Product Title Classification

11th April 2021

Saif Ul Islam, Muhammad Hassan Zahid
18K 0307, 18K 0208
Section D, H
Information Retrieval - CS 317

Overview

This project will attempt to solve and work on the "**Product Title Classification**" problem, taking the paper "Bagging Model For Product Title Quality With Noise" as a reference that was mentioned along with the project ideas in the semester classroom for *"Information Retrieval - CS 317"*. The project will attempt to go through the individual sections of the given data, in an attempt to try different approaches and achieve significant results with the process. The approach might be influenced by the original authors as the winners of the CIKM AnalytiCup 2017.

As far as the process and the contents of the project are concerned,

1. The idea is to identify the conciseness and clarity of a title when the given data consists of the descriptions of that SKU (Stock Keeping Unit) in an attempt to better label those titles according
2. Analyze Lazada's dataset, southeast Asia's number one online shopping and selling destination. The dataset is attached along with the paper itself, with training and testing data split and labeled on the ideas of being **concise** and **clear**
3. Explore the idea of "**Feature Engineering** / **Feature Extraction**" through the use of "**ensemble models**" *OR* "**bagging**" OR "**character n-grams**"
4. Experiment with **Light Gradient Boosting** (LGB) and **Extreme Gradient Boosting** (XGB) as potential model solutions

Goals

1. Provide a visual EDA (*Extrapolatory Data Analysis*) as the data is further analyzed
2. Extract features through the methods of **L1-Norm** and **SVM model(s)**

Specifications

The technological and technical requirements of the project have not yet been explored properly - but a quick go through of the contents of the tagged paper indicate a tool that will help us to,

1. Visually analyze the contents of the data and our evaluation against the desired results
2. Offers NLP tools to quickly ease the process of experimenting with a slew of different techniques very rapidly

In this case, a tool such as **Python** seems like an ideal choice. The most probable route the team will take to develop and finish the project. Maybe the entire implementation will on a tool such as **Google Colab** or **Kaggle**

Difficulties

The paper first needs to be broken down into different steps that are feasible for us both during the remaining time period of this semester. In a brief, the difficulties are,

1. Breaking all the requirements into modules that are easily divisible for both of us
2. Setting up a project repository which both can work on together
3. Creating the environment that will work on

Milestones - Todo

I. Preprocess data from the dataset

Constructing a basic pipeline for the dataset in order to allow for more elaborate feature extraction and to use more valuable fields

II. Start with a base model

Constructing a base model that uses the data from the previous stage and passes it for fitting on some basic model. The series of trials and techniques specifically has to be decided after the proposal and in some due time.