Springboard

Data Science Course: Capstone Project 2

Karolina Bzdusek

Opinosis Review Dataset

Source: http://kavita-ganesan.com/opinosis-opinion-dataset/#.XNnkDY5KiPo

This dataset contains sentences extracted from user reviews on a given topic. Example topics are “performance of Toyota Camry” and “sound quality of ipod nano”, etc. In total there are 51 such topics with each topic having approximately 100 sentences (on average). The reviews were obtained from various sources – Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics).

The general goal of a topic model is to produce interpretable document representations which can be used to discover the topics or structure in a collection of unlabelled documents.

The aim of this project is to learn techniques to determine these topics. It is useful to get a vital information from review (either what it is about, or whether user like it, etc.).

It is useful if you want to have an analysis of reviews at some service and you want to automate it.

We will use two (or three) techniques to determine these topics.

- Using TFIDF + PCA
- Word2vec
- (LSTM)

Our metric could be looking on False Positive rates/True Positive rates. We will look for other metrics as well and learn about them and decide which one is for us the best one.

The results will be presented via final report, with codes on GitHub and slide deck on GitHub as well.