

# Radius Collider Submission - The Approximators

Myles Scolnick, Shane Barratt and Alexander Danilychev Jr

## I. INTRODUCTION

The goal of this project was to determine the best North American Industry Classification System (NAICS) code for a business based on its name, address, description, and website. To solve this problem, we incorporated several techniques from a variety of fields: text cleaning, data augmentation, tf-idf [2], wordnet [3], neural word embeddings [4] and cross-validation. A pure supervised learner would be tough to model/train with less training data points than classes, so we went with what one could call a 'topic modeling' approach.

## II. DATA

To increase the amount of data on businesses and NAICS codes, we aggregated some data from public APIs:

- Google Places business type
- NAICS code titles
- NAICS code descriptions
- Frequency of NAICS 2-long industry codes

The respective scripts are listed in List 1 in the appendix.

We also created a function to remove stop words, lemmatize, lowercase, tokenize. We did this to normalize the text being compared. We also enhanced the richness of the text by adding synonyms obtained through the WordNet model.

## III. WEB APP

### A. Assisted Hand Classifier

To ease the painful process of hand-classification, we built a web application in Flask/Python that displays information about the business to be classified and a list of NAICS codes sorted by a preliminary algorithm's scoring function. This allowed more interactive classification - that is automatically written to a file. This greatly sped up hand-classifying 1,000 businesses. Refer to Figure 1 of the appendix.

### B. Database View and Code Comparison

To further understand the structure of the data and the quality of our classifier, we also built an additional endpoint to read the hand-classification and algorithm-classification databases. It displays a table of businesses with our hand-classification and algorithm-classification so that we could discern what we got it wrong and why. Refer to Figure 2 of the Appendix.

## IV. TECHNICAL MODELING DETAILS

Let the set of businesses be denoted by  $\{b_1, b_2, \dots, b_n\}$  and the set of NAICS codes by  $\{c_1, c_2, \dots, c_m\}$ . The algorithm attempts to construct a similarity matrix  $\mathbf{S}$  of dimension  $n \times m$  where  $\mathbf{S}_{i,j}$  denotes the similarity between business  $i$  and NAICS code  $j$ . The classification for business (row)  $i$  is  $\arg \max_{1 \leq j \leq m} \mathbf{S}_{i,j}$ , the most similar NAICS code to that

business. To construct this matrix, we first construct 9 similarity matrices,  $\{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_8\}$  with the same dimension. There are 4 ways to pair the business[.name .description] and NAICS[.title .description]. So the *tf-idf* cosine similarity and *word2vec* cosine similarity between these 4 combinations as well as a similarity matrix of 2-digit industry priors (independent of the business) constitute the 9 similarity matrices. The final similarity matrix is a convex combination of the preliminaries:

$$\mathbf{S} = \sum_{i=1}^8 w_i \mathbf{S}_i,$$

where  $\sum_{i=1}^8 w_i = 1$ . The weights were selected through random hyper-parameter search [1].

## V. CONCLUSION

### A. Results

After training, our algorithmic classifier scores 18.5 %. We hand-classified 1000 businesses.

### B. Confidence in our Predictions

- We are not particularly confident in the predictions our algorithm made, as it has lots of false-positives.

### C. Limitations

- The algorithm did not take into account the hierarchical structure of the NAICS codes.

### D. Challenges

- The data did not come perfectly formatted and sometimes the title, description, website did not actually match and was a mix of 2 businesses.
- Some of the businesses had names and descriptions in other languages.
- Some businesses have multiple obvious classifications making it difficult to assign a single NAICS code (e.g. a bakery which also is a sit-down restaurant).

## REFERENCES

- [1] Bergstra, James, and Yoshua Bengio. "Random Search for Hyperparameter Optimization." *The Journal of Machine Learning Research* 13.1 (2012): 281-305.
- [2] Brown, Peter F., et al. "A Statistical Approach to Machine Translation." *Computational Linguistics* 16.2 (1990): 79-85.
- [3] Miller, George A., et al. "Introduction to Wordnet: An On-line Lexical Database\*." *International Journal of Lexicography* 3.4 (1990): 235-244.
- [4] Mikolov, Tomas, et al. "Distributed Representations of Words and Phrases and their Compositionality." *Advances in Neural Information Processing Systems*. 2013.

APPENDIX

FIGURE 1

Name	Carolyn D Slawski CPA		
Type	accounting/finance		
Description	Carolyn D Slawski CPA PC is an entertainment company based out of 1757 Veterans Hwy Ste 60, Central Islip, New York, United States.		
Id	608b003b-41e3-48c1-b44a-9025e495ae5		
Website	http://www.carolynslawskipc.com		
Address	1757 VETERANS HWY STE 60, ISLANDIA NY 11749		
Code	Score		
Actual	541	27.4004441003	
Actual	561	18.1396537258	
Actual	333	17.3406181926	
Actual	423	15.7893421822	
Actual	311	14.7897617355	

Score	Code	Title	Description
Actual	0.934480297907	541211	Offices of Certified Public Accountants
Actual	0.858679797692	541219	Other Accounting Services
Actual	0.753702801839	541618	Other Management Consulting Services
Actual	0.750072904393	561499	All Other Business Support Services
Actual	0.746565862131	425110	Business to Business Electronic Markets
Actual	0.743015183456	541613	Marketing Consulting Services
Actual	0.721712458367	532420	Office Machinery and Equipment Rental and Leasing
Actual	0.710778290392	541120	Offices of Notaries
Actual	0.709972670687	541213	Tax Preparation Services
Actual	0.706452213122	926150	Regulation, Licensing, and Inspection of Miscellaneous Commercial Sectors

FIGURE 2

3	911111	811198	e4e09337-c084-42d6-a0f5-3aa8191a0498	Midas	store car repair	Midas is one of the world's largest providers of auto repair services, including brakes, oil changes, tires, maintenance, steering, and exhaust services. Visit your local Midas for additional services.	1.095384792558	811198	All Other Automotive Repair and Maintenance	333	26.579619758
6	441310	441310	b3b4b050-4a6b-45c7-9245-5b6413c9f335	CARQUEST Auto Parts	store car repair	CARQUEST Auto Parts is the premier supplier of replacement car parts, truck parts, products, accessories, supplies and equipment for virtually all makes and models.	1.02984011627	441310	Automotive Parts and Accessories Stores	333	25.330141832
-2	424460	114112	e95d5c98-389a-4870-a717-360194872056	Lansford Road Lobster & Fish		We are a retail fish market and have live or cooked lobsters. We have a huge selection of fresh fish. Located on picturesque Cape Poge State Harbor. We are open from May to the end of Sept. We ship lobsters overnight to UPS all year long.	0.758370702919	114112	Shellfish Fishing	311	22.270019883
-2	922320	813930	119d932c-8d73-4a70-b04a-3004681f8a04	Western Union	finance	Western Union - When you need to send money in person, Western Union®. Permanent agent clerks will help you send money quickly and easily.	1.17889039472	813930	Labor Unions and Similar Labor Organizations	561	13.5895060048
6	722513	722513	477d5a61-5911-4a21-b4c0-c0a31a3c7aa2	BOB EVANS RESTAURANT	meal takeaway restaurant	Goodness of the Farm® to you through quality food delivered in a friendly atmosphere at fair prices. Open for breakfast, lunch & dinner. As of 3/4/12, Bob Evans owns & operates 584 family restaurants in 18 states. For more info, please visit www.bobevans.com	0.777694635189	722513	Limited-Service Restaurants	424	15.4640352215
0	713110	0a1ff8de-13e9-49f5-895c-0a743077893		Laugh Out Loud Stations		celebration, and entertainment center where families, friends, and co-workers can run, play, eat, compete, and just Laugh Out Loud! Laser Tag, Ballroom, Solder Mountain, Land of Laugh-A-Lot Playground, Toddler World, Big Kid Bouncers, Climbing Wall, Game Arena, Kiddie Rides, &	0.584100950206	812930	Parking Lots and Garages	311	10.1807762252
0	821498	d178a62c-6c0b-4a2a-bc78-0806978695d7		Jing		At Jing, the premier holistic healthcare center in Louisville, Kentucky, we care about your health and well-being. Whether you're suffering from chronic pain, stress, or just want to improve your	0.577180289978	444190	Other Building Material Dealers	332	9.54718061495

0.706452213122	926150	Regulation, Licensing, and Inspection of Miscellaneous Commercial Sectors	333	26.579619758
0.709972670687	541213	Tax Preparation Services	561	13.5895060048
0.710778290392	541120	Offices of Notaries	423	15.7893421822
0.721712458367	532420	Office Machinery and Equipment Rental and Leasing	111	15.1711607068
0.743015183456	541613	Marketing Consulting Services	541	14.6032025959
0.746565862131	425110	Business to Business Electronic Markets	424	15.4640352215
0.750072904393	561499	All Other Business Support Services	111	13.3050205376
0.753702801839	541618	Other Management Consulting Services	275	10.9603619307
0.858679797692	541219	Other Accounting Services	311	10.1807762252
0.934480297907	541211	Offices of Certified Public Accountants	423	9.72432006755

LIST 1

- get\_business\_latlot.py: generates id\_to\_loc.pickle
- get\_business\_types.py: generates business\_types.pickle
- get\_naics\_data.py: generates naics\_list.json