CSSM502 Mt Report

My project aims to use text analysis for psychological constructs. Text analysis is used for other social sciences for dealing with social media data or newspaper articles, but psychological research rarely uses these methods, despite dealing with qualitative data often. Part of that is because studying psychological phenomena requires more intimate text data which is not quite present in big data sources, and qualitative data collected directly for psychology possesses this quality yet the size of the data is not sufficient to use NLP or ML models. For this reason, I decided to use the published journals of a real person, so that I can have text data of the desired quality and quantity. I found that Henry David Thoreau wrote extensive and mostly time stamped journals of 14 volumes ranging from 1837 to 1861. He also wrote his journals in English, thus any potential translation bias is prevented. In summary, my aim for this project is to conduct a methodical experiment to explore the use of computational text analysis methods for psychological science, by training a model that infers sources of meaning from the journals of Henry David Thoreau.

The construct of my project is a source of meaning. Sources of Meaning is a construct that aims to explain a person's way or ways of deriving meaning from their daily lives which I think is very suitable to study with journal data. This construct 26 dimensions collected under 5 categories: Self-Actualization (Achievement, Freedom, Creativity etc.), Order (Morality, Reason, Tradition etc.), Vertical Self Transcendence (Religion, Spirituality), Horizontal Self-Transendence (Unison with Nature, Social Commitment, Self-Knowledge etc.), and Wellbeing and Relatedness (Love, Fun, Community etc.). My main reference behind the theory will be "Schnell, T. (2009). The Sources of Meaning and Meaning in Life Questionnaire (SoMe): Relations to demographics and well-being. The Journal of Positive Psychology, 4(6), 483–499. https://doi.org/10.1080/17439760903271074". I will train a model that will label the 5 umbrella categories and I will use the sub-dimensions as labeling criteria.

The steps I foresee for this project are as follows. First I will find the pdf, epub, or html version of his entire journal catalog and extract the data from these files in Python. I will clean the irrelevant meta data like headers and footers, page numbers etc. I will split the data from each date entry with regex, and further split long entries (>250 words) with sentence awareness to bypass token limits. Then I will use active learning with SetFit (Sentence Transformer Fine-tuning) to train a model that will label my data probabilisticly based on each of the 5 categories of my construct. I chose this method because as little as 8-10 labeled data per category is enough to train an accurate initial model, since the model learns by contrasting the categories. Following that, I will label more data points that the model was uncertain about iteratively, in each iteration aiming for a more accurate model. Before the training phase I will lock away 200 random data points and label them myself, this data will not be included in the training, and will be used when the final model is trained to calculate f1 scores. I also plan to make simple negative vs positive valence analysis in addition to sources of meaning, in order to explore patterns between sources of meaning and sentiment. The main outcomes of the project are the model itself and the final augmented dataset suitable for various analyses like network and diversity analyses.