Toy to Enjoy:

An Exploratory Exercise on Text Analytics

DSP Research Project

July 3rd, 2018

Marleen van Kalmthout

Flavio de Oliveira

Jose Munoz

EVALUESERVE

POWERED BY MIND+MACHINE

Project Context

Context

You are working for a big store called “Toy to Enjoy” (TE) that sells baby and toddler toys. On the website and on the Facebook site of TE, customers are able to leave reviews on the products TE is selling. Unfortunately, something went wrong on the development side, and now only the comments of the reviews are visible, the overall ratings (1 star to 5 stars) are missing.

Your idea

Luckily, you are a driven data scientist with a brilliant idea that you share with your manager. We can scrape the reviews of our products from Amazon.com and use text mining techniques to categorize the comments from 1 to 5 stars. We can plot this model on the reviews on our website and we’ll have the problem solved!”

Your manager is really enthusiastic about this approach and asks you to start working on the program.

Deliverables

The deliverables of the project are:

- Categorized reviews (from star 1 to star 5)
- A presentation to show your approach and the difficulties you found
- Next steps / learnings / recommendations

Methodology

Scraping

The first step would be to build a web scraper to scrape our own data from amazon.com

Analyzing

The scraped data will be used for supervized learning (classification). You will create a train- and a testset and try both SVM as KNN models to see what model works best.

Our reviews

Once you have figured out the best model, it will be plotted on the Reviews from the website of Toy to Enjoy and you will present the results to the Marketing Manager

The team approach

One sub team will be working on building the scraper and scraping the website

In the meantime another sub team is using dummy data to write the query for the data preparation

While the third sub team is reading about the models and finding out how to query the models in R

The sub team that finishes first starts working on the presentation

To get an awesome team result in the end!

References

Some guidelines on HTML structure and web scraping:

- https://www.scrapehero.com/a-beginners-guide-to-web-scraping-part-1-the-basics/
- https://www.codingwithmax.com/blog/guide-to-web-scraping-walk-through
- https://www.w3schools.com/html/
- https://justrthings.com/2016/08/17/web-scraping-and-sentiment-analysis-of-amazon-reviews/

Prepare the dataset:

- https://machinelearningmastery.com/gentle-introduction-bag-words-model/
- https://www.rdocumentation.org/packages/tm/versions/0.7-3/topics/TermDocumentMatrix

Information on SVM models and KNN models:

- https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners
- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
- https://www.datacamp.com/courses/supervised-learning-in-r-classification (you can create a free account in Datacamp and folow this course, it is very descriptive and easy to understand!)

Presentation:

- https://blog.prezi.com/10-tips-for-making-a-persuasive-presentation/
- https://www.inc.com/geoffrey-james/7-ways-to-make-presentations-more-convincing.html

Thank you!