# Topic proposal for the Pattern recognition project

Thanh Hung Duong

January 31, 2021

## 1 Introduction

During recent years, data science has been described as one of the hottest fields with attractive salaries. however, when we search jobs in this fields, the results usually shows us various positions, such as data scientist, data analyst, data engineer, machine learning engineer. There are so many websites explained the differences between them and listed the requirement skill sets for each roles [1, 2]. Although those explanation are detailed and informative, they mostly are qualitative statement but not quantitative interpretation. Moreover, from my point of view, all those jobs are alive, which means their tasks may change time after time. As a consequence, an explanation might be correct at the time the author wrote it, but after a few months or years, it no longer is correct and need to be updated. Likewise, the other factors such as type of companies, job level may affect the requirement skill sets for each roles.

## 2 Goals

In this project, I would like to study the requirement skill sets for each of four positions: data scientist, data analyst, data engineer, machine learning engineer by analyzing the job postings on biggest job-search platform such as LinkedIn, Indeed, Glassdoor. My goal is to answer these following questions:

- What are the common qualifications for each roles
- What are the factors that can affect those requirements and what is the level of impact?
- Does those skill-sets change over year? if yes, what is the trend?
- Which jobs has more posts? Can we predict the number of job vacancies for each roles over year?
- For example, if I want to be a machine learning engineer, what skills should I prepare during school so that I can get a job when I graduate in the next few year?

## 3 Methods

My plan is:

1. Scrap the job descriptions from LinkedIn with 4 different keywords that corresponds to 4 selected position. BeautifulSoup and Selenium will be 2 main tools to carry out this task. Moreover, the other information, namely *Posted date, Job title, Company name, Location, Job level, Job type, Industry* , should be extracted too.
2. Clean the data and prepare it via steps like *Tokenization, Sentence Breaking, Part of Speech Tagging,..*
3. Visualize it to see trend or patterns.
4. If possible, build a classification model to predict role based on job descriptions

# References

[1] Data Flair. Data Scientist vs Data Engineer vs Data Analyst – What really differentiates them? https://data-flair.training/blogs/data-scientist-vs-data-engineer-vs-data-analyst/, 2019. [Online; accessed 30-Jan-2021].

[2] Aayushi Johari. Data Analyst vs Data Engineer vs Data Scientist: Skills, Responsibilities, Salary. https://www.edureka.co/blog/data-analyst-vs-data-engineer-vs-data-scientist/, 2020. [Online; accessed 30-Jan-2021].