Customer Reviews: A Conditional Generative Model on the Basis of Clustered Text Data

Nils Gandlau, Jean Maccury, Fabian Ulmer
Pattern Analysis and Machine Intelligence
Goethe Universität Frankfurt

March 2, 2020

Abstract

The goal of generative models is to find patterns in data, which can ultimately be recreated in a generated output. Recurrent Neural Networks (RNN) implementing a Long Short-Term Memory (LSTM) architecture have gained increasing popularity in this area, specifically in learning from and generating new text in accordance with the input data. In this project, our aim is to increase the influence one can take on the output of such a model. To this end, working with the Amazon Customer Review Dataset, we start by using novel Natural Language Processing techniques for feature extraction purposes. The resulting data is then clustered by means of an unsupervised Random Forest approach. Finally, we build a conditional model, consisting of multiple RNNs whose output will depend on the resulting cluster. The goal is to create a model which generates appropriate customer reviews based on the input it receives and the corresponding cluster.

This project serves as a final group submission for the course "Pattern Analysis and Machine Learning" at Goethe-University Frankfurt.

Keywords— Natural Language Processing, Unsupervised Clustering, Recurrent Neural Network, LSTM, Text Generation, Customer Reviews

Contents

Chapter 1

Introduction

When building generative models, the goal generally lies in creating a satisfactory output according to the problem at hand. If, for instance, we try to generate Shakespearean poetry, we may seek to be able to fool a literary enthusiast with our results. While very good results have been achieved in this area, machine learning models are often restrictive with regard to the control one can have over the output. When given a starting sequence, the model completes it with the most appropriate (i.e. probable) output based on what the model has learned. But what if we want to condition our output on more than a starting sequence? Our approach aims to give a more fine-tuned control over the generated content of the network by conditioning our RNN on self-defined clusters, which will nudge the model towards the desired output.
Of course, we need to take into account that this type of task is often of unsupervised nature. The textual input data is not split in classes, it is rather a corpus of texts which adhere to a single class. This also applies to our case, where the data consists of a set of customer reviews for products.

Chapter 2

Data

The dataset that underlies our approach is the Amazon Customer Reviews Dataset [1], a dataset published by Amazon in which they have aggregated over 130M customer reviews with a set of features describing every single one. In the following part, we will take a look at the structure of this dataset, specifically the features that will be relevant to our project.

Chapter 3

Models Tools and Frameworks

3.1 General

In this section, we will describe the models, tools and frameworks we make use of in our project, starting with more general information and followed by details with regard to the different aspects our project can be broken into.

3.2 Text Analysis

3.3 Unsupervised Clustering

3.4 Text Generation

Chapter 4

Text Analysis

References

[1] Amazon. Amazon customer review dataset. URL: https://s3.amazonaws.com/amazon-reviews-pds/readme.html.