JOBVITE

Job Supply and Demand Modeling

Joseph Ching, Sam Craig, Labiba Imdad, James Joko, Michael Keeley, Boris Lu, Elizabeth O'Connell, Jake Roach, Anish Tiwari, Sami Varadarajan

PURDUE UNIVERSITY

Introduction

Motivation: Model the supply and demand of skills in the job applicant pool, by extracting skills found in job descriptions.

Early Ideas:

- Job Description/Resume Parser
- Insight Tool to predict importance of skills in the future
- Insight Tool to improve user's skill set

Goal: Create a model to filter sentences in a job description that contains skills.

(Sentences containing skills will be classified as "has_skills", while sentences without skills will be labeled as "no_skill".)

Tools we used:

- Python for Natural Language Processing, Modeling, and Data Analysis
- Supervised + Unsupervised Random Forest Classification Model
- Unsupervised K-means Clustering Algorithm
- BERT-as-service (BaaS) to encode sentences

Methodology

BERT Encodings

- BERT uses pre trained models to turn sentences into sentence embeddings.
- Sentence embeddings are vectors that represent 1 sentence as 1024-dimension vector so a computer can understand it.

Classification Methods we tried to use to determine presence of skills, but had low accuracy or computationally intensive:

- Bi-directional LSTM
- Conditional Random Field
- Gradient Boosting
- Support Vector Machines

Random Forest model (Figure 1):

- Uses a multitude of decision trees to classify sentences
- has_skills vs. no_skills for descriptions and requirements

K-means Clustering (Figure 2):

- Unsupervised K-means clustering algorithm
- Used to cluster similar sentences
- Allowed for further exploration of the data

Figure 1: Random Forest Classifier

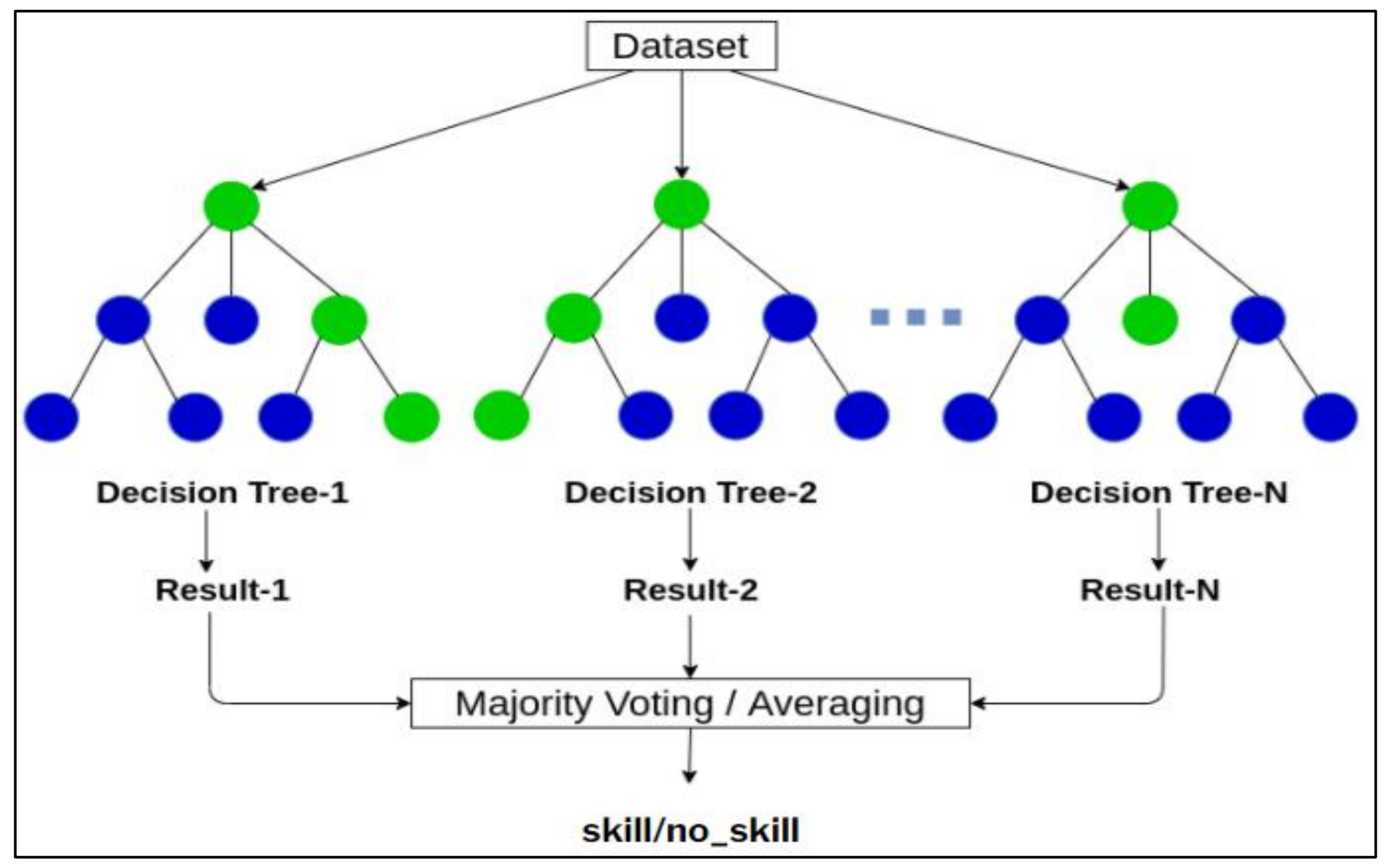
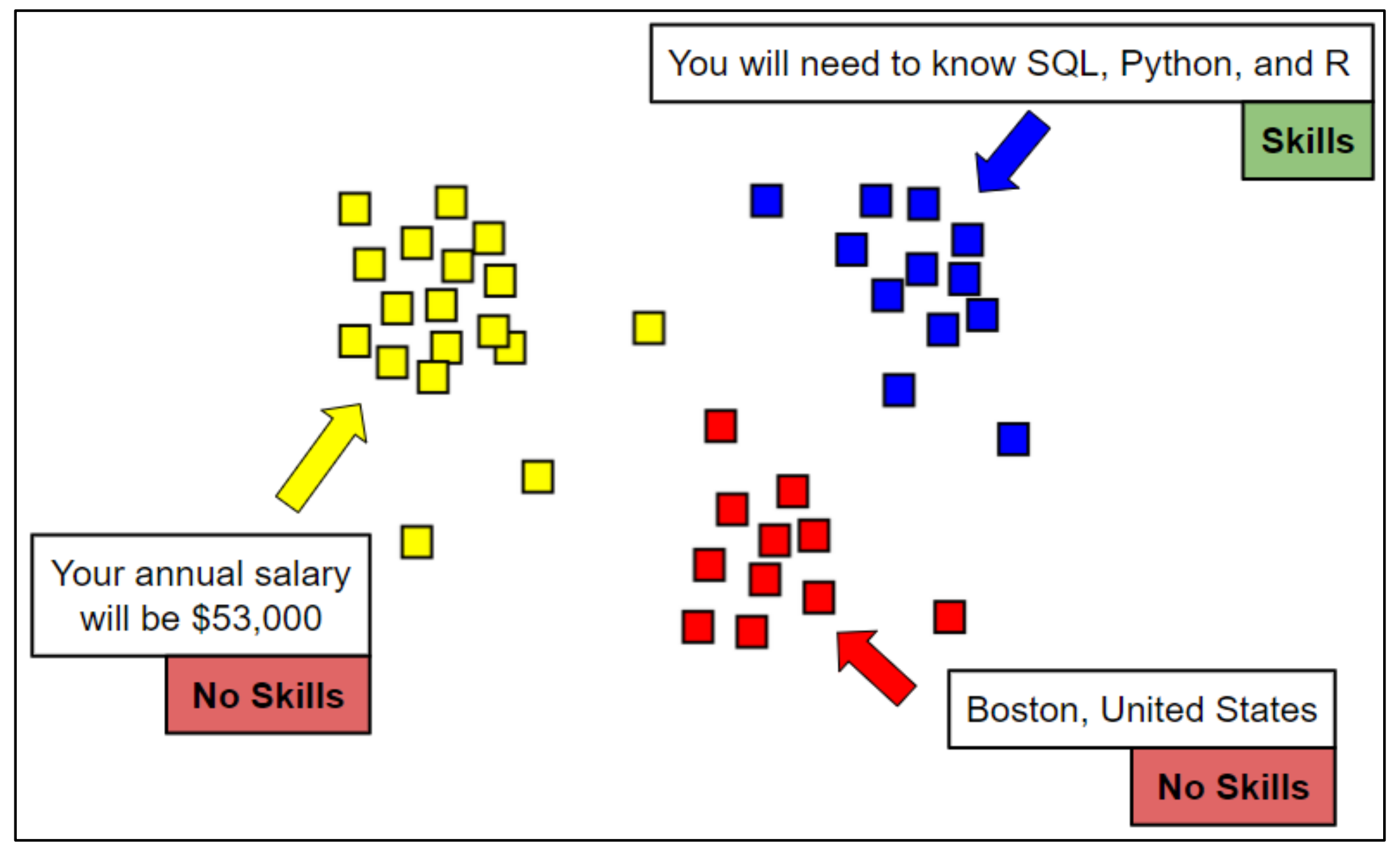
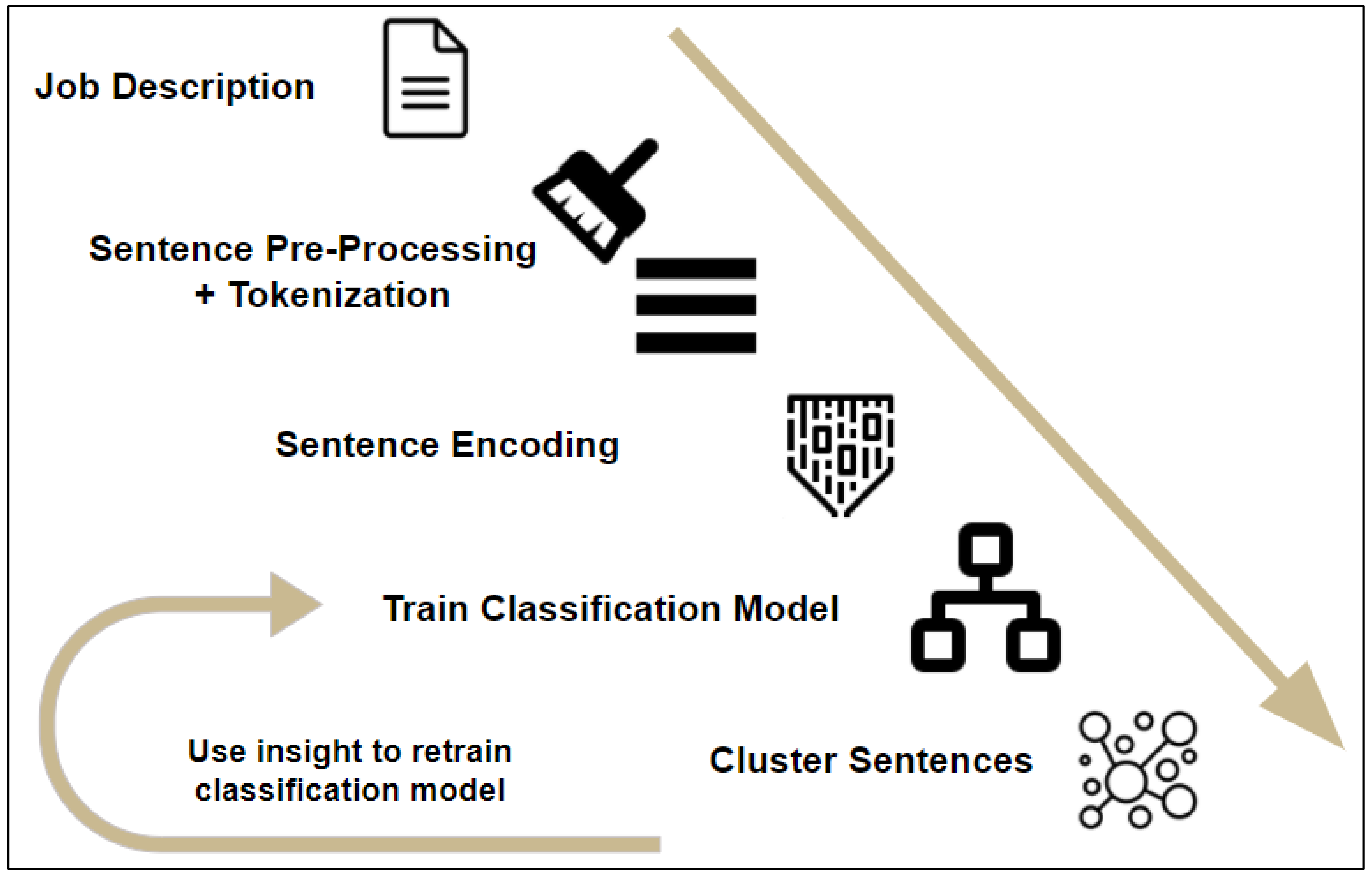


Figure 2: K-means Clustering with Sentences


You will need to know SQL, Python, and R
Skills
Your annual salary will be $53,000
No Skills
Boston, United States
No Skills



Job Description

Sentence Pre-Processing + Tokenization

Sentence Encoding

Train Classification Model

Use insight to retrain classification model

Cluster Sentences

Figure 3: Overall Algorithm Roadmap

Conclusion + Future Goals

Deliverable: Created a model that can identify sentences containing skills in a job description with 94% accuracy. Given a job description, we can identify "important" sentences.

Next Steps:

- Extract specific skills from the useful sentences we identified.
- Adjust the model to extract skills from resumes instead of job descriptions.

Skill extraction from "has_skills" (useful) sentences:

- Part-Of-Speech Tagging – grammar and word placement are used to determine where skills are in a sentence
- Extract phrases containing skills

Data

Kaggle Dataset

- 17,014 Job Descriptions from all fields
- Turned into ~216,448 sentences

Jobvite Dataset

- 20,000 Job Descriptions from IT-related jobs
- Page Sources (HTML)
- Turned into ~300,000 sentences

References

Special thanks to our mentors Dr. Morgan Llewellyn and Dr. Sasan Hashemi!

Kaggle Fake JobPosting Dataset
bert-as-service Documentation

The Data Mine Corporate Partners Symposium 2021