Dated: 07/07/2018

Project Ideas: Capstone Project

Objective: Capstone	**Understand the python codes & packages to use and understand the whole A-Z of performing data analyses, including importing, cleaning, wrangling and visualization of the data.**
Project Idea (I)	**Text mining in Python:** More than 70% of potentially usable data exists in unstructured form. This data is stored in text format. Text mining or Natural language processing provides us with techniques to gain actionable insights from these data. In this project, I propose to address HR analytics problem of understanding what the employees are saying about their company and how HR head can develop their strategy based on the insights we provide them with text mining of the reviews. **Data Set:** Amazon VS Google Reviews. (500+ and 500-ve reviews). Source (Data Camp) **Questions:** ● Which company has a better work-life balance? ● Which is better perceived to pay according to the online reviews?
Project Idea (II)	**HR Analytics/ People Analytics / Workforce Analytics:** For companies it is becoming more important every day, to hire the best talent and maintain a low turnover rate. At the same time, they want their existing employees to keep their productivity and engagement levels high and accidents a lowest. In this project, I propose to analyze several data sets to address these issues through statistical **Data Set:** a) HR data (2940, 4) b) Accident data (302,3), c) Survey data, d) Performance Data. **Questions:** ● When and where is the highest accident rate? ● Explaining the reasons for increase in accidents? ● Regression to analyze drivers. ● What is driving low employee engagement?
Project Idea (III)	**Market Analytics / Statistical Modeling:** Companies have to make important decisions every day. Statistical models help companies make hypothesis-driven business decisions based on the market data. In this project for the capstone project, I propose to explore the sales data and help them in deciding which customers are important for their business and growth. **Data Set:** Churn Data, Sales Data, Sales Data-months 2-4, survival data, Default data, News data. **Questions:** ● Which customers are valuable to your business? ● Which customers will leave your business? ● Survival Analysis. ● CRM data Analysis.

Text Mining for Data Science with Python

Natural Language processing or Texting Mining	**Introduction**: More than 70% of potentially usable data exists in unstructured form. This data is stored in text format. Text mining or Natural language processing provides us with techniques to gain actionable insights from these data.
Problem Statement	**Human Resources Analytics:** The HR department of Amazon wants to develop strategies for hiring appropriate talent and prospective employees who are more suited to their work culture. They also want to understand what distinguishes Amazon employees with Google employees, which is another technology giant. The Data Scientists from Springboard have been asked to explore online reviews and provide insights.
Datasets	**Amazon and Google Reviews**. (500+ and 500-ve reviews) Several employees from both companies submit online reviews. These are categorized into positive and negative reviews. Four data sets are available for analysis. Resource: Datacamp (Ted Kwartler)
Questions to be addressed	● What are employees saying about Amazon and Google? ● Which company has better work-life balance? ● Which is company is perceived better in compensating according to the online reviews? ● Can HR use the analyses in strategizing their talent hunt?
Methods	Steps involved in Analyzing the Text Data (NLTK library) 1) Organization & Text processing (raw text -> sentence segmentation -> tokenization -> corpus -> matrix -> word count) 2) Feature extraction and Analysis (TF, DTM, TF-ID, TF-IDF, DM -> Top terms & Word associations) 3) Machine Learning (clustering, classification, association rules, predictive modeling) 4) Data visualization: Dendrograms, Word cloud. Pyramid plots.
Significance	NLP can provide key business intelligence insights, which are not just limited to human resource management but can also help other departments including marketing to improve sales, accounts and finances, purchasing, customer service support, distribution, research, and development.

Text Mining, Tech Giant Reviews using Natural Language Processing

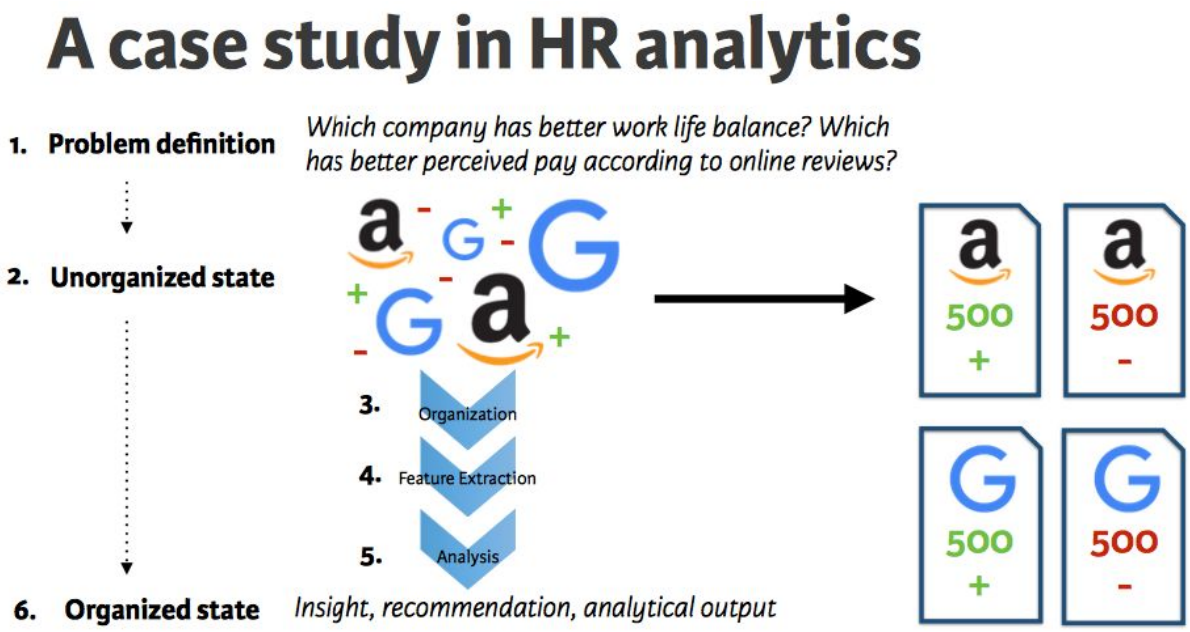
Introduction

More than 70% of potentially usable data exists in unstructured form. This data is stored in text format. Text mining or Natural language processing provides us with techniques to gain actionable insights from these data.

Problem Statement

The HR department of Amazon wants to develop strategies for hiring appropriate talent and prospective employees who are more suited to their work culture. They also want to understand what distinguishes Amazon employees with Google employees, which is another technology giant. The Data Scientists from Springboard have been asked to explore online reviews and provide insights.

Adapted from



Datacamp course

Questions to be addressed through text mining

- What are employees saying about Amazon and Google?
- Which company has better work-life balance?
- Which company is perceived better in compensating according to the online reviews?
- Can HR use the analyses in strategizing their talent hunt? Recommendations to the HR team?

Methods / Steps used to answer the above questions

- Organization & Text pre-processing (raw text -> sentence segmentation -> tokenization -> corpus -> matrix -> word count)
- Feature extraction and Analysis (TF, DTM, TF-ID, TF-IDF, DM -> Top terms & Word associations)
- Machine Learning (clustering, classification, association rules, predictive modeling) to predict reviews.
- Data visualization: Dendrograms for word relationships, Word clouds. Pyramid plots for comparison of words used in reviews.

Observations through Analyses Qs1:

What are employees saying about Amazon and Google?

Analyses of top 15 words (bigrams) from the Amazon and Google Reviews.

Amazon Pro

good pay	25
great benefits	24
smart people	20
place work	17
good benefits	16
fast paced	16
great place	12
learn lot	12
work environment	11
great pay	11
good work	10
pay good	10
pay great	10
people work	10
pay benefits	9

Amazon con

long hours	29
worklife balance	21
work life	21
life balance	20
not enough	9
peak season	8
management not	8
not good	8
high turnover	7
work environment	6
hard work	6
hours long	6
many people	6
work hours	6
not really	6

Google Pro

smart people	42
free food	41
place work	26
great benefits	22
great perks	20
great work	18
people great	16
work environment	16
great people	16
great place	15
great culture	14
worklife balance	12
people work	12
work life	12
work great	11

Google Con

not really	11
no cons	10
long hours	10
not think	9
not get	8
middle management	8
work life	8
life balance	8
worklife balance	8
get promoted	7
not many	7
but not	7
things done	7
get things	6
get lost	6

We could also visualize this as word clouds.

Amazon Pros and Cons

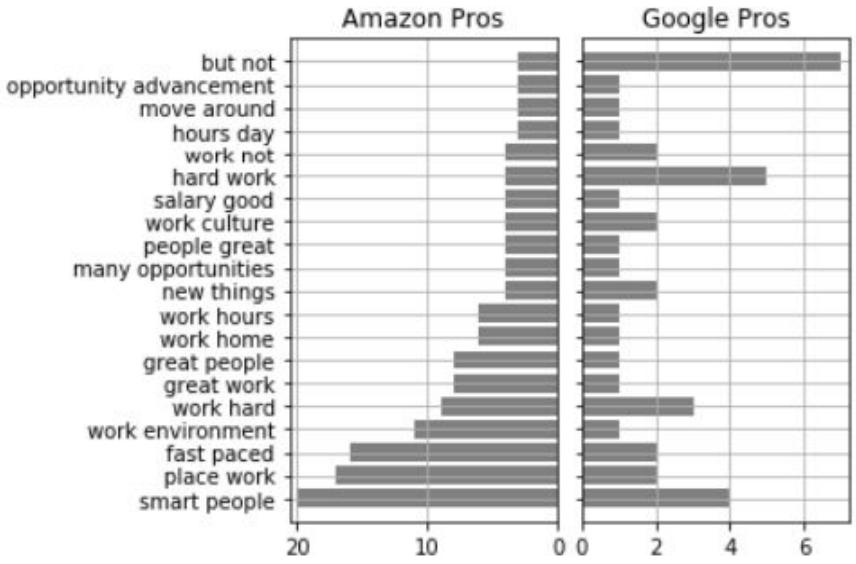
Google Pros and Cons

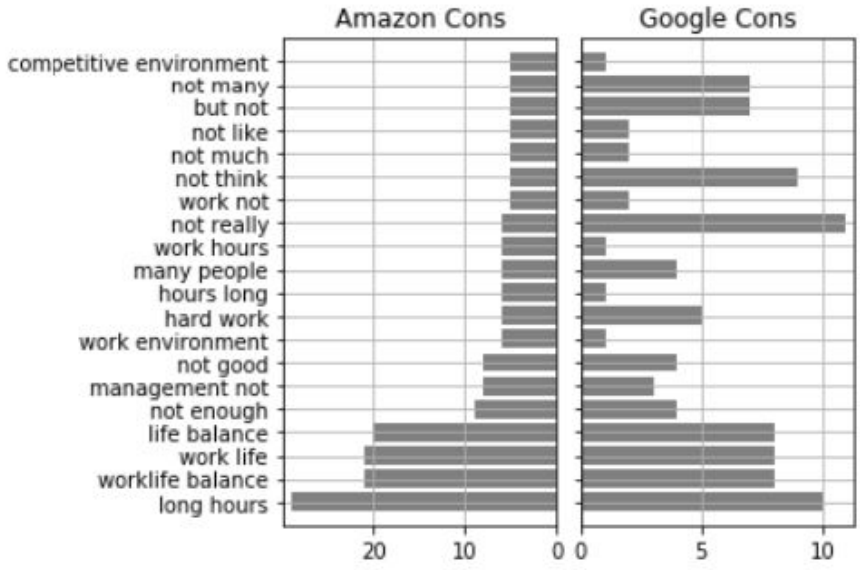
Q2: Which company has better work-life balance?

From the frequency table and the word cloud we see that work-life balance bigram appears in Amazon con reviews appears 21 times. Where as in Google reviews the word appears in both pros and cons 12 and 18 times respectively. This would suggest that Google has better work life balance ratings.

Q3: Which company is perceived better in compensating according to the online reviews?

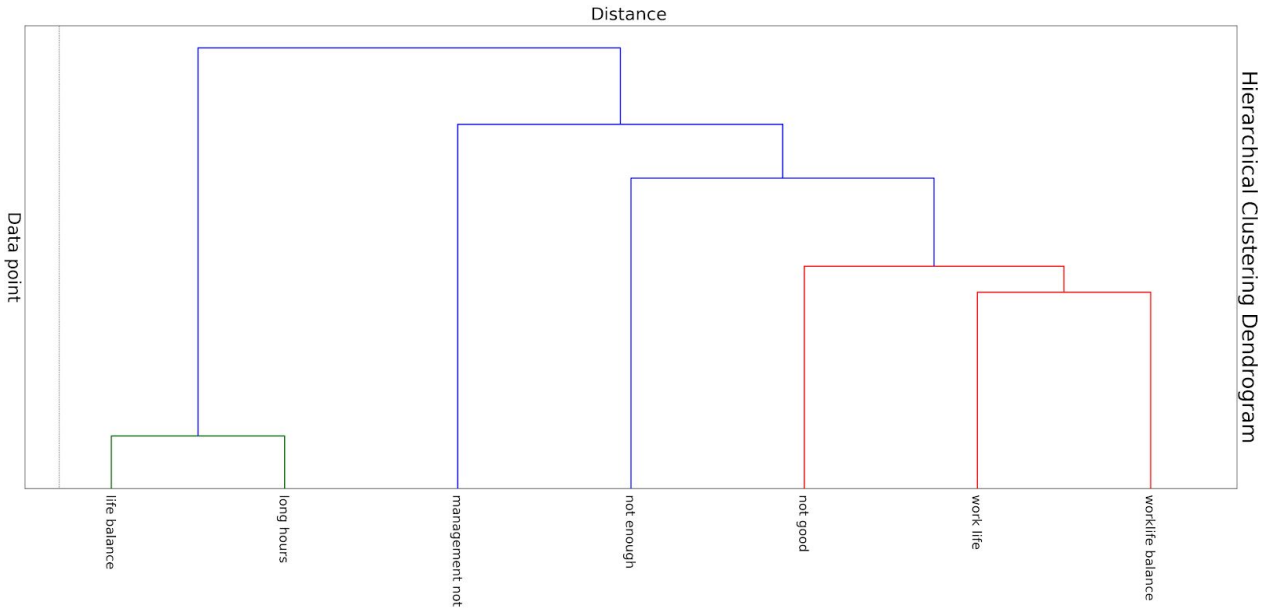
We can see that more employees in amazon are talking positive about the “good pay” and “benefits” as compared to google. This can be visualized by plotting pyramid plots.





Q4: Can HR use the analyses in strategizing their talent hunt? Recommendations to the HR team?

For this we can do more detailed analysis including generating TF-IDF and word associations to observe which words are being used by same employees more consistently. It seems there is a strong indication of long working hours and poor work-life balance in the reviews.



From Hierarchical Clustering we also observe that employees like fast paced work at amazon. This provides an indication to HR to include this criteria in their hiring process, and to find prospective employees which would like fast paced environment. Please turn over for the dendrogram.

As a part of the comprehensive report we will perform Machine Learning to create a model to predict if the a review is a pro or a con:

We will use logistic regression, random forest and multinomial naive bayes to train, test and predict whether the review is a pro or a con.

We will also make an assessment about the models works the best to classify pros and cons reviews.