KAI ANALYTICS

Qualitative Analysis 101

A BEGINNER'S GUIDE TO QUALITATIVE DATA ANALYSIS

TOPICS INCLUDE:

- Survey Question Strategy
- Coding Qualitative Data
- Segmentation in Excel
- Qualitative Analysis Pipeline
- Topic Modelling
- Sentiment Analysis

Table of Contents

Introduction

Welcome to Qualitative Analysis 101. From surveys to interviews, qualitative data can be a rich source of stakeholder data to guide decision making. However, this data can also carry great emotional weight, and many people struggle with implementing empirical methods for interpreting their qualitative data. This e-booklet covers the best practices for writing survey questions, collecting and processing qualitative data, as well as a couple of analysis tools you can get started using right away to effectively leverage your qualitative data.

We hope this booklet acts as a launching point for your journey into qualitative analysis. If you have any questions about the topics covered or wish to learn more, please feel free to reach out to us via our website. We care deeply about the success of our clients, and would love to chat about unlocking the value of your qualitative data.

Making Great Surveys and Questionnaires

Open Ended Questions

Open ended questions give the people responding to your survey a chance to give you feedback on areas you may not have thought to ask about. They can be an important source of rich qualitative data if they are done right.

Tips:

- **Be specific.** Ask: "What did you enjoy about today's experience?", rather than "Please tell us how you felt about today's experience." This will make your data easier to analyze.
- **Make questions optional.** This will allow respondents to focus on genuine feedback and avoid forced responses.
- **Set a limit of 3-5 open ended questions.** Answering open ended questions genuinely can feel exhausting, and overdoing it may actually lead to less completed surveys.
- **Be mindful of who is answering the survey.** If your respondents have a different culture than you, or live in another part of the world, consider how that culture might react to your questions.

Using the Net Promoter Score to Assess Satisfaction

One method for assessing satisfaction is the Net Promoter Score (NPS). The Net Promoter Score was developed by NICE Systems Inc. to help measure customer experience and predict business growth. Along with the Likert scale, NPS is considered best practice when creating surveys.

NPS style questions will sound like this: "please rate your experience with this call from 0-10, with 1 being the lowest score and 10 being the highest." Followed by: "please tell us why you gave us this score." NPS considers scores from 0-6 to be detractors, 7-8 passives, and 9-10 as promoters. Then the percentage of detractors is subtracted from the percentage of promoters to output the Net Promoter Score. The Net Promoter Score can be as high as 100 or as low as -100, where 0 is neutral. NPS scores are designed to measure customer experience, and are considered a key indicator of business growth.

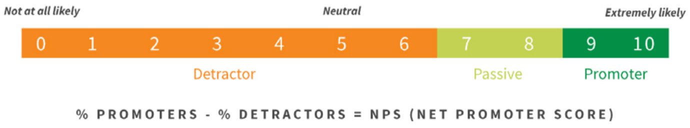


Figure 1 | The Net Promoter Score measured customer sentiment to predict growth

When creating your question, it's good practice to break it up into two parts. The first part will ask an NPS question and the second will be a follow up giving participants the opportunity to fully explain why they gave you that score. These explanations will allow you to dig deeper into specific areas, and the NPS will point you in the right direction.

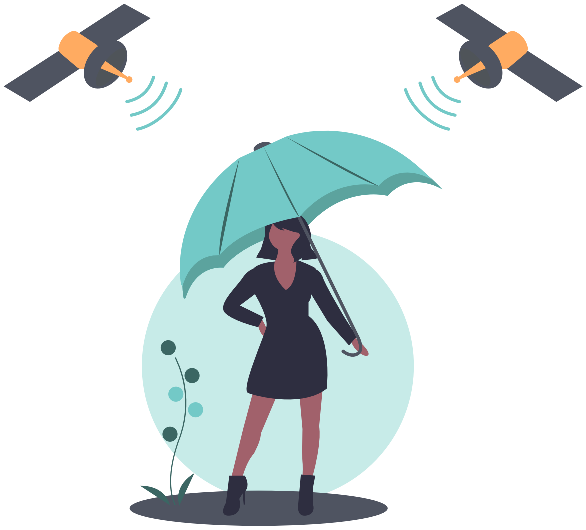
Data Compliance

Most institutions have specific data compliance policies, and some governments have special laws to protect the privacy of their citizens. Research the laws in your area and policies of your institution, and make sure you pick a survey platform that fits.

Confidentiality

The anonymity of a survey is mostly defined by a legal statement at the start of the survey, to be drafted by a lawyer, but there are steps you can take to make respondents feel more comfortable with giving you their data:

- Don't ask respondents to include their name, or any other uniquely identifiable information in the survey. This will help foster trust in your anonymous survey.
- Remind respondents not to include personal information in their responses. This may include their name, or the name of other customers or employees. Removing this information will help you to minimize the bias in your survey.
- Use an analysis tool that automatically recognizes and removes sensitive information.



Coding Qualitative Data

Coding data simply refers to a common method of sorting qualitative data into appropriate topics. These topics will be made up of 5-10 themes. These themes will usually be the general subject of the text, maybe a product, course, textbook, or experience; but they will be defined by the open-ended responses themselves. Simply read through each response, and assign it to a topic of similar subject. In the image below Topic 1 may refer to positive course experience, Topic 2 may refer to learning a lot, and Topic 3 refers to the intent to recommend.

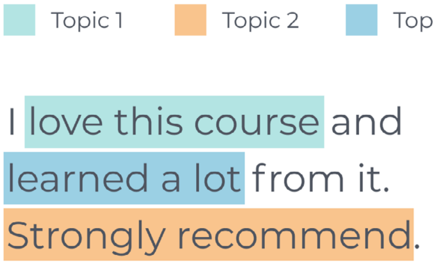


Figure 2 | Coding Qualitative Data, a comment tagged with the appropriate topics from a dataset

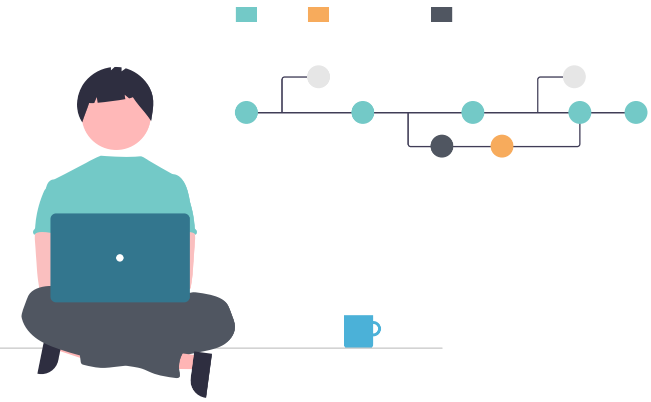
Once all your qualitative data has been tagged, you can measure the scale of each topic, and pair this information with other demographic and/or sentiment data to form a really detailed picture of your stakeholders.

Read or watch our example on how to code data through the link below.

LEARN MORE

Automatic Qualitative Data Analysis

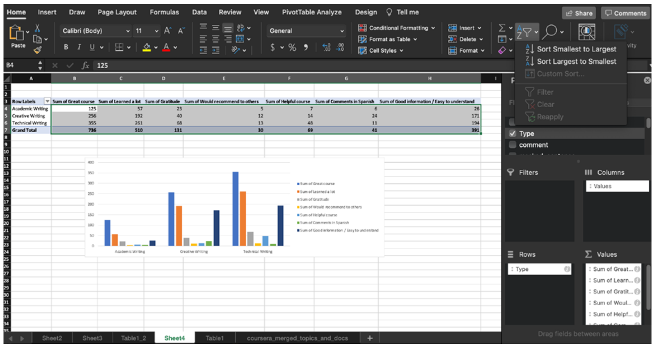
Manually reading through each line of text can be a very labor-intensive process. An alternative method exists through Natural Language Processing, or NLP. By using an NLP program, such as Unigrams, we can sort lines of text into their appropriate topic buckets very quickly. All that's left is for you to look at the network graph for each topic, and determine what each topic is about.



Using Excel Power Query to Segment Qualitative Data

A common objective for surveys is to observe how a specific segment of the population feels about a subject. Now that we've coded our qualitative data to determine which topics are talked about, we can use some of the powerful features in Microsoft's Excel to find which groups focused on what topics.

Note: *Mac users don't have access to Power Query yet. If this is your position, try using a pivot table organized like this to segment your data.*



B4 fx 125

Row Labels	Sum of Great course	Sum of Learned a lot	Sum of Gratitude	Sum of Would recommend to others	Sum of Helpful course	Sum of Comments in Spanish	Sum of Good information / Easy to understand
Academic Writing	125	57	23	5	7	6	26
Creative Writing	256	192	40	12	14	24	171
Technical Writing	355	261	68	13	48	11	194
Grand Total	736	510	131	30	69	41	391

Figure 3 | A Pivot Table in Excel segments coded survey data from Kai Analytics

Segmenting Data with Power Query

Power Query is Excel's built-in data transformation and preparation engine. It's a handy tool for connecting and reshaping datasets before performing analysis. We're going to use its the transformation tools on our dataset, so that we can present our data more intuitively.

As long as your data is formatted as a table, you can use the "Get Data, From Table/Range" function to access Power Query. Read more or watch our explanation here.

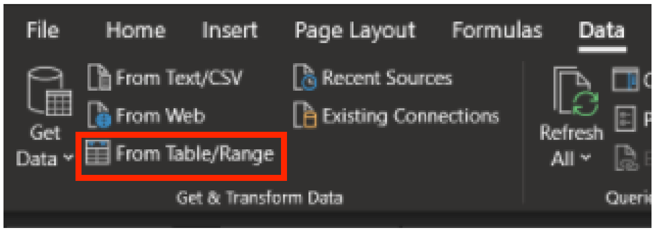
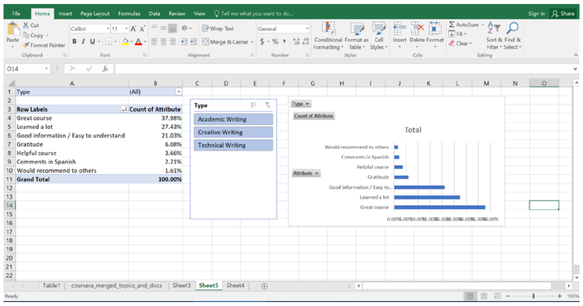


Figure 4 | Find Excel's Power Query Function in the top, left hand corner of your screen.

Once the Power Query Editor is open, you'll have access to a suite of data transformation tools. In this case we'll simply highlight the columns with our topics, and click "Unpivot" under the "Transform" tab.

Unpivot will transform your data so that you have two new columns: attributes and values. In the "Attributes" column will be the title of each topic. The "Values" column will contain a 1 or 0, depending on if the comment is matched to that topic. You'll want to go to the filter button and de-select the 0's to filter out null values. Once you've done that, go ahead and remove the values column. You can now close the Power Query Editor, choosing "Keep" if prompted.

Now that the data has been transformed, it's easy to create a pivot table based on the remaining "Attributes" column. This process will make it easier to sort and filter your data, which can translate directly into intuitive presentations.



Type	(All)
Row Labels	**Count of Attribute**
Great course	37.98%
Learned a lot	27.43%
Good information / Easy to understand	21.03%
Gratitude	6.08%
Helpful course	3.66%
Comments in Spanish	2.21%
Would recommend to others	1.61%
Grand Total	**100.00%**

Figure 5 | Excel's Power Query makes it easy to display coded survey data from Kai Analytics.

Building a Qualitative Data Analysis Pipeline for NLP

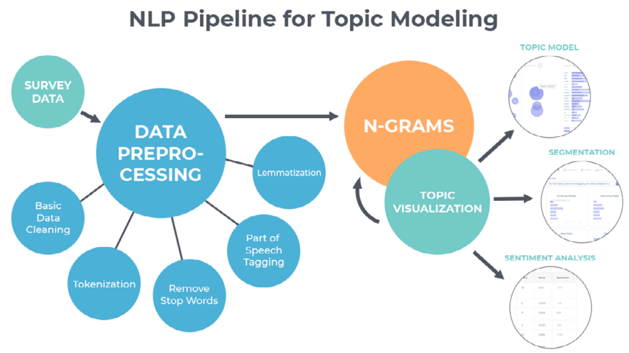


Figure 6 | A basic Qualitative data Analysis pipeline used at Kai Analytics.

The first part of and NLP data pipeline is the pre-processing phase. In this phase data is cleaned, broken apart, and reassembled into a format suitable for analysis.

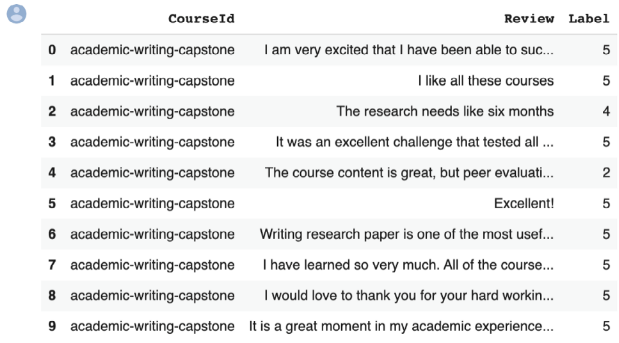
Step 1: Basic Data Cleaning

Before NLP can work its magic, you'll need to tidy up a bit.

- Account for blank responses by either deleting them or counting them
- Remove duplicate responses
- Spell check responses, either through a word processor or code package

Step 2: Tokenization

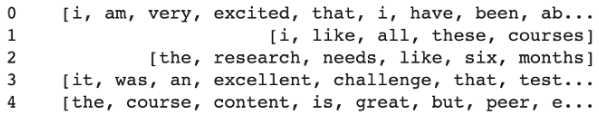
The first step is for the analyst to break down sentences into individual words (or tokens!). So, this list of responses:



	CourseId	Review	Label
0	academic-writing-capstone	I am very excited that I have been able to suc...	5
1	academic-writing-capstone	I like all these courses	5
2	academic-writing-capstone	The research needs like six months	4
3	academic-writing-capstone	It was an excellent challenge that tested all ...	5
4	academic-writing-capstone	The course content is great, but peer evaluati...	2
5	academic-writing-capstone	Excellent!	5
6	academic-writing-capstone	Writing research paper is one of the most usef...	5
7	academic-writing-capstone	I have learned so very much. All of the course...	5
8	academic-writing-capstone	I would love to thank you for your hard workin...	5
9	academic-writing-capstone	It is a great moment in my academic experience...	5

Figure 7 | Raw survey data.

Would be broken down into this:

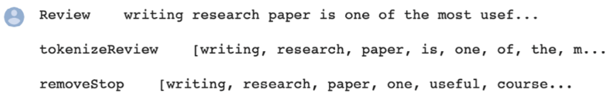


```
0    [i, am, very, excited, that, i, have, been, ab...
1                  [i, like, all, these, courses]
2           [the, research, needs, like, six, months]
3    [it, was, an, excellent, challenge, that, test...
4    [the, course, content, is, great, but, peer, e...
```

Figure 8 | Tokenized survey data.

Step 3: Removing Stop-words

Stop-words are words like "the", "a", "of"; we add them to our sentences to link words together and make language sound better. However, they're not really necessary to derive the meaning of a sentence. By using dictionary like ______ to remove stop-words, we're a step closer to analyzing our data efficiently.



```
Review     writing research paper is one of the most usef...

tokenizeReview    [writing, research, paper, is, one, of, the, m...

removeStop    [writing, research, paper, one, useful, course...
```

Figure 9 | Survey data before and after removing stopwords.

Not all stop words clutter the data, depending on the context; "not" and other negation words are crucial for sentiment analysis, but they would still clutter results. So, analysts must review the list of stop-words provided by the dictionary and add or subtract from it as necessary.

Step 4: Tagging Parts of Speech

Part of Speech tagging is where the analyst assigns "tags" to each word in the sentence, like in the example below. These tags are crucial in the next step of the preprocessing phase, Lemmatization.

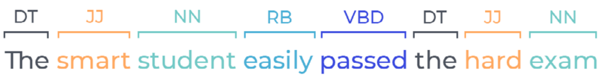


Figure 10 | Tagging the different parts of speech in a sentence.

Step 5: Lemmatization

Lemmatization is the process of getting to the dictionary root of each word, its Lemma. So the Lemma of "studies" is "study", "singing" is "sing"; but a problem exists with homonyms. For example, biking becomes bike, [he] bikes become bike, and so does [the]bikes. Bike, bike, bike. Enter Part of Speech Tags. "He bikes" would be given a verb tag, becoming "bikes (VBG)", and so on to differentiate between bikes. This allows the program to analyze all these words as single items while retaining the context for techniques like sentiment analysis.

FORM	STEM	LEMMA
studies	studi	study
studying	study	study

Figure 11 | An example of Lemmatization.

N-Grams: Theory

The goal of the pre-processing phase is to create something called an "N-Gram". An N-gram is simply a way of breaking down text data into manageable pieces. The N is equal to the number of entities in the gram. n = number of words in the gram.

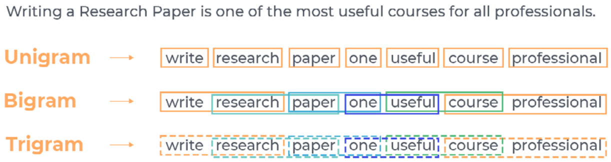
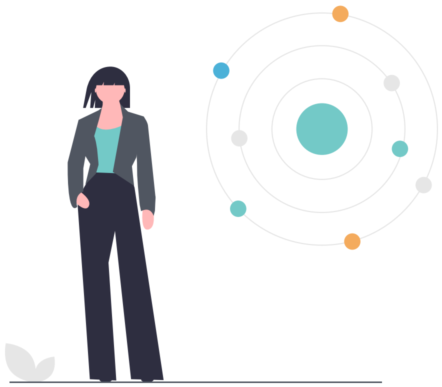

Writing a Research Paper is one of the most useful courses for all professionals.
Unigram →
write research paper one useful course professional
Bigram →
write research paper one useful course professional
Trigram →
write research paper one useful course professional

Figure 12 | Examples of an Unigram, Bigram, and Trigram.

N-grams are easily displayed through a network graph, like the one below. It shows which words (n-grams), appear together most often, and the frequency in the thickness of the connecting line. Alternatively, you could count the frequency of bi-grams within the data to uncover key themes.



Watch our Tutorial for a video walk-through explanation below.

WATCH NOW

Topic Modelling

Topic Modeling is a pillar of qualitative analysis. By using statistics to uncover the main topics/themes from a piece of text, we'll understand what the most prevalent ideas are in the text and which ones relate to one another most often. This information is useful on its own, but it can also be paired with other techniques like sentiment analysis.

One way to understand Topic Modeling is to take all the words in a document and measure how often they appear with every other word. This is a process called word embeddings, or vectorization (pictured below). The correlation matrix (on the left) created from this process generates a distance score between the words, representing how closely they are related in the text; words with a higher score appear together more often, words with lower scores appear less often.

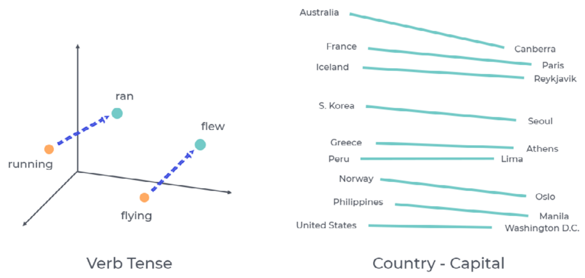
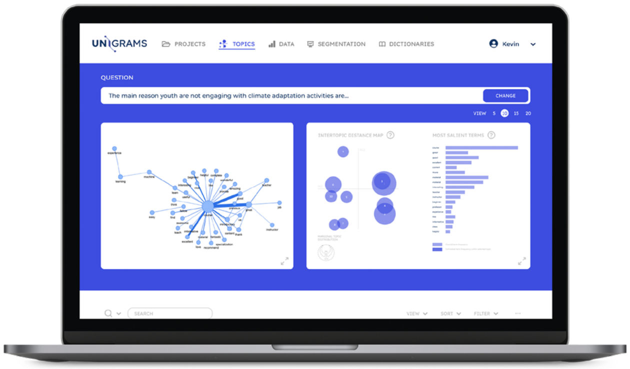


Figure 13 | How vectorization is used to store text data.

A popular way of conducting topic modeling with statistical tools, like Python, is to use the [Latent Dirichlet Allocation (LDA) package](). With this approach, you specify the number of topics and the machine will randomly assign every word in your document to every topic and asks, "do these words belong together or is there a better pairing of words"? The process continues over several iterations until all the words are classified into a topic with the best semantic pairing.

How to Interpret Results

By modeling our results on an Inter-Topic Distance map like the one below, we can see a couple of key results. The size of a circle represents the frequency of a topic, and the distance between circles is how related they are. If the circles overlap, they are closely related. Once you know what each topic is about, you can compare it to the graph to find focus points for further analysis.



The main reason youth are not engaging with climate adaptation activities are...

Figure 14 | Our Unigrams dashboard displays the results of Topic Modelling through an intertopic distance map a network graph.

Sentiment Analysis

With the Net Promoter Score

Understanding sentiment is one of the most powerful features of qualitative data analysis. Positive sentiment is a sign of growth, and by drilling down on areas of strong emotional reaction, it's easy to find high priority areas for strategic recommendation.

A simple way to measure sentiment is with the Net Promoter Score discussed earlier. Net promoter scores are great for measuring and benchmarking sentiment in high priority areas.

Net Promoter Scores are limited to measuring sentiment at a high level. Often, respondents will leave more detailed feedback in the open-ended questions. In order to efficiently summarize these responses, we'll use an NLP technique called opinion mining.

With Natural Language Processing

Sentiment analysis with NLP, sometimes called opinion mining, is a technique to evaluate how the writer of a text felt about a subject. This is an effective way to understand the sentiment of open-ended responses, like the kind found in customer reviews, or tweets.

Sentiment analysis works by using a dictionary that stores information on the feelings associated with a word (joy = positive, hate = negative, etc.= neutral), and then assigning that sentiment to the n-grams created earlier. You can then count the sentiment scores to model the overall feeling about a subject, producing a chart like Figure 15.

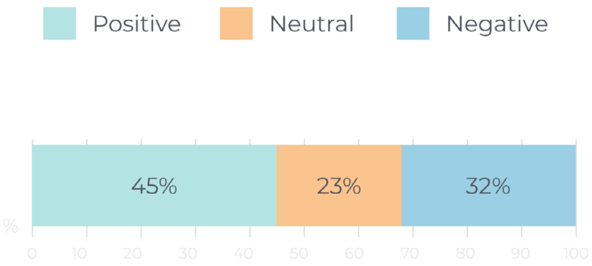
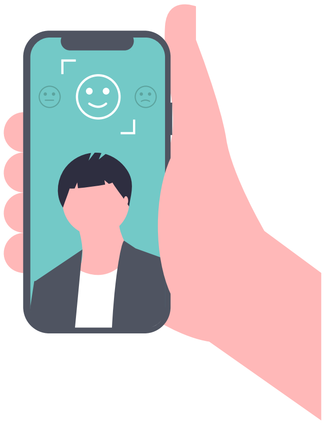


Figure 15 | How sentiment words come together to form a Sentiment Analysis Spectrum

[VADER (Valence Aware Dictionary and Sentiment Reasoner)](#) is one of the most popular Python packages to try out sentiment analysis. It is a rule-based model that has a large dictionary of words, mostly taken from social media posts, that are assigned a positive, neutral and negative score. The model can be applied to words, sentences, and even paragraphs to return a weighted average sentiment score.

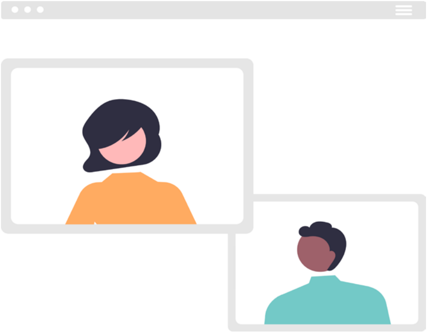
Combined with the pre-processing code shown above, you now have an entry level model for analyzing the sentiment of open-ended survey responses.



Natural Language Processing Workshop

Want to learn more about NLP?

Our NLP workshop covers all aspects of NLP and our customizable modules make it suited to a variety of learners. We can tailor our approach depending on your data sets, host Q+A sessions, and, as experts in the field, we have a wealth of knowledge to share! Join us to understand the secrets and benefits of using NLP to understand sentiment, transforming your relationship with your clients and stakeholders.



Our Workshops

1. Full-Day 6-hour Workshop
2. Two Workshops (2 hrs + 4 hrs)

Discount pricing is available for higher education and non-profit organizations.

Note: While we endeavour to accomodate all requests, our NLP workshops are subject to capacity. Should you be interested in attending, connect with us as early as possible!

What We Cover

NLP PRINCIPLES

Learn about NLP principles, as well as specific tools and Python libraries. Our workshop provides you with Python notebooks that show and explain common NLP libraries like NLTK and SpaCy.

NLP MODELS

Understand the processing done by these models, and what we can learn from the parts of speech, entity recognition, and lemmatization of your data.

THEMATIC ANALYSIS

The workshop also includes topic estimation models, which can be used to quickly group and understand the themes of your data.

REUSABLE CODE

Included in the workshop are interactive Python notebooks for you to try, follow along, and experiment with NLP at your own pace. This code is both intuitive enough for you to learn from, while being practical enough for real-world applications.

Reach out to us!

We hope that was helpful. To learn more about how Kai Analytics and qualitative data analysis can help you and your organization, book to speak with a member of our team here. If qualitative data analysis is an important part of your workday, sign up to our Newsletter "A Bag of Words", where we discuss important topics in the world of qualitative analysis, and share exciting company news!

Thank You!

CONTACT US

🌐 kaianalytics.com/

✉ hello@kaianalytics.com

☎ (800) 878 - 5214