Automated Question Response from FAQ using Word Embedding

Submitted in partial fulfillment of the requirements
of the degree of

BACHELOR OF ENGINEERING

In

COMPUTER ENGINEERING

By

Group No: 11

Roll Number	Name
1704051	Harsh Harwani
1704053	Chinmay Jadhav
1704055	Jenil Jain

Guide:

Prof. Jayant Gadge

Computer Engineering Department Thadomal Shahani Engineering College University of Mumbai 2020-2021

Chapter 1

Problem Statement

The purpose of this project is to answer user queries by automatically retrieving the closest question and answer from predefined FAQs when appropriate. In this project we examine the task of automatically retrieving a suitable response to customer questions from FAQs.

We will use a sample dataset of FAQs extracted from the site https://machinelearninginterview.com for this task. This dataset can be replaced with a more elaborate dataset as appropriate. Our basic strategy is to find the FAQ question that is closest in meaning to the user query and display it to the user.

Chapter 2

Input and Output

Sample Input	Sample Output
“What does a data scientist usually do”	(Most similar question) “How does a typical day of data scientist look like”

Chapter 3

Dataset Description

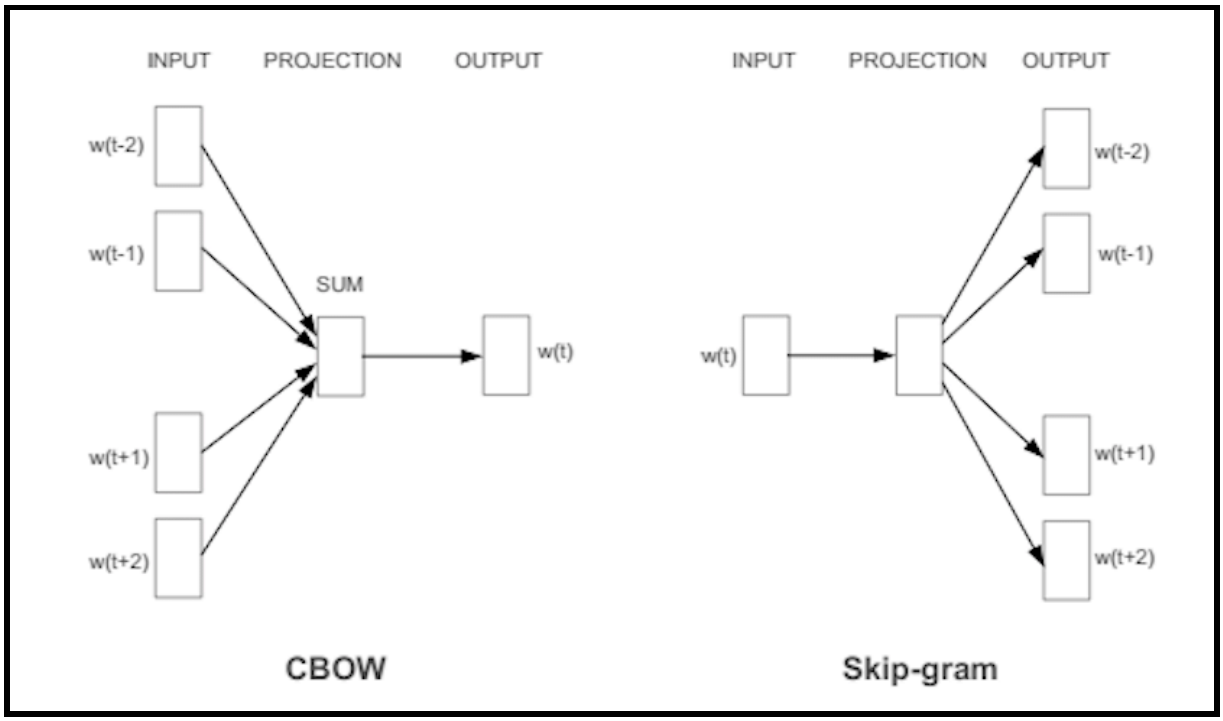
As you can see, the dataset contains two columns, first question and second, the corresponding answers. Here is the sample view of the dataset.
There are 10 records in the dataset.

```
import pandas as pd;

#Load dataset and examine dataset, rename columns to questions and answers

df=pd.read_csv("FAQ_NLP.csv");
df.columns=["questions","answers"];

df
```

	questions	answers
0	What does the job hunting experience look like ?	Job hunting experience involves networking to ...
1	Any insights you can offer about the DS job ma...	There are many kinds of roles, data scientist,...
2	What's the impact of Covid on hiring for DS ro...	Hiring is going to slow down. First in small c...
3	What skills and qualities do employers look fo...	The following are some skills employers usuall...
4	Do employers look for an advanced ML degree?	For more senior roles: People typically look f...
5	How does a typical day of a data scientist loo...	Here are some tasks in the typical day of a da...
6	Is preparation of algorithms and data structur...	Yes. In many data science interviews (ML Scien...
7	What is the mathematical background required t...	The following three are the basic building blo...
8	What are the various rounds in a data scientis...	Usually the data science interview has a subse...
9	What level of proficiency is needed for a data...	Needs to be reasonably proficient. Again, a da...

Chapter 4

Architecture Diagram

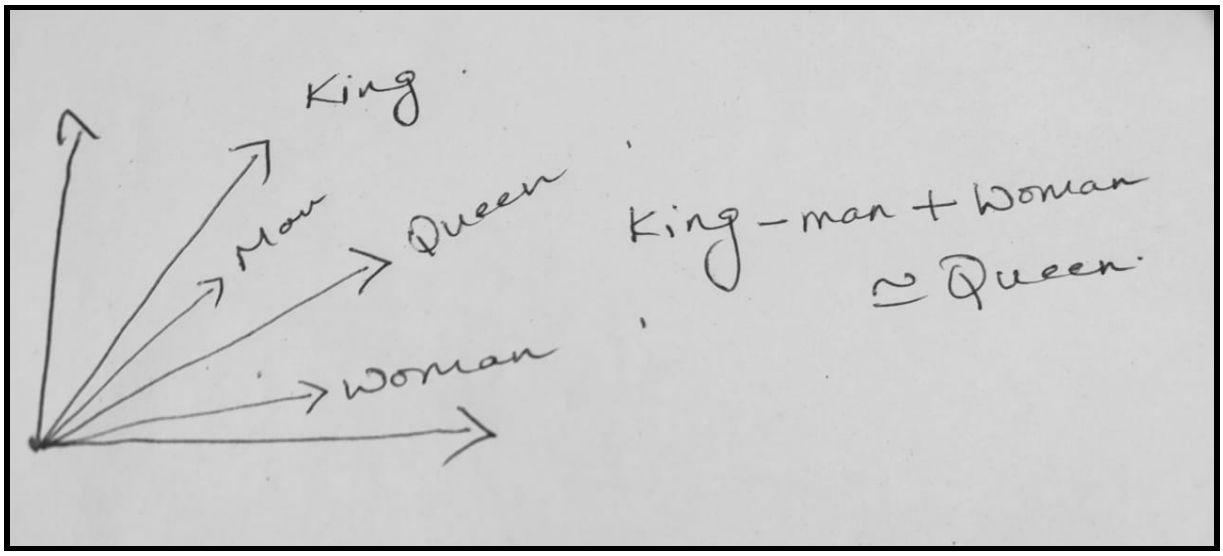


Chapter 5

Word Embedding Technique

Word2Vec Embeddings

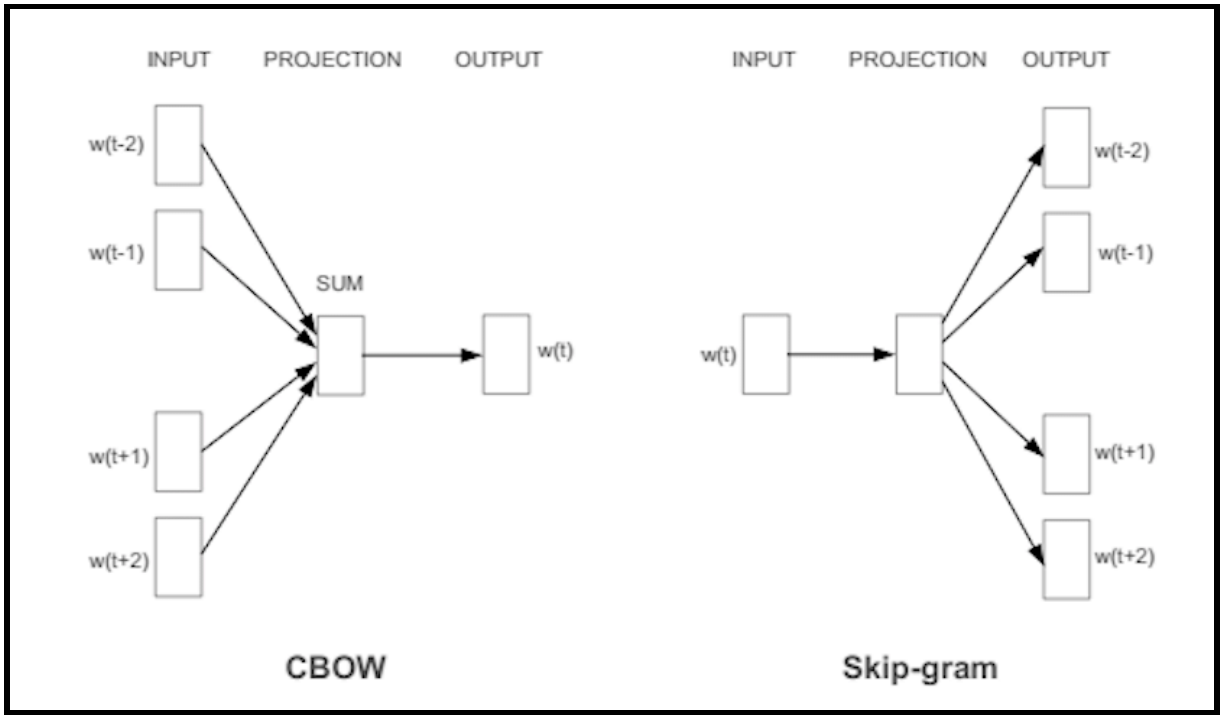
Word2Vec embeddings are popularly trained using the skipgram model. These embeddings are trained to take a word as input and reconstruct its context. As a result, they are able to take into account semantic similarity of words based on context information. The resulting embeddings are such that words with similar meaning tend to be closer in terms of cosine similarity.



Skipgram model

The most popular word2vec model is the skipgram model. Particularly, the most commonly used pre-trained model is based on the Google News dataset that has 3 billion running words and creates upto 300 dimensional embedding for 3 Million words.
Skip-gram is used to predict the context word for a given target word. It's reverse of CBOW algorithm. Here, target word is input while context words are output. As

there is more than one context word to be predicted which makes this problem difficult.



From the above diagram, we can see w(t) is the target word or input given. There is one hidden layer which performs the dot product between the weight matrix and the input vector w(t). No activation function is used in the hidden layer. Now the result of the dot product at the hidden layer is passed to the output layer.

Output layer computes the dot product between the output vector of the hidden layer and the weight matrix of the output layer. Then we apply the softmax activation function to compute the probability of words appearing to be in the context of w(t) at given context location.

Libraries Used

- Word2Vec
- Genism
- Pandas
- RE (regular expression)

Advantages

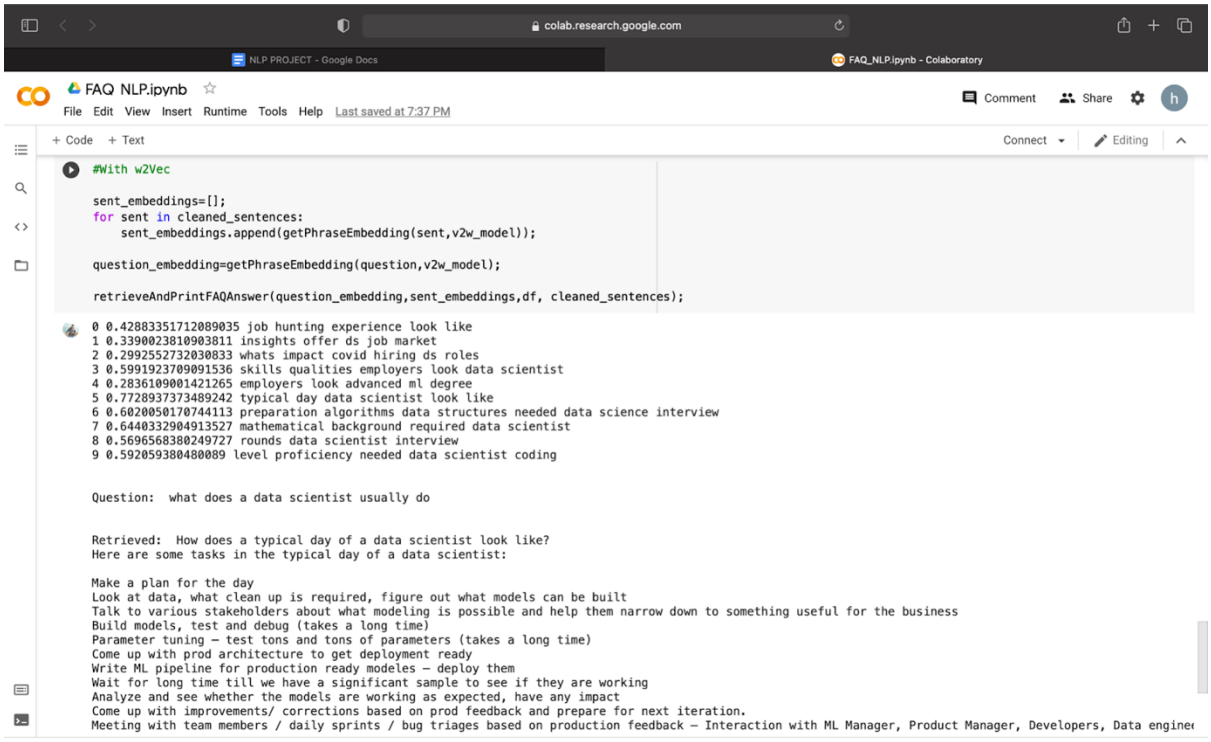
1. It is unsupervised learning hence can work on any raw text given.
2. It requires less memory comparing with other words to vector representations.

3. It requires two weight matrix of dimension [N, |v|] each instead of [|v|, |v|]. And usually, N is around 300 while |v| is in millions. So, we can see the advantage of using this algorithm.

Chapter 6

Results

Here, we can see when we provide input question to the user as,
“What does a data scientist usually do”.
We get the most similar question that is present in our dataset, viz. “How does a typical day of data scientist look like” as this question has the highest cosine similarity of 0.77


colab.research.google.com
NLP PROJECT - Google Docs
FAQ_NLP.ipynb - Colaboratory
FAQ NLP.ipynb
File Edit View Insert Runtime Tools Help Last saved at 7:37 PM
Comment
Share
+ Code + Text
Connect
Editing
#With w2Vec
sent_embeddings=[];
for sent in cleaned_sentences:
sent_embeddings.append(getPhraseEmbedding(sent,v2w_model));
question_embedding=getPhraseEmbedding(question,v2w_model);
retrieveAndPrintFAQAnswer(question_embedding,sent_embeddings,df, cleaned_sentences);
0 0.42883351712089035 job hunting experience look like
1 0.3390023810903811 insights offer ds job market
2 0.2992552732030833 whats impact covid hiring ds roles
3 0.5991923709091536 skills qualities employers look data scientist
4 0.2836109001421265 employers look advanced ml degree
5 0.7728937373489242 typical day data scientist look like
6 0.6020050170744113 preparation algorithms data structures needed data science interview
7 0.6440332904913527 mathematical background required data scientist
8 0.5696568380249727 rounds data scientist interview
9 0.592059380480089 level proficiency needed data scientist coding
Question: what does a data scientist usually do
Retrieved: How does a typical day of a data scientist look like?
Here are some tasks in the typical day of a data scientist:
Make a plan for the day
Look at data, what clean up is required, figure out what models can be built
Talk to various stakeholders about what modeling is possible and help them narrow down to something useful for the business
Build models, test and debug (takes a long time)
Parameter tuning – test tons and tons of parameters (takes a long time)
Come up with prod architecture to get deployment ready
Write ML pipeline for production ready modeles – deploy them
Wait for long time till we have a significant sample to see if they are working
Analyze and see whether the models are working as expected, have any impact
Come up with improvements/ corrections based on prod feedback and prepare for next iteration.
Meeting with team members / daily sprints / bug triages based on production feedback – Interaction with ML Manager, Product Manager, Developers, Data engine