

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)

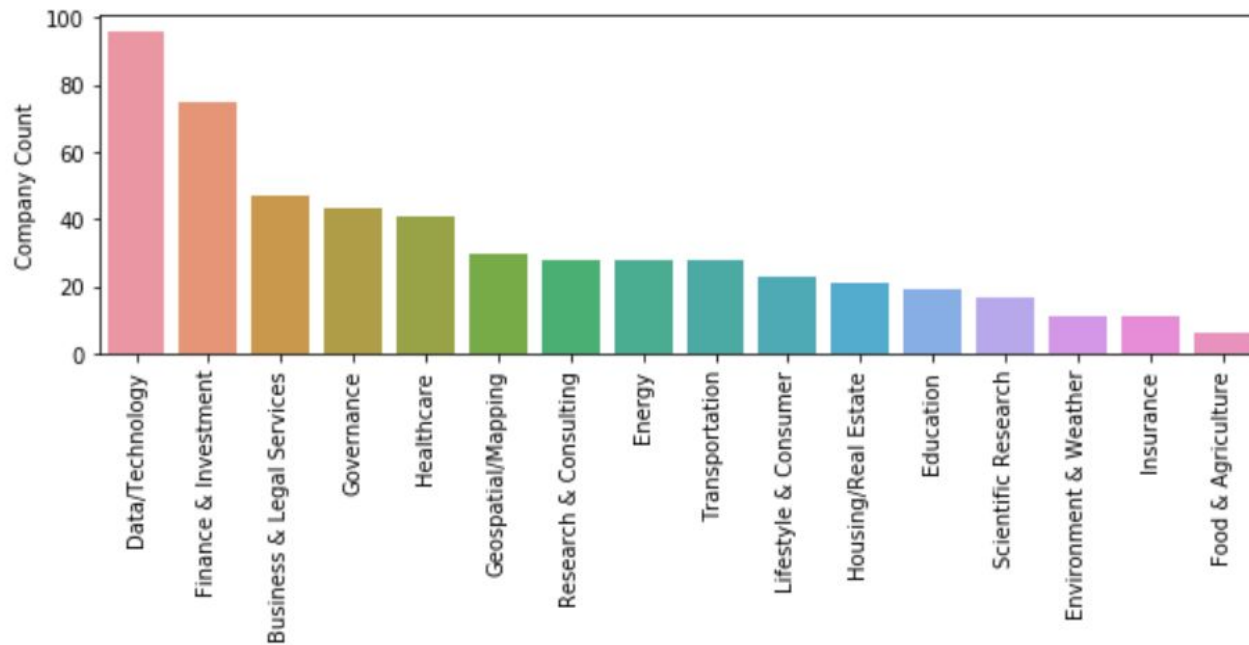
- Remove html tags, urls, numbers, accented characters, special characters, extra spaces
- Remove stopwords (a, to, and, etc.)
- Expand contractions (ex: we'll => we will, there're => there are)
- Lemmatize text (converting word to its lemma form)

Exploratory Data Analysis & Statistical Inference



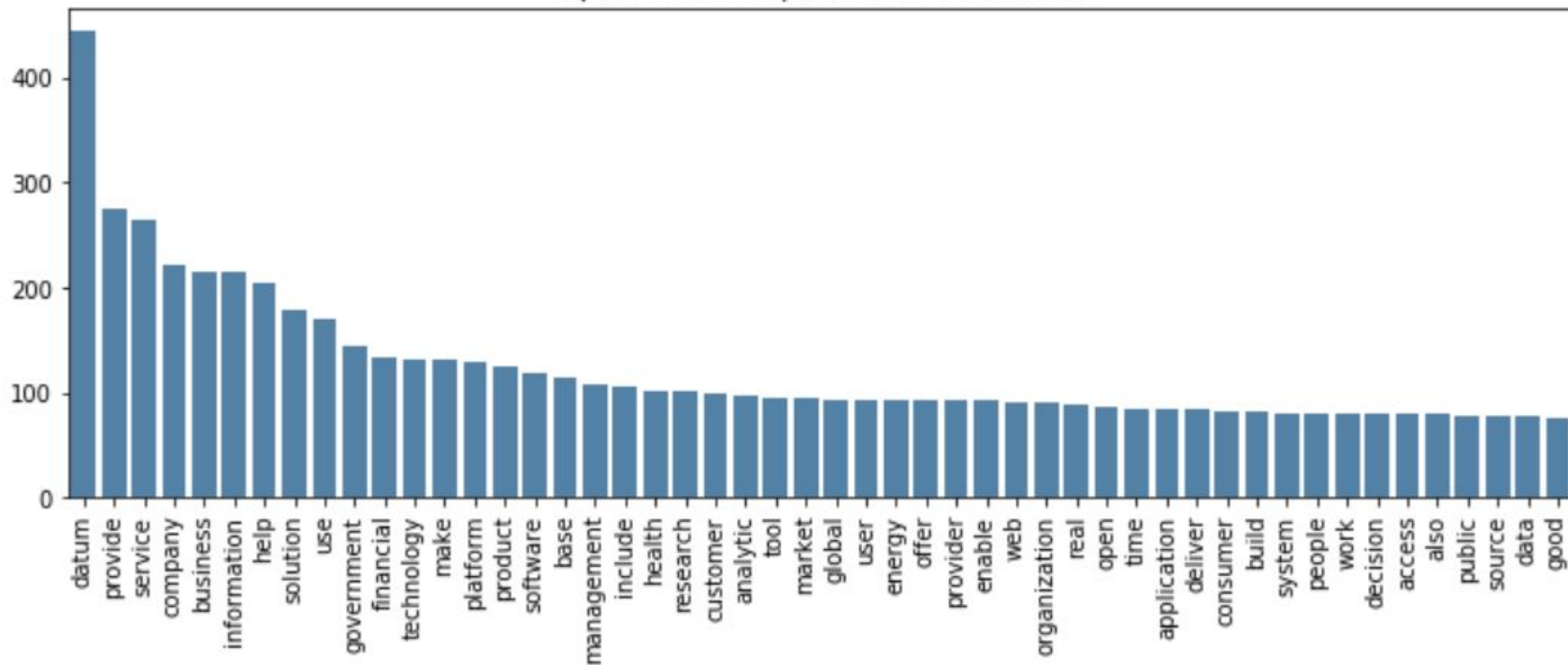
Number of Companies in Each Category

- Imbalanced Dataset



Top Word Frequencies

Top 50 Word Frequencies in the Dataset

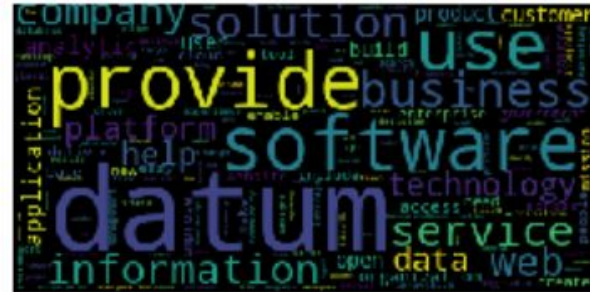


Sample Word Cloud of Each Industry

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

Multinomial Naive Bayes

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

Balanced Data

Multinomial Naive Bayes

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

Model Selection

Accuracy of Different Models

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

Model Improvements – Cont'd

Better Model

LinearSVC

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

New Word Corpus

LinearSVC

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

Best Model

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

Confusion Matrix

	Data/Technology	Finance & Investment	Research & Consulting	Governance	Environment & Weather	Business & Legal Services	Healthcare	Lifestyle & Consumer	Transportation	Insurance	Education	Energy	Scientific Research	Geospatial/Mapping	Housing/Real Estate	Food & Agriculture
Data/Technology	18	1	0	0	0	1	0	2	0	0	0	0	0	0	0	0
Finance & Investment	2	19	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Research & Consulting	1	0	6	0	0	1	0	0	0	0	0	0	0	0	0	0
Governance	0	1	0	13	0	1	0	2	0	0	0	0	0	0	0	0
Environment & Weather	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Business & Legal Services	2	2	0	1	1	10	0	1	0	0	0	0	0	0	0	0
Healthcare	0	1	0	0	0	0	11	0	0	0	0	0	0	0	0	0
Lifestyle & Consumer	1	0	0	1	0	1	1	1	0	0	1	0	0	0	1	0
Transportation	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
Insurance	0	1	0	1	1	0	0	0	0	4	0	0	0	1	0	1
Education	0	2	0	0	0	0	0	0	0	0	9	0	0	0	0	0
Energy	0	0	0	0	0	0	0	0	1	0	0	8	0	0	0	0
Scientific Research	2	0	1	0	0	1	0	0	0	0	0	0	3	0	0	0
Geospatial/Mapping	2	0	0	1	0	1	0	0	0	0	0	0	0	4	0	0
Housing/Real Estate	0	0	0	0	1	0	0	0	0	2	0	0	0	0	8	0
Food & Agriculture	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	Data/Technology (total 29)	Finance & Investment (total 27)	Research & Consulting (total 7)	Governance (total 17)	Environment & Weather (total 4)	Business & Legal Services (total 17)	Healthcare (total 12)	Lifestyle & Consumer (total 6)	Transportation (total 12)	Insurance (total 4)	Education (total 10)	Energy (total 8)	Scientific Research (total 3)	Geospatial/Mapping (total 6)	Housing/Real Estate (total 9)	Food & Agriculture (total 2)

Recall Score

	Data/Technology	Finance & Investment	Research & Consulting	Governance	Environment & Weather	Business & Legal Services	Healthcare	Lifestyle & Consumer	Transportation	Insurance	Education	Energy	Scientific Research	Geospatial/Mapping	Housing/Real Estate	Food & Agriculture
Data/Technology	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Finance & Investment	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Research & Consulting	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0
Governance	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0
Environment & Weather	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0
Business & Legal Services	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0
Healthcare	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0
Lifestyle & Consumer	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0
Transportation	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0
Insurance	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Education	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0
Energy	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Scientific Research	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Geospatial/Mapping	0	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0
Housing/Real Estate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0
Food & Agriculture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
	Data/Technology (total 29)	Finance & Investment (total 27)	Research & Consulting (total 7)	Governance (total 17)	Environment & Weather (total 4)	Business & Legal Services (total 17)	Healthcare (total 12)	Lifestyle & Consumer (total 6)	Transportation (total 12)	Insurance (total 4)	Education (total 10)	Energy (total 8)	Scientific Research (total 3)	Geospatial/Mapping (total 6)	Housing/Real Estate (total 9)	Food & Agriculture (total 2)

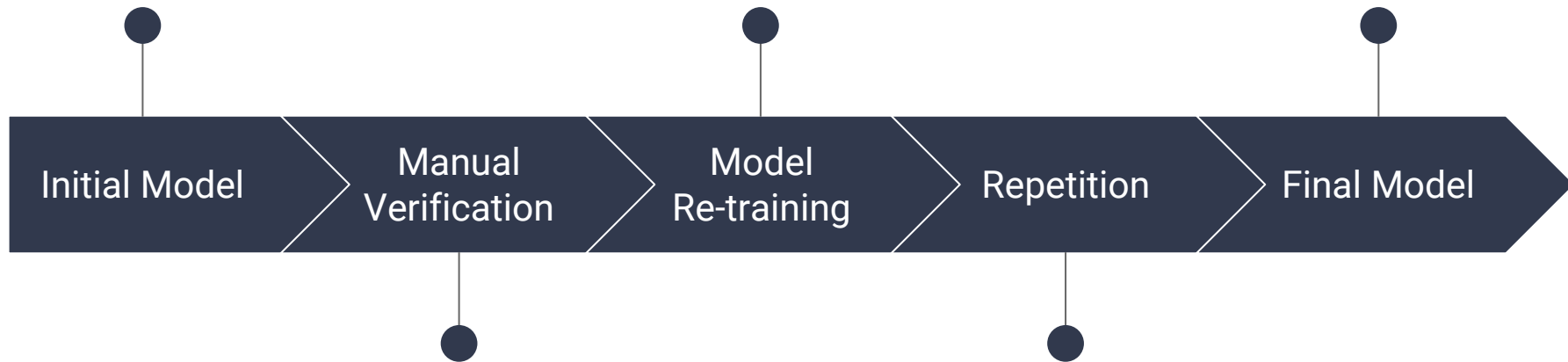
Recommendations



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