. Finally, the data in the RateMD corpus seems to be skewed in favor of male doctors, both in terms of number of reviews and in terms of average rating.

2) **Task#2:** **Exploratory Analysis of Corpus with LDA** [60 points]

You have to write a python program that takes as input the corpus (i.e., your RateMD 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, in this procedure you have to consider a number of steps:

**Step 1: Clean the corpus**

The cleaning I applied consists of the following:

* convert reviews to lowercase
* tokenize reviews using nltk.word\_tokenize
* filter out stop words (the basis list of stop words in the English language was augmented with a couple of words found in the text)
* filter out infrequent words (words that appeared less than 10 times in the whole corpus)
* filter out Doctor’s last names and digits/numbers

**Step 2:** **Create the dictionary**

After the cleaning step, the vocabulary consisted of XXXXXX words.

**Problem#1 (no lemmatization)**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Problem#2 (with lemmatization)**

**Problem#3:**

Compare your program’s output with and without lemmatization. Which of these settings generates better topics? Is lemmatization worth doing? (compare the goodness of the topics with and without lemmatization). Explain.

**Extra-credit:** [35 points]

**Problem#1**: [10 points]

Repeat Task#2 but this time with k = 20 topics.

What do you notice? Are the results with 20 topics better than those with 10 topics (under both with and without lemmatization scenarios)? Again, ‘better results’ here refers to the goodness of your topics.