**Preliminary Report**

# **Project: Understanding the Determinants of Health**

Team members: Gagan Kaushal, Chengyuan E, Tiam Moradi

**1. Introduction**

BU School of Public Health has access to clinician notes collected during routine medical visits at the Boston Medical Center. These notes are written free-text reports of visits capturing the clinician’s findings, observations, and diagnoses.

We need to predict a diagnosis of substance abuse, alcohol use and tobacco smoking from the data in the notes by using techniques such as NLP and machine learning methods.

**2. Descriptive Statistics**

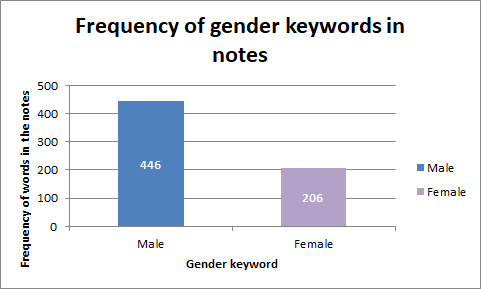
* 1. **Number of Logician Notes**: 393
  2. **Number of Epic Notes:** 1824
  3. **Number of patients:** 22 (Based on the number of unique IDs)
  4. **Target labels:**
     1. alcohol use
     2. current smoking status
     3. drug use

**3. Unique Patient IDs in the notes**

|  |  |
| --- | --- |
| **Logician Notes** | **Epic Notes** |
| 2642 | 2654 |
| 2614 | 2638 |
| 2665 | 2642 |
| 2667 | 2683 |
| 2646 | 2597 |
| 2597 | 2672 |
| 2593 | 2656 |
| 2645 | 2715 |
| 2710 | 2602 |
| 2683 | 2646 |
| 2656 | 2662 |
| 2602 | 2614 |
| 2682 | 2714 |
| 2654 | 2682 |
| 2662 | 2665 |
| 2714 | 2667 |
| 2715 | 2699 |
| 2699 | 2698 |
| 2698 | 2710 |
| 2672 | 2593 |
| 2638 |  |
| 40000 |  |

**4. Visualizations**

1. **The bar graph shows the frequency of different labels / unique features**



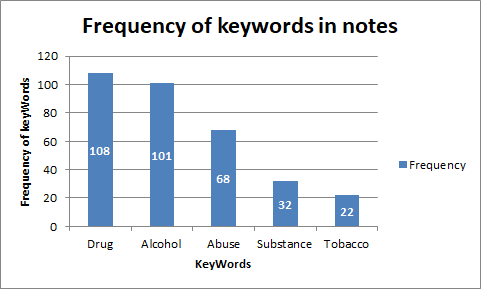
|  |  |
| --- | --- |
| **Gender** | **Frequency** |
| Male | 446 |
| Female | 206 |

**Observations from above Bar graph**

From the above visualization, we observe that the ‘male’ keyword occurs almost 2 times more as compared to ‘female’ keyword.

So, as per our preliminary investigation, it seems that more male patients are addicted to substance abuse as compared to females.

1. **The bar graph shows the frequency of different labels / unique features**



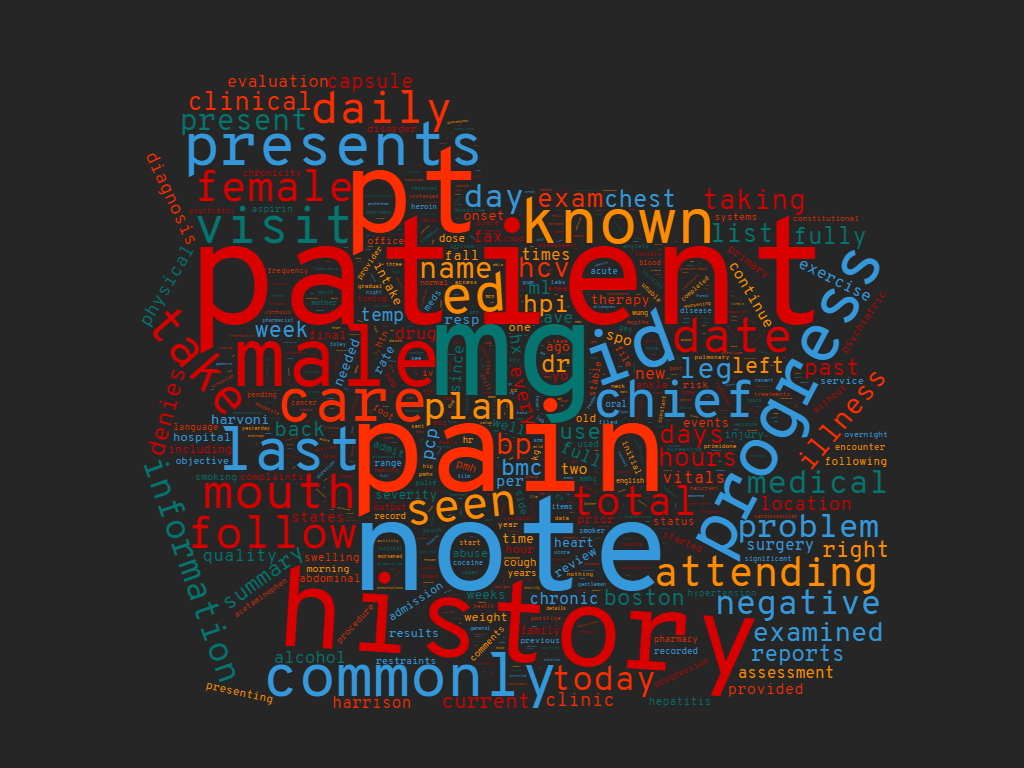
|  |  |
| --- | --- |
| **KeyWord** | **Frequency** |
| Drug | 108 |
| Alcohol | 101 |
| Abuse | 68 |
| Substance | 32 |
| Tobacco | 22 |

**Observations from above Bar graph**

From the above visualization, we can observe that the **‘drug’** keyword occurs the most in the notes. So, as per preliminary investigation, we suspect that there are more patients affected by ‘**drug**’ as compared to **‘Tobacco’**

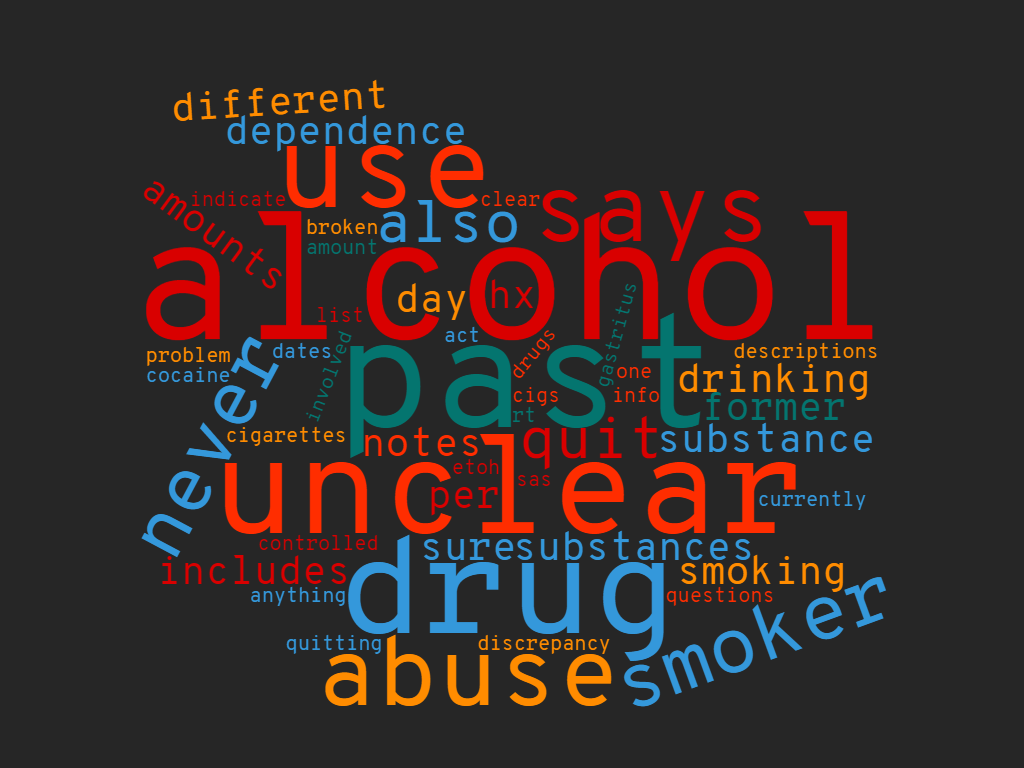
1. **Word Cloud to showcase the most frequent words in the notes**

* **Epic Notes**



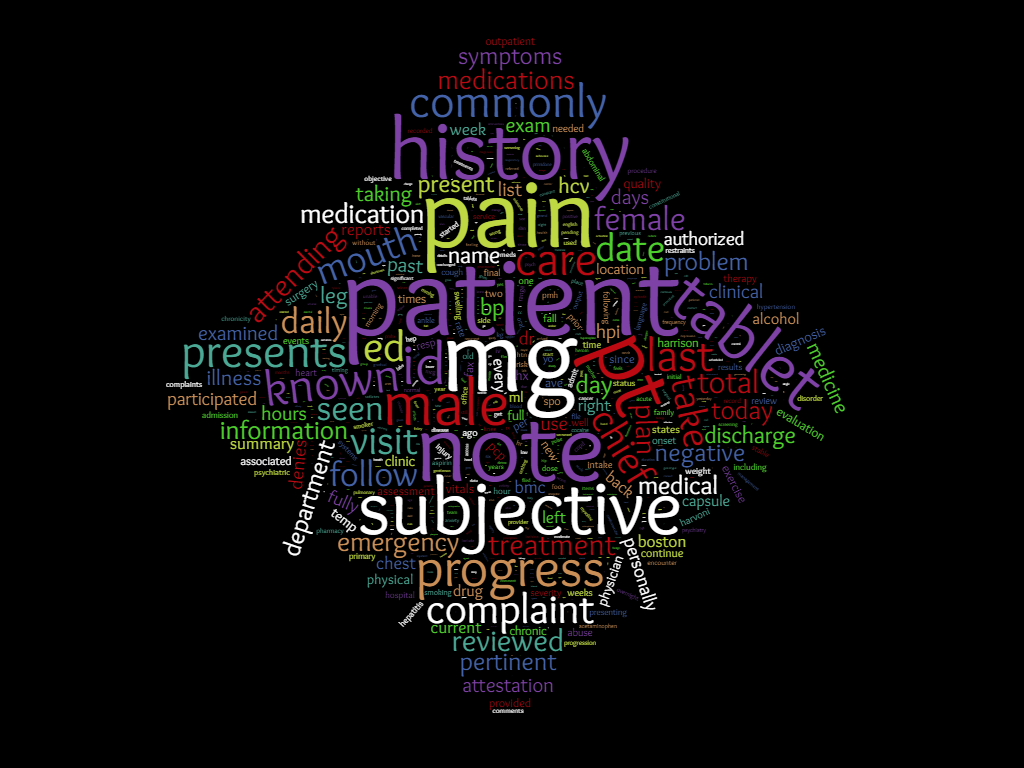
**Top 10 frequent words (with frequencies) in Epic Notes**

* ('patient', 1536)
* ('mg', 635)
* ('pain', 613)
* ('note', 610)
* ('pt', 551)
* ('history', 548)
* ('subjective', 513)
* ('tablet', 453)
* ('male', 446)
* ('id', 426)
* **Logician Notes**



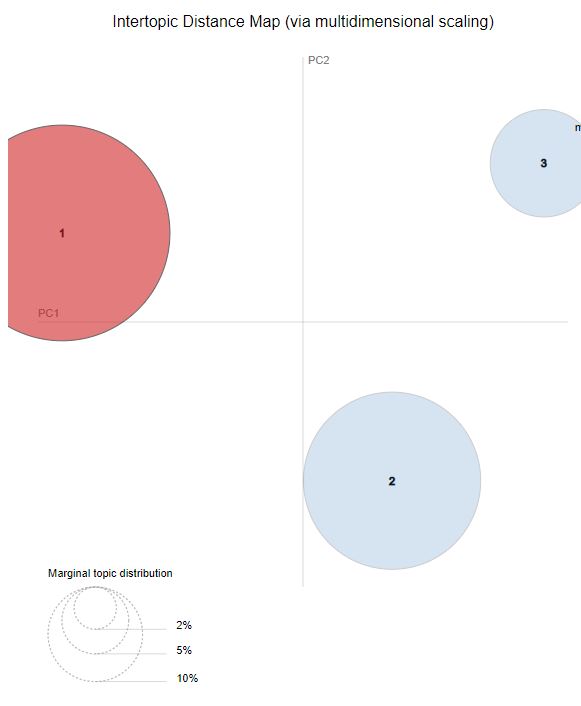
**Top 10 frequent words (with frequencies) in Logician Notes**

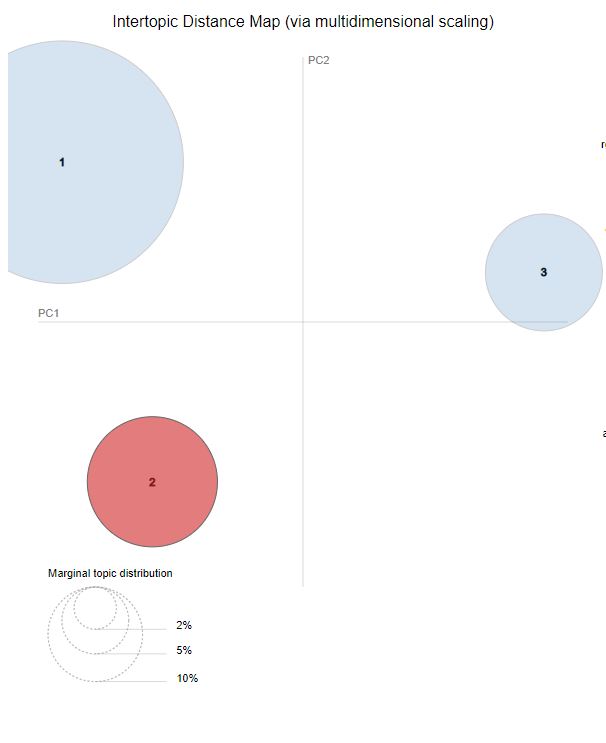
* ('alcohol', 18),
* ('past', 17),
* ('unclear', 16),
* ('drug', 15),
* ('use', 11),
* ('present', 10),
* ('says', 9),
* ('abuse', 9),
* ('current', 7),
* ('smoker', 7)
* **Logician and Epic Notes**

****

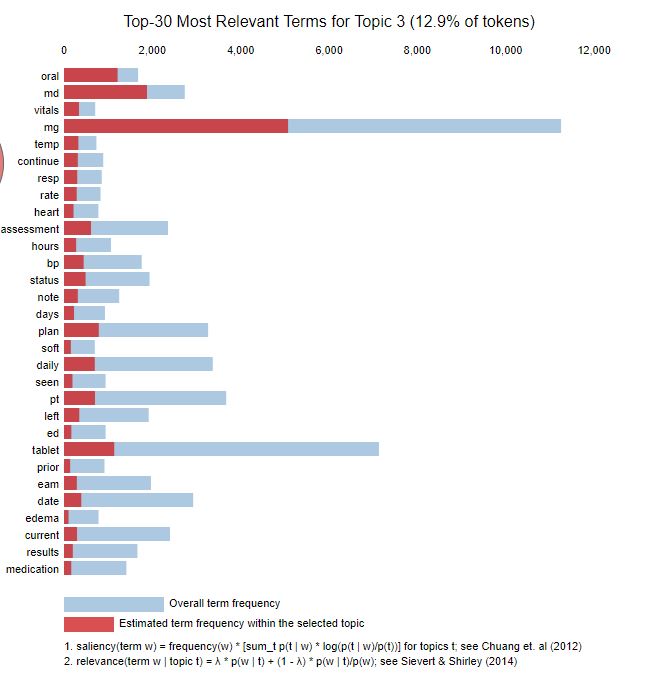
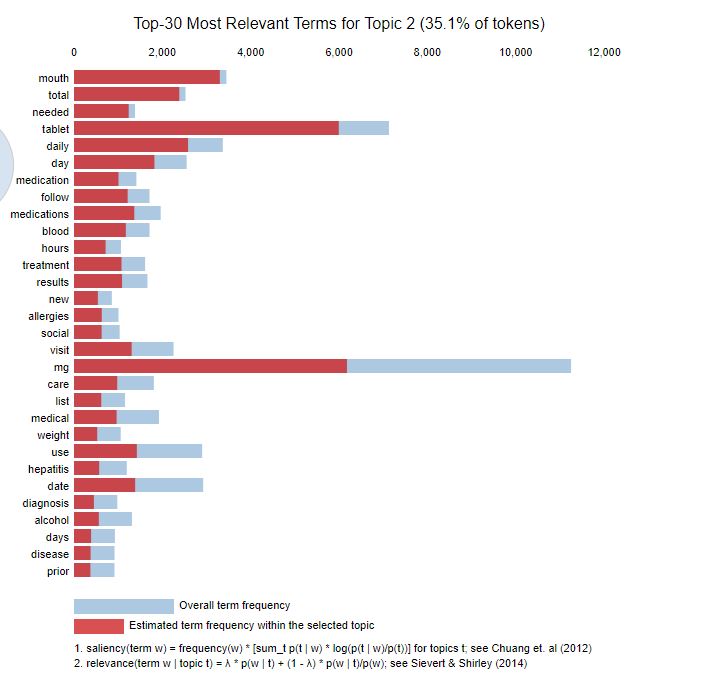
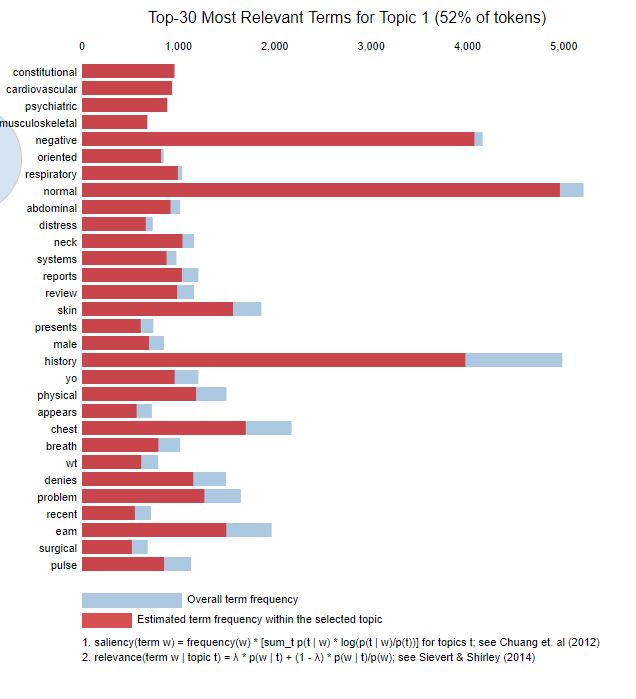
**Top 10 frequent words (with frequencies) in Epic and Logician Notes (combined)**

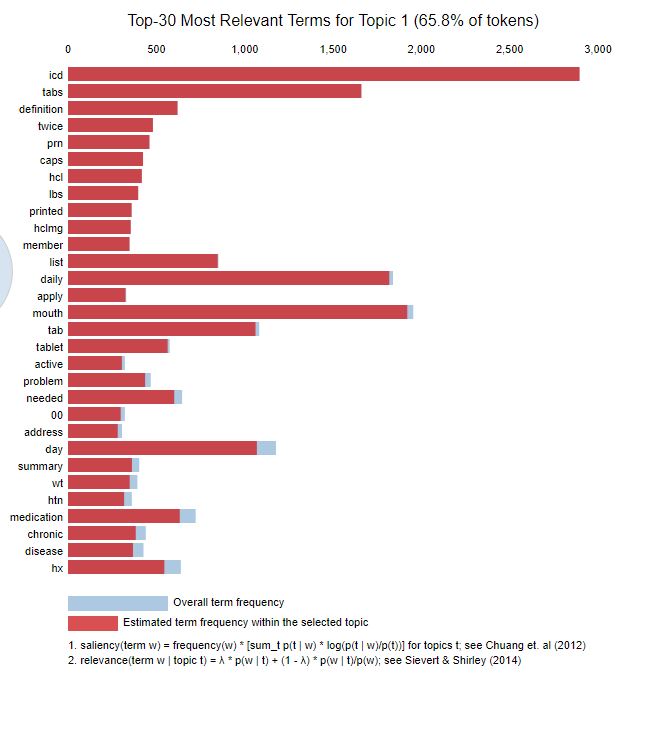
* ('patient', 1536)
* ('mg', 635)
* ('pain', 613)
* ('note', 610)
* ('pt', 551)
* ('history', 548)
* ('subjective', 513)
* ('tablet', 453)
* ('male', 446)
* ('id', 426

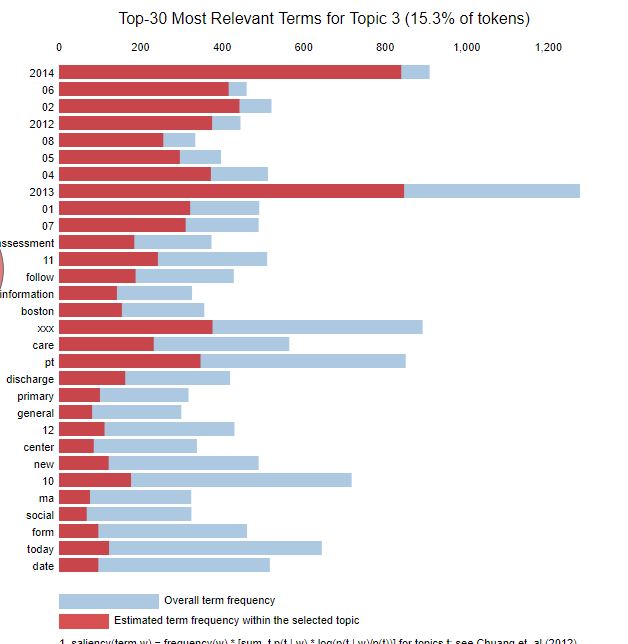
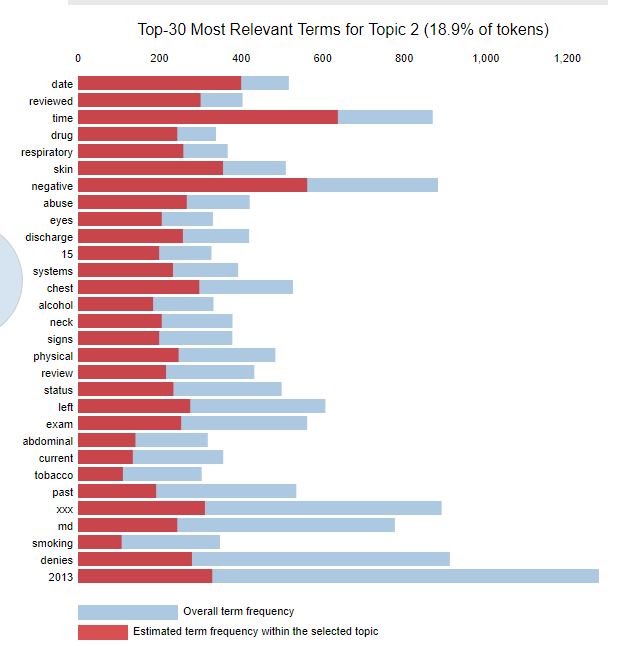
We also used Topic Modeling to try to understand the underlying structure of the Epic and the Logician notes. Initially we set the number of topics to be equal to 3 because we wanted to see if the distribution of words was different for the different determinants of health : drugs, alcohol, and smoking. However, what ended up occurring was that our groups described the following: Symptom Topics, Medication / Treatments, and Vitals and Measurements. Based on our figures, we can see that both sets of notes were able to distinguish between these topics well and their word distributions are different ( the first figure is the clustering from the Epic notes, the second clustering was from the Logician notes). From here we may want to investigate further vocabulary used within medical sciences to hopefully be able to distinguish these determinants of health more easily when it comes to classification. We are also going to show the rankings of the terms shown, and these rankings are going to be based on their lift, which is the ratio of a term’s probability within a topic to its margin probability across the corpus.



Here are also the types of words that we found for the different Clusters: the first 3 rankings would be from epic notes and the second three rankings will come from the Logician Notes:







Our next steps would begin to look into creating the labels for our data to be able to utilize Recurrent Neural Networks to be able to predict the multi-classification. We will also need to have another meeting with Sara and Rajat in order to get answers to any questions regarding the data.