

Maestría En Ciencia de los Datos

Programación II

Analysis of comments on Glassdoor

Andrés Felipe Vázquez Valdés 224808017

andres.vazquez0801@alumnos.udg.mx



Description Project:

The project analyzed a Database from Glassdoor. This DB shows a list of employees with some opinions regarding the company they are working for. (Link on last page)

The objective of this project is to be able to categorize their comments into XOV (Bad, Neutral, Good). To do that, I took the information on Pros, Cons, and Headlines compared with the general opinion of those comments (column Recommend).

I had to merge the comments of the three columns, remove link words, and symbols.

Then pull the unigrams and bigrams (I made two models, one of each)

The Bigrams model gave an accuracy of 0.605 vs the accuracy of Unigrams 0.505.

Conclusion:

The model, with Bigrams worked much better than Unigrams. The thing here is that I had to drop a lot of the information to work with fewer employees because my computer wasn't able to process all the information.

It could be interesting to create another model using trigrams or more.

MLflow:

mlflow2.20.2

Experiments

Models

+

+

Search Experiments

☐ Default

☒ Deteccion de opiniones...

☐ WIDS2025

Deteccion de opiniones Empleados

Provide Feedback

Add Description

Share

Runs

Evaluation

Experimental

Traces

Q

metrics.rmse < 1 and params.model = "tree"

Time created

+

New run

State: Active

Datasets

Sort: Created

Columns

Group by

<input type="checkbox"/>	Run Name	Created	Dataset	Duration	Source	Mode
<input type="checkbox"/>	Model_Bigram	1 hour ago	-	33ms	ipykern...	-
<input type="checkbox"/>	Model_Unigram	1 hour ago	-	27ms	ipykern...	-
<input type="checkbox"/>	Model_Unigram	1 hour ago	-	20ms	ipykern...	-

Link to Kaggle to get the information: Glassdoor Job Reviews

<https://www.kaggle.com/davidgauthier/glassdoor-job-reviews>

Q

Search

76

<>

Code

Download

76

Code

Download

Glassdoor Job Reviews

A large dataset of job reviews with textual features and numerical targets

Data Card

Code (11)

Discussion (5)

Suggestions (0)

About Dataset

This large dataset contains job descriptions and rankings among various criteria such as work-life balance, income, culture, etc. The data covers the various industries in the UK. Great dataset for multidimensional sentiment analysis.

This data set complements the Glassdoor dataset located [here].

(<https://www.kaggle.com/davidgauthier/glassdoor-job-reviews-2>)

Usability

10.00

License

CC BY-SA 4.0

Expected update frequency

Never

Tags