**Interviewee -** Subject 06, 07/24/2019 @ 2pm

**Demographics**

Male

Product Manager

Company 1

BSc & MS in CS and MBA

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

Years experience in review analysis: 6 months

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

*Goal: Extract items (e.g., presence of facilities, services, type of food, attractions, etc.) from hotel reviews.*

*Pipeline: (1) Data prep / cleaning (done by someone else) → (2) Labeling (labeled data is converted to a standard format, CoNLL ) → (3) Train the model*

*Uses a Japanese crowdsourcing platform/service called IR Alt (advanced linguistic technologies) and generally happy about the quality of labeling gathered.*

*Before submitting labeling tasks, manually labels a few hundred tags himself to test the task design.*

*After getting the labels, validates them by eyeballing a small sample of them.*

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

*Reviews are from [proprietary source] (~7 M), stored in a BigQuery DB.*

**How do you get them?**

*By running SQL queries*

**Is scalability a problem?**

*Works on a sampled subsets because both NLP computation and particularly human labeling are expensive (i.e., costs money & time). While the original dataset has 7M reviews, he (or his group) labels ~10K reviews using crowdsourcing. Items are frequent and labeling them is laborious task, takes 10~20mins to label a single review text.*

It seems though the opportunity cost of not analyzing or using complete dataset is not perceived to be high or significant enough to be a priority or a concern, as you can be seen in the answers to the bottleneck and desired features questions.

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

*Brat for rapid text annotation. Jupyter Notebook for checking/validating intermediate results data and performing statistical analysis (typically, over tags)*

**What are the downstream applications?**

*Many hotel descriptions are partial (incomplete) specifications. So, the item extraction aims to use the consumer provided information embedded in reviews to fill in the missing information in hotel listings and enrich these listings further.*

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

*Bottleneck -*

*Designing crowdsourcing tasks*

*Time consuming -*

*Labeling and model development*

The interviewee gets the data already cleaned, which might explain why we didn’t get the most frequent answer (data prep + labeling) here.

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

*Integrated tools supporting regex & NLP functionalities together and enabling the execution of the basic data curation pipeline (data cleaning /preparation → exploration → labeling) in an end-to-end fashion.*

*Tool for interactive labeling rule / function generation*

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

*Collaborates using Slack, Excel and Jupyter Notebook. Uses Zoom for remote meetings.*