5/26/2022

Machine Learning Assignment

Smit Kanjariya

smithkanzariya@gmail.com

Table of Contents

Introduction

Storyline

In a parallel universe, Amanda embarks on a journey to find a team of highly skilled software engineers for her startup. Mr. Robot and Silicon Valley were two of Amanda's favorite television series. But in this universe, Elliot Alderson works for the government, and Richard Hendricks does not destroy Pied Piper in the end. A world where a few lines of code could solve problems or create menace. She was intrigued when she got a chance to enroll in her first programming class. Unlike some other girls from her class, she was convinced that this was the path she wanted to take. During college, she completed multiple internships in Machine Learning, enhancing her skills and expertise. After her college, with a mission to build custom software for their clients, she started Amanda Coding LLC. She wants to build a team of highly skilled software engineers. For this, she wants to find teammates who meet all the technical skills as per the job requirement

1.1 Problem Statements

The Objective of this case study is to clean the data and extract Technical (Hard) Skills.

1.2 Dataset

The Dataset consist of more than 30k datapoints at which contain technical skills and a lot of jargon mixed in.

We need to construct a code to extract the Hard (Technical) Skills by clean the data.

Variable Name	Description
RAW DATA	Text data consist of hard skill and jargons

The sample of the data is shown below.

	RAW DATA
1	What ifs
2	seniority
3	familiarity
4	functionalities
5	Lambdas

Methodology

By observing the dataset we conclude that this is an Unsupervised Machine Learning Problem as there is no labeled data. To analyze the data missing values analysis was performed. As this is a text data NLP is used to process the data. To clean the data and to create a corpus Text Preprocessing was used.

2.1 Text Preprocessing

In the text preprocessing, removing of punctuation marks, splitting of text to make a list of each data point is done and after that stopwords were removed and finally Lemmatization is used to lemmatize the data and a corpus was made

The sample of the corpus was

```
corpus[:10]
```

```
['What ifs',
 'seniority',
 'familiarity',
 'functionality',
 'Lambdas',
 'Java Streams',
 'Object Oriented analysis',
 'Relational Databases',
 'SQL',
 'ORM']
```

2.2 Word Vectorization

As the model won't understand a text data, we use vectorization technology to map the words from the vocabulary into a corresponding vectors of real number which will be used by the model.

```
X[0:5]
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

Clustering

After the vectorization of data a matrix was formed. Now for collecting the technical skills from the dataset K-Means Clustering is used the cluster the points.

3.1 K- Means Clustering

The K-Means is used to find the similarity of the words and places them near each other. After that finally the hard skills were collected

```
Hard_Skills[:20]
```

```
['AWS',
 'SQL',
 'Python',
 'JavaScript',
 'framework',
 'Azure',
 'development',
 'deployment',
 'skill',
 'environment',
 'data',
 'automation',
 'system',
 'Kubernetes',
 'documentation',
 'Familiarity',
 'team',
 'HTML',
 'software',
 'experience']
```

Alternative way to find the Technical skills

4.1 Topic Modeling

```
It is a type of statistical modeling for discovering the abstract "topics"
 that occur in a collection of documents. Latent Dirichlet Allocation (LDA
) is an example of topic model and is used to classify text in a document
to a particular topic. It builds a topic per document model and words per
topic model, modeled as Dirichlet distributions.
```

We have used topic modeling to find the technical skills from the data. The vectored data (X) is fed to LDA to find the 'topic '. This topic contain similar type of data, in this case Hard Skills.

```
Technical_skills[:15]
```

```
[('AWS', 373.7676221811341),
 ('SQL', 356.49832853808175),
 ('Python', 336.48634749872014),
 ('development', 308.9891759852366),
 ('JavaScript', 302.2460114566521),
 ('Azure', 297.56644768065735),
 ('skill', 279.5393449047699),
 ('framework', 274.82930339077967),
 ('automation', 251.7639089575617),
 ('environment', 245.93848957394692),
 ('deployment', 238.25451246107554),
 ('degree', 218.76115263414744),
 ('HTML', 215.6382518101045),
 ('Familiarity', 210.97375243566142),
 ('Git', 201.87427254281567)]
```

Conclusion

The Hard Skills (Technical skills) was extracted from the given dataset using two methods.

1. K-Means Clustering

2. Topic Modeling (LDA)