Project Proposal: Jobs to Skills

24 October 2021

Nguyen Hoang Giang (HG)	gianghn2@illinois.edu	Captain
Nguyen Thu Giang (TG)	tgn3@illinois.edu	Member

Background & Motivation

With the current ubiquitous use of online job platforms as a main channel of recruitment, it is increasingly easier to identify in-demand skills. This information availability allows potential job seekers to be informed of what skills they need to master in order to gain a headstart in their career.

In this project, we would propose the usage of text mining and text analytics methods to increase the efficiency of identifying relevant skills in any given job description and gain updated insights of popular skills in recent job descriptions.

Task Description

In this project, we are aiming to

1. Identify main skills explicitly stated in English job descriptions.
2. Identify implicit skills related to the main requirements in job descriptions.
3. Identify highly demanded skills in the market based on a collection of recent job descriptions.

Planned Approach

General Approach

We are planning to approach the skill extraction tasks via 2 routes:

1. Topic modelling: As we assume that topics in job descriptions would mainly consist of keywords describing skills needed, topic mining is a possible approach to extract major skills explicitly stated in the job description.
2. Text clustering:
    a. We would identify several cluster job descriptions. The cluster would help to identify job postings having similar contents in terms of skill requirements and job scopes.

b. With a set of skill descriptions, we would identify relevant clusters that these job descriptions belong to, which would help in interpreting the meaning of each cluster.
c. Topic modelling could also be applied to a group of job descriptions in each cluster in order to identify related skills.

Dataset

1. Job description dataset: We would crawl around 2500 online job descriptions on job portals such as Linkedin, Indeed, etc.
2. Skill description:
    a. First, we would identify a list of top 50 common skills by mining LinkedIn and Angel (https://angel.co/skills).
    b. Based on skills listed in step a, we would then obtain the descriptions of each skill on Wikipedia. For example, description for project management skill could be based on the following article on Wikipedia.

Algorithm

The following algorithms proposed are as follows but not exclusive to:

Task	**Text representation**	**Algorithms**
Topic modelling	Bag-of-word	• Latent Dirichlet Allocation (LDA) • Probabilistic Latent Semantic Analysis (PLSA)
Topic clustering	• Word embeddings: We would explore a suitable word embedding technique such as word2vec, BERT, etc. • TF-IDF	K-means clustering

Systems and Programing Language

We would utilise Python 3 and relevant packages as the main programming language for this project.

Task	Package
Dataset mining	phantom, beautifulsoup, selenium
Model development	BERT, scikit-learn, gensim

Expected Outcomes

- A model that will take in a job description and output a number of relevant skills.
- A list of trending skills as reflected by the collection of recent job descriptions that we crawl.

Evaluation Methods

As both supervised and unsupervised learning algorithm are used, there is no simple and direct approach to evaluate the system. However, several methods can be combined to validate the outcome. Metrics such as elbow method and silhouette analysis are used for text clustering evaluation. A small dataset including job descriptions annotated with related skills will be used to evaluate the preciseness of extracted skills. Metrics such as mean average precision (MAP) or normalized discounted cumulative gain (NDCG) can be used to validate job matching or skill rank.

Work Allocation

No	Task	Hours allocated	Member in-charge
1	Crawl job description	5	HG, TG
2	Dataset preparation and cleaning	5	HG, TG
3	Topic modelling	10	HG
	a. LDA	(5)	
	b. Plsa	(5)	
4	Topic clustering	10	TG
5	Analysis of result	5	HG, TG

6	Report writing	5	HG, TG
Total hours		**40**	