Inferring Popularity of WikiHow Articles

Nikhil Kokra, Qingyun Zeng, Li Zhang

November 2020

1 Introduction

`wikihow.com` is a website consisting of more than 110,000 professionally-edited how-to articles spanning a surprisingly wide range of domains. While there are many resources with a large audience, not all wikiHow articles are received equally well. Some articles have millions of views and extremely high rating, while some are rather invisible or rated abysmally.

The goal of our project is to machine learning and natural language processing technologies to investigate what are possible factors for the popularity of articles, and whether a model can accuracte predict such popularity.

2 Data sources

We will be using the wikiHow corpus [1], which is a result of scraping and processing wikiHow and contains 112111 how-to articles. Each article is stored as a JSON file, with all essential information such as title, sections, steps, ratings, views, and so on.

3 Project plan

We want to use the wikiHow corpus to study the rating of different wikiHow articles. The goal is to learn which wikiHow article get higher rating and more views. If time permits, we may try some more interesting work like generating the answers in the Q&A secitions. We will first try using some basic machine learning techniques like linear regression/logistic regression/decision trees as a baseline, and then use state-of-the-art NLP techniques and neural networks. For task assignments, Li Zhang will do the data processing, Qingyun Zeng and Nikhil Kokra will build the machine learning models, and Li Zhang will further assist the tate-of-the-art NLP techniques and provide references.

4 Interesting Elements

The most interesting part of this project will be the feature extraction, since we will try to use state-of-the-art NLP techniques that only require raw text input from the article. We hope to come up with interesting visualizations for these features before feeding them into our predictive model as well.

5 Anticipated Challenges

Some challenges we expect to face include data cleaning. The wikihow website is notoriously malformatted, in that not all articles have the same features. For example, the QA section is present in some but not all. Beyond cleaning the data, actually implementing the NLP techniques we hope to use will be complicated, especially in Spark, because they are new and there aren't many existing libraries that will help to implement them.

References

[1] Li Zhang, Qing Lyu, and Chris Callison-Burch. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online, November 2020. Association for Computational Linguistics.