ANLY-580 project proposal

Yifeng Chen, Changwen Li, Jingyi Qi

The ultimate goal of this project is to create an automatic classification model for text posts. With the emergence and quick adoption of smartphones, people are spending tremendous amounts of time interacting with each other online using social media. However the system is not so smart yet where people have to choose what category they want to post to. By adopting such system created by using multiple ensemble classification, neural networks or nlp models, users can decide where they want to post from the smart categories suggestion or even automatically send messages without even thinking what categories the messages belong to which could help people to post things faster, connect with other people more easily and make people more addicted to the platform.

The topic of this project is “The Reddit Self-Post Classification”. Reddit is a social news site with the slogan: A voice ahead of the news, a voice from the Internet. Users (also called redditors) can browse and submit links to content on the Internet or post their own original or relevant user-submitted text. Other users can vote on the links posted with high or low scores, and links with outstanding scores will be placed on the front page. In addition, users can comment on posted links and reply to other commenters, thus creating an online community. Reddit users can create their own section of the argument, both informal for those posting links and comments, and formal for those posting links and comments, like Reddit user submissions, and formal for associations. In this project, we will focus on self-posts, with a title and a body of markdown text written by users on Reddit.

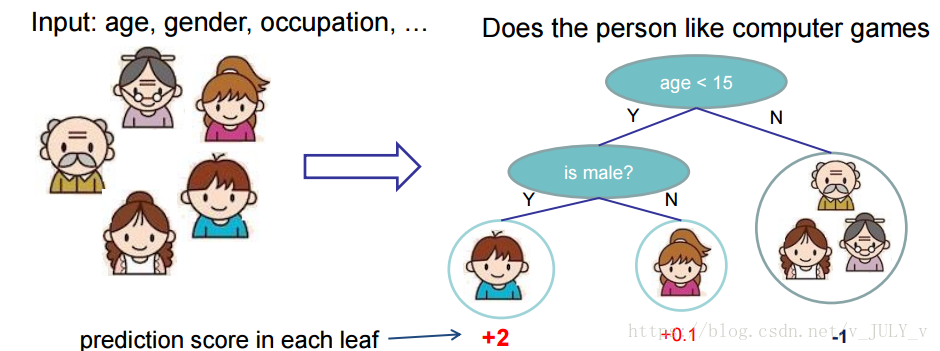
Self-posts are divided into various “subreddits” based on the types of posts being submitted, for example “r/politics” or “r/MachineLearning”. However, subreddits are usually created by users themselves rather than the administrator of Reddit. Thus, sometimes redditors may not know which community (or subreddit) their self-posts belong to and they may need recommendation from the website. Therefore, the purpose of our project is to create a text classifier which groups users’ posts into different topics so that the website can provide recommendations for users based on the results from our text classifier.

We will apply three models in this project, compare their training performances and select the best model. The models are as follow:

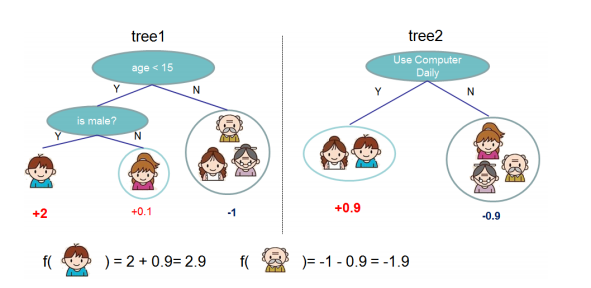
* XGBoost

XGBoost is called eXtreme Gradient Boosting. It is often used in some competitions with remarkable results. The algorithm used in XGBoost is a modification of GBDT (gradient boosting decision tree), which can be used in both classification and regression problems. Because XGBoost is still essentially a GBDT, XGBoost strives to maximize speed and efficiency. Both of them are boosting methods.

XGBoost is combined with a categorical regression tree (CART tree). For example, we want to predict how much a family likes video games. We would first consider that young is more likely to like video games compared to old; and male is more likely to like video games compared to female; we will first distinguish between children and adults according to their age, and then distinguish whether they are male or female by gender, scoring each person one by one in terms of their video game preference.



We trained two trees, tree1 and tree2 based on age and gender. Similar to the principle of gbdt, we need to add up the conclusions of the two trees to reach the final conclusion. The predicted score of the child is the sum of the scores of the nodes where the child falls in the two trees: 2 + 0.9 = 2.9. The predicted score of the grandfather is the same: -1 + (-0.9) = -1.9.



Extending the above integrated learning approach to the general case, the prediction model is known to be：

Where K is the total number of trees, denotes the Kth tree, and indicates the prediction result of the sample of .

* Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Bayes’ theorem states the following relationship, given class variable y and dependent feature vector X1 through Xn: P(y|x1,....,xn) = P(y)P(x1,...,xn|y)/P(x1,...xn)

Using the naive conditional independence assumption that P(xi|y,x1,...xi-1,xi+1,....xn) = P(xi|y), for all i, this relationship is simplified to P(y|x1,...,xn) = P(y)Since P(x1,...,xn) is constant given the input, we can use the following classification rule: P(y|x1,...,xn) P(y)= argmax(y)P(y)

* Hugging Face

Hugging Face is an NLP-focused startup with a large open-source community, in particular around the Transformers library. Transformers is a python-based library that exposes an API to use many well-known transformer architectures, such as BERT, RoBERTa, GPT-2 or DistilBERT, that obtain state-of-the-art results on a variety of NLP tasks like text classification, information extraction, question answering, and text generation. Those architectures come pre-trained with several sets of weights.

The dataset will be downloaded from Kaggle, which is called “The reddit self-post classification task’. The data consists of 1013M self-posts, posted from 1013 subreddits (1000 examples per class) from 2016/06/01 to 2018/06/01. For each post, there will be the subreddit, the title and content of the self-post. It is worth noting that the subreddits here are not the subreddits shown on the Reddit website. Instead, the author of the dataset builds a taxonomy of subreddits --- classifying each subreddit into categories, and subcategories since classifying into the original subreddits is often not feasible due to massive overlap between the topics of different subreddits. As the author of this dataset has already organized the last 20% of the data as a random, stratified sample of all the data, we will split out the first 80% of the data as a train set and the last 20% as a test set.

Accuracy is our most important audit metric. After testing data using three algorithms XGBoost, Naive Bayes and Hugging Face, we will get predicted subreddits, compare them with the original subreddits and generate accuracy scores by using the confusion matrix. The algorithm with the highest accuracy score will be chosen.

The main difficulty that may appear in our project is that the dataset is so large that it may take a long time to train. Most of our computer CPUs are i5 processors. It will take us a long time to process some data over 100MB. In this project, we will be dealing with a database of 1000MB or more. This data also contains 1000s of subreddits, which will further increase the strain on our computer CPU.

As for the final result, We will present it in a full academic report. For now, we only have a very rudimentary idea, there may be constant questions and changes when we utilize NLP. The academic report will be a very clear representation of our whole project running process.