Text Classification using NLP techniques

Olga Shiligin

13/02/2020

Today's machines can analyze more language-based data than humans, without fatigue and in a consistent, unbiased way. Considering the staggering amount of unstructured data, automation will be critical to fully analyze text efficiently. Natural Language Processing is important because it helps to resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as text analytics.

Problem

I propose to use the MS Research Project as an opportunity to solve Natural Language Processing task - identify duplicate questions using Quora data set. The project paper will focus on NLP feature engineering, utilising NLP techniques, applying Machine Learning algorithms, tuning them and selecting the best one based on their performance.

Quora is an American question-and-answer social network where questions are asked, answered, and edited by users, either factually or in the form of opinions. Quora is a place to gain and share knowledge. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. Millions of people use Quora every month, and many of them ask similar questions, so finding similar questions can significantly increase the efficiency of using this site.

I would like to make especial emphasis on feature engineering process for this particular project. The importance of feature engineering can hardly be overestimated for unstructured, textual data because we need to convert free flowing text into a numeric representation which can then be understood by machine learning algorithms.

In this project I will classify questions on similar and not similar (binaty classification problem). Because the correct answer would be associated with a group of questions regarded as similar the accuracy of predictions for Quora is quite important. To address this I need to set up high probability threshold.

Project Structure

The project includes following main parts: data pre-processing, feature engineering and modeling.

There will be multiple ways of cleaning and pre-processing the data: accented characters, expanding contractions, removing special characters, stemming, lemmatization, removing stopwords, tokenization etc.

Feature engineering

I expect that many derived columns will need to be created which will be numeric representation of the text data. Basic (Bag of Words, Bag of N-Grams, TF-IDF Model etc.) and advanced feature engineering strategies will be explored and applied for extracting meaningful features from text data. This process will create a large number of derived columns and then their importance will be evaluated.

Model building

Logistic Regression, SVM, XGBoost and Deep Learning models will be trained and tuned in order to solve this classification problem. The project will explore Python NLP and Deep Learning Libraries such as sklearn, nltk, re, spacy, textblob, Keras. Jupyter Notebook will be used for presentation purpose and the whole project will be stored on the github.

Data

For this project I propose to use Quora Question Pairs data set from Kaggle website.

Data fields:

id - the id of a training set question pair

qid1, qid2 - unique ids of each question (only available in train.csv)

question1, question2 - the full text of each question

is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

Data Source link:

https://www.kaggle.com/c/quora-question-pairs/data

References

[1] - S.Tong, Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research, 2011

[2] - M. IKONOMAKIS, S.KOTSIANTIS, Text Classification Using Machine Learning Techniques, University of Patras, Greece, 2005

[3] - X. Zhang, J.Zhao, Y.LeCun, Character-level Convolutional Networks for Text Classification, Courant Institute of Mathematical Sciences, New York University, 2015

[4] - M. Rogati, Y.Yang, High-performing feature selection for text classification, International conference on Information and knowledge management

[5] - K.Kowsari, D.Brown, Hierarchical Deep Learning for Text Classification, International Conference on Machine Learning and Applications (ICMLA), 2017