Review Analysis Interview:

Subject02

**7/23/2019 11:00am**

**Demographics**

**Title** Research Scientist

**Company** Company 2

**Education** Phd

**Experience in data analysis** 4-5 years (mostly text)

**Experience in review analysis** 1.5 yr

**Gender** M

# Summary

### [Project 1] project walkthrough

*The interviewee is a research scientist working primarily with text. He has more text analysis experience than any other subjects thus far, and most of his exploration is in the command line using grep and regular expressions.*

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

*Goal: Extracting features of apartments from reviews (e.g., using often keyword search, regex matching, etc.)*

*This is to automatically populate a predefined schema.*

Pipeline: (1) Extract / match patterns → (2) Refine / filter / interpret initial matching → (3) Schema filling (feature extraction)

*The reviews came from real estate agents and didn’t have data quality issues*

*Manually validates the quality of extractions .*

*Many extractions are easy to find with simple heuristics, and he spent most of his time looking for such heuristics and determining if they were sufficient.*

*If not, project was potentially going to include ML aspects (but never progressed to that stage)*

#### Preparing/debugging the training data

* manually reviewing heuristic results  
  regular expression searches for relevant phrases
* Manual reviews of small samples
* debugging the model

#### What is the goal of the analysis?

* Create an automatic labeler for [other company’s] data

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

* Received cleaned data from [other company]
* worked with small sample of the data, ~1000 examples
* .csv files

#### Is scalability a problem?

* Always worked with a small sample (1000s)-- every review had enough information to be useful individually.

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

* Python scripts
* Spacy
* used grep and preferred it

#### How do you present your results? (To whom?)

#### How do you collaborate with others on your work?

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

* bouncing back and forth between scripting and regex searches
* Often lost the record of what regular expressions had been tested, and parsing a long regex is complicated.

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

* wish that you could do more from command line, didn’t like constantly bouncing back and forth between the file and scripting
* annoyed at file I/O in python
  + grep runs faster
* semantic matching for regular expression
* history for regular expression searches
* history for other extraction searches

#### Do you have anything else to add? (Wrap-up)

* enjoyed the project

### Hotel review project walkthrough

* very similar according to the subject

#### What are the downstream applications?

* establish if pattern matching is enough for question answering tasks in the domain

#### How do you collaborate with others on your work?

* sharing data and subsets has been a nightmare for the subject

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

* same as Project1 + having mismatches between data sets

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

* better file management system for data files (and subsets thereof)
* combine lightweight NLP and set operations into one toolkit
* sentence partitioning and tokenizing that is faster than spacy
* better sharing of data, functions, progress, between collaborators
* bundling comments with regex history