# CS 410 Final Project Proposal: Topic Mining Healthcare Data

## Team Members


- Satish Reddy Asi- sasi2@illinois.edu
- Srikanth Bharadwaz Samudrala - sbs7@illinois.edu
- Raman Walwyn-Venugopal *(Project Coordinator/Team Leader)* - rsw2@illinois.edu


## Motivation

TimeDoc is a telemedicine company that focuses on ensuring patients receive proactive healthcare to improve the treatment of their chronic diseases. Since 2015, TimeDoc has accumulated roughly 1.8 million unstructured text documents created by licensed healthcare professionals that summarize telemedicine encounters with patients. Out of that total dataset there are 13,496 unique documents that have been labelled as a positive outcome for the patient by their author. A positive outcome is very important for a patient as it indicates that their health was improved which also translates into them valuing the telemedicine service.

Utilizing TimeDoc's data, **our primary goal is to identify more patients that had these positive outcomes and identify patients that are more likely to have a positive outcome.** To accomplish this our team plans to primarily use python and open source tools as described in our solution below.

## Solution

Our assumption is that there is a relation between a patient's profile and their likelihood of having a positive outcome. If this assumption is correct we can begin recommending patients with certain profiles for healthcare professionals to focus on.

### Curate Dataset

Expected Duration: 20 hours
Expected Completion Date: November 15th

To ensure we're adhering to HIPPA [1] guidelines. The telemedicine encounter document data and the patient profiles need go through a Patient Health Information [2] (PHI) de-identification process. We will automate the PHI de-identification for free text fields on the patient profile and the telemedicine documents utilizing a De-Identification Software Package [3]. All structured fields of a patient profile known to contain PHI will be replaced as well.

### Topic Mining and Analysis on Dataset

Expected Duration: 20 hours
Expected Completion Date: November 24nd

We plan on mining the topic models using Latent Dirichlet Allocation [4] for both the de-identified encounter telemedicine and patient profiles from the curated dataset. We plan on using gensim [5] and nltk [6] in python to accomplish this task.

[1] HIPPA Privacy Guidelines, https://www.hhs.gov/hipaa/for-professionals/privacy/index.html
[2] Patient Health Information, https://www.hipaajournal.com/what-is-protected-health-information/
[3] De-Identification Software Package, https://www.physionet.org/content/deid/1.1/
[4] Latent Dirichlet Allocation (LDA), https://arxiv.org/pdf/1711.04305.pdf
[5] gensim: python library for topic Modeling, https://pypi.org/project/gensim/
[6] nltk: Natural Language Toolkit python package, https://pypi.org/project/nltk/

After completing the topic modeling we will investigate if there is a significant difference between telemedicine encounters that were labelled as a success story versus ones that were not. We will also be investigating if there is a significant difference between the patient profiles that had success stories versus the patient profiles that do not have any success stories.

**Recommendation**

Expected Duration: 20 hours
Expected Completion Date: December 5th

After we have identified the topics of telemedicine documents and patients that had positive outcomes, we would like to identify other patients that may fit this criteria. We can accomplish this by indexing all the patient profiles by their mined topic data, the likelihood of the topic will be used as the score for the vector. We will then recommend patients that have topic profiles similar to the topics of patient profiles that had positive outcomes. The similarity score will be calculated using one of the newer variations of Okapi BM25.

In addition to this, we can also design a similar recommendation system to attempt to identify telemedicine documents that may contain a positive outcome for out patient.

The utility for both of these systems will be evaluated by licensed healthcare professionals at TimeDoc.

**Miscellaneous**

Other libraries that may be used but not discussed are listed but not limited to; pyspark [7], scipy [8], metapy [9], pandas, [10], numpy, [11] and whoosh [12].

[7] pyspark, https://www.gangboard.com/blog/what-is-pyspark
[8] scipy, https://www.scipy.org/scipylib/index.html
[9] metapy, https://github.com/meta-toolkit/metapy
[10] pandas, https://pandas.pydata.org/
[11] numpy, https://numpy.org/
[12] whoosh, https://whoosh.readthedocs.io/en/latest/intro.html