Amazon Review Helpfulness Classification

Word2Vec and logistic regression

Alex Morris
Viktor Shaumann

Objective

Predict usefulness measure for Amazon reviews

Top Customer Reviews

★★★★★ **No more winning for you, Mr. Banana!**

By SW3K on March 3, 2011

Size: 10â€ | Item Package Quantity: 1

For decades I have been trying to come up with an ideal way to slice a banana. "Use a knife!" they say. Well...my parole officer won't allow me to be around knives. "Shoot it with a gun!" Background check...HELLO! I had to resort to carefully attempt to slice those bananas with my bare hands. 99.9% of the time, I would get so frustrated that I just ended up squishing the fruit in my hands and throwing it against the wall in anger. Then, after a fit of banana-induced rage, my parole officer introduced me to this kitchen marvel and my life was changed. No longer consumed by seething anger and animosity towards thick-skinned yellow fruit, I was able to concentrate on my love of theatre and am writing a musical play about two lovers from rival gangs that just try to make it in the world. I think I'll call it South Side Story.

Banana slicer...thanks to you, I see greatness on the horizon.

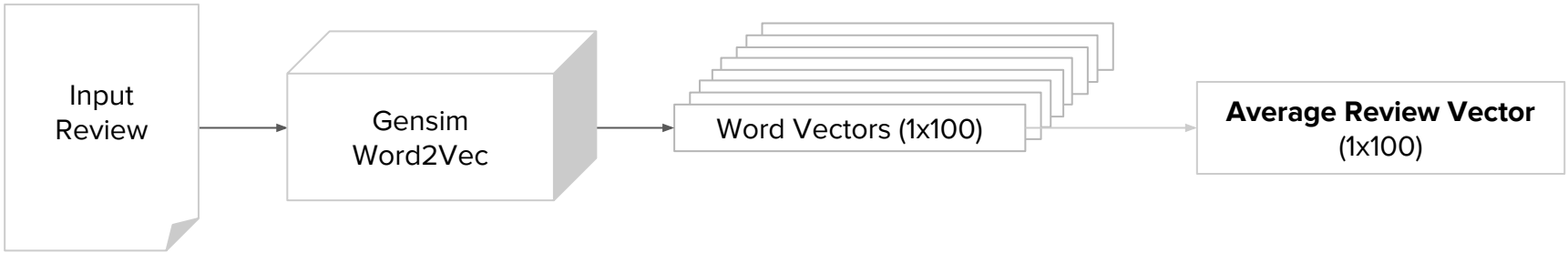
477 Comments | 55,013 of 55,803 people found this helpful. Was this review helpful to you? Yes No Report abuse

Dataset

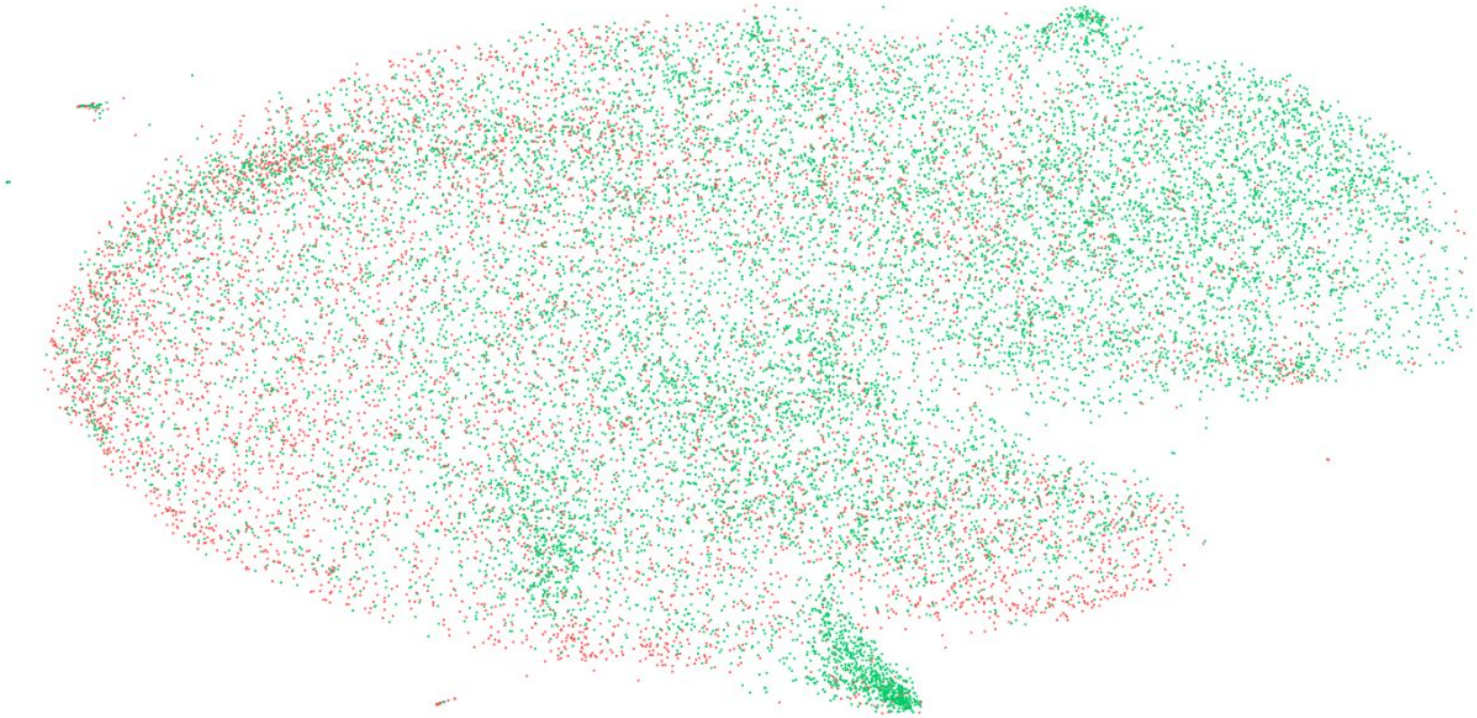
- Movies and TV subset
- 4.6 Million Amazon product reviews
- 3.6GB
- Word2Vec
  - Gensim library
  - Heavily C optimised

Word2Vec

- “Shallow” neural network
- Maximize the conditional probability of context given the word
- Every word is mapped to n-dimensional feature vector
- Trade off: Model complexity for bigger dataset



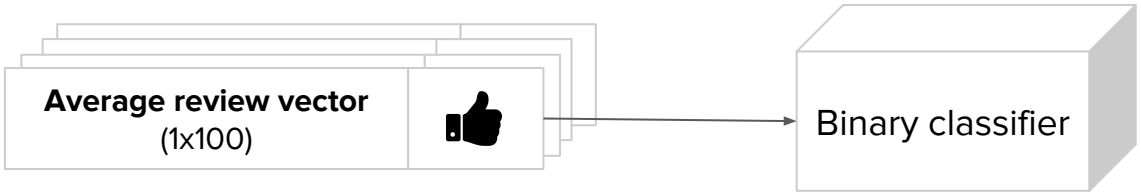
t-SNE: Review Vectors Reduced to a 2 Dimensional Space



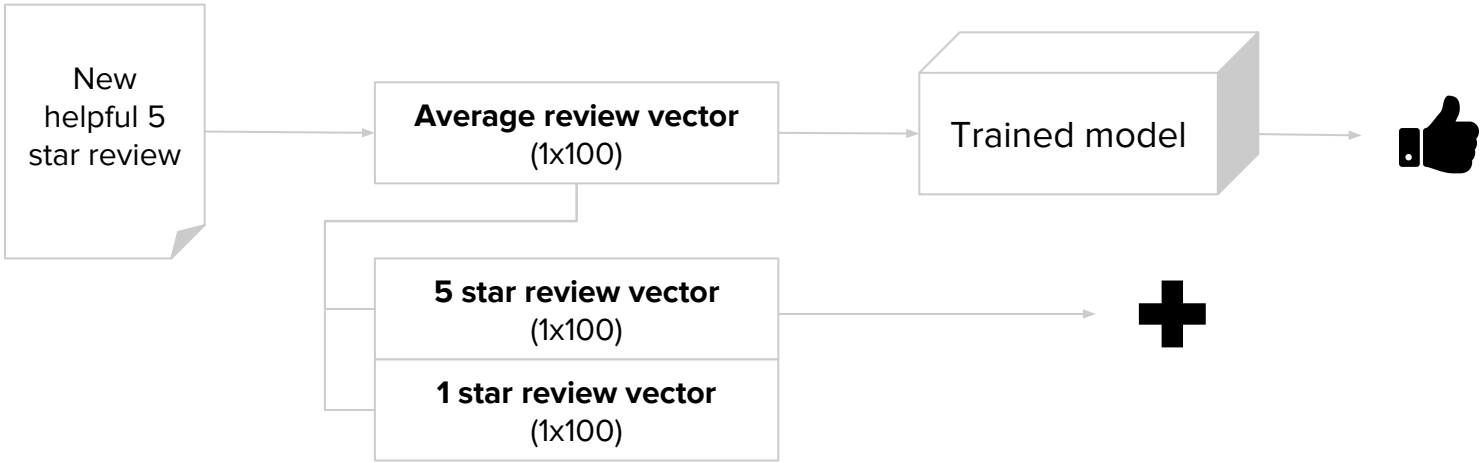
Overview

1. Helpfulness classification
    a. Random forest
    b. Linear SVM
    c. Logistic regression
2. Sentiment classification with cosine distance
    a. Compare to average review of all extreme reviews (1 or 5 star)

Training Our Model



Classification



Future

- Increase Word2Vec vector size
- Use Doc2Vec
- Balance data / adjust logistic regression intercept
- Train SVM with using a kernel function
- Regression
- Use entire dataset

Demo