CS 410 Course Project

10.24.2021

Ratan Bajpai
MCS-DS Student
University of Illinois at Urbana-Champaign
Champaign, IL

1. Overview

This document contains the project proposal for the course project for CS 410.

2. Team Captain / Member

1. Ratan Bajpai - rbajpai2@illinois.edu

3. Free Topic - Sentiment Analysis

This course project attempts to implement and analyze sentiment analysis for the "Movie Reviews" (from NLTK corpus) dataset using two approaches. The first approach is by using Amazon Comprehend NLP service and the second is the python based Natural Language Toolkit (NLTK). We will analyze the results from both and compare / contrast them.

4. Project Details

1. Project Task

The project task is:

a. Implement sentiment analysis on movie reviews dataset using Amazon Comprehend.
b. Implement the above sentiment analysis task using NLTK.
c. Analyze and compare the results using both approaches.
d. List any practical considerations to improve results.

This project task is interesting as it will compare the results of Amazon Comprehend NLP service which is a much newer system for NLP (launched in November 2017) as compared to NLTK (launched in 2001). If available we will try to list the NLP algorithms used by both systems.

2. Planned Approach

As listed in the project task, the planned approach is to implement sentiment analysis using both Amazon Comprehend and NLTK, and compare the results. Also,

we will try to see if there are any parameters in both systems that can be fine tuned to give better results.

3. Tools, Systems or Datasets

The following tools, systems and datasets will be used:

a. Amazon Comprehend, Amazon S3
b. Boto3
c. NLTK
d. Movie Reviews Dataset (http://www.cs.cornell.edu/people/pabo/movie-review-data/)

4. Expected Outcome

Both the systems should be able to classify movie reviews into positive and negative categories. The training and test accuracies for both systems will be compared. If possible parameters on which the accuracy depends will be identified.

5. Project Evaluation

Project will be evaluated upon successful implementation of both systems, and any interesting insights on comparing the results from both implementations.

6. Programming Language

Python and it's libraries as well as some additional tools will be used for implementing this project.

Project Estimate

Task	Estimated Time
Analyze, massage movie reviews dataset	2 hours
Design, implement, test, tune and troubleshoot sentiment analysis using Amazon Comprehend.	10 hours
Design, implement, test, tune, and troubleshoot sentiment analysis using NLTK.	10 hours
Analyze and compare results for both	5 hours

systems. Look at any changes based on any system parameters.	
Prepare the project report.	4 hours
Demo setup.	4 hours