[DATA-10P] Mining car reviews

Miguel-Angel Canela, IESE Business School

October 7, 2016

Introduction

Marketing managers of the automotive industry collect information about the customers' feelings from different sources. One of them is the customer reviews that are available in the Internet. CarReview (`www.carreview.com`) is a website where we can find information on car models, including reviews by customers and experts.

Reviews posted by users in CarReview come in three parts: Strengths, Weaknesses and Summary. The users also give to the model reviewed an Overall Rating and a Value Rating, in a 1-5 scale. These reviews provide a very useful information to sentiment analysis experts. These experts use lists of positive and negative words, whose occurrences they count in order to produce a polarity measure. Various lists of positive and negative words are available, and have been used for research in many fields, such as finance and marketing, but a specific list for a particular or segment is prefered.

An interesting question, related to the search for "signed" words, is the extent to which the customer ratings can be predicted from the occurrence of these words. The rating itself can be tried, or a binarized form can be preferrable. For instance, RATING = 5 can be taken as a definition of good rating, or RATING < 3 as one of bad rating.

The data sets

You are provided with two data sets, extracted from the reviews posted by users in CarReview, in the Luxury segment. The file `reviews.txt` contains the Summary part of the reviews and the file `ratings.csv` contains the name of the car model and the two ratings.