THE FUTURE OF FEEDBACK

USING SENTIMENT ANALYSIS FOR CONSULTATIONS

A LITTLE BIT ABOUT US AND ME

CSPS·EFPC
Digital Academy
Académie du numérique

THE DIGITAL ACADEMY

- Twitter: @DigiAcademyCAN
- GitHub: CSPS-EFPC-DAAN , DIS-SIN
- A teaching organization at the Canada School of Public Service. Our mandate is to build digital acumen across the public service.
- We also act as a catalyst to create proof of concepts, demo projects and early government adoption of digital best practices.

A LITTLE BIT ABOUT US AND ME

OMAR NASR

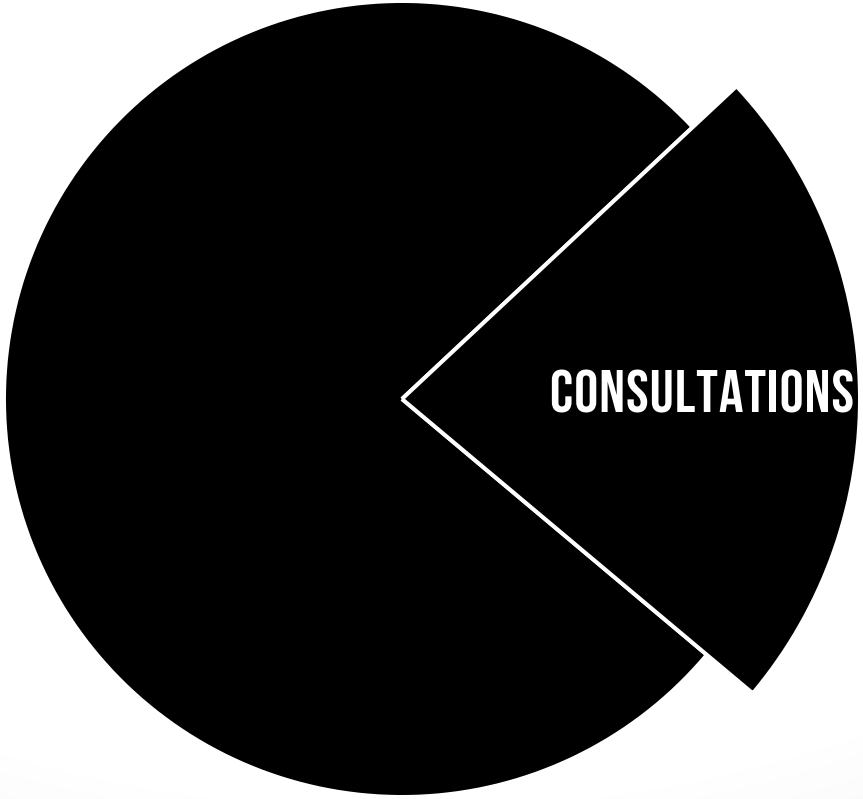
- Data Scientist
- Twitter: @thenextmusk
- GitHub: Moro-Code
- Background: Bioinformatics and Comp Bio

“THE TRUE PROMISE OF THE INFORMATION AGE ISN'T TONS OF DATA BUT DECISIONS AND ACTIONS THAT ARE BETTER BECAUSE THEY ARE BASED ON AN UNDERSTANDING OF WHAT'S REALLY GOING ON”

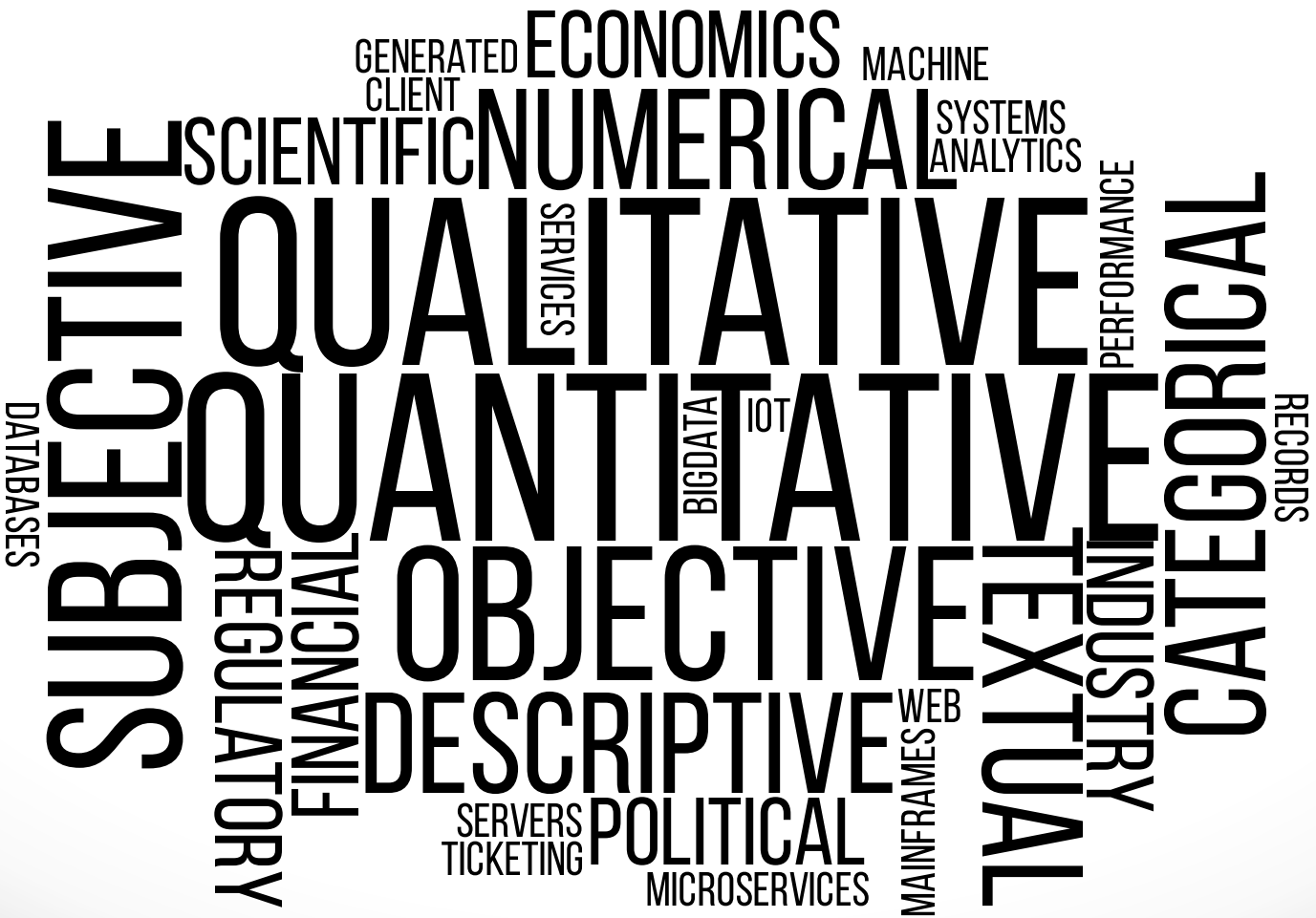
STEPHEN FEW - SHOW ME THE NUMBERS

THE GOVERNMENT OF OF CANADA IS THE BIGGEST DATA SOURCE IN CANADA AND ONE OF THE BIGGEST IN THE WORLD

A significant portion of this data is from feedback, or consultations



DATA ISN'T JUST ONE TYPE



THE PROBLEM WITH CONSULTATION AND SURVEY DATA

THE SHEER AMOUNT OF DATA WE HAVE

The school alone has 90k comments from 2015 to present from thousands of surveys from learning events

HOW WE STORE OUR DATA

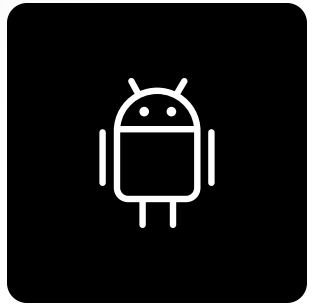
Currently surveys are stored in static schemas in large and messy data warehouses making it difficult to access

THE DATA ITSELF

Textual data is often subjective and with current analytical methods we use would require a lot of manual labor

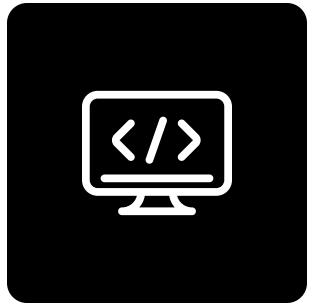
LETS LOOK AT AN EXAMPLE

AN ATTEMPT TO SOLVE THIS PROBLEM



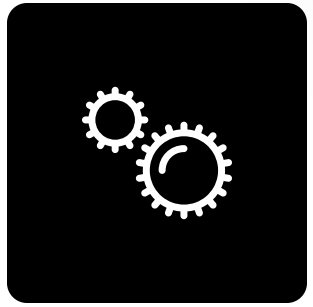
NATURAL LANGUAGE PROCESSING (NLP)

- Natural Language Processing is one of the oldest applications of machine learning and one of the most commonly used today
- Machines have the ability to consume data at an extremely fast pace
- This allows us to analyze things like emotion in text in a more objective way (e.g. Sentiment Analysis)



APPLICATION PROGRAM INTERFACES (API)

- Allows us to decouple application to application communication from application implementation
- This means that when technical implementation changes we do not need to update dependent applications
- API's allow a more dynamic way of delivering machine readable data



AUTOMATION

- Manual preparation of data is time consuming and requires a lot of effort
- Bottleneck around capacity of employees to prepare data and store it cleanly
- Surveys can be treated as objects with defined attributes thus their processing can be automated

DEMO

SURVISTA

- PROOF OF CONCEPT
- INTEGRATED SENTIMENT ANALYSIS
- KEYWORD EXTRACTION BASED ON SELECTED TEXT
- USER MANAGEMENT AND ACCESS LEVELS (COMING SOON)


LETS TALK ABOUT THE TECH

See me after the presentation if you would like to learn more on the technologies used

THE IMPORTANT STUFF

PYTHON

- Python is one of the most horizontal programming language today.
- A utility language at heart it is among the most popular programming language with the advancements in the availability of machine learning tools.
- You can think of it as a Swiss army knife of programming

THE IMPORTANT STUFF

NATURAL LANGUAGE TOOLKIT (NLTK)

- NLTK is an essential package for python for natural language processing.
- It provides utilities to quickly, easily and efficiently clean text to make it ready for machine learning.
- We use NLTK to clean text and Normalize

THE IMPORTANT STUFF

SCIKIT LEARN

- A simple and efficient machine learning package for python built on the SciPy stack (numpy, pandas and much more).
- Scikit learn enables the user to build simple regression, classification models.
- It also contains a suite of tools for preprocessing such as CountVectorizer which is used for feature extraction (more details to come).

THE IMPORTANT STUFF

GOOGLE NATURAL LANGUAGE API

- Google sits on mounds of textual data and thus are able to build extremely accurate language models
- The google natural language API is used to extract sentiment and magnitude scores as well as detect language

THE IMPORTANT STUFF

FLASK

- My preferred web development (micro) framework in python
- Flask does not define strict templates for how to do things rather gives you simple to use utilities and awesome docs.
- It takes more time to do the basics then using frameworks like django with lots of boilerplates.
- If you are just starting out in web development I really recommend Flask

THE IMPORTANT STUFF

CELERY

- This is an awesome python asynchronous job/ task queue
- Currently I have a celery worker listening to a redis instance which acts as the job queue.
- A task is sent to the job queue and it's then executed asynchronously(or synchronously if specified) by the worker

KEYWORD (FEATURE) EXTRACTION

CLEAN TEXT

This involves removing punctuation and special characters from text. Here a variety of tools may be used depending on preference. I have implemented this with regular expressions however NLTK provides simple straight forward tools for this as well.

NORMALIZATION

In language there are many words we use frequently but do not have any use outside a context like the word “the”. These are known as stop words and in this process is considered as non-useful noise. We must remove these stop words in order to extract relevant keywords.

LEMMATIZATION

Lemmatization is the process of getting the root of a word or the lemma . This is done by using the WordNet database developed by Cambridge through the NLTK package. There is an alternative method called stemming not discussed here.

TOKENIZATION

Tokenization is the process of splitting text into an array or a list of individual words. This is essential for the following step.

VECTORIZATION

Machine learning algorithms are not very well suited to handle textual data but are able to consume a high volume of numerical data. Here we transform our text to a large vector of word frequencies with sentences as our index and the tokenized words as our columns. This is done through scikit learn.

SUMMATION

From the vector we then sum the frequencies of the columns (the words) and sort based on summed frequency. We now have a list of keywords.

I will be providing a jupyter notebook on github with the implementation of this process

EXAMPLE OF FREQUENCY MATRIX

	Hello	Name	Omar	My	Is	Dog	He	Likes	To	Eat
Hello my name is omar	1	1	1	1	1	0	0	0	0	0
My dog is good	0	0	0	1	1	1	0	0	0	0
He likes to eat	0	0	0	0	0	0	1	1	1	1

Note that stop words here have not been removed and no lemmatization carried out to simplify presentation. This is a very simple example and often times these frequency matrices can have thousands of columns! This is an intense computational process, however the scikit-learn package has extremely optimized algorithms to calculate these matrice.

WHY AM I NOT TRAINING MY OWN SENTIMENT MODEL

An age old debate... well, not that old

1

SIMPLICITY

When developing a product one should remove as much complexity as possible. With the google natural language API it cannot get anymore simpler to implement sentiment analysis in your products.

2

COST

For me to implement my own models means hours of work. In comparison, the google natural language API costs 50 cents per 1 million units.

3

ACCURACY

While a government specific model may be more accurate, today there is a lack of readily available data for training, essential to build any machine learning models (let alone a good one). Google sits on an enormous amounts of readily available data and employ the best in the field to build their models.

4

SPEED

When building a web accessible application, speed is of the essence. As with the point above, organizations that provide these ML consumables employ the best in the field. Thus delivery of these services are optimized for maximum efficiency and speed.

5

LANGUAGE SUPPORT

The google natural language API supports both English and French languages out the box as well as six other languages.

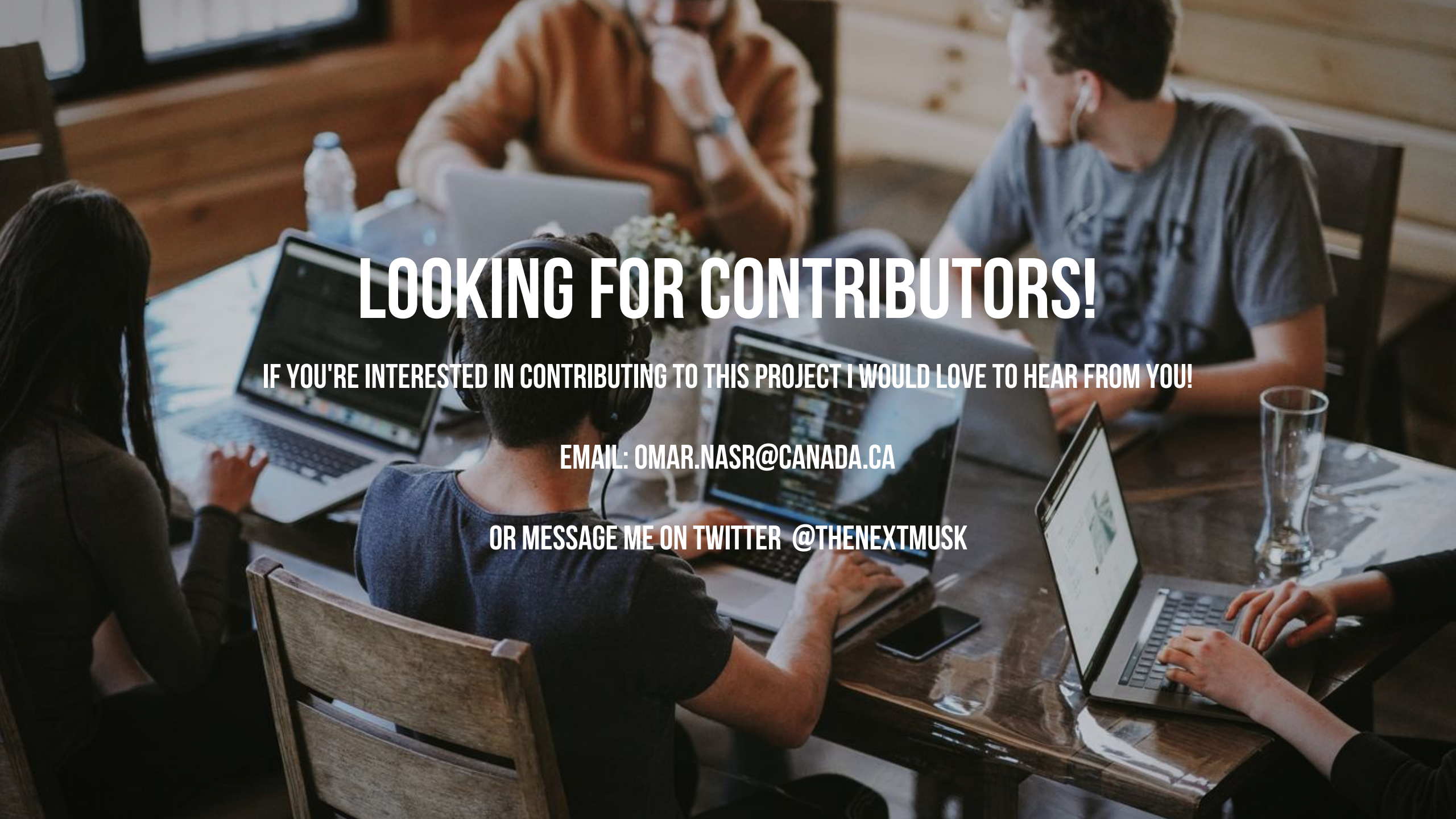
NEXT STEPS

1 FEATURES SUCH AS:

- Keyword based search for comments
- Bigrams and trigrams as well as current unigram keywords
- Admin panel, user management and access levels
- Connectors to popular survey tools to make processing survey data as simple as submitting a link

2 FINDING MOTIVATED CONTRIBUTORS TO GROW THIS PROJECT

- Currently this project is geared towards the schools learning however, there is a government wide implication


LOOKING FOR CONTRIBUTORS!
IF YOU'RE INTERESTED IN CONTRIBUTING TO THIS PROJECT I WOULD LOVE TO HEAR FROM YOU!
EMAIL: OMAR.NASR@CANADA.CA
OR MESSAGE ME ON TWITTER @THENEXTMUSK