Radius Collider Competition

Instructions and Guidelines

Table of Contents

The Challenge

Your challenge is to build a natural language model to algorithmically determine the best industry classification code for a given American small business. This contest will use the NAICS (*North American Industry Classification System*) system for industry classification codes; for more information about the NAICS system, see the NAICS Codes section below.

Data

The challenge dataset (challenge_set.json) is a json-structured text file which contains 10,000 records of business information. Each record contains the following fields:

- ‘name’: the name of the business
- ‘address’: the address of the business
- ‘description’: a text description of the business
- ‘website’: a list of URLs related to the business
- ‘unique_id’: a unique identifier for this record

Each record is associated to a “ground truth” NAICS code of between 3 and 6 digits, which we will use to score your prediction.

Conditions and Requirements

1. You must use Python or Scala to complete the challenge
   a. You may use any publicly available packages or modules; please cite packages used in your written report.

    b. You may also use Spark; with only 10,000 examples it is likely overkill, but a well-implemented Spark pipeline would certainly please the judges!
2. You may manually label up to 1000 examples for the purposes of creating a labeled training set, but you must algorithmically label the remainder.
    a. If you choose to manually label some examples, please include that in the report that you submit.
    b. You may ***not*** conscript a mechanical turk service or any other individual to label the examples. Only team members may apply labels.
3. You should be able to complete the challenge using only the textual information in the dataset itself, the contents of the websites provided for each example, and the static NAICS files on census.gov. You *may* seek out additional sources of information to inform your knowledge of the NAICS system, but using an external file or service that maps keywords to NAICS codes would be against the spirit of the competition.

Submission

You should submit the following files in a single archive ("collider_YOURNAME.zip", for example):

1. Your predicted NAICS codes in a csv file ("predictions.csv")
    a. Each line should simply consist of: `unique_id,prediction`
2. All code you wrote to produce your model and predictions.
3. A written report (see details below).

Scoring

Your final score will be a weighted combination of three scores, each on a scale from 0.0 to 1.0:

1. Prediction Score (60%)
2. Report Score (30%)
3. Code Score (10%)

Prediction Score (60%)

Our scoring system seeks to balance rewarding you for correct predictions and penalizing you for incorrect predictions. You are *not* required to predict a NAICS code for every business in the challenge set. For each example that you *do* predict, you should submit a code between 2 and 6 digits. **Your final Prediction Score will be the ratio of your total points to 50426 (the total number of possible points).**

Details of our scoring model

- Correct "industry" (first 2 digits): **+2 points**
    - Each additional correct digit beyond the first two is worth **+1 point**
    - Predicted digits beyond the extent of the "ground truth" code will not be considered
    - If your prediction is correct up to *n* digits but diverges from the ground truth beyond that, you will receive ***n-1*** **points**

- Incorrect “industry”: **-2 points**
    - If your industry is incorrect, you will be penalized exactly 2 points, no matter how long your predicted code is
- No Prediction: **0 points**

Example Scoring Table

Prediction	**Ground Truth**	**Score**	**Reason**
123456	123456	+6	6 correct digits
123456	123	+3	3 correct digits, 3 extraneous digits
123456	123876	+2	3 correct digits but incorrect “tail”
123456	777777	-2	Incorrect Industry
123	123456	+3	3 correct digits
None	123456	0	No prediction

Report Score (30%)

The report is an opportunity for you to show your understanding of the problem and the approach you pursued to solve it. You should consider the report to be an technical summary of your work. We require no particular structure or format, except that **the report must be limited to one page**. Some questions that might help you to prepare your report include:

1) What were the major challenges of this problem? What data did you use? How did you process or transform it?
2) Explain your modeling approach. Why did you choose it? What are the underlying assumptions and inherent limitations?
3) How did you generate your final predictions? How confident are you in your predictions?

Code Score (10%)

The primary reason that we require your code is to ensure that your predictions were indeed generated by a model and not simply by manual labeling. We reserve the right to run your code locally but will only do so if have a hard time interpreting it. If we are not convinced that your code *actually generated your predictions*, you will be disqualified from the competition.

Your codebase will be subjectively evaluated and scored based on two primary criteria:

1. Is it well-documented? (5%)
2. Is it well-structured? (5%)

Background Information

NAICS Codes

According to the US Census Bureau, “The North American Industry Classification System (NAICS) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy.” NAICS codes range from 2 to 6 digits and are encoded in a hierarchical tree structure. The first two digits indicate the broad industry of the business, and each subsequent digit encodes more detailed information. For example:

- 72 - Accommodation and Food Services
- 722 - Food Services and Drinking Places
- 7225 - Restaurants and Other Eating Places
- 72251 - Restaurants and Other Eating Places
- 722514 - Cafeterias, Grill Buffets, and Buffets

You may note that many “parent” codes have a child which appears to be identical, like 7225 and 72251 above. This *does indeed* happen from time to time throughout the system, but should not be a point of concern. When in doubt, predict the deeper of the two codes.

The most useful files for your purposes will likely be found here in the Downloadable Files section. Each of the 2012 files may provide some useful insights and raw data.

The Radius Business Graph

NAICS codes are only one of the hundreds of fields of firmographic signals that are modeled, imputed, and compiled each week into the Radius Business Graph. This graph of over 15M US businesses provides an industry-leading foundation for predictive marketing efforts. By matching the Radius Business Graph with our customer’s CRM (i.e. Salesforce), we empower marketers to discover new markets, acquire customers, and measure their performance.