**Interviewee -** Subject04, 07/25/2019 @ 3pm

**Demographics**

Male

Research Scientist

Company 2

PhD

Years experience in data analysis: ~2-3 years (mostly text)

Years experience in review analysis: 6 months

**Walk us through a recent review analysis task**

*Goal: Building a commonsense knowledge base (hotelsense) for improving review insight modeling (e.g., developing better question answering model)*

*Pipeline: (1) Aspect extraction → (2) Labeling → (3) Modeling*

*Has two types of labeling tasks*

*Manually verifies the quality of labels from the crowdsourcing tasks*

**What kinds of data sources and formats do you use?**

*TripAdvisor and proprietary datasets in CSV format. They are already preprocessed and cleaned.*

**Is scalability a problem?**

*Generally works with a small set of reviews (in 100s or 1000s). Working with larger datasets would be slow and expensive*

Looks like, the ability to scale the analysis would be nice to have but not a bottleneck

**What tools and languages do you use?**

*Python scripts, SpaCy (clean, minimal api) for computing statistics,*

*Emacs for eyeballing sample results, keyword search;*

*Figure 8 for crowdsourcing*

*Jupyter Notebook for visualization (sentiment histograms, graph visualizations, etc.)*

*Google NLP API for sentiment analysis*

*Tensor board for model visualization / debugging*

**What are the downstream applications?**

*Improve question answering (e.g., chatbot like systems)*

**What are the bottlenecks you run into or things you spend the most time on? (Are they different?)**

*Bottleneck --*

*Getting a quick, general sense of reviews is somewhat challenge. Doesn’t even attempt many exploratory or statistical summarization tasks due to the difficulty (lack of tools) of performing them*

*Time wasting --*

*Designing crowdsourcing tasks*

*When you try to collect subject opinion, quality control becomes difficult*

*Writing clear & concise instructions is also part of the challenge*

**What features/tools do you wish you had?**

*Review cleaning tools*

*Quick access to descriptive statistics over review text*

*Visualization of how sentiment changes and the boundaries of the change per sentence in reviews*

I.e., Visualization of the second derivative of a fine-grained sentiment function

**How do you present your results? (To whom?) and collaborate with others on your work?**

*Pyplot, Google Spreadsheets, NeworkX (a Python package for graph analysis and visualization)*