This is the quantitative part of my research. I was feeling desperate when I relied completely on qualitative analysis but then it hit me that I can actually use all the text data I already collected to do some basic NLP analysis.

*Here is what did not make it into the report:*

**Methodology**

However, I included a small portion of quantitative research both as an example of a Data Science project to illustrate key findings and meaningful analysis to reinforce key ideas.

Although I was unable to procure any structured data that comes from databases in the form of spreadsheets, my qualitative analysis already yielded a large amount of unstructured data, namely the transcripts of all interviews and the text data from all blog posts and online discussion. It is precisely what is called unstructured data in Data Science and I will be analyzing word frequency on the corpus.

**Example & Quantitative Analysis**

Here, I would like to showcase a small Data Science project. It will serve both as an illustrative example to better understand the key findings in qualitative analysis and a supplementary analysis that further proves the integrity of those findings.

The idea of a possible quantitative analysis burgeoned when Dr. Goldberg advised me to look for repeated themes in different interviews. In the process of conducting qualitative research, I procured the transcripts of all interviews. These transcripts, together with blogs and Quora discussion, constitute a corpus that can serve as unstructured data for a Data Science project. Therefore, I set out to find the most frequent words in the corpus as an attempt to find the repeated themes. By then, I have finished problem framing and set the goal.

Diagram

Description automatically generated

The data preparation process takes a lot of time. I downloaded transcripts from YouTube if possible, but not all YouTube videos have transcripts. For those that don’t and the interview I conducted on my own, I typed the transcripts word-by-word. And for those download transcripts, I went through all of them to fix the typos. I explored the data a little by counting the number of works in total, etc.

Then, for the analysis, I would say that the baseline modeling would have been simply splitting each word by spaces. That would be a terrible model because it does not take into account punctuations, plural and single forms, -ing forms, -s and -es forms, etc. The word frequency would be way off, so I did not even attempt to do this. I know that the nltk package (a very popular natural language processing Python module) has the whole process covered so I went straight to using nltk. It helps me get rid of all the stopwords (meaningless words such as I, me, you, which and so on) and lemmatization (converting rocks into rock and better into good, etc). However, the first try yielded terrible results because the top ten most frequent words were punctuations. Therefore, I created my own list of meaningless words and cleaned the corpus more thoroughly.

The newer version with an extra list of meaningless words was comparatively better than the older version, so I stopped here.

Chart, bar chart

Description automatically generated

Given the top 10 most frequent words in my data, I went back to the corpus and see their context.

These frequent words are consistent with the themes I identifies earlier in my qualitative research. Data Scientists are constantly learning. A large part of their job centers around building models. They work in different teams. They value their time and think about how to optimize their time. They are creators and they make decisions. Their goal centers around business. Their work usually comes in the form of projects. They are constantly dealing with different methods and ideas. Data Science is an extremely new field with a lot of new things happening every day. Data Scientists solve problems.

The result of the quantitative analysis is in line with the qualitative analysis and I record my findings in this document.

There is definitely room for improvement here. For example, using the bag-of-words method (not considering word sequence and word context) loses a lot of information. Further analysis could be phrase analysis or sentiment analysis. For example, the learning in Machine Learning should not be count the same as any other learn that is a verb. Sentiment analysis can give us some idea of how positive or negative do Data Scientists think of their jobs. These ideas can serve as the problem framing stage of another round of analysis.

Some examples:

**Learn** in context

* learn new techniques
* learn methods
* learning there how to work as a team member sometimes
* learn how to behave
* learn some new techniques
* learn some new things
* learn more difficult things
* learn things almost continuously
* happy learning
* learn the architecture of BERT
* just learned what a LSTM was
* you can never learn enough
* learn things you are excited about
* learn how to implement things
* machine learning
* learn to design tests
* learn on the way
* I love learning

**time** in context

* time management
* sacrifice a quite amount of time and effort
* spend more time on analysis
* time series forecasting
* have enough time to learn new things
* there's not enough time
* what should I be spending my time on

**make** in context

* make data easy to use
* make sure you are implementing it correctly
* make that easy
* makes a company a great place for interns
* make you learn
* make work pleasant
* make do with the skills available
* make sense
* make it useful
* make decisions using data they have
* make reports and dashboards
* makes everyone involved smarter
* make deliberate decisions
* make better, more informed business decisions
* makes the cooperation easier and more fruitful
* make a crucial decision
* make your life a lot easier
* make things better
* make things more and more difficult
* make an impact
* make a high-level impact
* makes it really fun
* make sure
* make decisions

**business** in context

* business model
* business insights
* grow the business
* business need
* informed business decisions
* business intelligence
* business logic
* impact the business
* business problems
* business values