News Analysis

Team Name: Three Bunnies
Members: Aihan Yang, Qiuyu Mei, Zhenghao Yu

Motivating Background

With the development of network technology, online news is becoming an indispensable part of people's daily life. How to efficiently analyze huge volume of news becomes a problem. In our project, we aim to apply advanced technology to do news text analysis.

Problem Statement

Based on the news content and tag information on each news portal, we try to divide the news into exact categories so that readers can accurately find the news of their favorite category and view the relevant content. When news can be accurately classified, readers can save a lot of time browsing all the news and can spend a limited amount of time reading the content they are interested in. In short, we help readers quickly find all the news content of the sections that attract them.

For whom are you solving this problem?

We are solving this problem for News Website and News Application editors to reduce their workload. Ultimately, our solution will impact the readers of the News Website and News Application, therefore they could find the news, which they are interested easily and clearly.

Why is this problem challenging?

The content of daily news is usually very messy and it would be impossible for us to manually classify the news. Also, if we use machine algorithms, the garbage vocabulary in the news text will affect the judgment of the computer on the classification task. Meanwhile, how to convert messy text data into data that can be used by computers will be the most challenging part of this project.

Datasets and Evaluation

We call Alibaba Cloud's API interface to query popular headlines and apply some preprocessing like text format conversion. According to the news contents and the category labels, we divide the data into training dataset, validation dataset and test dataset. After training the model with the training dataset, the model accuracy is measured by precision, recall and F-1 score. The model is ultimately applied to predict the category labels for subsequent news. Finally we will do results visualization.