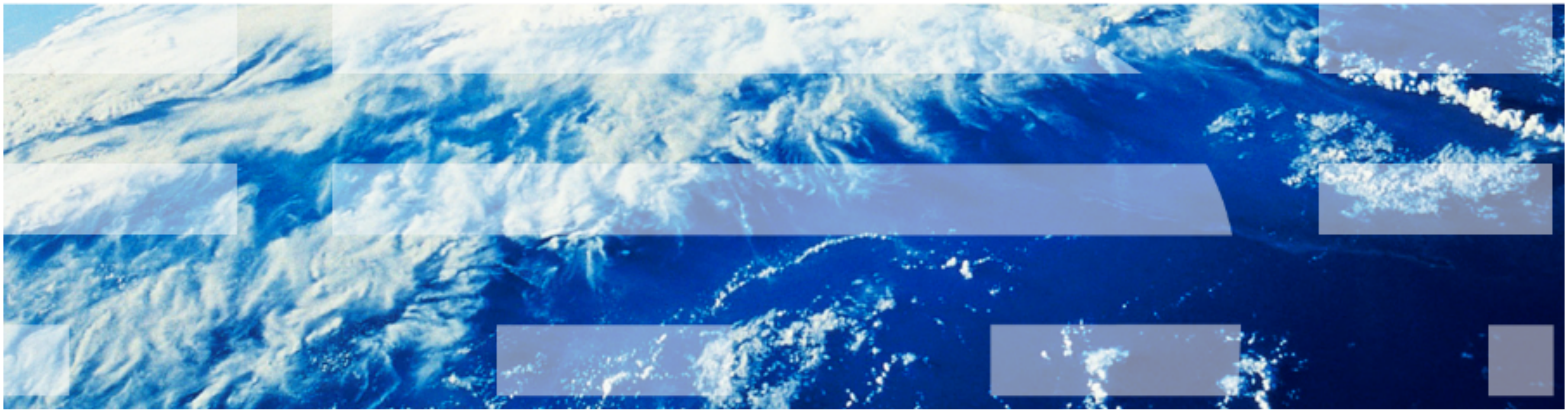
E6893 Big Data Analytics:

Amazon Co-purchasing Network Analysis and Prediction

Team Members:
Xinwei Li, Qi Chen, Ke Ma

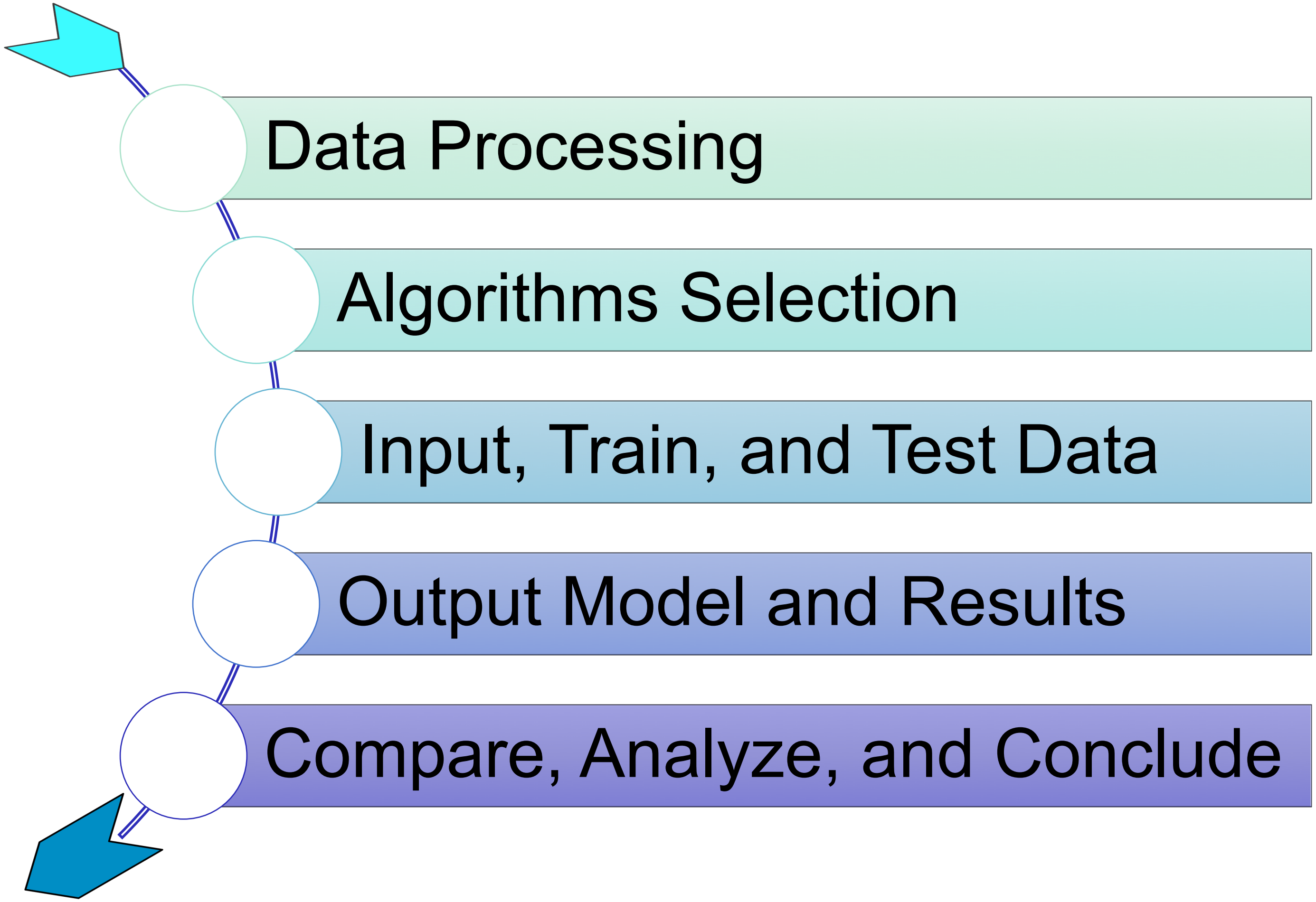


December 15, 2016

Overview

According to our analysis of Amazon Co-purchasing meta dataset, we build a model to evaluate the ‘similarity’ between two products (a pair of items). We calculate three correlation factors to describe the similarity relationship; then we combine them with co-purchasing history to generate the train/test data to be used for Naïve Bayes and Decision Tree classification. Finally, the model we build can predict if two products will be purchased together with 87% accuracy.

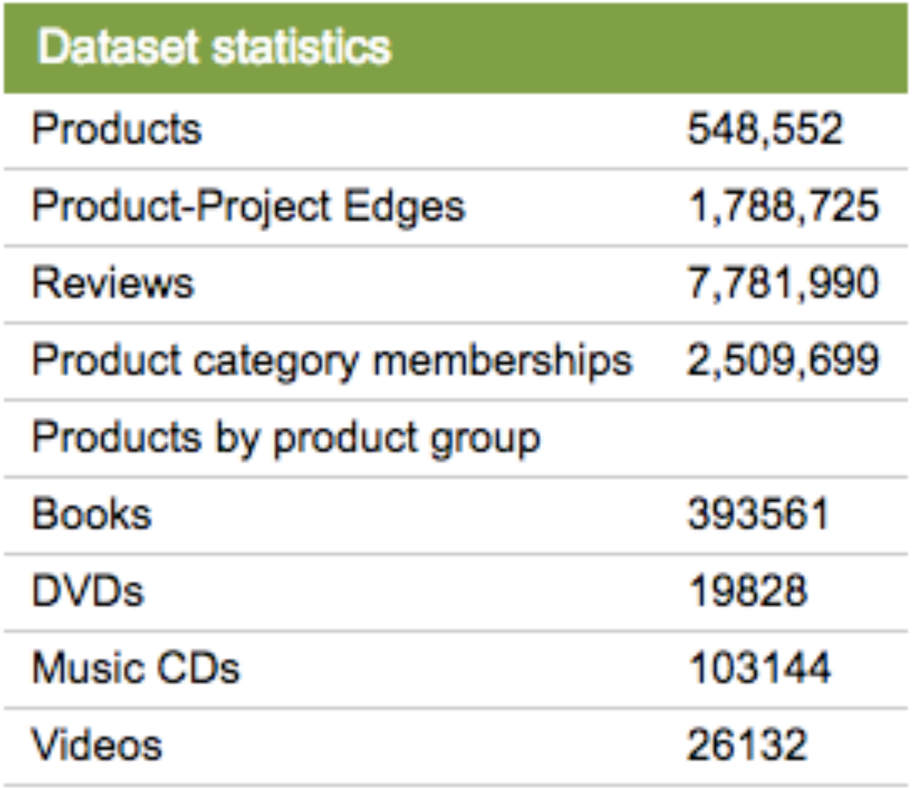
Steps



- Data Processing
- Algorithms Selection
- Input, Train, and Test Data
- Output Model and Results
- Compare, Analyze, and Conclude

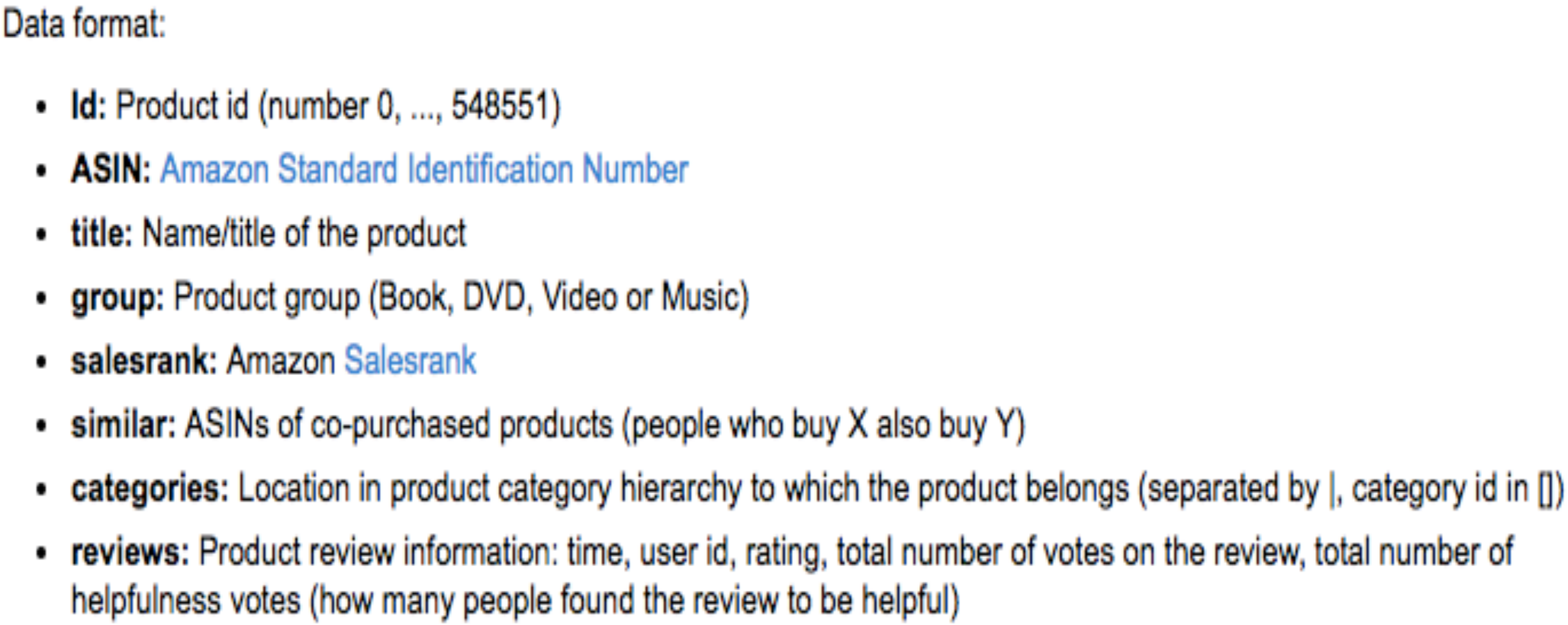
Dataset overview

Dataset: Amazon product co-purchasing network metadata
Source: Stanford Large Network Dataset Collection



Dataset statistics	
Products	548,552
Product-Project Edges	1,788,725
Reviews	7,781,990
Product category memberships	2,509,699
Products by product group	
Books	393561
DVDs	19828
Music CDs	103144
Videos	26132

Dataset overview



Data format:

- **Id:** Product id (number 0, ..., 548551)
- **ASIN:** Amazon Standard Identification Number
- **title:** Name/title of the product
- **group:** Product group (Book, DVD, Video or Music)
- **salesrank:** Amazon Salesrank
- **similar:** ASINs of co-purchased products (people who buy X also buy Y)
- **categories:** Location in product category hierarchy to which the product belongs (separated by |, category id in [])
- **reviews:** Product review information: time, user id, rating, total number of votes on the review, total number of helpfulness votes (how many people found the review to be helpful)

Data Processing

Obtain three factors to evaluate correlation between item A and item B.

Factor1: title similarity = $\frac{|wordsintitle(A) \cap wordsintitle(B)|}{|wordsintitle(A) \cup wordsintitle(B)|}$

Factor2: category similarity = $\frac{|category(A) \cap category(B)|}{|category(A) \cup category(B)|}$

Factor3: **Type I:**

Rating similarity = (rating(A) + rating(B))/2

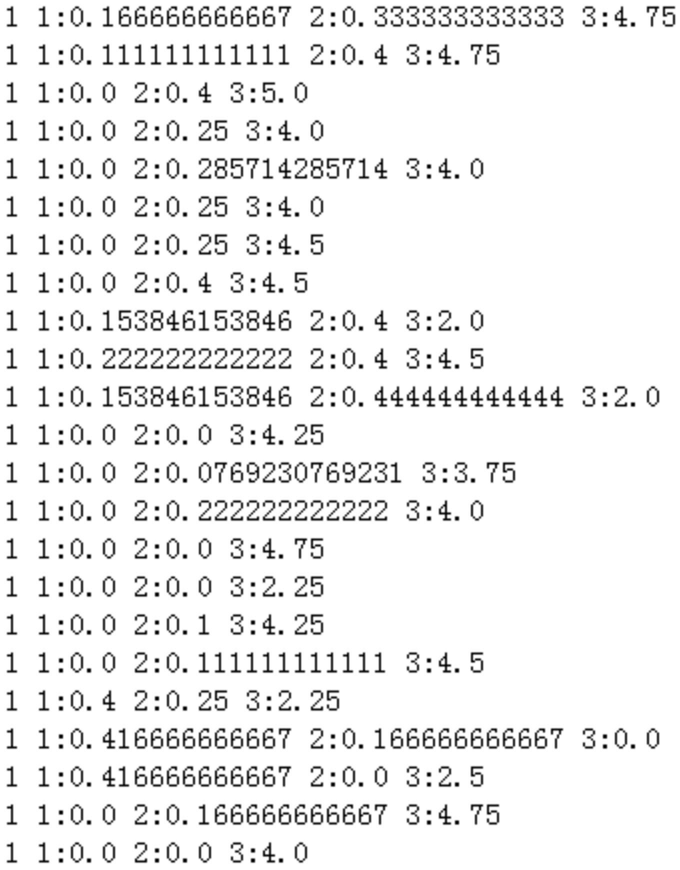
Type II

If (rating(A) >2.5 and rating(B) > 2.5) or
(rating(A) <2.5 and rating(B) < 2.5),
rating similarity = 1
Else: rating similarity = 0

Data Processing

Data format: Labeled point

Label index1:titile similarity index2:category similarity index3:rating similarity



```
1 1:0.166666666667 2:0.333333333333 3:4.75
1 1:0.111111111111 2:0.4 3:4.75
1 1:0.0 2:0.4 3:5.0
1 1:0.0 2:0.25 3:4.0
1 1:0.0 2:0.285714285714 3:4.0
1 1:0.0 2:0.25 3:4.0
1 1:0.0 2:0.25 3:4.5
1 1:0.0 2:0.4 3:4.5
1 1:0.153846153846 2:0.4 3:2.0
1 1:0.222222222222 2:0.4 3:4.5
1 1:0.153846153846 2:0.444444444444 3:2.0
1 1:0.0 2:0.0 3:4.25
1 1:0.0 2:0.0769230769231 3:3.75
1 1:0.0 2:0.222222222222 3:4.0
1 1:0.0 2:0.0 3:4.75
1 1:0.0 2:0.0 3:2.25
1 1:0.0 2:0.1 3:4.25
1 1:0.0 2:0.111111111111 3:4.5
1 1:0.4 2:0.25 3:2.25
1 1:0.416666666667 2:0.166666666667 3:0.0
1 1:0.416666666667 2:0.0 3:2.5
1 1:0.0 2:0.166666666667 3:4.75
1 1:0.0 2:0.0 3:4.0
```

Spark: algorithms

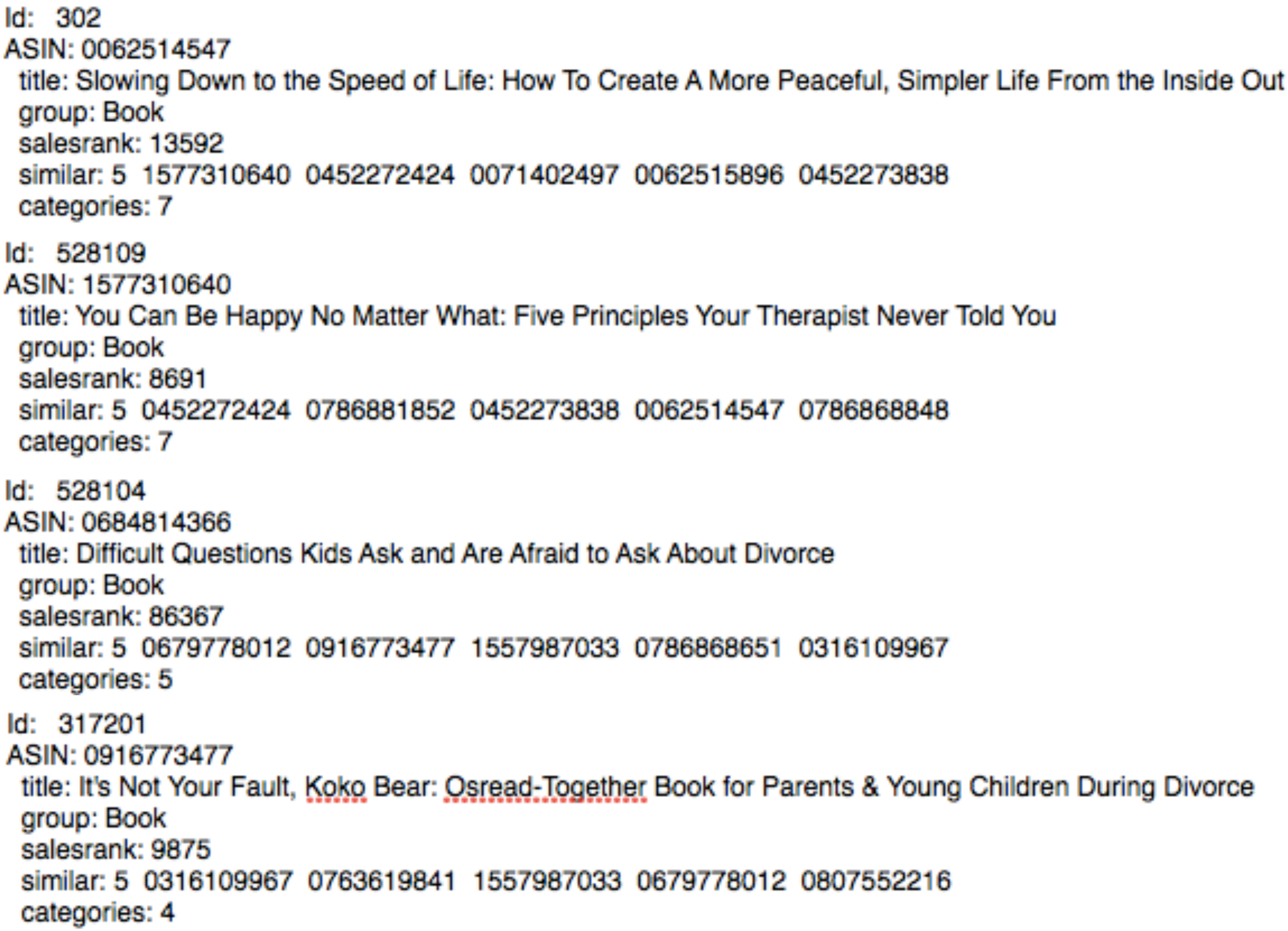
Classification:

1. Naïve Bayes

1. Decision Tree

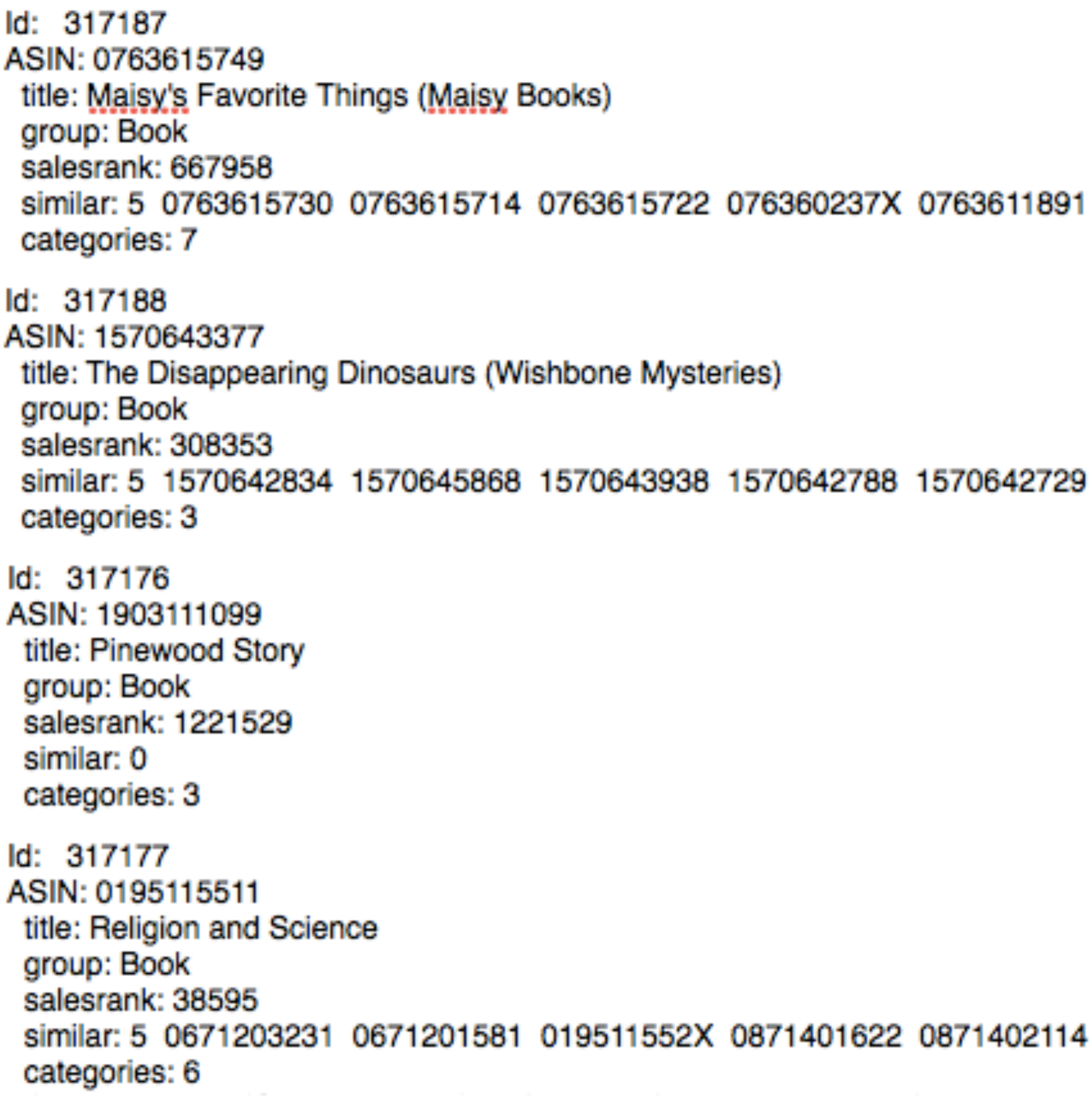
Demo

Step 1: we randomly select 2 pairs of items that are co-purchased and 2 pairs of items that are NOT co-purchased. Here is the sample information:



```
Id:   302
ASIN: 0062514547
  title: Slowing Down to the Speed of Life: How To Create A More Peaceful, Simpler Life From the Inside Out
  group: Book
  salesrank: 13592
  similar: 5  1577310640  0452272424  0071402497  0062515896  0452273838
  categories: 7

Id:   528109
ASIN: 1577310640
  title: You Can Be Happy No Matter What: Five Principles Your Therapist Never Told You
  group: Book
  salesrank: 8691
  similar: 5  0452272424  0786881852  0452273838  0062514547  0786868848
  categories: 7

Id:   528104
ASIN: 0684814366
  title: Difficult Questions Kids Ask and Are Afraid to Ask About Divorce
  group: Book
  salesrank: 86367
  similar: 5  0679778012  0916773477  1557987033  0786868651  0316109967
  categories: 5

Id:   317201
ASIN: 0916773477
  title: It's Not Your Fault, Koko Bear: Osread-Together Book for Parents & Young Children During Divorce
  group: Book
  salesrank: 9875
  similar: 5  0316109967  0763619841  1557987033  0679778012  0807552216
  categories: 4
```

Demo



```
Id:   317187
ASIN: 0763615749
  title: Maisy's Favorite Things (Maisy Books)
  group: Book
  salesrank: 667958
  similar: 5  0763615730  0763615714  0763615722  076360237X  0763611891
  categories: 7

Id:   317188
ASIN: 1570643377
  title: The Disappearing Dinosaurs (Wishbone Mysteries)
  group: Book
  salesrank: 308353
  similar: 5  1570642834  1570645868  1570643938  1570642788  1570642729
  categories: 3

Id:   317176
ASIN: 1903111099
  title: Pinewood Story
  group: Book
  salesrank: 1221529
  similar: 0
  categories: 3

Id:   317177
ASIN: 0195115511
  title: Religion and Science
  group: Book
  salesrank: 38595
  similar: 5  0671203231  0671201581  019511552X  0871401622  0871402114
  categories: 6
```

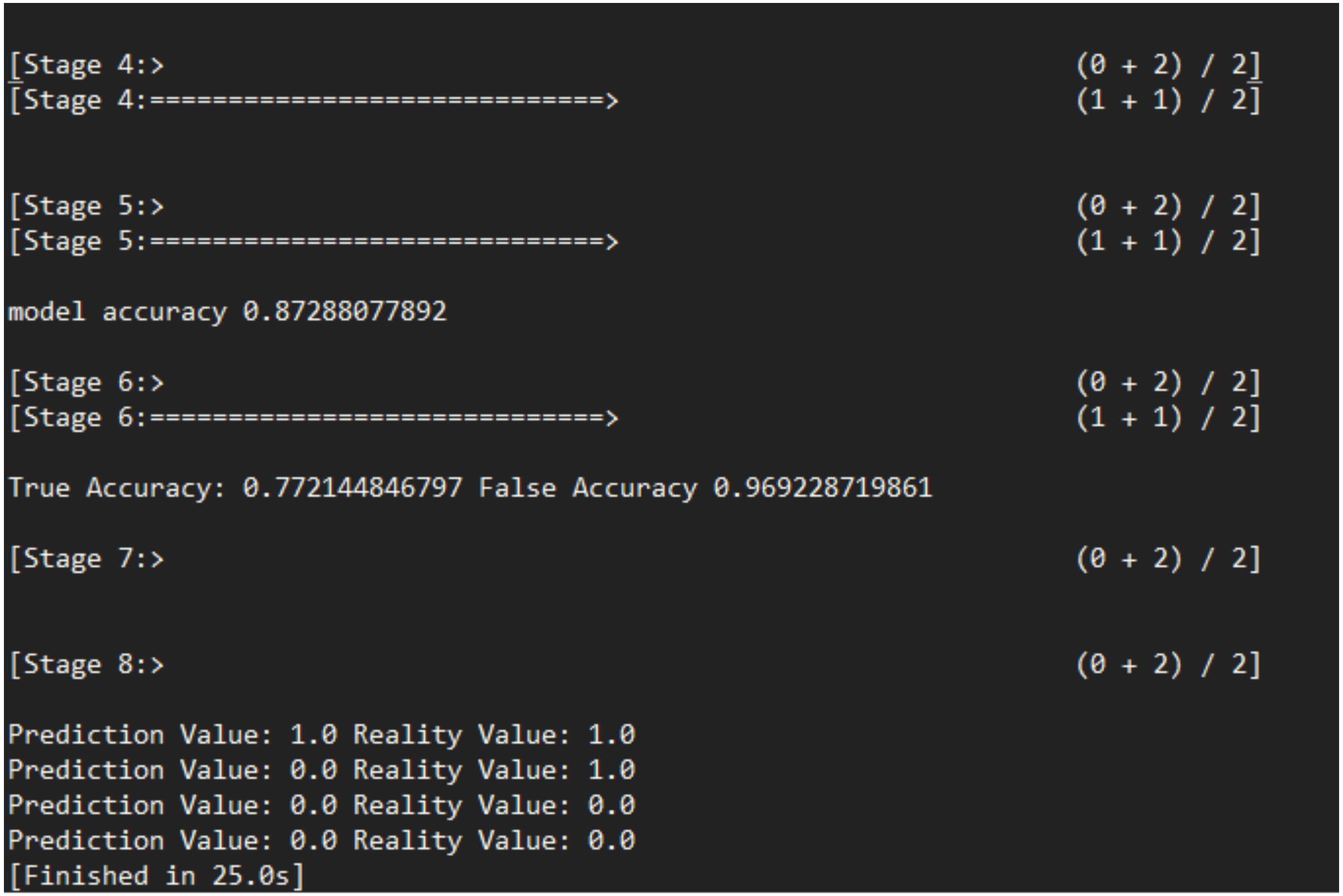
Demo

Step 2: We use our model to compute the correlation factors and parse the data file.

```
1 1:0 2:0.214285714286 3:1
1 1:0.037037037037 2:0 3:1
0 1:0 2:0 3:0
0 1:0 2:0 3:0
```

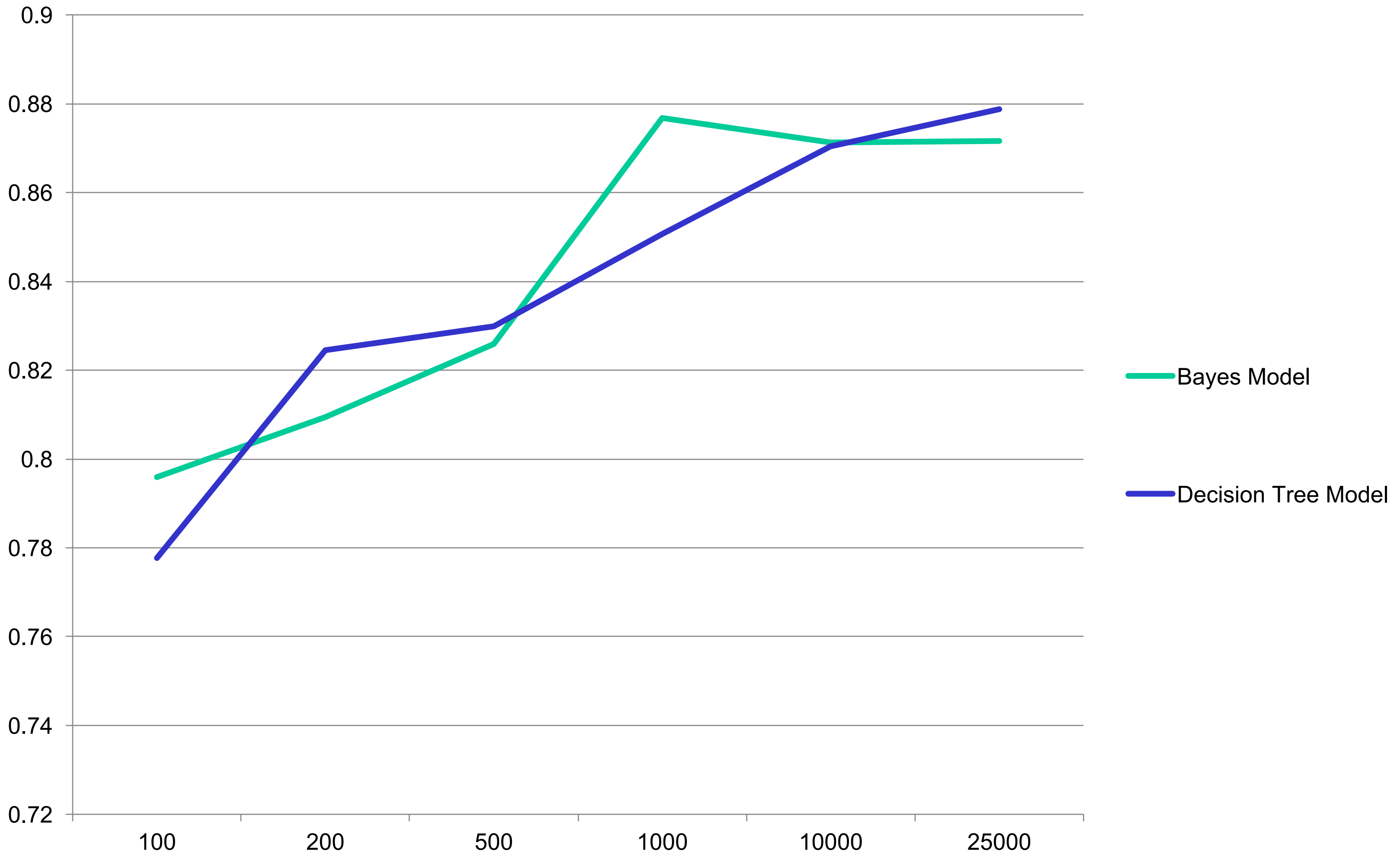
Demo

Step 3: we use the model we train to predict if these paired items will be co-purchased. Here are the sample results.



```
[Stage 4:>                                                          (0 + 2) / 2]
[Stage 4:=============================>                             (1 + 1) / 2]


[Stage 5:>                                                          (0 + 2) / 2]
[Stage 5:=============================>                             (1 + 1) / 2]

model accuracy 0.87288077892

[Stage 6:>                                                          (0 + 2) / 2]
[Stage 6:=============================>                             (1 + 1) / 2]

True Accuracy: 0.772144846797 False Accuracy 0.969228719861

[Stage 7:>                                                          (0 + 2) / 2]


[Stage 8:>                                                          (0 + 2) / 2]

Prediction Value: 1.0 Reality Value: 1.0
Prediction Value: 0.0 Reality Value: 1.0
Prediction Value: 0.0 Reality Value: 0.0
Prediction Value: 0.0 Reality Value: 0.0
[Finished in 25.0s]
```

Findings: Data size VS Accuracy



Findings: Model and Algorithms

Accuracy (Type I)	Naïve Bayes	Decision Tree
Total Accuracy	0.8629	0.8739
True Accuracy	0.7609	0.9695
False Accuracy	0.9608	0.8120

Accuracy (Type II)	Naïve Bayes	Decision Tree
Total Accuracy	0.8715	0.8747
True Accuracy	0.7694	0.9486
False Accuracy	0.9699	0.8238

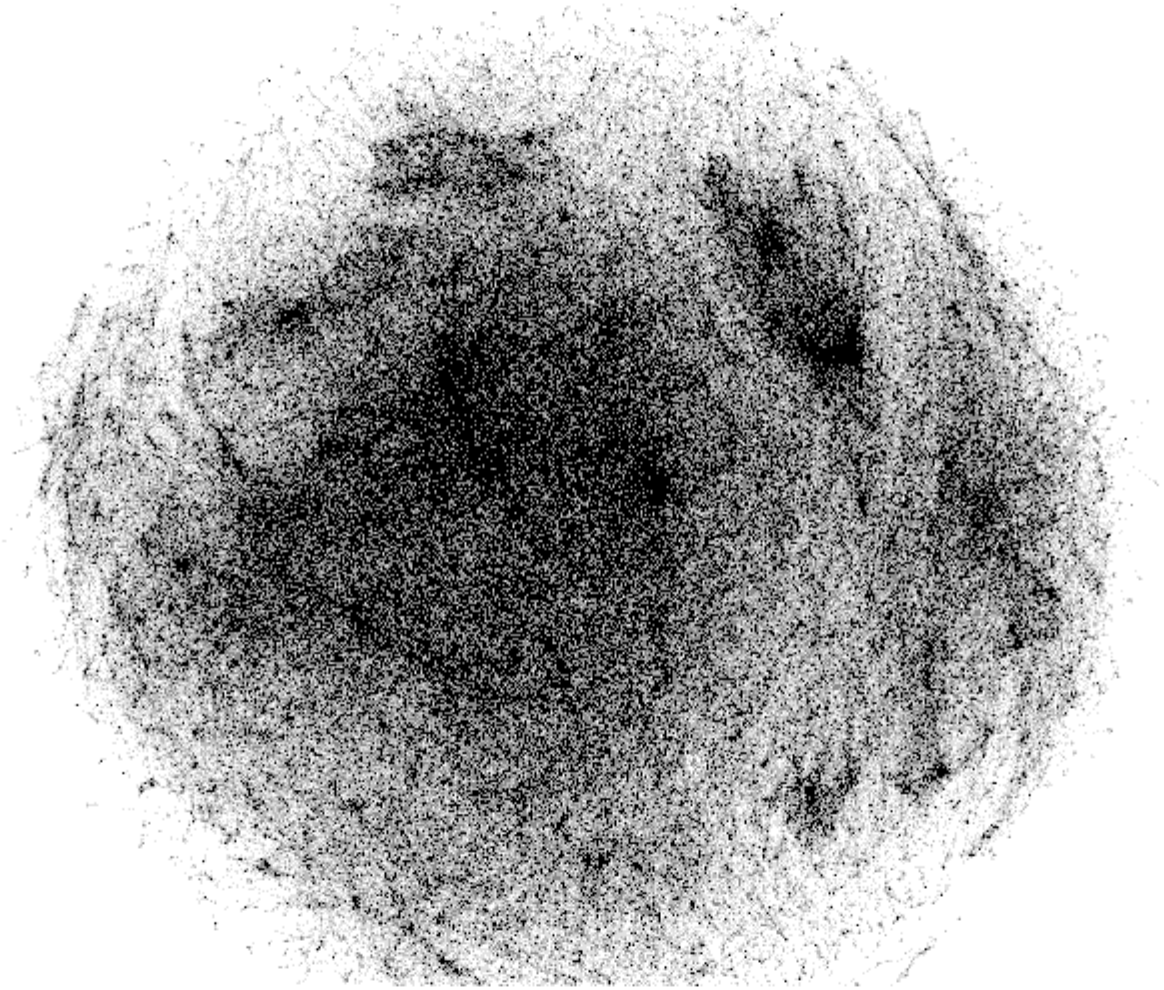
Findings: Importance of factors

Accuracy	Naïve Bayes	Decision Tree
Category Missing	0.6495	0.6605
Rate Missing	0.8236	0.8787
Title Missing	0.8562	0.8565

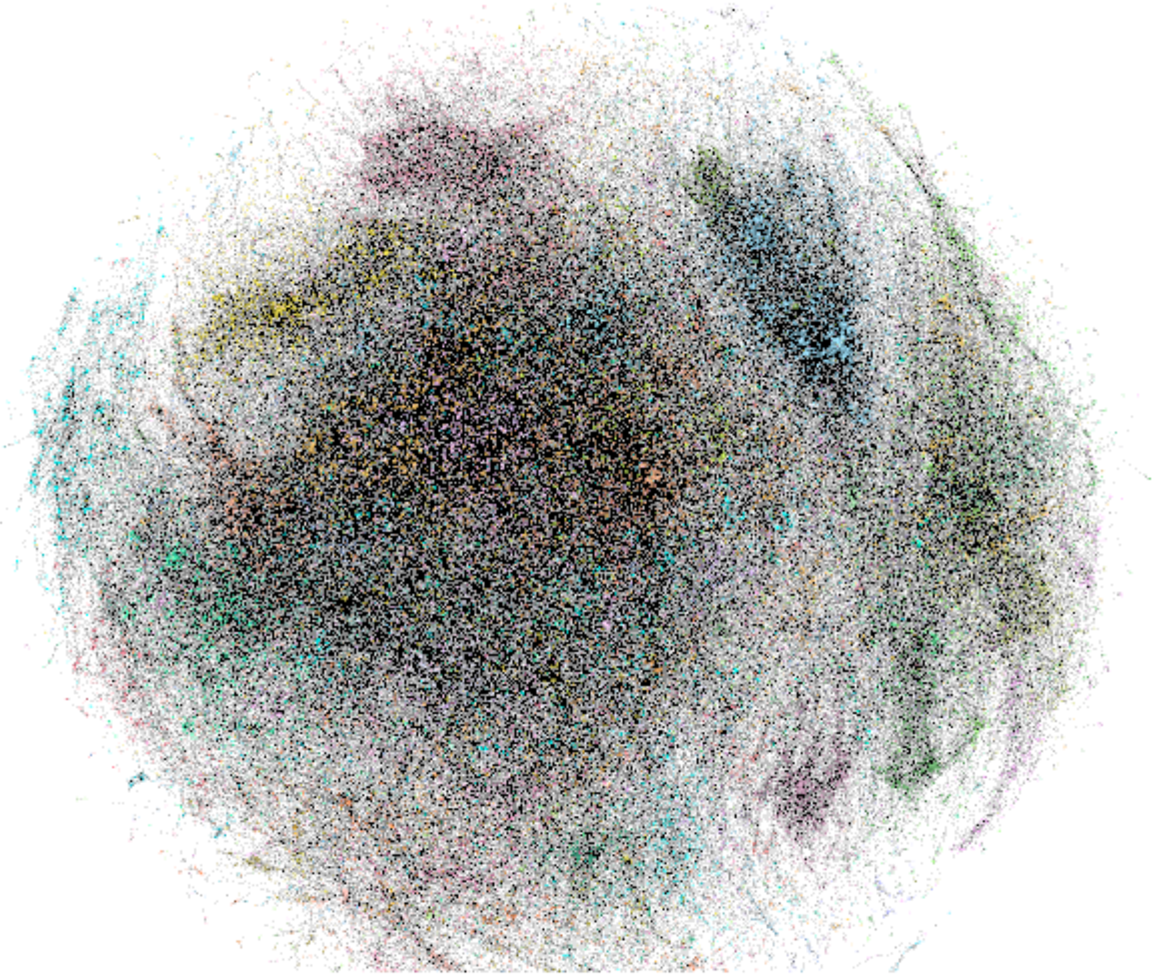
Conclusions

1. Title, rate, and category similarities of items are indeed factors that will influence customers when they have the co-purchasing options.
2. In our project, Naïve Bayes model and Decision Tree model both generate impressive performance.
3. The more data we use to train the model, the more accuracy we will achieve.
4. Category-similarity is the most powerful among the three factors.

Visualization: Cluster



Visualization: Cluster



Improvement

- Improve formula of factors
- Recommendation
- NLP