INNOMATICS RESEARCH LABS

Journal Post Comments Prediction

7th March 2020

OVERVIEW

This data originates from Journal posts. The raw HTML-documents of the Journal posts were crawled and processed. The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. In order to simulate this situation, we choose a basetime (in the past) and select the blog posts that were published at most 72 hours before the selected base date/time. Then, we calculate all the features of the selected Journal posts from the information that was available at the basetime, therefore each instance corresponds to a blog post. The target is the number of comments that the blog post received in the next 24 hours relative to the basetime.

GOALS

1. The task associated with the data is to predict how many comments the post will receive.
2. Good Readable Code with interpretations.

MILESTONES

RMSE

About Data

1...50: Average, standard deviation, min, max and median of the Attributes 51...60 for the source of the current blog post with source we mean the blog on which the post appeared.
For example, myblog.blog.org would be the source of the post myblog.blog.org/post_2010_09_10
51: Total number of comments before base time
52: Number of comments in the last 24 hours before the base time
53: Let T1 denote the datetime 48 hours before base time, Let T2 denote the datetime 24 hours before base time. This attribute is the number of comments in the time period between T1 and T2.
54: Number of comments in the first 24 hours after the publication of the blog post, but before base time
55: The difference of Attribute 52 and Attribute 53
56...60: The same features as the attributes 51...55, but features 56...60 refer to the number of links (trackbacks), while features 51...55 refer to the number of comments.
61: The length of time between the publication of the blog post and base time
62: The length of the blog post
63...262: The 200 bag of words features for 200 frequent words of the text of the blog post
263...269: binary indicator features (0 or 1) for the weekday (Monday...Sunday) of the base time
270...276: binary indicator features (0 or 1) for the weekday (Monday...Sunday) of the date of publication of the blog post

277: Number of parent pages: we consider a blog post P as a parent of blog post B, if B is a reply (trackback) to blog post.
278...280: Minimum, maximum, average number of comments that the parents received
281: The target: the number of comments in the next 24 hours
(relative to base time)

Time

9:30 AM to 1:30 PM
7th March 2020

All the best from team Innomatics.