**Assignment 3**: Topic Models for Healthcare

Issued: 11/12/2018

Due: 11/23/2018

Total points: 100

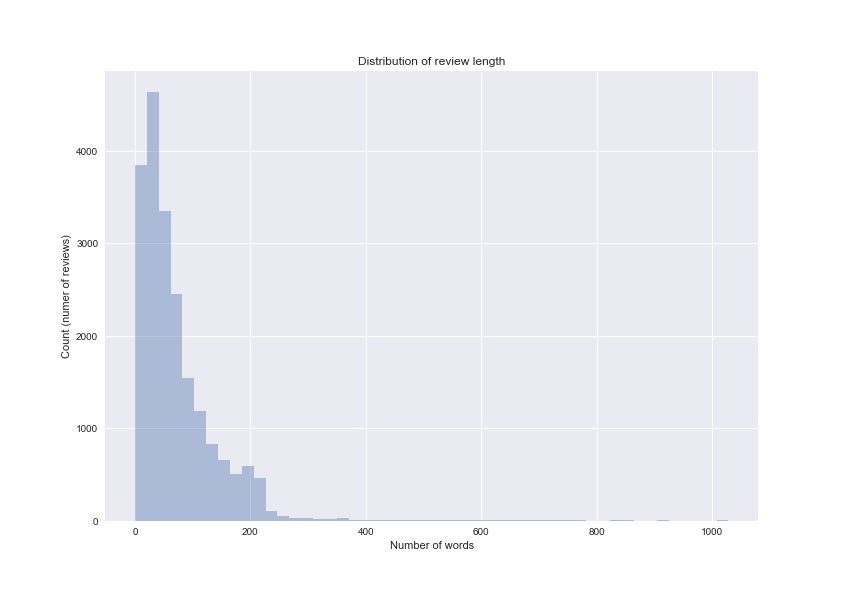
1) **Task#1: Corpus collection and Corpus Descriptive analysis** [40 points]

**Problem#1:**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

Minimum size is 0, i.e. no review. If we discard 0-length reviews, 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:**

|  |  |  |  |
| --- | --- | --- | --- |
| **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) |

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]

**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 4227 words without lemmatization and 3096 with lemmatization.

**Problem#1 (no lemmatization)**

|  |  |
| --- | --- |
| Unknown | “son”, “cancer”, “skin”, “breast”, “face”, “child”, “one”, “look”, “nose”, “body” |
| Staff (negative) | “staff”, “office”, “rude”, “like”, “never”, “go”, “would”, “get”, “patients”, “ever” |
| Physician (ER?) | “patients”, “patient”, “care”, “medical”, “health”, “time”, “treatment”, “years”, “one”, “physician” |
| Appointments | “time”, “appointment”, “see”, “told”, “room”, “called”, “office”, “get”, “said”, “would” |
| Staff (positive) | “time”, “staff”, “always”, “questions”, “feel”, “great”, “helpful”, “takes”, “office”, “friendly” |
| Billing | “insurance”, “pay”, “office”, “dentist”, “bill”, “medical”, “done”, “company”, “money”, “billing” |
| Surgery (good) | “recommend”, “surgery”, “would”, “staff”, “highly”, “great”, “excellent”, “surgeon”, “experience”, “anyone” |
| General physician (good) | “years”, “family”, “would”, “recommend”, “caring”, “great”, “care”, “ever”, “always”, “one” |
| Surgery (bad?) | “pain”, “surgery”, “years”, “back”, “life”, “knee”, “severe”, “ago”, “able”, “months” |
| Actions | “told”, “went”, “said”, “would”, “surgery”, “could”, “hospital”, “back”, “got”, “go” |

**Problem#2 (with lemmatization)**

|  |  |
| --- | --- |
| Bedside manners (good) | “recommend”, “would”, “highly”, “manner”, “care”, “anyone”, “bedside”, “excellent”, “great”, “staff” |
| Staff/nurses (good) | “time”, “care”, “staff”, “take”, “question”, “patient”, “great”, “answer”, “always”, “explain” |
| Actions | “go”, “tell”, “get”, “say”, “would”, “take”, “know”, “see”, “want”, “problem” |
| Severe conditions | “pain”, “life”, “year”, “back”, “surgery”, “save”, “help”, “thank”, “problem”, “go” |
| Appointments | “office”, “time”, “wait”, “get”, “call”, “appointment”, “go”, “see”, “tell”, “never” |
| Staff (positive) | “staff”, “office”, “go”, “great”, “nurse”, “nice”, “experience”, “rude”, “skin”, “look” |
| Unknown | “time”, “care”, “see”, “patient”, “like”, “feel”, “always”, “make”, “year”, “go” |
| Health insurance | “patient”, “medical”, “insurance”, “care”, “treatment”, “physician”, “test”, “condition”, “pay”, “need” |
| Family | “husband”, “u”, “son”, “daughter”, “family”, “mother”, “heart”, “hospital”, “cancer”, “year” |
| Surgery | “surgery”, “procedure”, “surgeon”, “perform”, “result”, “do”, “go”, “breast”, “would”, “dentist” |

**Problem#3:**

The output is not drastically better with or without lemmatization. It seems that without lemmatization we get slightly more defined topics. For this reason, it doesn’t seem like applying lemmatization pays off.

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

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

The top 10 words for the 20 topics can be found in the Jupyter Notebook. With 20 topics we get very similar results in terms of goodness of the topics.