q NLP experience?

Yes, I have finished a couple of nlp project like sequence classification, sentiments analysis. In each of these projects, I went through general procedures like gathering data, cleaning data, representing data, inspecting data and classification.

q Difference between constituency parser and dependency parser?

A constituency parse tree breaks a text into sub-phrases. Non-terminals in the tree are types of phrases, the terminals are the words in the sentence, and the edges are unlabeled.

A dependency parse connects words according to their relationships. Each vertex in the tree represents a word, child nodes are words that are dependent on the parent, and edges are labeled by the relationship.

q Stemmers vs Lemmatizes?

Lemmatization is the process of converting the words of a sentence to its dictionary form. For example, given the words amusement, amusing, and amused, the lemma for each and all would be amuse.

Stemming is the process of converting the words of a sentence to its non-changing portions. In the example of amusing, amusement, and amused above, the stem would be amus.

q Is there an algorithm that tells the semantic similarity of two phrases?

I often check their cosine similarity.

q experience with RASA?

Rasa NLU’s job is to interpret messages, and Rasa Core’s job is to decide what should happen next. It is an open-source natural language processing tool for intent classification and entity extraction in chatbots.

q XGBoost experience?

XGBoost is the leading model for working with standard tabular data (the type of data you store in Pandas DataFrames, as opposed to more exotic types of data like images and videos).  XGBoost is an implementation of the **Gradient Boosted Decision Trees** algorithm (scikit-learn has another version of this algorithm, but XGBoost has some technical advantages.)

q experience with tables in spark?

Yes, I have used spark sql to deal with structured data. We can also query, cache, filter, and perform any operation with spark sql.   
  
q what is embedding mean in general? what about with respect to NLP? Experience?

Vectors for representing words, word is embedded in a particular vector space, vector models of meaning are now the standard way to represent the meaning of words in NLP.

q what are you steps involved in a typical NLP pipeline? specifically text preprocessing?

* Gather data
* Clean data
* Find a good data representation (one-hot encoding, embedding)
* Classification
* Inspection
* Accounting for vocabulary struction
* Leveraging semantics
* Leveraging syntax using end-to-end approaches

q how to deal with mis-spelled words in the nlp model?

Peter Norvig, director of Google research, has written a terrific [spell checker](http://norvig.com/spell-correct.html).

it would be straightforward to extend it to word bigrams or trigrams if you had a corpus of correct text in the language of choice.

q how do you implemented a text summarizer? Architecture? Approach? Alternatives? what would the

training samples look like for this model? purpose of text summarizer - business use case?

Text summarization is the problem of creating a short, accurate, and fluent summary of a longer text document. Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster.

1. Summaries reduce reading time.
2. When researching documents, summaries make the selection process easier.
3. Automatic summarization improves the effectiveness of indexing.
4. Automatic summarization algorithms are less biased than human summarizers.
5. Personalized summaries are useful in question-answering systems as they provide personalized information.
6. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

I usually use a txt summarization api in mashape marketplace, it returns as many sentences to summarize the text your feed in as you want.

q languages you know the best? Python? Scala? R? why?

I like python, but I am also familiar with Scala and R as well.

q tech stack your most comfortable with?

Getting started with Data Engineering – Richard Taylor – Medium



