# **Assignment 3**: Topic Models for Healthcare

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.

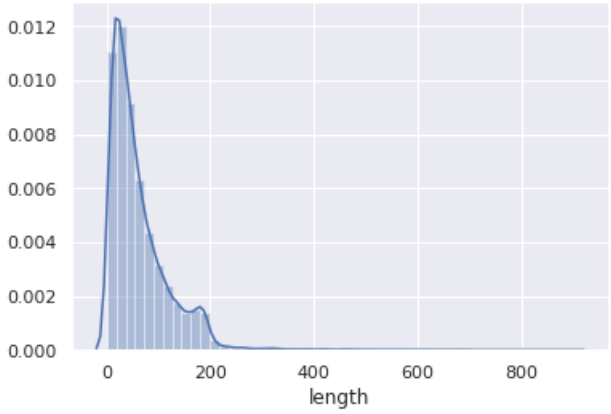
| **Gender** | **Sentiment** (count and %) | | **Total (count and %)** |
|---|---|---|---|
| | **Positive** | **Negative** | 20421 100% |
| **Female** | 2686<br>13.15% | 2120<br>10.38% | 4806<br>23.53% |
| **Male** | 9877<br>48.37% | 5738<br>28.10% | 15615<br>76.47% |

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).

Length of smallest review is 0 while the max is 900; (it shows 1 because we getting rid of null values will result more trouble)
Descriptive statistics as fellow:

| | rating | length |
|---|---|---|
| **count** | 20421.000000 | 20421.000000 |
| **mean** | 3.741528 | 62.998531 |
| **std** | 1.480634 | 60.164758 |
| **min** | 0.750000 | 1.000000 |
| **25%** | 2.250000 | 22.000000 |
| **50%** | 4.500000 | 45.000000 |
| **75%** | 5.000000 | 85.000000 |
| **max** | 5.000000 | 899.000000 |

The distribution of review text length is as fellow:



**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)? | Public narratives by patients. | 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, specialization, overall rating, review text | Doctor's name, gender, clinic location; review sentiment |
| 4 | What socio-demographic information is provided about the patients who wrote the reviews? | Nothing | Gender, age, economic and educational status |
| 5 | Is the corpus balanced along the meta-data dimensions considered? (look only at sentiment and gender) | No; neither gender or sentiment is not distributed evenly. | 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).
Hint: Think of who is writing the reviews for RateMD. How does this compare with the healthcare company's data (i.e., who wrote of the reviews there).

The data is definitely relevant to this project, yet decent but not as good compared with the company's data. The company has larger text corpus and, more importantly, they have more detailed info about the users who wrote the review. However we have most of the doctors information available and that should be enough for many analysis, so it's not as good but still descent.