# NLP analysis of a small online chat group

## Introduction

Natural Language Processing offers the possibility to analyse free flowing conversations and generate insights over large texts and transcripts. This project aims to investigate the potential of NLP to provide insights into the conversations on a small chat group.

## Aims

I am keen to discover the potential of NLP to show user intent, interests, pain points and behaviours. This project aims to create an analysis model that assesses the messaging type and participation levels in a small slack groups. Aims include:

* Create a model that categorises interaction types, ranks users, assesses value and spots repetition.
* Leverage my familiarity with the channel to help create a model that is appropriate to it – while coding in flexibilities so that the model can be applied elsewhere.
* Produce recommendations for more efficient and successful slack interactions for organisations and individuals.

## About the IronHack slack group

The dataset is the complete record of the communications between students of IronHack on their group Slack channel between March 18th and April 27th 2021. As the group was relatively small (29 user profiles) and the conversation tended to happen principally in one channel, I decided to focus just on this ‘general’ channel.

I have the advantage of knowing much of the content and context of the conversations that have happened in this channel already. While this could be seen as a problem as it potentially biases my approach to the data, it is also an advantage. My familiarity with the content allows me to build a model that is shaped to the characteristics of group conversation. This should allow me to make more accurate model, know where to look and what to expect.

Also, it is not unrealistic to expect to have a considerable amount of prior knowledge about the nature of a conversation in other Slack channels I might want to analyse in the future.

Here are some expected characteristics of the IronHack group chat:

* Largely positive sentiment. To my knowledge there have been virtually no disputes within the group. It is a small supportive group of people come together to achieve a common purpose and only for a short time and working remotely. All this means there is little chance for tensions or disputes and the mood is overwhelmingly positive
* While there is some ‘chatter’ on the channel, it’s generally a practical channel for information sharing and course organisation. I expect the core message types to be:  
  - Information sharing  
  - Question/Problem posing  
  - Solution/Answers  
  - General encouragement  
  - Humour  
  - Gratitude

## Defining Success

One challenge will be to define what success is. If we are to make useful recommendations then we some success criteria. Here are some suggestions of what success is or isn’t:

For individuals:

* No of likes
* No of replies
* Expressions of gratitude
* Contributor v. Problem poser score (net contributor score?)

For admins:

* Repetition in or across channels (less repetition = good)
* Issues resolved/unresolved (lost conversations?)
* Channel clarity/discipline score
* Overall activity and engagement rates

For organisation:

* Overall sentiment
* List of brand mentions and links
* Flagging problems

## Process

1. Importing and joining json files into one dataframe in Python (data came in daily json files)
2. Cleaning and Wrangling the data.
3. Processing in the DataFrame for attributable ‘scores’ and characteristics per post.
4. Combining all text into one TextBlob in order to run TextBlob functions such as n-gram and noun-phrase extraction
5. Returning two csv files
6. Realising insights visually in Tableau.

## Basic stats about the Slack Conversation

597 Messages  
81 Threads  
132 Questions  
153 emojis and reactions

<Put in Days of the week chart>

<Put in emoji chart> - add emoji images in the story.

## Links and Brand Mentions

<Put in Link stem chart>

## Emoji usage and expressions of gratitude.

<Emoji graph> append emoji images if you have time

## Sentiment

Ran a basic sentiment analysis that with a spot check seemed to accurately grade the varies in positivity of language used. This could be very useful for analysing other chat groups, but was not that relevant for this example. The group conversation was almost entirely free of negative comments, so even those with the lowest polarity rating were really fairly neutral comments (only went as negative as -0.2 on a scale that stretches to negative 0.5). The sentiment ranking was for this reason in three categories that were not negative: Positive, Neutral, Less Positive.

<insert sentiment rank chart>

<Put in days of week sentiment chart>

## User anaylsis

Easy to score users for participation, contribution and response rate. Such an analysis might be considered a bit intrusive, but is potentially very useful for individual users who wish to see how much they contribute and how visible they are within a online community like this. It could also be used by administrators to track whether individuals were taking part, if they were contributing or if they were dominating.

<insert chart>

## Topic analysis

Despite road-testing several models, I couldn’t get results that produced much useful information on the type of topics covered using the unsupervised tools commonly applied to similar formats (such as twitter).   
There are several likely reasons for this. One is the disjointed nature of the conversations and difficulty of grouping the conversations effectively. Another is that there was a lot of noise in the text that I was only partially successful in cleaning up. The lists of topics and words that were generated by these tools were, as such, often peppered with non-words and characters that inhibited a proper functioning of the algorithm. What was produced made little sense.

The best indicators of topic are given by analysis of questions, brand meantions (urls) and counts of phrasal nouns. Here is the question count

## Repetitions

As part of the topic analysis I ran the texts through N-gram functions and noun-phrase extractors. These provided some information about topic, but were most useful for generating lists of repeated issues. Here is some useful information for channel admins, who can see the frequency with which a topic is repeated. This might prompt them to open a new channel on the topic or to provide the information in a different format. A good example in this chat is the number of times that calls to complete the student survey were made. Admins might consider automating this weekly call to action. Another is the repeated posting of the zoom call credentials. – again these could be shared in a more permanent medium. Here is a list of repeated topics:

<see repeated csv – make a chart in Tableau>

## Brand Analysis

<see dataframe.csv> for counted mentions of brands and for repeated links

* Need to realise this in Tableau

## The shape of the conversation

The model does allow for some analysis of the shape of even if not the exact content of the conversation. We can see how spread-out user participation was: whether it was dominated by a few individuals or widely used by all. We can see if there were individuals who had low participation rates. We can also see what kind of initial message prompted most of a reaction and of course we can track the sentiment of a conversation (Although not so relevant in an harmonious group, this could be a very important issue in another context).

<dashboard> illustrates all of this.

## Conclusion

Although a little disappointed with the noun-phrase and n-gram analysis methods as tools to categorise topic, it is easy to see how this model could be effectively applied to a larger set of data to monitor the health of the conversation against a range of different factors. These include:

* Participation rates
* Overall mood of a group
* Repeat topics that could be handled elsewhere
* Brand mentions

## Links and further reading

1. <https://medium.com/stratifyd/how-to-use-data-analytics-on-chat-sessions-a56bd2da72b0>
2. Analysing a non-committal relationship on WhatsApp. Good for some basic KPI ideas: <https://medium.com/swlh/chat-analysis-on-whatsapp-part-1-text-analysis-and-data-visualization-with-r-3a7e4e8362f2>
3. The ‘science’ of conversation analysis. How to open ‘slots’ for conversation: [Opening ‘slots’ for people to make productive contributions](https://www.sciencefocus.com/science/a-scientists-guide-to-life-how-to-be-a-better-conversationalist/)
4. Ted Talk by Elizabeth Stokoe. – Conversational analyst