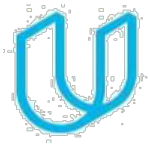
# Capstone Project Proposal



*<Ying Shan>*

---

## Business Goal

### Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?

I would like to create a smart recruitment system, which aim to solve the problem of person-job Fit. The system will automate the recruitment process for a company, help employer to make better and faster decision

ML and NLP (natural language processing) will be the key technologies for the system, which will automatically match candidates and jobs, prompting the efficiency of corporate recruitment

Based on machine reading technology, the system will identified the job seeker's resume, and generate the structured. The system will also extracted the key information in the resume, and match the resume information with the job requirements of open position.

### Business Case

Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.

The company will invest a lot of labor forces and time costs in the recruitment of staff. Among them, the process of resume screening is the most time consuming section. Resume screening is a simple repetitive work, but it is very important for company recruitment. Manual screening is often time-consuming and labor-intensive, and it is easy to miss out on good candidates. If a company can save their work input in the resume screening process and put more energy into the interview of suitable candidates.

The such system can help company to save a lot of resources, and will not missing every suitable job seeker that  gives the company a higher probability of getting the ideal employee.

### Application of ML/AI

What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?

I will complete two tasks through ML/AI, which are resume data reading and text information extraction.

Data reading technology enables a resume document to become structured data. Text information extraction technology can obtain key information in the resume, which will match the job requirements.

A company only need to write a job description and upload the resumes to the system. The system will automatically calculate the scores of job matching for each candidate, and select the appropriate candidates for the company.

## Success Metrics

### Success Metrics

What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.

The business metric for the smart recruitment system: increasing the speed of resume screening.

The base line will be the total monthly time consume (hrs) for HR team to working on the resume screening.
Compare the time cost based on the same resume screening quantity.

## Data

### Data Acquisition

Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?

The data can be found in the talent pool of company, as well as online open source resume & CV database.

There are some paid resume database resources, which cost around 1$ per resume.
I may also write a crawler to crawl data on the job website (indeed, job.net, hires, glassdoor, linkedin etc.). In addition, there are some open source CV database are free to access.

There may have some sensitively issue, because resume data may contains candidates personal information (name/date of birth/phone number etc.) However, the model training will not require those information, which only focus on work experience, academic background, job keywords, skill keywords, job year, time, location, etc. Also, the database will only be used for model training.

Data information needs to be constantly adjusted according to the demand of the company's job responsibilities. As professional requirement change and particular skills change, data needs to be constantly updated and adjusted.

## Data Source

### Consider the size and source of your data; what biases are built into the data and how might the data be improved?

In the beginning, I would like to solve only the matching problem of AI companies, so my product will be designed for the position information of AI related companies. I will pick 4 of the most popular job position that offer from AI company, and collect resume data based on the 4 positions. (Full-stack Engineer; PM; ML Engineer; Data Engineer; ) I will find 500 resume for each position (totally 2000 data for modal training)

First of all, I only have data for 4 positions, so the resume that the model can handle will have certain limitations. Also, I only have 2000 training data, which may not enough for training an ideal model.
In order to improve the data and get rid of the biases, I should get more data from more positions.

## Choice of Data Labels

### What labels did you decide to add to your data? And why did you decide on these labels versus any other option?

For the label task, contributors will annotate tokenized text by assigning class labels from an ontology. I will set 5 classes which are job title, schools, previous work companies, key skills and main duties.

Since the main function of my product is the extraction and matching of key information of text data, I will create an data label work that annotate entities of text. (The template “Annotate Text by Tokens or Spans” on the Figure Eight is suitable for the task)

# Model

## Model Building

### How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?

I will complete the annotation of the resume data through the text annotation tool. The model is trained locally through my own GPU. I want to build a model from an internal team and deploy it privatized.

Since my training dataset is not very large, I may use the GPU to complete the training. In addition, the data content is sensitive, so it's better to complete the model training locally. Finally, because my product is for business users, the internal human resources data needs to be kept private, so I need to privatize training and deployment.

## Evaluating Results

### Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?

My model should be able to successfully extract the five key information I defined from the resumes of the four types of positions.

I would like to use 20% of my dataset for testing the model. (100 data for each position and 400 in total)

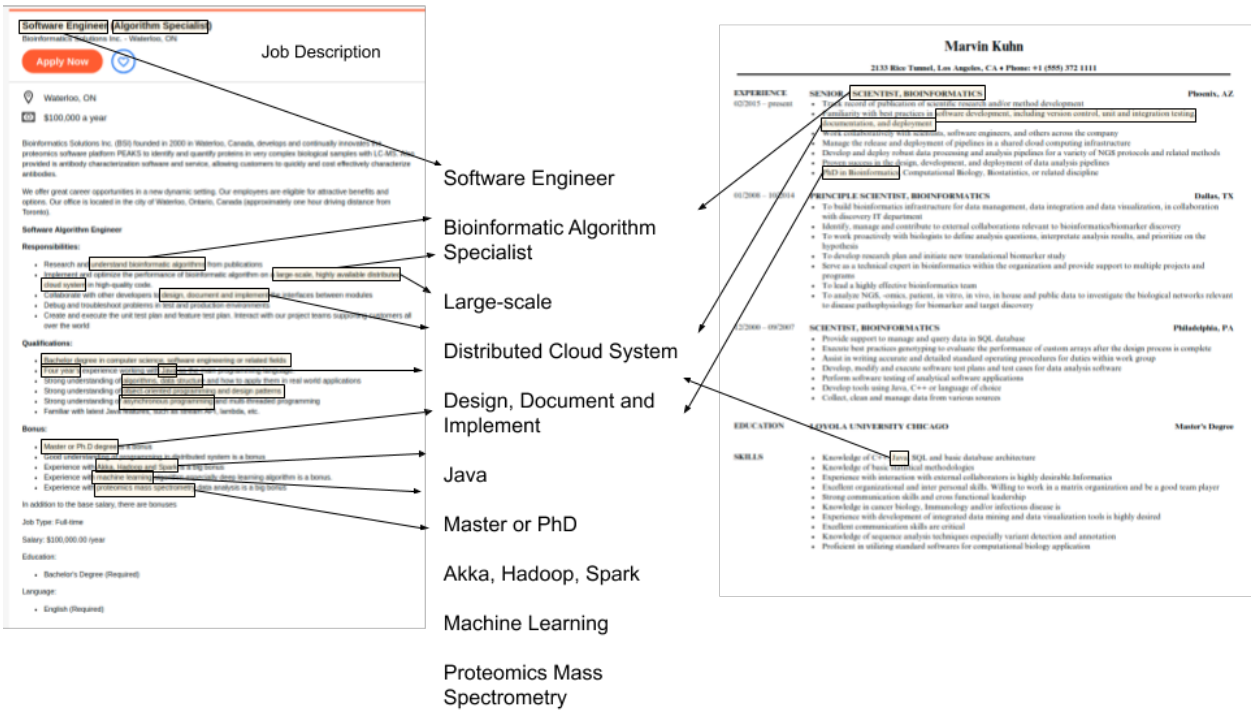
I will calculate the F1 score for the five data class (5 key information on resume) .
The F1 score should be higher than 0.75, otherwise, the model is not qualified.

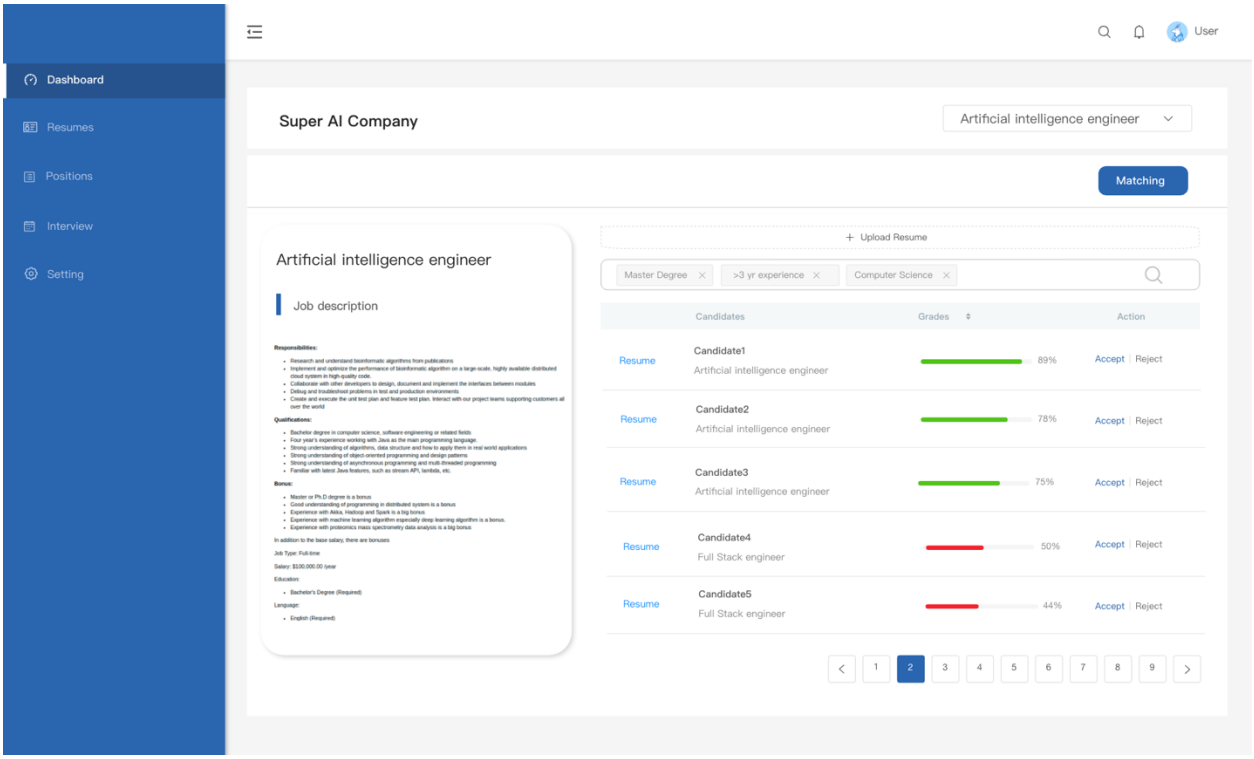
I will also classify the evaluation result based on the 4 job positions. The average result for each class should be similar, which means the model has consistently performance in different use cases.

# Minimum Viable Product (MVP)

## Design

### What does your minimum viable product look like? Include sketches of your product.



## Use Cases

### What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?

The target user of the product is the company's human resources and talent recruiting person, who is responsible for collecting the resume of the job seeker and screening the resume. The target company is a startup (between 50 and 200 people) that engages in artificial intelligence and machine learning related businesses. It has just completed the A round or B round of funding, with the goal of rapidly expanding its business and scale of company.

Use case:
1 Company writes job descriptions and admission requirements in the system.
2 The job seeker selects a position of interest and passes the resume to the system.
3 Companies get a list of resumes from different job seekers in the system.
4 Press the “Match” button in the system, the system automatically scores each job seeker and sorts the scores from high to low.
5 Companies can customize the score limit and filter the resume of job seekers.
6 Companies make the decision based on the matching results (reservation interview / elimination)

Users can log in to the system by entering the account password through the front end of the web app on a desktop. The product can be customized for privatized deployment

## Roll-out

### How will this be adopted? What does the go-to-market plan look like?

The Smart recruitment System is a product for the enterprise. Therefore, I plan to promote it in the company that I current worked in, as well as some of my friend’s company that in the relevant field.

The system is a web app that is deployed on the local server. It also has a mobile app, a chatbot that collects resumes from candidates.

Through the chatbot to understand the basic information of candidates, and guide candidates to upload their own resumes. The resumes will be automatically stored in the system's resume database. When the company has recruitment requirements, it can publish a QR code of the chatbot, candidates will scan the QR code and upload their basic profile and resume/CV.

# Post-MVP-Deployment

## Designing for Longevity

### How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?

In the long term, I need to collect more resume data to provide smart matching services for more job positions. In addition, I need to dig deeper into the enterprise scenario, familiar with the relevant processes of enterprise human resource management and recruitment, understand user requirements, and provide valuable functions.

In the real world, the content and format of resumes can be varied. (pictures instead of documents) It's necessary to have more training data, including as many resume formats as possible. Also, It may require to convert image content into text information using techniques such as OCR.

As more candidates upload their resume into the system, the system will be familiar with more resume type and contents. The system will be able to learn popular information in the industry based on new resume datasets.

I can employ an A/B testing, which focus on different key elements on the resume.
For instance, based on the same conditions, System A matches the position, education, work year, skill keywords, and technical keywords; System B adds a school name based on System A.
I can also try different way to collect candidates information.
For instance, System A: all candidate upload their resume into the system; System b: all candidate need to fill a digital table. (similar as Linked in profile)

## Monitor Bias

### How do you plan to monitor or mitigate unwanted bias in your model?

I would like to mitigate unwanted bias of my project in three aspects.
Data Bias: Collect data from diverse place, meanwhile, narrow down the target industry and job type.
Model Bias: Set the model and each parameter carefully. Though diversity model development process to get rid of unwanted bias
Annotation Bias: Make the rule of annotation task clear and simple, only focus on objective contents, and particular key words.

It impossible to avoid bias completely, but it is possible to reduce unwanted bias by proper design and rules. Collect user requirement through their feedback, and adjust data and models in a targeted manner to continuously improve product performance and user experience.