NLP Sentiment Analysis Project Proposal

1. General:

Sentiment analysis, which is a subtopics of Natural Language Processing (NLP), has been gradually becoming more and more popular. It is a contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

Sentiment Analysis has many applications ranging from ecommerce, marketing, to politics and any other research to tackle with text or unstructured text data. Companies, especially in e-commerce, also do sentiment analysis to collect and analyze customer feedback about their products. Besides that, potential customers prefer to review the opinions of existing customers before they purchase a product or use a service of a company. As seen here, there are two parts in e-commerce; one is the online retailer, which wants to maximize e-commerce sales or services, and the other is the consumers, who want to have the best product or service over alternatives.

2. Problem:

In this project, Amazon is our client. The company wants to develop a software tool that will identify the positive and negative words which customers use when they write reviews for the beauty products as their purchase inclination. For that, they gave their 9 years beauty products' reviews between 2005-2014 and asked us to develop a model which will identify positive and negative words used in the reviews as a component of customer's sentiment towards to the company's beauty products.

According to the customer request, we will build a sentiment analysis model as part of natural language processing, based on their reviews on the beauty product online purchases. Our dataset consists mainly of customers' reviews and ratings.

3. Description of the Data Set:

Beauty dataset revolving around the reviews written by customers. This is a real commercial data.

This data includes 28798 rows and 9 feature variables. Memory usage is 2.2+ MB.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A6VPK7X53QNAQ	B0000CC64W	AmazonDiva "Keep Calm and Carry On."	[5, 5]	I am a devotee to this serum, it does wonders ...	5.0	If I had to choose only one product to take ca...	1245283200	06 18, 2009
1	A3CHMHGSJSQ02J	B0000CC64W	Anon. A. Non	[2, 2]	As a woman nearing 50, I need all the help I c...	5.0	Makes my skin lovely and smooth	1358467200	01 18, 2013
2	A1V1EP514B5H7Y	B0000CC64W	asiana	[0, 0]	I've used this regenerating serum for more tha...	5.0	Works well at a reasonable price	1322524800	11 29, 2011
3	A1X2LENOF84LCQ	B0000CC64W	D "D"	[62, 75]	I have tried so many products to just be total...	4.0	This does work ladies	1113350400	04 13, 2005
4	A2PATWWZAXHQYA	B0000CC64W	Farnoosh Brock	[1, 1]	I love Oil of Olay. My primary moisturizer is ...	1.0	Did not like the feel/texture of this serum	1387584000	12 21, 2013

Each row corresponds to a customer review, and includes the variables:

reviewerID : ID of the reviewer, e.g. A2SUAM1J3GNN3B - type: object
asin : ID of the product , e.g. 0000013714 – type: object
reviewerName : name of the reviewer – type: object
helpful : helpfulness of the review, e.g. 2/3 – type: object
reviewText : text of the review – type: object
overall : Rating – type: float64
summary : summary of the review – type: object
unixReviewTime : time of the review (unix time) – type: int64
reviewTime : time of the review (raw) – type: object

The data was in Standford Analysis Project webpage. The original data was in a JSON format there. In order to analyze the data, I should change the data format. For that, I import JSON and decode JSON file with using query in order to convert JSON file to csv file format .

Data Source:
http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Beauty_10.json.gz

4. Approach to solving the problem

I will approach this NLP Sentiment Analysis project by following the steps below:

a. Understand the business problem
b. Create a repository
c. Gather the data from Amazon review link and load it into Jupyter notebook.
d. Analyze the data to determine the data quality
e. Preprocessing
   (1) Data Set Basic Formatting
   (2) Missing Values
   (3) Cleaning the text feature
   (4) Creating a new column consists of the classification of the ratings
f. Data Storytelling
g. Apply feature extraction and NLP techniques
h. Selecting Evaluation Metric
i. Modeling
j. Selecting Best Model
k. Prepare a report

5. Project Deliverables

My deliverables will be a milestone report, a PowerPoint presentation, and a Jupyter notebook associated with my project.