|  |  |
| --- | --- |
| **Date:** | 2-December-16 |
| **Time:** | 4:00pm to 5:00pm |
| **Venue:** | Skype |
| **Attendees:** | Yong Lee Jia, Lim Wei Fa, Low Jianhao Jonathan |
| **Absent:** |  |
| **Agenda:** | 1. Determine scope of project 2. Clarify doubts of the project |

1. **Initial Analysis sample data and metadata provided?**

* We are hoping to get the data from MTI by Jan, not confirmed yet.
* It is not too difficult to scrape the jobsbank site for data, have existing scripts to scrape
* Require only job description and job requirements field
* Want to build a classifier to extract skills (previously using whitelist/blacklist)
* Have 1 month worth of data for Jan 2016

1. **Working on top of the prototype?**

* Nope, it’s a total rework (want to build a more production application). Building a classifier requires labelled data which is not available, have only raw data

1. **Languages and software that we will have to use for the project?**

* Python, basic machine learning - coursera and udacity
* Tensorflow - <https://www.tensorflow.org/>
* Maybe can use DNN to do the prediction or can use other models too
* Jupyter notebook: <http://jupyter.org/> - best IDE for data science
* Standard machine learning lib in python - <http://scikit-learn.org/> look thru the various technique available and focus on a few u think most applicable
* Standard text processing lib in python - <http://www.nltk.org/> read thru some examples to learn general concept on NLP
* Focus on tensorflow or scikit-learn. 1 learn tensor flow, 1 learn scikit learn
* Tensorflow is deep learning. scikit-learn is the std machine learning
* Priority - general python first, nltk and pandas <http://pandas.pydata.org/>

1. **For one of the subtasks - generating labels to train the classifier, could you further elaborate more on what we are supposed to do?**

* Labelled data means the data is annotated with the target u want to predict.
* Want to predict whether a word in the job description is a skill relevant to the job
* Need to annotate a set of job descriptions with the correct words that are skills
* For each word in the job description, we label a true or a false, then use this annotated data to train a model
* Before u train a model, u need to select the features.
* as in what features are good predictors that whether a word is a skill or not.
* so, the step to do in a typical data science project is  
  1. clean and prepare the data  
  2. feature engineering aka select good features  
  3. train a model with the feature set, and test against the validation set to see how good your model is  
  4. repeat 2 & 3 until u get a good enough model
* Labelling of data is a bit manual.
* What I proposing is to scrap online for some standard skills, like python, javascript, etc.
* then use that as a whitelist to label some of the text.
* then see that is enough to train a good model
* if not, need to manually go thru the text, and improve the annotation.
* first half of the project will mostly be preparing the data.
* if you can get the model up fast enough, we can go a step further, and use these skills to cluster similar jobs together

1. **Objectives of the project is to build a model to extract skills from job postings in** [**jobsbank.gov.sg**](http://jobsbank.gov.sg)**?**

* Dump a job posting text into the model, the model dump out a list of skills

1. **We are supposed to:**

* come up with that model that we will have to train and validate to increase its accuracy that’s the main objective and if we could finish the model in time, then we may proceed on to the clustering of jobs using the skills generated by our model

1. **Motivation for the project as we would have to include in the proposal?**

* the motivation is to identify skill mismatch and emerging skills
* basically, we know jobs that no one is applying and the pay is high
* so, a possible reason is Singaporean dun have the skill
* so, we want to identify what sort of skills that are need by industry but Singaporean do not have
* so, we can come up with policy or training to build those skills.
* similarly, we also want to monitor the jobs over time
* to see if there are any emerging skills
* for e.g. reactjs was developed 2,3 years ago
* it became popular like 1,2 years ago
* u start seeing more jobs requiring reactjs
* so, we want to monitor and detect new skills.
* that is why a whitelist approach is not enough
* because whitelist will never catch new or unknown skills.

1. **“We”: do u mean govtech or MTI?**

* by a lot of parties
* we are building an application
* MTI is the product owner
* but the users are a lot of different agencies
* mostly the various manpower planning units
* so, it is going to be part of a cloud service to various government agencies

1. **For the time being, we will come up with a rough proposal for the project and we will send to u to validate when we are done**