# UBC PAIR - Text Analysis on Student Survey Data

***Team Members:*** *Amrith Anand, Himabindu Joopally and Wei Tang – UBCO MDS*

The project involved creating a solution that automates analysis of text responses in UES and NUBC surveys conducted by the PAIR department in UBC. The solution provides the missing piece in the current survey report that captures the cumulative response for rating questions but not the evaluation of textual responses. The manual evaluation of thousands of textual responses can be a highly time consuming and laborious task which the solution now automates. The solution also allows the flexibility to perform sentiment analysis, topic categorization and key-phrase analysis based on the nature of the question. The results of the analysis are presented in graphs that are similar to the graphs that are used for the rating questions and hence provides a seamless report. The evaluation of textual responses is performed using sound text mining and natural language processing techniques that are the most recent and hence provide reliable results. Summarizing, text-based questions allow students to express their experience that may not be captured by rating questions alone and evaluation of these text responses provides a value addition that encapsulates the intent of these surveys.

## Key-Phrase Analysis

Key-Phrase Analysis is performed to identify a few words or phrases that best summarize the responses. TextRank algorithm was used to identify the most important words by calculating similarity scores between words. Two-word phrases with the highest frequencies in the responses are also identified.

## Topic Categorization

Responses to the open-ended questions in the survey are categorized into one or more of the ten selected topics: academics, career, community, commute, diversity, extracurricular, facilities, finance, housing, and health & wellness. It helps us to identify the dominant aspects of student life at UBC and also assists in topic-wise sentiment analysis. The provided datasets are used to train the self-supervised word2vec algorithm and word-stocks related to each topic are obtained. The topic categorization is done by comparing the words in the responses with these word-stocks.

## Sentiment Analysis

Sentiment Analysis is the process of identifying the writer's attitude towards a particular topic from a piece of text. It was done by implementing an ensemble approach which exploits the structure of the sentences inherent to the language. Some phrases in sentences used to express sentiments are identified by matching certain patterns of parts of speech, then the semantic orientation is calculated. The strength of the sentiments is derived by querying SentiWordNet lexicon that provides a score for the intensity of the semantically oriented words. Additionally, negation expressions, contradicting words and positive words are identified for a comprehensive sentiment score for each response.

## Results

We have achieved good results in all three types of analyses. We have manually tagged 300 responses to evaluate our algorithms. The results of topic categorization show that around 82% of the responses have at least one topic tagged correctly. We have achieved 84.2% accuracy for the sentiment analysis of responses. In addition to accuracy, we have achieved good recall and precision values. These results show that our algorithm performed well compared to the state-of-the-art unsupervised methods available for text analysis.

## Conclusion

The tool performs an analysis of text-based responses and produces reports that can be easily consumed. The analysis is based on different techniques in the fields of text mining and Natural Language Processing. The tool is user-friendly, extensible and produces fairly good results. It allows PAIR team to derive insights from the students' responses to survey questions to a greater extent.