1 Overview

In this write-up, I review the natural language processing ("NLP") analysis of the sample responses provided by XYZ (""). The core analysis covered the following aspects:

- Basic exploration of response word counts and n-gram distributions.
- Unsupervised exploration of the responses using topic modeling.
- Supervised modeling of the data using Maximum Entropy classification (logistic regression).

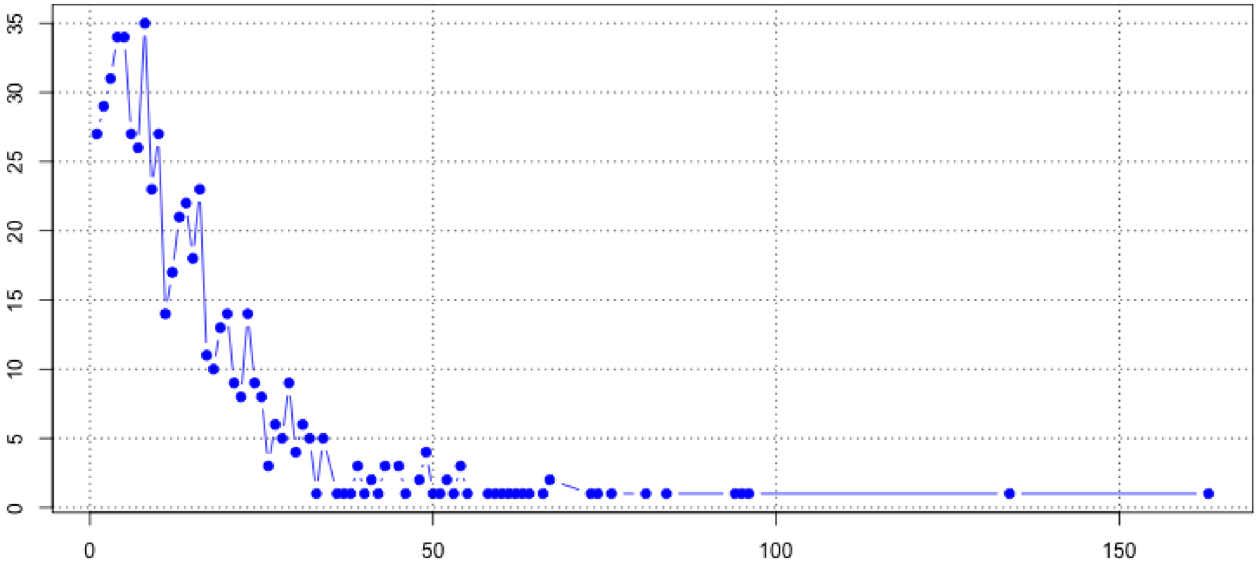
For this initial proof-of-concept, no time was spent on part-of-speech ("POS") tagging and other more complicated feature engineering algorithms.

2 Basic Analysis

The XYZ data was provided on a non-curated basis across 599 sample responses describing why one had decided to stay at a particular company. All references to the employer had been scrubbed as *YOURCO*, which greatly aided in the uni-gram analysis of the data. Additionally, in second tab of the spreadsheet, the 10 Clarabridge sentiment states were included which ranged across Benefits, Career Path, Compensation, Culture, Education/Training, Independence, Job Scope, Satisfaction, Teamwork and Treatment/Respect.

Given the relatively terse nature of the responses, I first developed a distribution of the number of words per response as shown in Figure 1. As one can see, as much as one quarter of the dataset contains responses with less than 5 words, which creates a number of NLP challenges in terms of the limited number of degrees of freedom in fitting regression models to classify the data.

Figure 1: **Distribution of Words per Response**



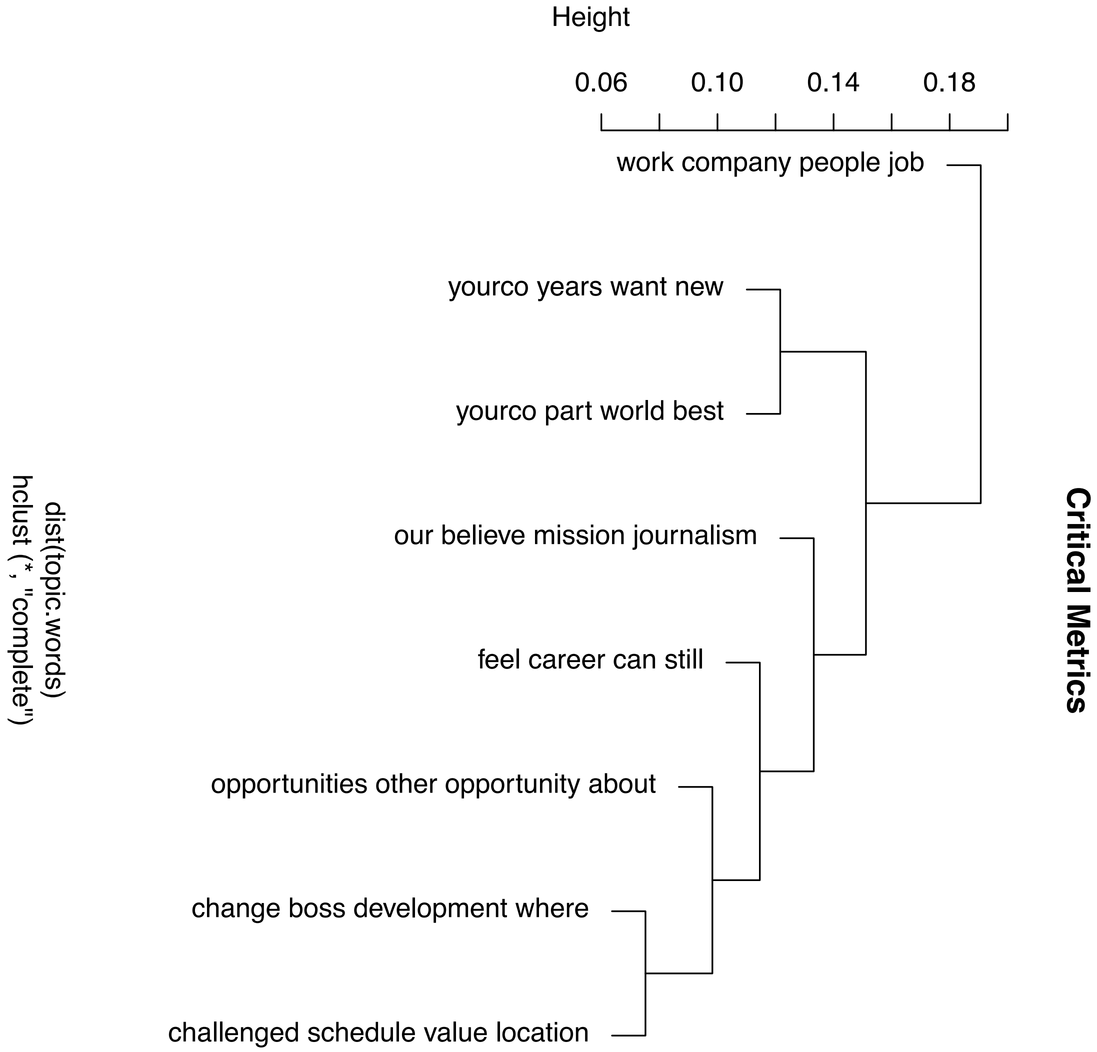
Given this sparseness of the data and a typical machine learning split of raw cases into 80% for training and 20% for testing, it seemed to me that a training dataset of only 479 cases across 10 classes would be far too sparse to achieve stable test dataset results. Hence, my intuition was to create a smaller set of labels depending on how the unsupervised learning on the data worked out.

3 Unsupervised Analysis

In order to develop a statistical sense of the various topics being discussed in the data sample, I employed Latent Dirchlet Analysis to develop a model of the various topics manifest within the data. For this analysis, I employed the open source University of Massachusetts' package know as Mallet[1]. Within this approach one can control both the number of n-grams and total number of topics, in order to arrive at the best probabilistic understanding of the data. Owing to the corpus sparsity, I decided to exclude a hand-curated list of stop words endemic to the response data. As such, I experimented with 2-5 grams and 5-15 topics. The best segmentation of the data seemed to be with 4-grams for topics with about 6-10 topics.

A useful aspect of topic modeling is that the linguistic distance of the topics can be nicely visualized using hierarchical cluster analysis. As one can see in Figure 2, the *work company people job* cluster is dominant and a number of sub-clusters fall out with the mentions of *yourco* grouping together[2].

Figure 2: **Distribution of Words per Response**



[1] Machine Learning for Language Toolkit - http://mallet.cs.umass.edu/

[2] Most NLP analytics convert all words to lower case and remove punctuation as part of the tokenization process.

Given the topic modeling results, I decided to develop my own six-state classification model versus using the Clarabridge labels, though, of course, they were related. When I sorted the responses by their topics, I noticed that there was an approximate hierarchy of Company− >Team− >Job Itself− >Opporunity− >Benefits− >Stability. Hence, I collapse Topic 3, 5 and 6 into Company and used this semi-supervised labelling approach to seed the supervised learning in the following section.

Table 1: **Potential Clustering Factors**

Topic	Phrase	Count	Class
1	love career more department	73	Stability
2	opportunity new reason years	71	Opportunity
3	great company product very	71	Company
4	benefits opportunities feel other	74	Benefits
5	company can time still	66	Company
6	yourco brand our proud	73	Company
7	working team environment flexibility	68	Team
8	work people job enjoy	104	Work

4 Supervised Analysis

The six collapsed states from the topic modeling were then use as the initial seed parameter for the Maximum Entropy modeling, where I opted to use only a uni-gram approach owing to the sparsity to the linguistic data. As mentioned earlier, the first 80% of the dataset was used as training and the remaining 20% as test. As brief test was conducted to ensure that there was no obvious sample bias in the the data ensuring that the class frequencies were roughly equal in both training and test datasets.

4.1 Curation

The initial model on all data was fitted and then compared to the raw responses. The responses were then sorted alphabetically and the validity of both the initial topic model guess and then the MaxEnt classification were examined (class_guess, class_model fields in the MySQL database). Then, I provided a final class_golden label for the response reflecting human judgement based on the text and the two estimates.

4.2 Agreement

The inter-annotator agreement between the topic model and the human estimates was 59% with a kappa score of 52%, which seemed reasonable given such a sparse dataset. (Human annotators generally agree at no more than a 60-80% rate.)

4.3 Training

The accuracy of the training dataset was 97.3%, which is quite high, but training is always susceptible to overfitting.

Table 2: **Potential Clustering Factors**

Class	**Company**	**Team**	**Work/Job**	**Opportunity**	**Benefits**	**Stability**
Company	168	0	0	0	0	0
Team	0	45	2	0	0	0
Work	7	0	127	0	1	0
Opportunity	0	0	1	40	0	0
Benefits	0	0	0	0	37	0
Stability	0	1	1	0	0	49
Accuracy :	466	97.3%				
False Pos:	10	2.1%				
False Neg:	3	0.6%				

4.4 Test

The accuracy of the test dataset was 83.3%, which is also quite high. As one will notice, the majority of False Positives occurs in the Benefits and Stability classes, so if one were to collapse those together, accuracy would likely rise by another 14 observations or to 95%, which is extremely high.

Table 3: **Potential Clustering Factors**

Class	**Company**	**Team**	**Work/Job**	**Opport.**	**Benefits**	**Stability**
Company	63	0	2	1	0	0
Team	1	5	0	0	0	0
Work	1	0	21	0	0	0
Opportunity	1	0	0	3	0	0
Benefits	4	0	4	0	3	0
Stability	3	0	3	0	0	5
Accuracy :	100	83.3%				
False Pos:	17	14.2%				
False Neg:	3	2.5%				

5 Conclusion

The basic results of the analysis of the 599 sample responses were very encouraging. Among the various possible improvements to the model would be:

- Use of Part of Speech ("POS") tags to track Adjective and Adverb frequencies.
- Use of temporal methods to detect length of time at the company, which is clearly a factor in staying at one's job.
- Use of positional determinations, i.e. position of YOURCO and other qualifiers in the string like JOB, WORK, TEAM, etc.
- Explicit use of negation in bi-grams, i.e. NOT_GOOD, etc.

The next steps would be to determine whether a 5-6 state model for determining why employees stay at a given company is a valid way to proceed. If the Clarabridge 10 classes are preferred, then I would simply train on a larger dataset with the Clarabridge labels and develop the feature engineering until we approached similar results with this simpler model. The main question in my mind is whether the reasons for staying are the same as going. Training two different models definitely makes sense, as the words used in each situation will likely be different, but MaxEnt could certainly handle it as a single model, if need be.