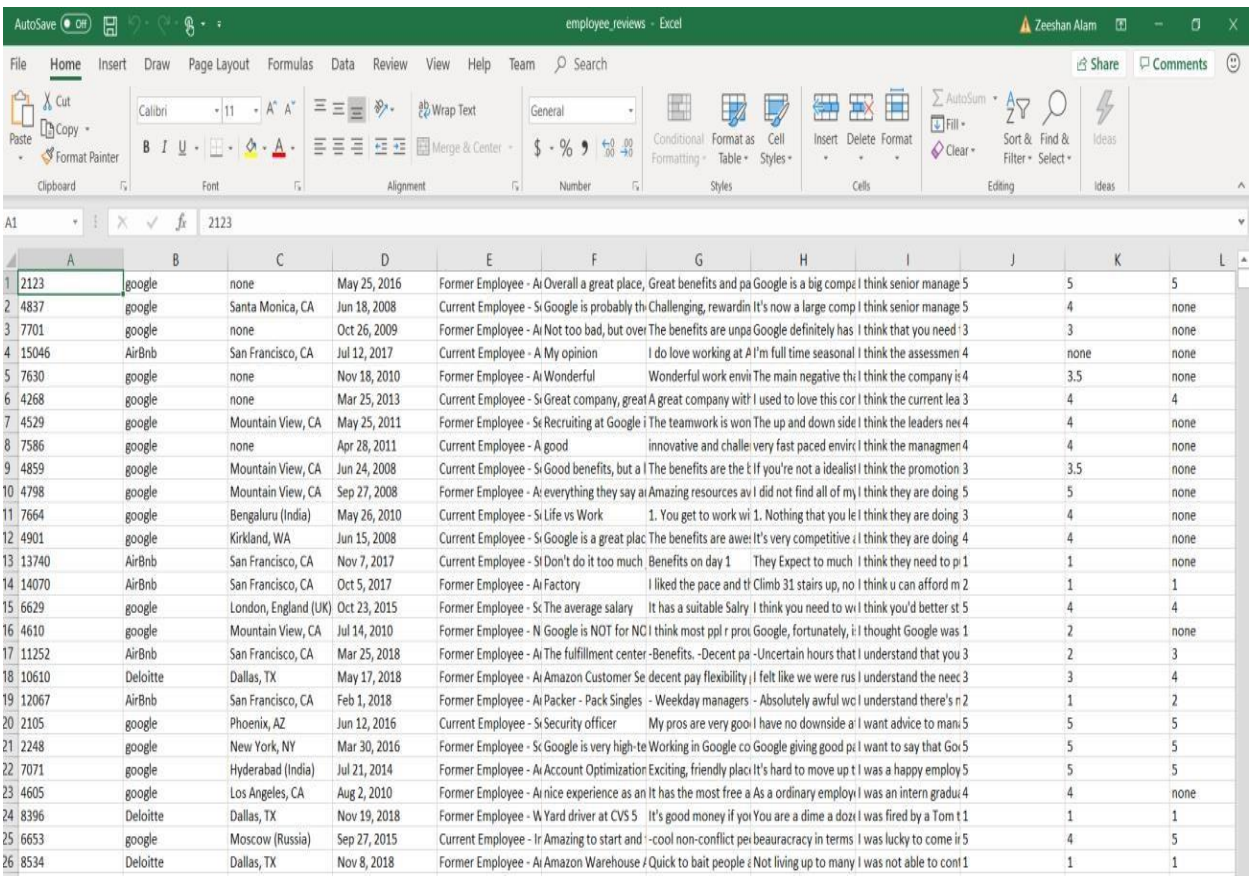
Zeeshan Alam
1001577559

Development Phase I: Search

https://dmsearch.herokuapp.com/

Dataset:



	A	B	C	D	E	F	G	H	I	J	K	L
1	2123	google	none	May 25, 2016	Former Employee - A	Overall a great place,	Great benefits and pa	Google is a big compa	I think senior manage	5	5	5
2	4837	google	Santa Monica, CA	Jun 18, 2008	Current Employee - S	Google is probably th	Challenging, rewardin	It's now a large comp	I think senior manage	5	4	none
3	7701	google	none	Oct 26, 2009	Former Employee - A	Not too bad, but over	The benefits are unpa	Google definitely has	I think that you need	3	3	none
4	15046	AirBnb	San Francisco, CA	Jul 12, 2017	Current Employee - A	My opinion	I do love working at A	I'm full time seasonal	I think the assessmen	4	none	none
5	7630	google	none	Nov 18, 2010	Former Employee - A	Wonderful	Wonderful work envi	The main negative th	I think the company i	4	3.5	none
6	4268	google	none	Mar 25, 2013	Current Employee - S	Great company, great	A great company with	I used to love this cor	I think the current lea	3	4	4
7	4529	google	Mountain View, CA	May 25, 2011	Former Employee - S	Recruiting at Google i	The teamwork is won	The up and down side	I think the leaders nee	4	4	none
8	7586	google	none	Apr 28, 2011	Current Employee - A	good	innovative and challe	very fast paced envir	I think the managmen	4	4	none
9	4859	google	Mountain View, CA	Jun 24, 2008	Current Employee - S	Good benefits, but a l	The benefits are the b	If you're not a idealist	I think the promotion	3	3.5	none
10	4798	google	Mountain View, CA	Sep 27, 2008	Former Employee - A	everything they say a	Amazing resources av	I did not find all of m	I think they are doing	5	5	none
11	7664	google	Bengaluru (India)	May 26, 2010	Current Employee - S	Life vs Work	1. You get to work wi	1. Nothing that you le	I think they are doing	3	4	none
12	4901	google	Kirkland, WA	Jun 15, 2008	Current Employee - S	Google is a great plac	The benefits are awe	It's very competitive a	I think they are doing	4	4	none
13	13740	AirBnb	San Francisco, CA	Nov 7, 2017	Current Employee - S	Don't do it too much	Benefits on day 1	They Expect to much	I think they need to p	1	1	none
14	14070	AirBnb	San Francisco, CA	Oct 5, 2017	Former Employee - A	Factory	I liked the pace and th	Climb 31 stairs up, no	I think u can afford m	2	1	1
15	6629	google	London, England (UK)	Oct 23, 2015	Former Employee - S	The average salary	It has a suitable Salry	I think you need to w	I think you'd better st	5	4	4
16	4610	google	Mountain View, CA	Jul 14, 2010	Former Employee - N	Google is NOT for NO	I think most ppl r pro	Google, fortunately, i	I thought Google was	1	2	none
17	11252	AirBnb	San Francisco, CA	Mar 25, 2018	Former Employee - A	The fulfillment center	-Benefits. -Decent pa	-Uncertain hours that	I understand that you	3	2	3
18	10610	Deloitte	Dallas, TX	May 17, 2018	Former Employee - A	Amazon Customer Se	decent pay flexibility	I felt like we were rus	I understand the neec	3	3	4
19	12067	AirBnb	San Francisco, CA	Feb 1, 2018	Former Employee - A	Packer - Pack Singles	- Weekday managers	- Absolutely awful wo	I understand there's n	2	1	2
20	2105	google	Phoenix, AZ	Jun 12, 2016	Current Employee - S	Security officer	My pros are very goo	I have no downside a	I want advice to man	5	5	5
21	2248	google	New York, NY	Mar 30, 2016	Former Employee - S	Google is very high-te	Working in Google co	Google giving good pa	I want to say that Go	5	5	5
22	7071	google	Hyderabad (India)	Jul 21, 2014	Former Employee - A	Account Optimization	Exciting, friendly plac	It's hard to move up t	I was a happy employ	5	5	5
23	4605	google	Los Angeles, CA	Aug 2, 2010	Former Employee - A	nice experience as an	It has the most free a	As a ordinary employ	I was an intern gradu	4	4	none
24	8396	Deloitte	Dallas, TX	Nov 19, 2018	Former Employee - W	Yard driver at CVS 5	It's good money if yo	You are a dime a doz	I was fired by a Tom t	1	1	1
25	6653	google	Moscow (Russia)	Sep 27, 2015	Current Employee - Ir	Amazing to start and	-cool non-conflict pe	beauracracy in terms	I was lucky to come i	5	4	5
26	8534	Deloitte	Dallas, TX	Nov 8, 2018	Former Employee - A	Amazon Warehouse /	Quick to bait people a	Not living up to many	I was not able to con	1	1	1

Implementation:

TF-IDF:

TF-IDF is a weighting factor for feature, the more weight the more that type of term occurs in the document which is offset by the number of times the words appear in the entire document which helps removing really common words (stop-words) in the language.

tf-idf(t,D) = tf(t,d) * idf(t,D)

tf(t,d_ = f(t|d)

the number of time single words appear in a given specific documents

idf(t,D) = log(N/|{dED = tEd}|)

the log is used to dumpen the effects of IDF function

<u>Smoothing for IDF</u> :log (1 + N/|{dED:tEd}|)

The smoothing factor is used for introducing the lower bound of log(2) so that nothing will ever be multiplied by 0 by the IDF

```
def index(Query):
    data = pd.read_csv('employee_reviews.csv')
    text = data['Pros']
    stop_words = stopwords.words('english')

    def process_text(text):
        text = str(text)
        text = re.sub('[^a-z\s]', '', text)
        text = [w for w in text.split() if w not in set(stop_words)]
        return ' '.join(text)

    data['Pros'] = data['Pros'].apply(process_text)

    english_stemmer = SnowballStemmer('english')
    analyzer = CountVectorizer().build_analyzer()

    def stemming(text):
        return (english_stemmer.stem(w) for w in analyzer(text))

    count = CountVectorizer(analyzer=stemming)

    count_matrix = count.fit_transform(data['Pros'])

    tfidf_transformer = TfidfTransformer()
    train_tfidf = tfidf_transformer.fit_transform(count_matrix)

    def get_search_results(query):
        query = process_text(query)
        query_matrix = count.transform([query])
        query_tfidf = tfidf_transformer.transform(query_matrix)
        sim_score = cosine_similarity(query_tfidf, train_tfidf)
        sorted_indexes = np.argsort(sim_score).tolist()
        return data['Company'].iloc[sorted_indexes[0][-10:]]

    working = get search results(Query)
```

Challenges:

There were many challenges in my path to complete the search feature. First and foremost was to learn and understand properly about the search algorithm used. Initially my idea was to go for a mobile application but later I found that Windows 10 Home does not support Hyper-V and I was not able to use emulator for testing and building purposes and on the top of that not having an android phone made it even tougher to work on the mobile application Therefore; I switched to web application using python Django framework.

I had to learn the framework by following different tutorials and the structure for the Django framework was pretty unique to all the other frameworks I have used in the past. After learning and understanding about the Django framework I was done with my web application front-end and I used libraries (python) and followed tutorials on the internet to finish my search feature.

Final step in the process was to find the public server to deploy my application for that I choose Heroku as it is one of the well-known free option available in the market. There were many issues when deploying the web application to the server and I had to check the log files every time to check errors in my web application.

References:

https://medium.freecodecamp.org/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3
https://towardsdatascience.com/tfidf-for-piece-of-text-in-python-43feccaa74f8
https://github.com/Heetmadhu/Movie-Recommendation/blob/master/MovieSearch.ipynb
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html