Word2Vec

Alex Clavel, Alex Ruchti

Problem Statement

Given an Amazon product title and description, what category would it be filed under?

e.g. Meat Slicer 200W Electric Deli Food Slicer with Removable 7.5’’ Stainless Steel Blade, Adjustable Thickness Meat Slicer for Home Use, Child Lock Protection, Easy to Clean, Cuts Meat, Bread and Cheese

Data Collection

http://jmcauley.ucsd.edu/data/amazon/links.html

9.4 M entries

Stored in JSON

Collected between 1996 - 2014

Data Preprocessing

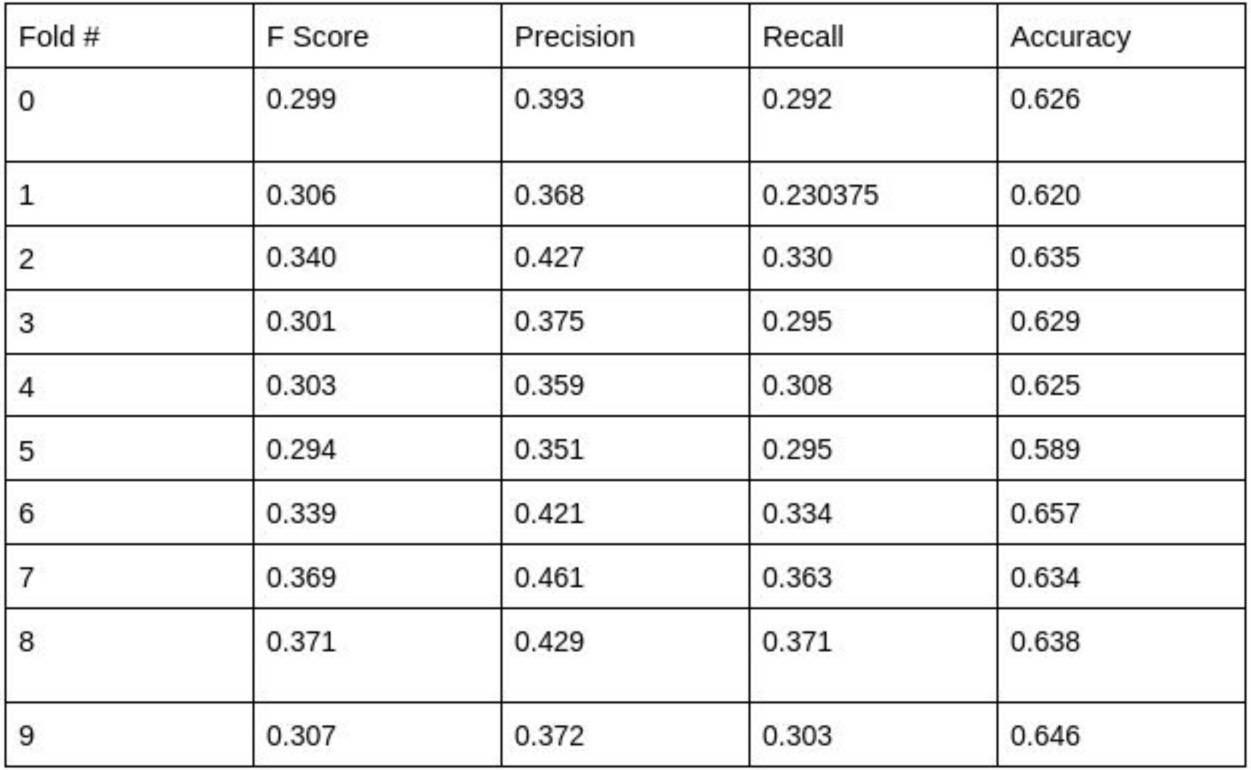
Cannot load all the data at once into ram (10gb)

Split the data into 10 parts

Clean each part individually

- Extract title, category, and description
- Convert to lowercase
- Remove Punctuation

Cross Fold Validation



Fold #	F Score	Precision	Recall	Accuracy
0	0.299	0.393	0.292	0.626
1	0.306	0.368	0.230375	0.620
2	0.340	0.427	0.330	0.635
3	0.301	0.375	0.295	0.629
4	0.303	0.359	0.308	0.625
5	0.294	0.351	0.295	0.589
6	0.339	0.421	0.334	0.657
7	0.369	0.461	0.363	0.634
8	0.371	0.429	0.371	0.638
9	0.307	0.372	0.303	0.646

Model Results

Average F-Score: 0.3229

Average Precision: 0.3956

Average Recall: 0.3121

Average Accuracy: 0.6299

These numbers are slightly worse than the out of the box solution, but are still very good

Next Steps

Issues in the implementation

- No word normalization
  - “Dog” and “Dogs” are seen as distinct
- Not enough descriptions
  - Most products are missing descriptions, and there is not much info in the title
- Computation issues
  - Very cpu intensive
  - KNN implementation uses a lot of memory

Lessons Learned

- Don’t use the whole data when making prototypes
    - Use a subset, this will allow for faster iteration
- People who write libraries know what they’re doing
    - There is no need to reinvent the wheel. Libraries are going to be easier to work with, and faster, most of the time
- Intermediate results take time
    - Do not ignore the time you need to make intermediate products and deliverables. Factor these in to your planning.

Questions?