Project Proposal for Data1030

Yue Zhuang

September 30, 2019

Abstract

In this project, we would like to dig into the popularity of news in multiple social media platforms. With the timing of publishment an important feature that affects the popularity of news, we will develop a real time predicting system for certain piece of news. The result of popularity prediction will be helpful for the existing news recommend system.

1 Problem description

In this section, we will describe the problem we want to solve.

1.1 traget variables

The target variable of our project would be the popularity rate of a certain piece of news.

1.2 problem type

The problem would be regression since we will predict the popularity rate of news.

1.3 interest & importance

Unlike most regression problems which has well organized features as the input of a ML model, the popularity rate of a piece of news is effected by both the content of the news and the timing of publishment. We will need to dig into the time series data to investigate the popularity trend of a certain topic. The way we select features for the trend model will greatly affect the performance of our prediction model, bringing challenge to the problem.

The prediction of popularity rate of news can be of significant importance for the recommend systems. Social media platforms will be able to push the most popular piece of news to consumers by introducing the real-time prediction model to their existing recommend system.

2 Dataset description

In this section, we will describe the dataset – News Popularity in Multiple Social Media Platforms Data Set. This is a large data set of news items and their respective social feedback on multiple platforms, including: Facebook, Google+ and LinkedIn. The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: economy, microsoft, obama and palestine.

2.1 dataset size

We have multiple csv files storing data in two distinct formats.

2.1.1 news final sheet

In this csv file, each row represents a certain piece of news, with total 93240 rows.

The columns includes 11 features: its IDLink, title, headline, source, topic, publishdate, sentimenttitle, sentimentheadline and the final popularity on different social media platforms (Facebook, GooglePlus, LinkedIn).

2.1.2 times series sheet

In these csv files, each file includes data for the time series popularity rate for news on a certain social media platform and with certain topics.

Each row represents a certain piece of news, with approximatel 3000 rows in each csv file. The columns are the popularity rates in each time slice, with a total of 144 slices. The time series starts from the time of publishment and each time slice has a range of 20 minutes.

2.2 documentation

2.2.1 news final sheet

Table 1: News Final Sheet

Column Name	Type	Brief Description
IDLink	numerical	Unique identifier of news items
Title	string	Title of the news item according to the official media sources
HeadLine	string	Headline of the news item according to the official media sources
Source	categorical	Original news outlet that published the news item
Topic	categorical	Query topic used to obtain the items in the official media sources
Publishdate	numerical	Date and time of the news items' publication
SentimentTitle	numerical	Sentiment score of the text in the news items' title
SentimentHeadline	numerical	Sentiment score of the text in the news items' headline
FaceBook	numerical	Final value of the popularity according to Facebook
GooglePlus	numerical	Final value of the popularity according to Google+
LinkedIn	numerical	Final value of the popularity according to LinkedIn

2.2.2 times series sheet

Table 2: Time Series Sheet

Column Name	Type	Brief Description
IDLink	numerical	Unique identifier of news items
TS1	numerical	Level of popularity in time slice 1 (0-20 minutes upon publication)
TS2	numerical	Level of popularity in time slice 2 (20-40 minutes upon publication)
...	...	...
TS144	numerical	Final level of popularity after 2 days upon publication

2.3 public projects

In this section, we will give short descriptions for 2 public projects where the data has been used, and talk about how the features were used.

2.3.1 public project 1

Multi Source Social Feedback of Online News Feeds
Author: Nuno Moniz, Luis Torgo
Date: Jan 2018

In this project, the authors provided a smoothed approximation of the amount of news per day, for each topic from both Google News and Yahoo. They also concerns the evolution of news items' popularity in the various social media sources, finding that news items obtain close to half of their final popularity in a short amount of time. The authors used R to analyze the evolution of available information in each time slice for all topics and social media sources, coming to statistical results.

2.3.2 public project 2

The Utility Problem of Web Content Popularity Prediction
Author: Nuno Moniz, Luis Torgo
Date: May 2018

In this project, the authors provided a study of numerical web content popularity prediction approaches and their ability to forecast and rank highly popular web content. Kernel regression and k nearest neighbour methods are used to timely predict the popularity of a certain piece of news and to provide a ranking for those news. The authors did not release which features they have chosen for the model.

2.3.3 our project

In our project, we will focus on predicting the popularity for each piece of news on a real-time basis. We will try multiple advanced machine learning models and pay more effort to data preprocessing and feature selection. What's more, our training will focus on the most popular pieces of news rather than all available news.

3 Data Preprocessing

Table 3: Data Preprocessing for News Final Sheet

Column (Feature) Name	Type	methods	reasoning
Source	categorical	OneHotEncoder	categorical data with no specific order
Topic	categorical	OneHotEncoder	categorical data with no specific order
Publishdate	numerical		used later to get access to another sheet
SentimentTitle	numerical	StandardScaler	continuous data without exact range
SentimentHeadline	numerical	StandardScaler	continuous data without exact range

Table 4: Data Preprocessing for News Final Sheet

Column (Feature) Name	Type	methods	reasoning
TSi	numerical	Devide by the final popularity TS144	get percentage data
TS144	numerical	/	final level of popularity

Now we have 5 features for sheet1 and 144 features for sheet2. We may make extra progress to get those two sheets connected, providing additional convincing features for our model.

Github url: https://github.com/zysophia/News_Rank_Prediction.