WikiClassify

Brian Chu
github.com/

Brian Faure
github.com/bfaure

Nathan Kjer
github.com/nathankjer

Adam Massoud
github.com/amassoud20

Viswanatha Subramanian
github.com/

Wanshu Sun
github.com/waynesun95

Luke Wielgus
github.com/lukewielgus

Project Description

Problem Domain

Wikipedia is a crowdsourced online encyclopedia founded in 2001. With the goal of centralizing all of human knowledge, it attempts to be a neutral, global, and uncensored source of free information. As of 2016, Wikipedia is ranked number seven on the worlds most popular websites, holds a top priority on search engines, and h as been estimated to be worth hundreds of billions of dollars. Furthermore, articles on Wikipedia h ave been cited in hundreds of court cases. Reliable or not, Wikipedia is heavily used as a result of its highquality content across a broad range of subjects.

Anyone can edit Wikipedia, which is the primary reason why it is one of the largest online encyclopedias in the world. However, this has numerous tradeoffs. Destructive editing is a common problem that can go unnoticed for extended periods of time. In addition, articles are frequently cited with unreliable sources, or not cited at all. Furthermore, Wikipedia has editor bias; 90% of Wikipedians are male, 40% of the worlds population has access to the internet (Wikipedia can only be edited from a selection of 40% of the worlds population), editors are often from white collar backgrounds, have stronger technical abilities, are more frequently from the Northern Hemisphere, and common languages and Western culture are more heavily represented. With such an editor bias, there is a larger probability of the content being misrepresented or not being of acceptable quality.

Thus, despite being one of the most popular and heavily accessed online encyclopedias, the biggest issue with Wikipedia is that not all of its content may be of satisfactory quality.

Specific Problems

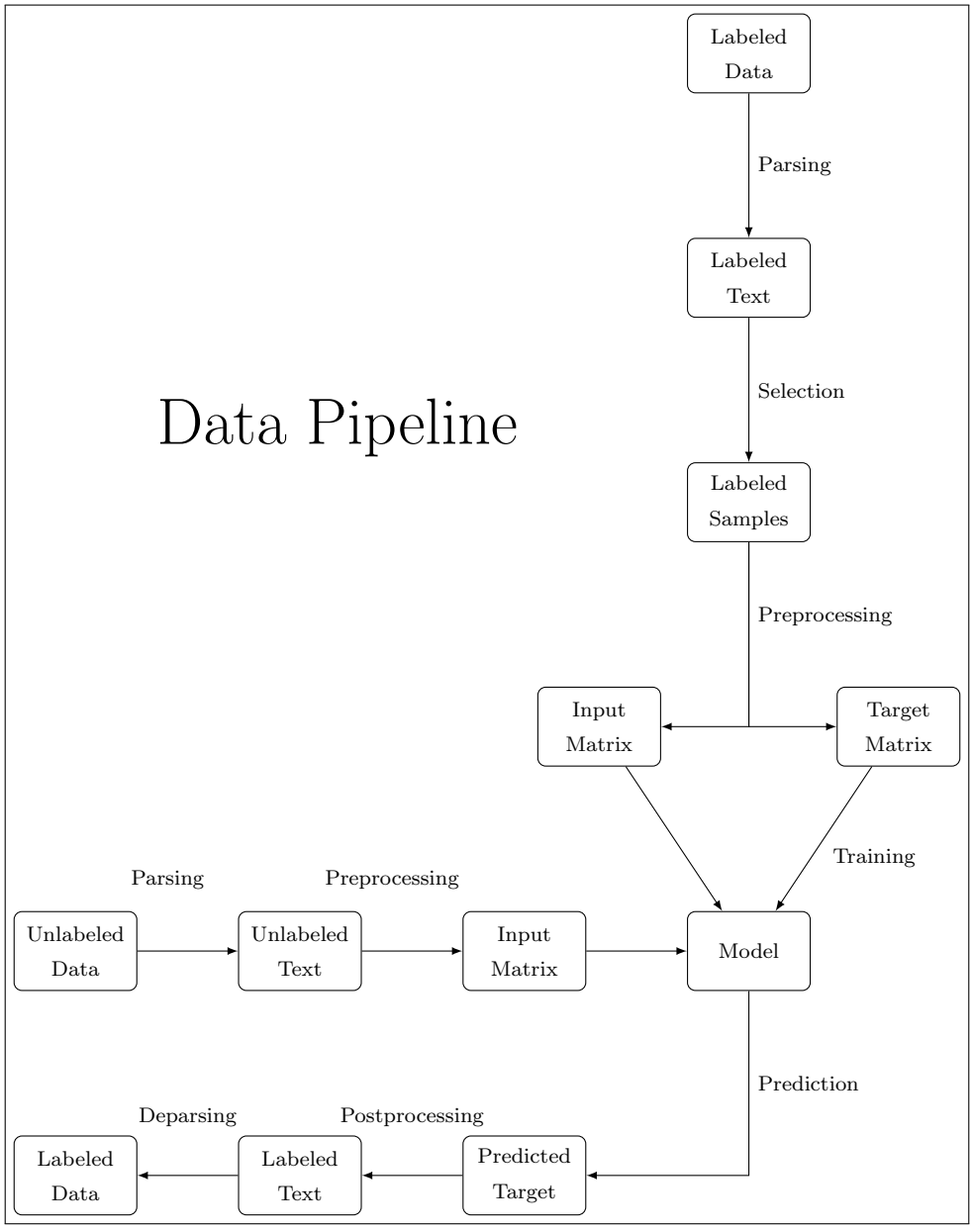
Wikipedia's quality is not consistent. It is too large for a human to proofread and too hard for a computer to edit. Therefore, a text classifier is desired so that content needing the most attention can be found. If there was to be some sort of software or system in place that could rate Wikipedias quality autonomously, a user could easily determine which Wikipedia articles have satisfactory quality for viewing. Thus, if this solution was successfully implemented, any user who would like to access Wikipedia can get high quality content. Furthermore, besides typical everyday Wikipedia viewers, Wikipedia itself could benefit from such a system, as they would have an easy solution to determining whether or not articles are of acceptable quality.

Problem Summary

- Input text and receive output parameters that classify the text based on our machine learning algorithm (classifiers here******).
- Uncover biases hidden in Wikipedia rating schema.

- Compare accuracy of multiple input text articles.
- Create procedurally generated articles on the fly.
- Check text for historical accuracy/bias.
- Compare different portions of Wikipedia based on article quality.

Plan of work



The figure above gives a general outline to the process that we will follow to achieve our solution. The following describes each iteration of the process in further detail.

Raw Data: The direct Wikipedia data dump. A monthly snapshot of Wikipedia is available from https://dumps.wikimedia.org/, saved as a gzipped tarball of an XML file.
Target Data: Classified text sequences that have been parsed from the XML file. The classifications are from articles containing tags such as those listed below.
Preprocessed Data: Classified text sequences (strings) are truncated into fixed lengths and converted into a numeric matrix format characterwise and preprocessed.
Transformed Data: Dimensions of the preprocessed matrix are reduced
Patterns: The machine learning will create a model that best reconstructs the targets from the inputs. This model will be used to predict values where targets do not yet exist. We will know our program has succeeded when the models cost, or error is sufficiently low.
Knowledge: The backtested results of classifications are visualized and delivered to the user for interpretation.
The table below holds the pairs we split into as well as our general starting tasks. The way that we wish to work on the project is more so of helping out on each part, with our starting points being the ones listed below.