Sentiment Analysis of IMDB Movie Reviews

Data Science Capstone Project
Launch Report

Date: 01/20/2021

Team Members: Lawrence Love, Gustavo Ferreira, Frank Zhao, Yan Li

The System/Product

System/Product Name: Sentiment Analysis of IMDB Movie Reviews

Introduction:

Our team is interested in sentiment analysis by using word2vec technology. So we choose the IMDB dataset which has 50K movie reviews. The dataset is simple with just two columns: reviews and sentiment(positive/negative). First, we will process the reviews to make it more concise. Then, we will convert it to vectors by using word2vec. Lastly we will build some classification models to predict whether a given review is positive or negative.

Highlighted Features:

- Text data cleaning
- Utilizing word2vec
- Sentiment Analysis
- Word Cloud
- Classification: Logistic, Naive Bayes, Random Forest, and LSTM

Issues:

Since we are not familiar with word2vec technology, we may need to put in some effort to study it first.

The Team

Team Members and their specialties:

Lawrence Love:

- I know Python very well.
- I know data-frame manipulation and processing with Pandas very well.
- I know null-hypothesis, z-test, t-test, Anova-test and their applications.
- I know data visualization with matplotlib and seaborn in Python very well.
- I completed a Graduate Certificate in Applied Statistics in 2016, I could use some brushing up on his knowledge.
- I know the Keras package

Gustavo Ferreira:

- I know Python very well.
- I know data-frame manipulation and processing with Pandas very well.
- I know null-hypothesis, z-test, t-test, Anova-test and their applications.
- I know data visualization with matplotlib and seaborn in Python very well.

Frank Zhao:

- I know Python very well.
- I know data-frame manipulation and processing with Pandas very well.
- I know null-hypothesis, z-test, t-test, Anova-test and their applications.
- I know data visualization with matplotlib and seaborn in Python very well.
- I have used Python for almost 3 years, but I would say my skills would be 7/10. I am interested in Deep Learning and have some practices by using Tensorflow 2.0

Yan Li:

- I know Python very well.
- I know data-frame manipulation and processing with Pandas very well.
- I know null-hypothesis, z-test, t-test, Anova-test and their applications.
- I know data visualization with matplotlib and seaborn in Python very well.
- I am also good at SQL and Tableau. And I am interested in Deep Learning by using Tensorflow.

Team Communication:

We will use Slack to communicate and meet on Zoom. We plan to meet twice a week on Tuesday and Sunday, and will share the code through Github and documents through google Docs.

Team Issues:

Right now, our team is good since this is the second time that we work together.

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Project**	Lawrence Love, Gustavo Ferreira, Frank Zhao, Yan Li	Yan Li
2	**Team**	Lawrence Love, Gustavo Ferreira, Frank Zhao, Yan Li	Yan Li
3	**Plan**	Frank Zhao, Yan Li	Yan Li

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.