**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**Work Integrated Learning Programmes Division**

**Post Graduate Program in**

**Artificial Intelligence and Machine Learning**

**NLP system for creating Relevant Question and Answer based on job role**

CAPSTONE PROJECT

Submitted in partial fulfillment of the requirements of the Post Graduate Certification Program

in

Artificial Intelligence and Machine Learning By

|  |  |  |
| --- | --- | --- |
| SL No. | Student Name | BITS ID |
| 1 | Dheeraj TV | 2019AIML623 |
| 2 | Kayamkhani Taher Abbasbhai | 2019AIML599 |
| 3 | Komal Batra | 2019AIML649 |
| 4 | Pavan IV | 2019AIML626 |

Under the supervision of Shubhendu Jain

Project work carried out at

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Rajasthan) INDIA (April, 2021)

## Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of our project. All that we have done is only due to such supervision and assistance and I would not forget to thank them.

We would like to thank all the professors and faculty, for providing us an opportunity to do the project work in BITS and giving us all support and guidance, which made us complete the project duly. We are extremely thankful to Mr. Shubhendu Jain for providing such a nice support and guidance, although he had a busy schedule still managed to provide enough time for us.

We owe my deep gratitude to our project guide prof. Sudarshan S Deshmukh, who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system along with his feedback and industry expertise

We would not forget to remember all the TAs for their encouragement and more over for their timely support for any required information and guidance till the completion of our project work.

We are thankful and fortunate enough to get constant encouragement, support and guidance from all Teaching staff of WILP AIML course which helped us in successfully completing our project work. Also, we would like to extend our sincere esteems to all sporting staff for their timely support.

Thanks & Regards, Student Group 32

## BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

**CERTIFICATE**

This is to certify that the Capstone Project entitled NLP system for creating Relevant Question and Answer based on job role and submitted by Mr./Ms. Dheeraj TV, Kayamkhani Taher Abbasbhai, Komal Batra and Pavan IV with ID No. 2019AIML623, 2019AIML599, 2019AIML649 and 2019AIML626 in partial fulfillment of the requirements of PCAM ZC321 Capstone Project, embodies the work done by him/her under my supervision.

Place : Bangalore Signature of the Mentor

Date : 31/05/2021 Name : Shubhendu Jain

## BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI SECOND SEMESTER 2019-20

**PCAM ZC321 CAPSTONE PROJECT**

Project Title : NLP system for creating Relevant Question and Answer based on job role

Name of Mentor : Mr. Shubhendu Jain

Name of Student : Dheeraj TV, Kayamkhani Taher Abbasbhai, Komal Batra and Pavan IV

ID No. of Student : 2019AIML623, 2019AIML599, 2019AIML649 and 2019AIML626

# Abstract

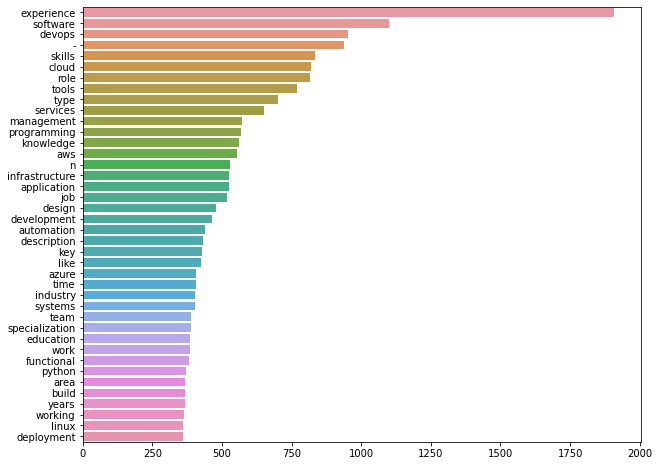
We present a methodology, based on Natural Language Processing, to cluster the input source data as per the standard feature engineering and algorithm thereby ensuring the optimum performance. As part of solution, underlying steps like Data cleaning, Language translation, Ranking of Organization feature, Creating dictionary of Organization feature, Ranking of University feature, Creating dictionary of University feature, Binning of education feature, Creating dictionary of Education feature, Data Visualization ( Scatter Plots, Line Plots), Model Building & Evaluation. Also, different clustering methods have been exercised like Density-Based, Hierarchical Based (Agglomerative and Divisive), Partitioning, Grid-based.

The project aims at providing more reliable way for evaluating the rank of a user across given collection of users by analyzing the features like education, experience and employment history. The input data has been refined and data dictionaries have been created to extract organization, university and education features. The input dataset has been referred against these master lookup data and the respective rankings are derived for Companies, University. Also, few additional features like certificate completion count, etc have been added to improve the output accuracy of model. This paper analyzes the performance aspect with three algorithms being implemented and exercised for optimum correctness. This thesis will walk you through the detailed insights on the use case and implementation thereby keeping into consideration the performance aspects.

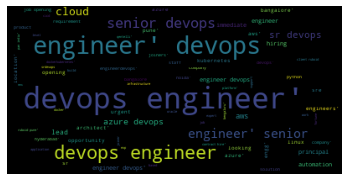
## List of Data Dictionaries Used

## List of Figures-S

Bar Plots:



Word Cloud:



## Table Of Contents

|  |  |  |
| --- | --- | --- |
| **Sl No.** | **Chapter** | **Page No.** |
| **1** | **Synopsys** | **11** |
| **2** | **Business Understanding/Problem statement** | **11** |
| **3** | **Objective Of The Project** | **11** |
| **4** | **Conclusions / Recommendations** | **12** |
| **5** | **Background of previous work done in the chosen area** | **12** |
| **6** | **NLP Approach Used** | **12** |
| **7** | **Resources needed for the project** | **14** |
| **8** | **NLP process flow & Solution Architecture** | **15** |
| **9** | **Resources needed for the project** | **16** |
| **10** | **Potential data challenges & risks in doing the project** | **17** |
| **11** | **Detailed Plan of Work** | **17** |
| **12** | **NLP Modeling & Techniques Applied** | **18** |
| **13** | **Code Screenshot** | **20** |
| **14** | **Interpretation** | **35** |
| **15** | **Future Work & Extension or Scope of improvements** | **36** |
| **16** | **Checklist of items for the Final report** | **37** |

**Synopsis**

Natural Language Processing is becoming an integral part of most of the organizations, where NLP techniques are used to simplify the task for the employees.

Talent acquisition is one of the most critical part of any organization, since it provides the correct resource for particular job, so we have developed a system using NLP technique to simplify the task where interviewer can get Q&A based on the job description.

In this project Word Embedding techniques are used to classify the right set of questions which interviewer can use while screening a candidate. This eliminates manual task of searching for relevant questions based on the job description.

Basically this model can be used by the organization for hiring new resources or during internal job posting of an existing employee.

## Business Understanding/Problem statement

Job description is important to understand the requirement. Inconsistent job descriptions lead to improper selection which affects the ROI. Based on Role and years of experience, an AI based solution is expected to write the job description so that it will avoid the above-mentioned problem.

In today’s parlance, IT managers and hiring leads are finding it tough to hire the right talent swiftly and consistently. There seems to be a lot of gap between what the applicants for a specific job has put in resume vs. what they know. Because of this there is a lot of time wasted in interview process.

## Objective Of The Project

The objective of this project is to build a model which can effectively parse candidates understanding of subject by instantaneously posting questions given a job description.

## Conclusions/Recommendations

Hiring team will be easily able to filter out the right question for the candidate based on the job description.

## Background of previous work done in the chosen area

Currently the SQuAD Dataset is the only dataset which is being used in this line of work. SQuAD stands for Stanford Question Answering Dataset ([SQuAD](https://rajpurkar.github.io/SQuAD-explorer/)) is a new reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. With 100,000+ question-answer pairs on 500+ articles, SQuAD is significantly larger than previous reading comprehension datasets.

There has been a rapid progress on the SQuAD dataset with some of the latest models achieving human level accuracy in the task of question answering!

## NLP Approach Used

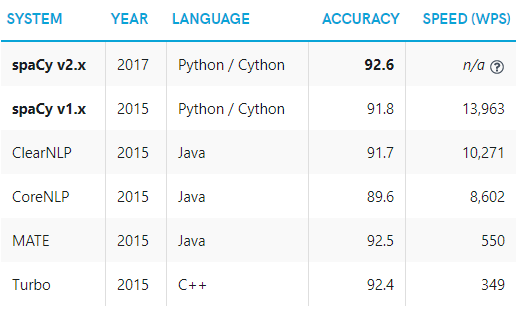
1. **Named Entity Recognition for Entity Extraction:**

Named entity recognition (NER) assigns a named entity tag to a designated word by using rules and heuristics. The named entity, which presents a human, location, and an organization, should be recognized. Named entity recognition is a task that extracts nominal and numeric information from a document and classifies the word into a person, an organization, or a date category. NER classifies all words in the document into existing categories.

Name entity recognition is very important in language processing of job description texts, especially in extracting information of job description from text corpus. Finding named entities of job description from texts is a very important and difficult task. Finding a description in texts corresponds to finding a company name or a human name in newspapers. Recognizing specific field related named entities seems to be more difficult than recognizing normal named entities. Numerous research studies have recognized named entities by using supervised learning algorithms based on many rules.

Named entity recognition has three approaches dictionary based, rule based, and machine learning based. A dictionary-based approach stores as many named entities as possible in a list called a gazetteer. This approach seems to be very simple, but at the same time has limitations. The NER is difficult because the target words are mainly proper nouns or unregistered words. In addition, new words can be generated frequently, and even the same word stream could be recognized as diverse named entities in terms of their current context. The second approach of the NER is a rule-based approach. This approach ordinarily depends on the rules and patterns of named entities appearing in real sentences. Although rule-based approaches can use context to solve the problem of multiple named entities, every rule should be written before it is actually used. The third approach, the machine learning-based approach, tags the named entities to words even when the words are not listed in the dictionary and the context is not described in the rule set. For these approaches, support vector machines (SVMs), Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), and conditional random fields (CRFs) are mainly utilized.

One way to get more relevant job recommendations is to classify words into categories such as Skills, Experience, Degree, etc. (entities) instead of searching for static keywords. Once entities are extracted from job descriptions, we can perform similarity analysis to the job description to get more relevant Q&A recommendations. For this project, we have used NER machine learning tool to extract relevant entities from job postings. There are several NER tools available such as Stanford NER, NLTK, Spacy, etc. I chose the open source Spacy library because it is fast and has the highest accuracy as shown in the table below:



NLP models accuracy and speed comparison.

## NER Algorithms:

* **Conditional random fields**

A CRF is a statistical sequence modeling framework. CRFs are a class of statistical modeling methods often applied in pattern recognition and machine learning, where they are used for structure prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to “neighboring” samples, a CRF can take context into account. The reason why CRFs are more effective than HMMs is that CRFs use the conditional probability property instead of the independence assumption mainly used in HMMs. CRFs also avoid label bias problems and avoid the weaknesses of other Markov models derived from MEMMs and graphic models. CRFs show better performance than MEMMs and HMMs in bioinformatics, computational linguistics, and voice recognition. CRFs are also used for the prediction and analysis of labels for data in natural language writing. Features can be chosen randomly, and they are to be normalized to obtain solution.

* **Word embedding**

Word embedding is also called word representation or distributed representation. It learns vector representation for every word appearing in the corpus. Previous research studies represented a word as a one-hot representation. The one-hot representation uses a vocabulary-sized vector, and takes a 1 when the word appears in the document and 0 when it does not. Word embedding reduces the dimensions and sparseness of the original vector and fills the vector with real numbers. Figure [1](https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0573-6#Fig1) shows the difference between one-hot representation and word embedding.

* **Word2Vec**

Word2Vec assumes that the words sharing the same context could have similar meanings. It classifies words near to the given word into related words and learns the words using artificial neural networks. Word2Vec has two structures: Continuous Bag of Words (CBOW) and skip gram architectures. Figure [2](https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0573-6#Fig2) shows the Word2Vec architecture.

* **Global vector (GloVe)**

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. GloVe considers the global context as well as the local context.

* **Canonical correlation analysis (CCA)**

Canonical correlation analysis (CCA) was introduced by Hotelling. CCA is a statistical method to investigate the relationship between two variable sets, and it can concurrently examine the correlation of variables belonging to different sets. CCA finds correlations between two variable sets (X, Y), and also finds parameters that maximize the correlation coefficients. CCA can be calculated directly from the data set, and can also be calculated after transforming the data sets into covariance matrices. These two methods are represented based on singular value decomposition. If CCA is used to predict labels in data, string theory guarantees the correspondence to lower-dimensional embedding. CCA tries to find two projection vectors to maximize the correlation. Using random variables (X, Y∈ R), where X is a word representation and Y is its related context representation, CCA tries to find k-dimensional projection vectors that maximize the correlation between two variables.

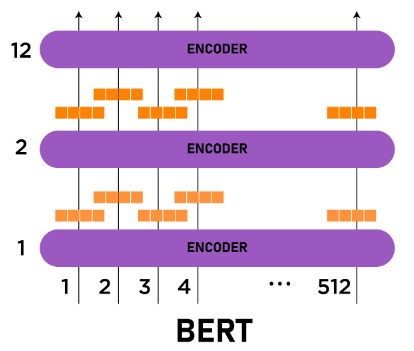
* **Recurrent neural network**

In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks that are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. ANNs work well in nonlinear functions and pattern recognition. Many researchers working in data mining, artificial intelligence, and bioinformatics have been interested in ANNs for its diverse applications.

* **BERT**

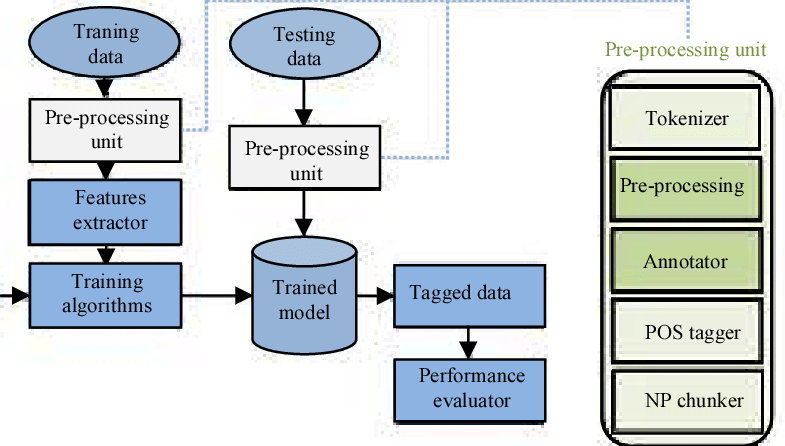
BERT (Bidirectional Encoder Representations from Transformers) is a recent [paper](https://arxiv.org/pdf/1810.04805.pdf) published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering, Natural Language Inference (MNLI), and others.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT’s goal is to generate a language model, only the encoder mechanism is necessary.



## NLP process flow & Solution Architecture

1. **Overview of the Workflow of NLP:**

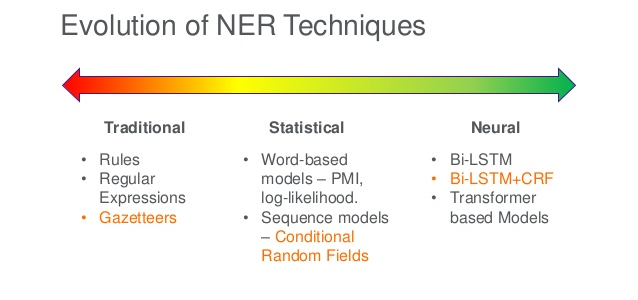
****

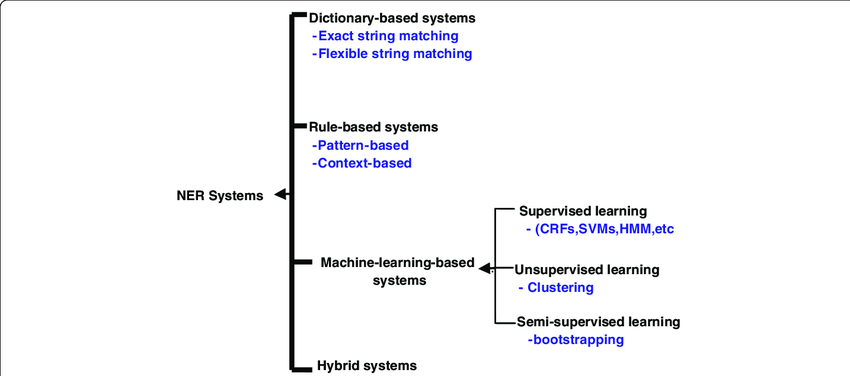
1. **We can define the NLP workflow in 3 stages**
2. **Data Collection:** The data set can be collected from various sources such as a web portals, file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.
3. **Data pre-processing:** These are some of the pre-processing techniques that can be used to convert raw data

* **Conversion of data**: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.
* **Ignoring the missing values**: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient, but it shouldn’t be performed if there are a lot of missing values in the dataset.
* **Filling the missing values**: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

## Researching the model that will be best for the type of data

Finding the use case and correct approach from NLP techniques is the key to best model selection.





1. **Overview of model**

* For training a model we initially split the model into 3 three sections which are ‘Training data’, ‘Validation data’ and ‘Testing data’.
* You train the classifier using ‘training data set’, tune the parameters using ‘validation set’ and then test the performance of your classifier on unseen ‘test data set’. An important point to note is that during training the classifier only the training and/or validation set is

available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

1. **Evaluation**

* Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.
* To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also look at the confusion matrix to try to increase the number of true positives and true negatives.

**Resources needed for the project**

|  |  |  |
| --- | --- | --- |
| Sl No. | Resource Type | Detail |
| 1 | People | 4 (Team Members) |
| 2 | Software | Python3 (scikit learn), MS Excel |

## Potential data challenges & risks in doing the project

Most of the real-world data is messy, some of these types of data are:

1. Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
2. Noisy data: This type of data is also called outliers; this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
3. Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.
4. There are a several problems with clustering, few are listed below:
5. Current clustering techniques do not address all the requirements adequately (and concurrently);
6. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
7. The effectiveness of the method depends on the definition of “distance” (for distance-based clustering);
8. If an obvious distance measure doesn’t exist, we must “define” it, which is not always easy, especially in multi-dimensional spaces.
9. The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

## Detailed Plan of Work

Overall project having three stages, each step is explained below

## Data cleaning and preprocessing of raw data

1. Using label encoding & google translator libraries, all the other languages data changed to English
2. Column name correction
3. Special characters removing





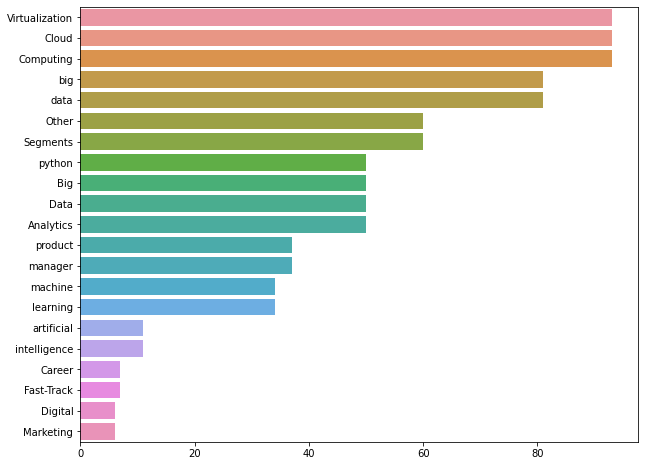
## Code Screenshots

**Data Cleaning Code Screenshot**

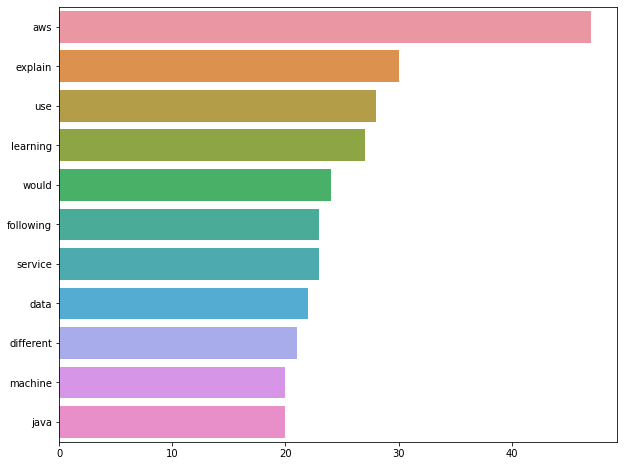


**Exploratory Data Analysis**

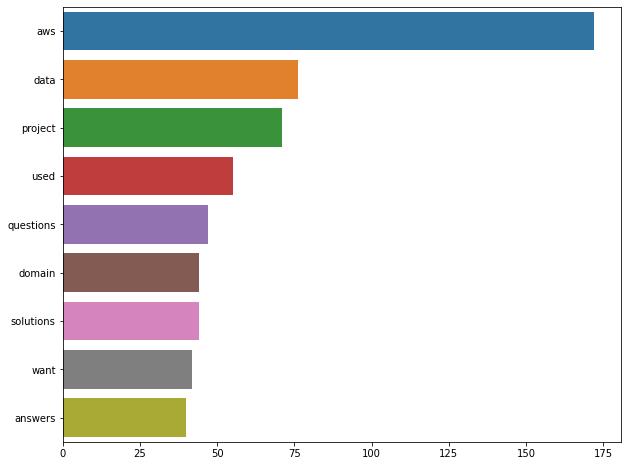
Top Skills Bar Chart



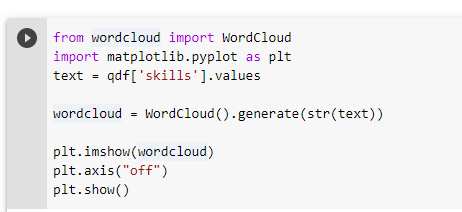
Top Questions Bar Chart



Top Answers Bar Chart



Word Cloud Code

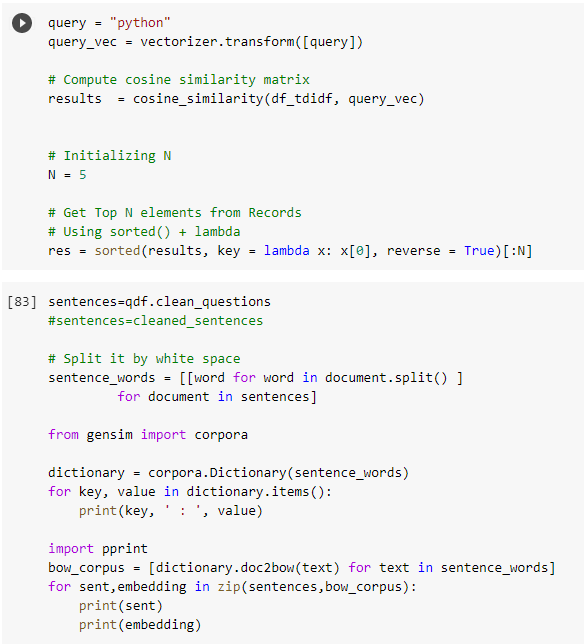




## Directions for future work

1. Complete tag identification of the queries

We have used BOW (Bag-Of-Words) for word or sentence embeddings.



1. Build front end UI - WIP
2. Containerize the application - WIP

## Containerization:

Containerization allows developers to create and deploy applications faster and more securely. With traditional methods, code is developed in a specific computing environment which, when transferred to a new location, often results in bugs and errors. For example, when a developer transfers code from a desktop computer to a virtual machine (VM) or from a Linux to a Windows operating system. Containerization eliminates this problem by bundling the application code together with the related configuration files, libraries, and dependencies required for it to run. This single package of software or “container” is abstracted away from the host operating system, and hence, it stands alone and becomes portable—able to run across any platform or cloud, free of issues.

Put simply, containerization allows applications to be “written once and run anywhere.” This portability is important in terms of the development process and vendor compatibility. It also offers other notable benefits, like fault isolation, ease of management and security, to name a few.

It is typically used in DevOps and agile CI/CD.

Containerization offers significant benefits to developers and development teams. Among these are the following:

* **Portability**
* **Agility**
* **Speed**
* **Fault isolation**
* **Efficiency**
* **Ease of management**
* **Security**

In a nutshell, virtualization eliminates the need for an entire server for one application. Containerization eliminates the need for an entire OS for each application. Portability, agility, fault isolation, ease of management, and security are among the advantages of utilizing containerization technology.

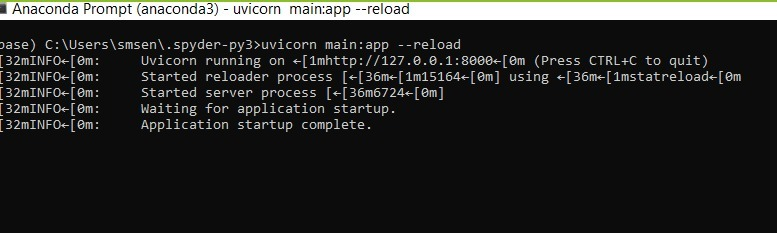
 Build containerized applications using continuous integration and continuous delivery (CI/CD) tools. Orchestrate containers using managed Red Hat OpenShift or Kubernetes services. And modernize existing applications one can use open source components in IBM Cloud Paks.

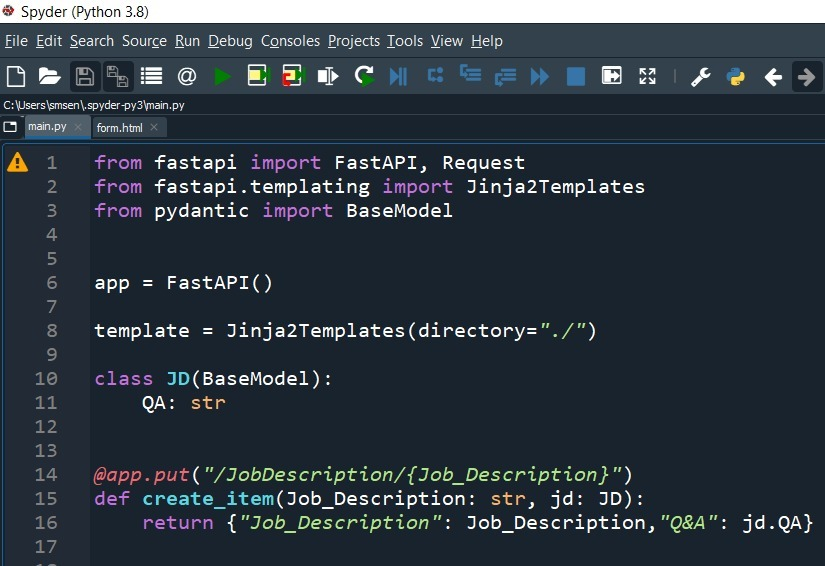
## How to deploy this project using microservice:

## Make your model for which you want to create the API ready

## Create API using FastAPI framework







## Checklist of items for the Final report

1. Is the Cover page in proper format? Y / N
2. Is the Title page in proper format? Y / N
3. Is the Certificate from the Mentor in proper format? Has it been signed? Y / N
4. Is Abstract included in the Report? Is it properly written? Y / N
5. Does the Table of Contents page include chapter page numbers? Y / N
6. Does the Report contain a summary of the literature survey? Y / N
   1. Are the Pages numbered properly? Y / N
   2. Are the Figures numbered properly? Y / N
   3. Are the Tables numbered properly? Y / N
   4. Are the Captions for the Figures and Tables proper? Y / N
   5. Are the Appendices numbered? Y / N
7. Does the Report have Conclusion / Recommendations of the work? Y / N
8. Are References/Bibliography given in the Report? Y / N
9. Have the References been cited in the Report? Y / N
10. Is the citation of References / Bibliography in proper format? Y / N