Springboard Data Science Career Track

Capstone Project 2 - Company Industry Classification

- Natural Language Processing

Irene Yao
January 29, 2019

Introduction

Rapid Growth in Venture Capital Investment

- Venture Capital investment in 2018: $138b in 9,216 deals (source: EY)

Importance of Information Tracking

- Accurate data collection
- Efficient data analysis
- Automated data cataloging

Project Summary

- Cataloging of company industry segment based on business description
- Model selected: LinearSVC

Data Source & Data Wrangling

Data Source: The Open Data 500 by The GovLab

Data Wrangling Steps:

- Removing records with missing value
- Encoding of categorical features
- Outliers inspection
- Feature extraction

Text Pre-processing

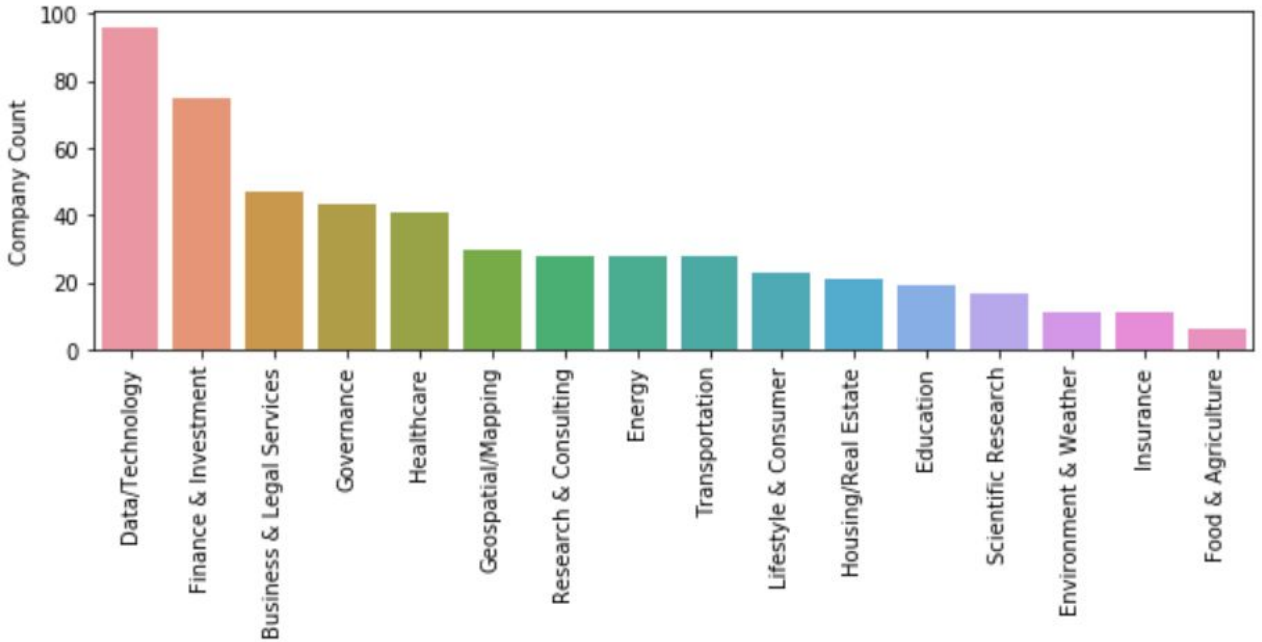
Text pre-processing of company description (free text)

- Remove html tags, urls, numbers, accented characters, special characters, extra spaces
- Remove stopwords (a, to, and, etc.)
- Expand contractions (ex: we'll => we will, there're => there are)
- Lemmatize text (converting word to its lemma form)

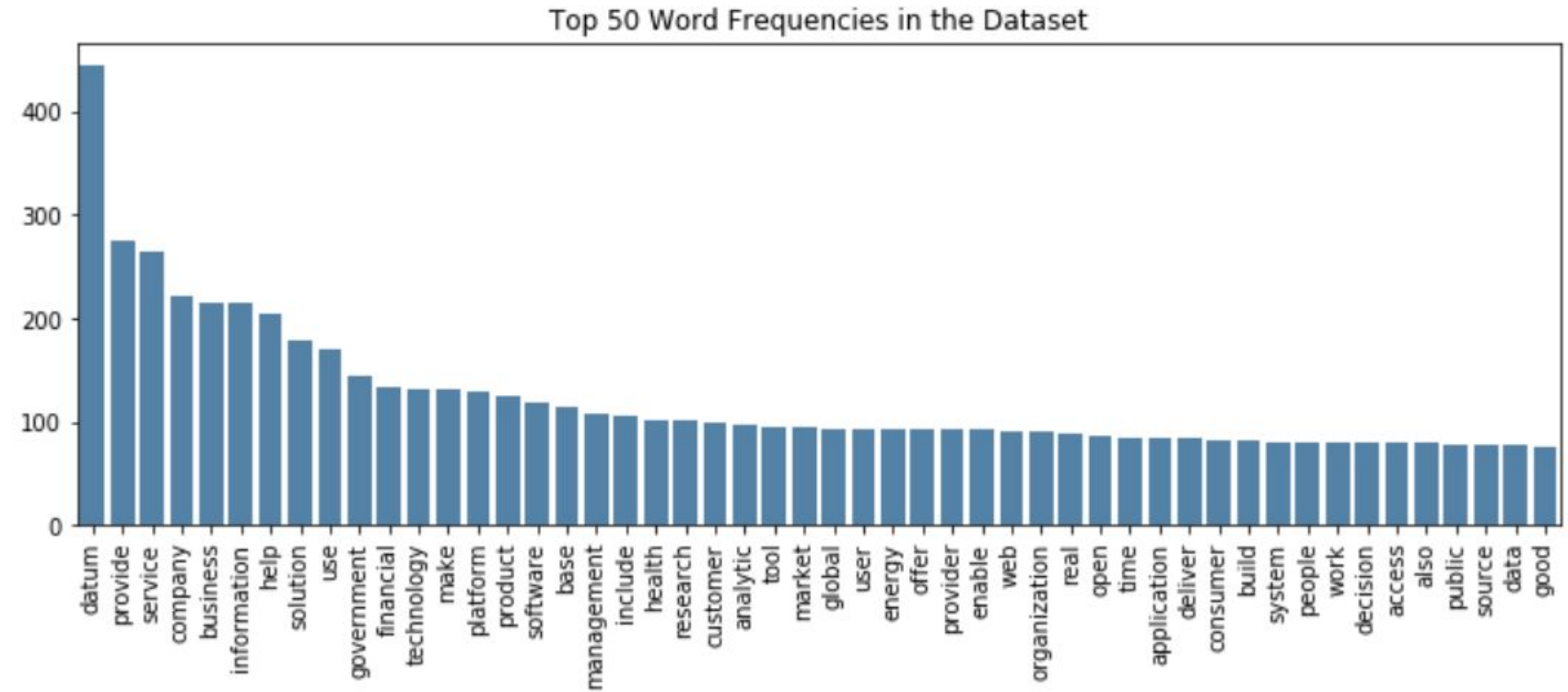
Exploratory Data Analysis
&
Statistical Inference

Number of Companies in Each Category

- Imbalanced Dataset



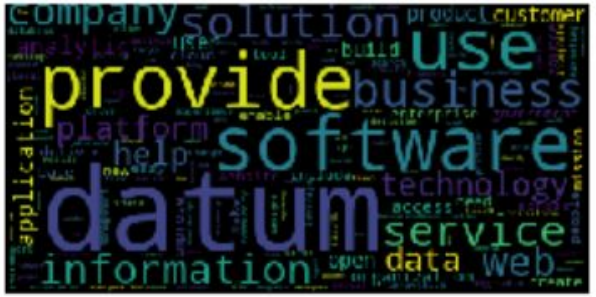
Top Word Frequencies



Sample Word Cloud of Each Industry

Finance & Investment

Data/Technology



Governance

Research & Consulting

Most Relevant Words in Each Category

Business & Legal Services

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	legal	23.169131	individual business	9.046163
1	lawyer	12.846582	customer need	4.508171
2	patent	6.770733	background check	4.442675

Data/Technology

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	datum	5.181395	business intelligence	3.637256
1	demographic	4.376913	complex datum	3.262470
2	software	3.898663	enterprise datum	3.205333

Education

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	student	70.875412	student college	12.910004
1	college	58.340190	help parent	12.012365
2	aid	26.508216	school teacher	11.038416

Energy

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	energy	100.064870	energy datum	13.45330
1	solar	35.900085	energy management	13.22840
2	utility	19.881284	energy efficiency	12.73332

Machine Learning

Model Improvements

Baseline Model → Balanced Data → Model Selection

Multinomial Naive Bayes

- Corpus built on training set
- Corpus size: 2505
- Train accuracy: 0.45
- Test accuracy: 0.27

Multinomial Naive Bayes

- Balanced classes with 67 samples each
- Train accuracy: 0.97
- Test accuracy: 0.64

Accuracy of Different Models

- LinearSVC: 0.60
- LogisticRegression: 0.45
- MultinomialNB: 0.31
- RandomForest: 0.25

Model Improvements – Cont'd

Better Model → New Word Corpus → Best Model

LinearSVC

- Corpus built on training set
- Train accuracy: 1.0
- Test accuracy: 0.68

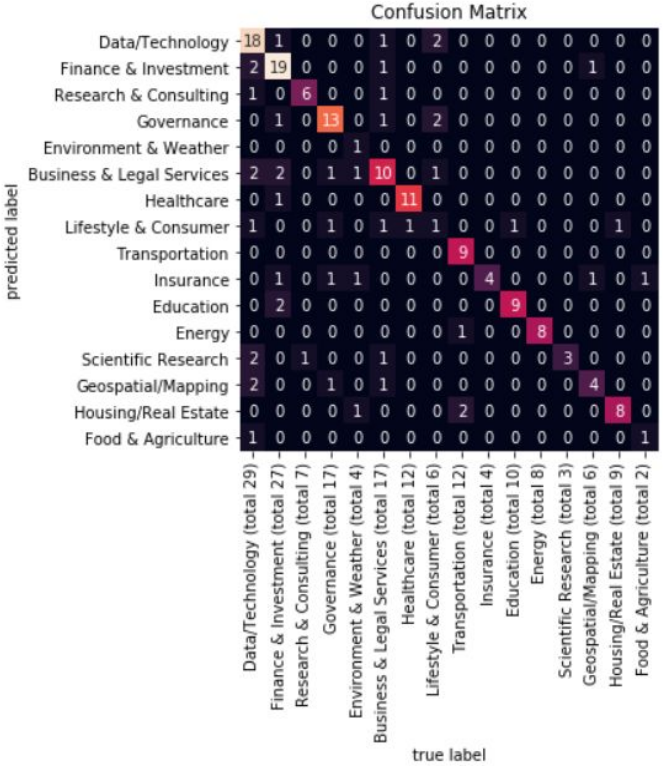
LinearSVC

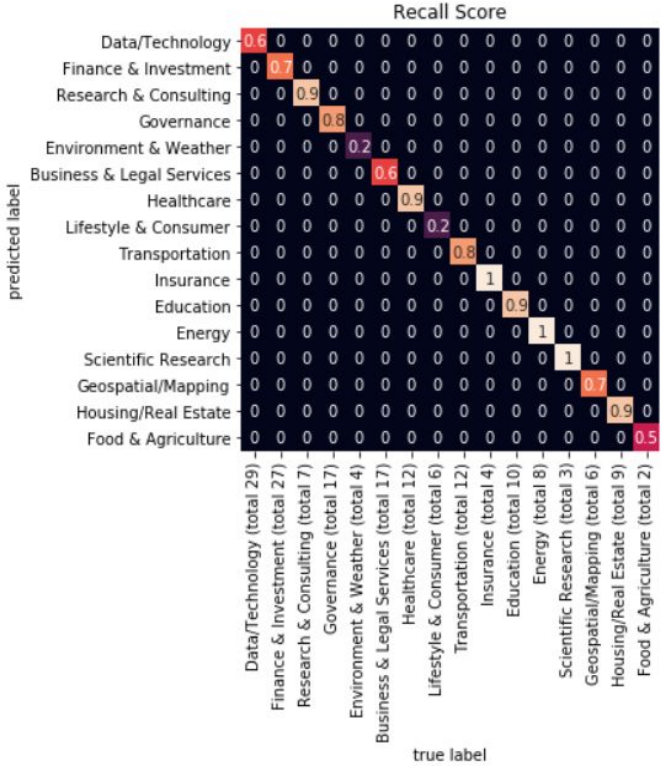
- Corpus built on entire document
- Corpus size: 3624
- Train accuracy: 1.0
- Test accuracy: 0.71

LinearSVC with Tuned Features

- SMOTE: k_neighbors=1
- TfidfVectorizer: min_df=2, max_df=0.4
- LinearSVC: C=0.7
- Train accuracy: 0.98
- Test accuracy: **0.73**

Final Prediction Result



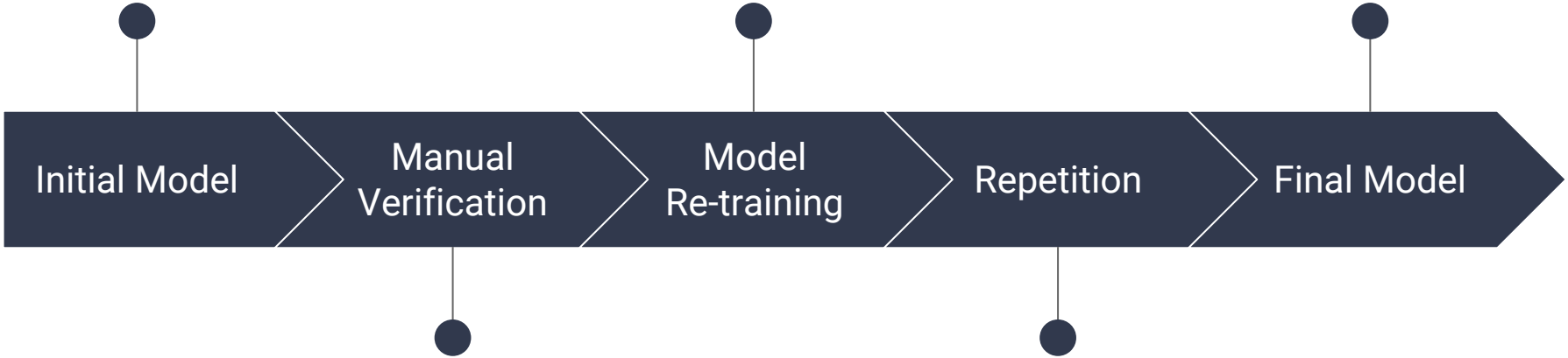


Recommendations

Train the model with large dataset

Re-train the model with new data and new word dictionary

Use the optimal model for automation without human involvement



Evaluate the prediction and modify the corpus

Repeat the manual verification and re-training until optimal result

Future Work

Possible Enhancement

- Obtain more data, especially for companies in the minority classes
- Expand feature set to include other variables other than company description
- Add curated taxonomies of words associated with each class to improve prediction
- Apply Latent Dirichlet Allocation (LDA) for topic modeling