Capstone Project-2 Proposal
NLP Sentiment Analysis

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.

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 home and kitchen products as their purchase inclination. For that, they gave their 14 years home and kitchen products' reviews between 2000-2014 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 home and kitchen products.

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

3. Description of the Data Set:

Home and kitchen dataset revolving around the reviews written by customers. This is a real commercial data.

This data includes 25445 rows and 9 feature variables. Memory usage is 1.9+ MB.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A1115ST6F5CWYP	B00000JGRT	Amalfi Coast Girl	[29, 33]	I have had one of….	4.0	good for a first ice cream machine	1148256000	05 22, 2006
1	A188JOXWF4EY1R	B00000JGRT	Ann B. Hibbard "anbee"	[4, 4]	We actually found…	4.0	Wonderful Product!	1282176000	08 19, 2010
2	AUAX1QWUCYKSX	B00000JGRT	Ashley S	[1, 1]	This product works…	5.0	Works as expected	1243555200	05 29, 2009
3	A2C27IQUH9N1Z	B00000JGRT	audrey	[12, 13]	After trying other…	5.0	this will be one of your favorite small applia...	1043712000	01 28, 2003
4	A2PN65B6BSTIYZ	B00000JGRT	B. A. Chaney	[1, 1]	I bought this ice cream…	5.0	You'll be addicted to homemade ice cream!	1214179200	06 23, 2008

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://seotest.ciberius.info/seo--snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Home_and_Kitchen_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.