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

- 82.5 Million Amazon Product Reviews
- 54.2 GB in total
- PySpark on EMR
- MLlib Word2Vec

Dataset

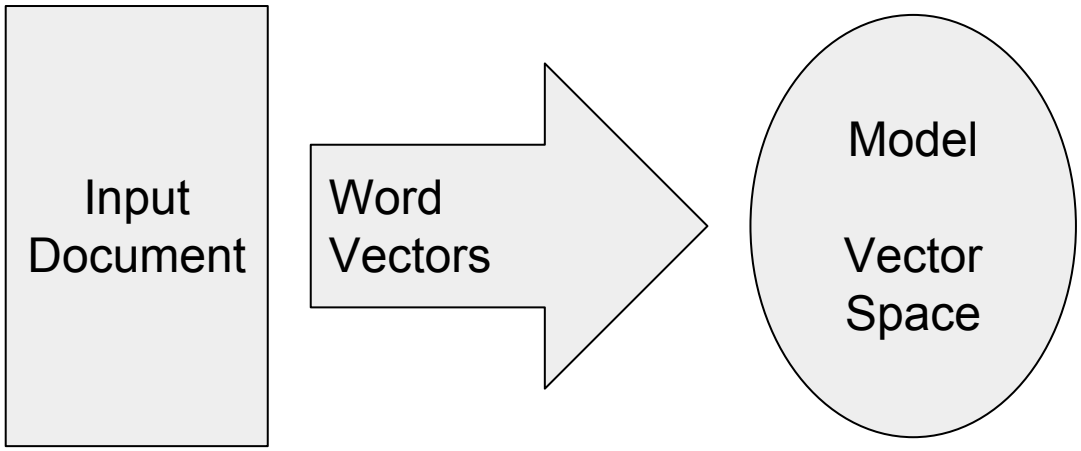
- Movies and TV Subset
- 4.6 Million Amazon Product Reviews
- 3.6 GB
- Gensim Library

Overview

1. Extract Features with Word2Vec
2. Helpfulness Classification with Logistic Regression / Random Forest
3. Sentiment Classification with Logistic Regression
4. Flask Application

Word2Vec

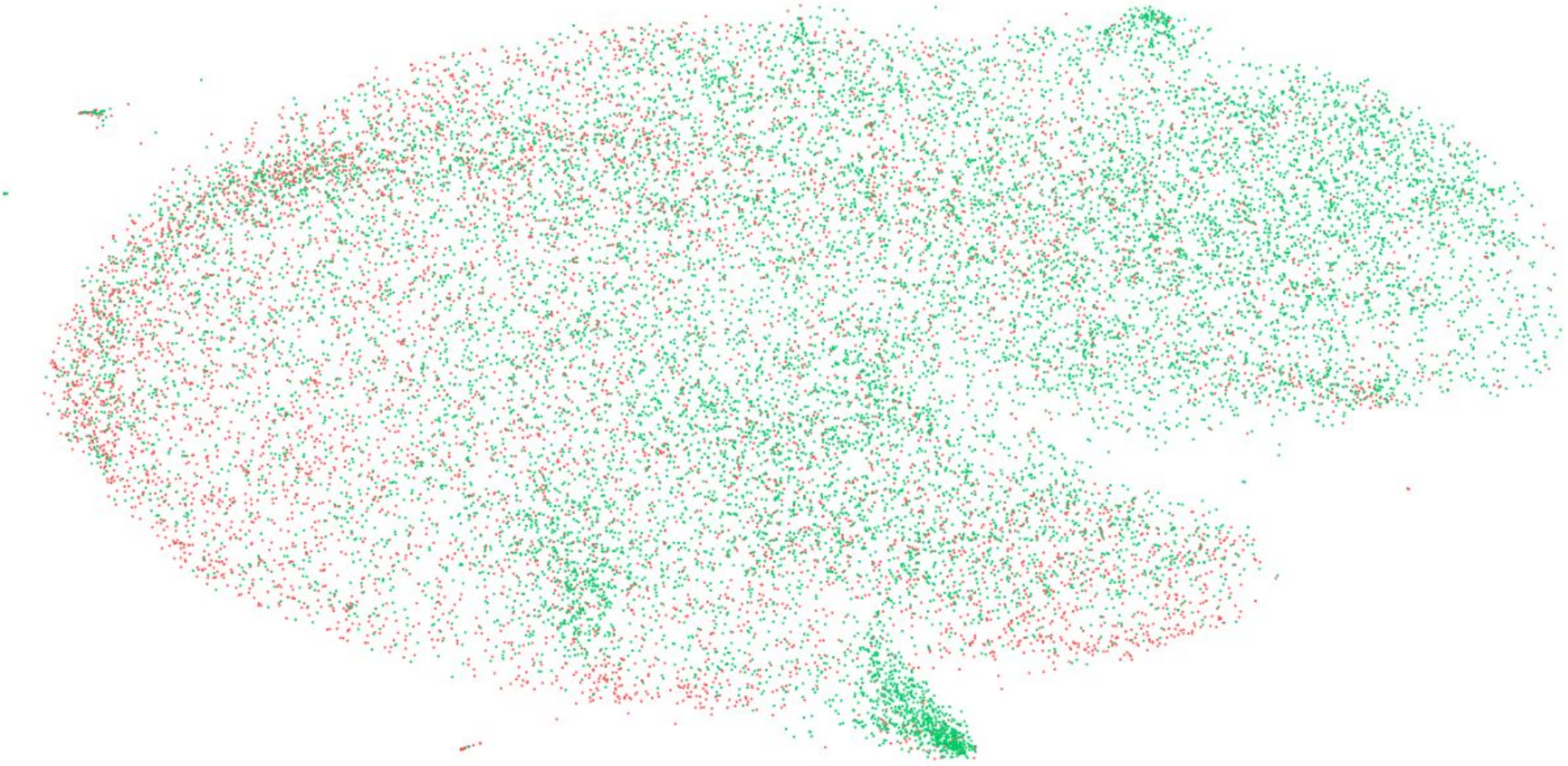
- Topic Modeling
- Every Word is Mapped to N-Dimensional Feature Vector
- Trade Off: Model Complexity for Bigger Dataset



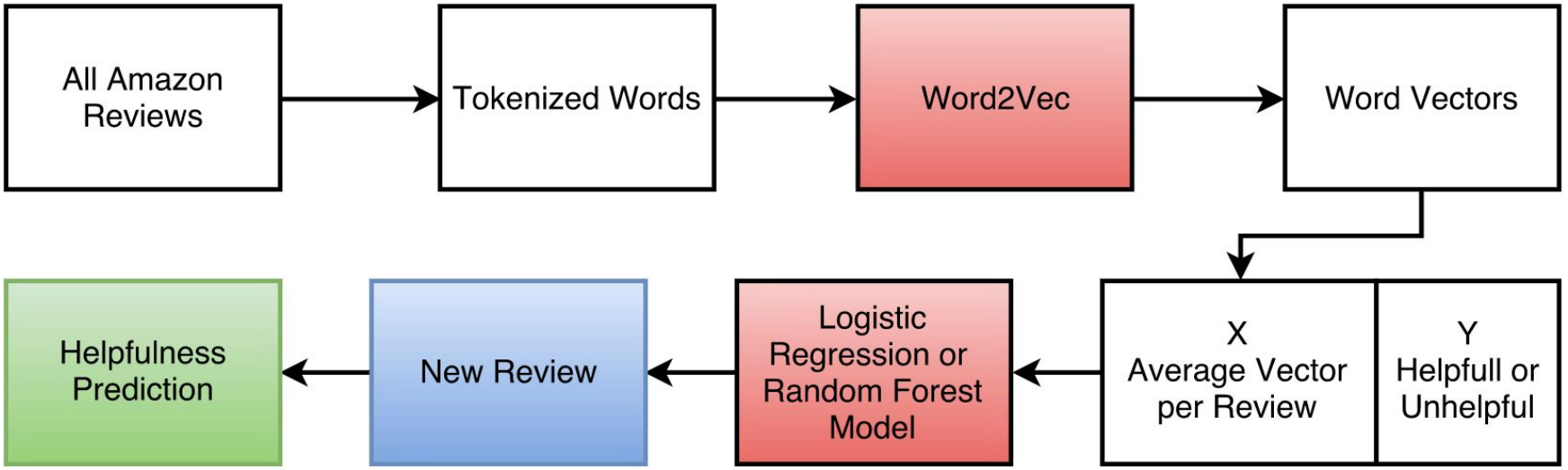

```
In [20]: w2vmodel.most_similar('funny')

Out[20]: [(u'hilarious', 0.8079836368560791),
          (u'humerous', 0.782819926738739),
          (u'humorous', 0.781631588935852),
          (u'hillarious', 0.7653717994689941),
          (u'comical', 0.7578990459442139),
          (u'cute', 0.7476027011871338),
          (u'hilarous', 0.7453163862228394),
          (u'funy', 0.7319284677505493),
          (u'halarious', 0.7270227670669556),
          (u'amusing', 0.723440408706665)]
```

t-SNE: 2D Vector Representation



Helpfulness Classification



To Do:

- Increase Word2Vec Vector Size
- Balance Data / Adjust Logistic Regression Intercept
- Run on the entire dataset

Flask App Demo