Review Analysis Interview Subject08

**Demographics**

**Title** Data scientist

**Company** Company 4

**Education** BS and MS Computational linguistics

**Experience in data analysis** 4 years

**Experience in review analysis** 4 years

**Gender** M

## Summary

* **Walk us through a recent review analysis task**
  + Expanding to different domain (working with employee reviews of companies)
* **Goal:** Display of review scores for every attribute for a given business with scores and relevant excerpts from reviews, as well as classification of the business entities into different types to inform the businesses of their strengths and weaknesses

**Pipeline**

1. **Figure out main topics** (schema/categories, this is an iterative process which includes input from the client)
2. **Data cleaning – unsupervised**

* Started with huge dump of reviews.
* Language detection
* Spellchecking
* Tokenization (w/ punctuation), case conversion
* Anonymization of people’s names

1. **Extraction processing – unsupervised**

* Identify substrings that express relevant opinions, detect negation, THEN calculate sentiment
* Rule-based dependency parsing with a SpaCy backend
* Match extractions against topic schema using word embeddings
* Classify each extraction as positive, neutral, or negative
* Aggregate extraction scores and convert to a human-readable 5-point scale for each category

1. **Debugging the model**

* Sample 1000 sentences
* Manual multi-annotator annotation
* Measure precision and recall using annotations as a test set
* Many unit tests, built iteratively for known complex constructions and edge cases
  + Client also reports mistakes, which the company follows up on

**Data and scalability**

Client dependent, largely xml files provided by the client.

Process XML and store it in a database (Hive, Mongo)

Only need to look at raw data points for labeling a subset for training and testing

60 million reviews in multiple languages, 70% is in English

Currently only supporting English, will expand

Try to use as much data as possible for topic analysis

Extraction of sentiments is done on a random sample to fine tune the algorithm, then run on the full set

**Tools**

Python

Hadoop

SpaCy

NLTK

Spark

Gensim for word embeddings

In-house lexicon for sentiments

Transitioning from NLTK to SpaCy for better parsing quality, performance, and interface

**Downstream applications**

Generate quality assessments for client businesses’ services

**Sharing and presenting results**

Team of 4 computational linguists

Collaborate with team using Github pull requests, keep track of accuracy metrics. Some data visualization

Send clients final analysis write ups, includes some visualizations

**Bottlenecks**

Clients changing requirements

Difficult tasks:

Rule based approach does not work well with certain expressions, particularly negation

Word-sense disambiguation

Heavily context-dependent matches

Most time spent:

Topic exploration. Once a schema is set it cannot be changed

**Feature requests**

Anything to help with topic exploration and schema generation