ADVANCED (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/ADVANCED/)

MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

NLP (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/NLP/) PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/PYTHON-2/)

TEXT (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/TEXT/)

TOPIC MODELING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/TOPIC-MODELING/)

UNSTRUCTURED DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/UNSTRUCTURED-DATA/)

Topic Modelling in Natural Language Processing

YAMINI5 (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/YAMINI5/), MAY 1, 2021 LOGIN TO BOOKMARK THIS ARTICLE (HTTPS://ID...

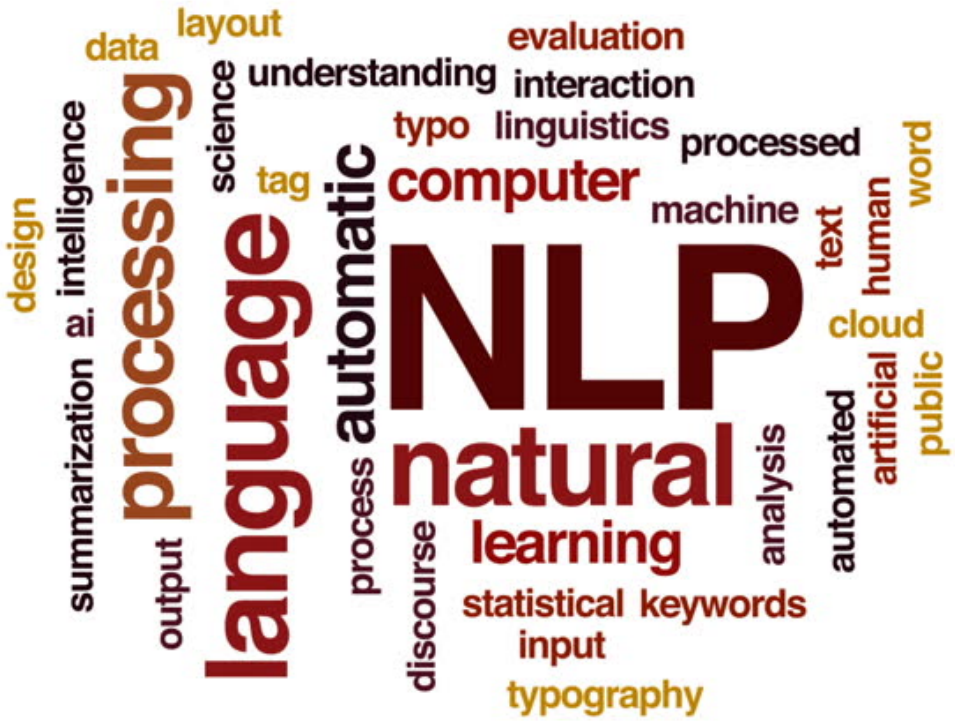
Article Video Book

This article was published as a part of the Data Science Blogathon (https://datahack.analyticsvidhya.com/contest/data-science-blogathon-7/).

Introduction

Natural language processing is the processing of languages used in the system that exists in the library of nltk where this is processed to cut, extract and transform to new data so that we get good insights into it. It uses only the languages that exist in the library because NLP-related things exist there itself so it cannot understand the things beyond what is present in it.

If you do processing on another language then you have to add that language to the existing library. For example, NLP is used in Email Spam filtering where when such data is given then it converts to new data which is understandable by the system and a model is built on it to make predictions on spam or no spam mails. NLP is used in text processing mainly and there are many kinds of tasks that are made easier using NLP. Eg: In chatbots, Autocorrection, Speech Recognition, Language translator, Social media monitoring, Hiring and recruitment, Email filtering, etc.



Topic Modelling:

Topic modelling is recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than

extracting them from topics present in the document. For example, there are 1000 documents and 500 words in

each document. So to process this it requires 500*1000 = 500000 threads. So when you divide the document containing certain topics then if there are 5 topics present in it, the processing is just 5*500 words = 2500 threads.

This looks simple than processing the entire document and this is how topic modelling has come up to solve the problem and also visualizing things better.

First, let's get familiar with NLP so that Topic modelling gets easier to unlock

Some of the important points or topics which makes text processing easier in NLP:

- Removing stopwords and punctuation marks
- Stemming
- Lemmatization
- Encoding them to ML language using Countvectorizer or Tfidf vectorizer

What is Stemming, Lemmatization?

When Stemming is applied to the words in the corpus the word gives the base for that particular word. It is like from a tree with branches you are removing the branches till their stem. Eg: fix, fixing, fixed gives fix when stemming is applied. There are different types through which Stemming can be performed. Some of the popular ones which are being used are:

1. Porter Stemmer

2. Lancaster Stemmer

3. Snowball Stemmer

Lemmatization also does the same task as Stemming which brings a shorter word or base word. There is a slight difference between them is Lemmatization cuts the word to gets its lemma word meaning it gets a much more meaningful form than what stemming does. The output we get after Lemmatization is called 'lemma'.

For example: When stemming is used it might give nav cutting its affixes whereas lemmatization gives have.

There are many methods through which Stemming and Lemmatization can be performed. Some of the popular ones are WordNetLemmatizer, TextBlob, Spacy, and Tree Tagger. (https://www.analyticsvidhya.com/terms/).

lemmatization. Lemmatization can be applied from the mentioned libraries.

Topic modelling is done using LDA(Latent Dirichlet Allocation). Topic modelling refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent). And one popular topic modelling technique is known as Latent Dirichlet Allocation

(LDA).

Topic modelling is an unsupervised approach of recognizing or extracting the topics by detecting the patterns like clustering algorithms which divides the data into different parts. The same happens in Topic modelling in which we get to know the different topics in the document. This is done by extracting the patterns of word clusters and frequencies of words in the document.

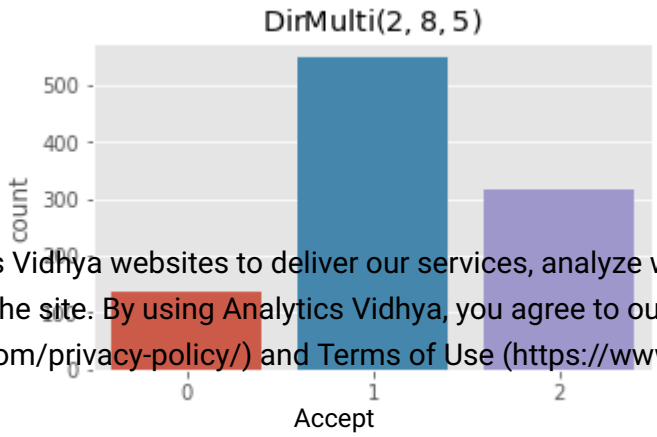
So based on this it divides the document into different topics. As this doesn't have any outputs through which it can do this task hence it is an unsupervised learning method. This type of modelling is very much useful when there are many documents present and when we want to get to know what type of information is present in it. This takes a lot of time when done manually and this can be done easily in very little time using Topic modelling.

What is LDA and how is it different from others?

Latent Dirichlet Allocation:

In LDA, latent indicates the hidden topics present in the data then Dirichlet is a form of distribution. Dirichlet distribution is different from the normal distribution. When ML algorithms are to be applied the data has to be normally distributed or follows Gaussian distribution. The normal distribution represents the data in real numbers format whereas Dirichlet distribution represents the data such that the plotted data sums up to 1. It can also be said as Dirichlet distribution is a probability distribution that is sampling over a probability simplex instead of sampling from the space of real numbers as in Normal distribution.

For example,



Normal distribution tells us how the data deviates towards the mean and will differ according to the variance

Normal distribution tells us how the data deviates towards the mean and will differ according to the variance present in the data. When the variance is high then the values in the data would be both smaller and larger than the mean and can form skewed distributions. If the variance is small then samples will be close to the mean and if the variance is zero it would be exactly at the mean.

Now when the LDA is clear than now the Topic Modelling in LDA? Yes, it would be, let's look into this one.

Now when topic modelling is to get the different topics present in the document. LDA comes to as a savior for doing this task easily instead of performing many things to achieve it. As LDA brings the words in the topics with their distribution using Dirichlet distribution. Hence the name Latent Dirichlet Allocation. The words assigned(or allocated) to the topic with their distribution using Dirichlet distribution.

Implementation of Topic Modelling using LDA:

```
# Parameters tuning using Grid Search
from sklearn.model_selection import GridSearchCV
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.manifold import TSNE
grid_params = {'n_components' : list(range(5,10))}
# LDA model
lda = LatentDirichletAllocation()
lda_model = GridSearchCV(lda,param_grid=grid_params)
lda_model.fit(document_term_matrix)
# Estimators for LDA model
lda_model1 = lda_model.best_estimator_
print("Best LDA model's params" , lda_model.best_params_)
print("Best log likelihood Score for the LDA model",lda_model.best_score_)
print("LDA model Perplexity on train data", lda_model1.perplexity(document_term_matrix))
```

LDA has three important hyperparameters. They are 'alpha' which represents document-topic density factor, 'beta' which represents word density in a topic, 'k' or the number of components representing the number of topics you want the document to be clustered or divided into parts.

To know more parameters present in LDA click sklearn LDA (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html).

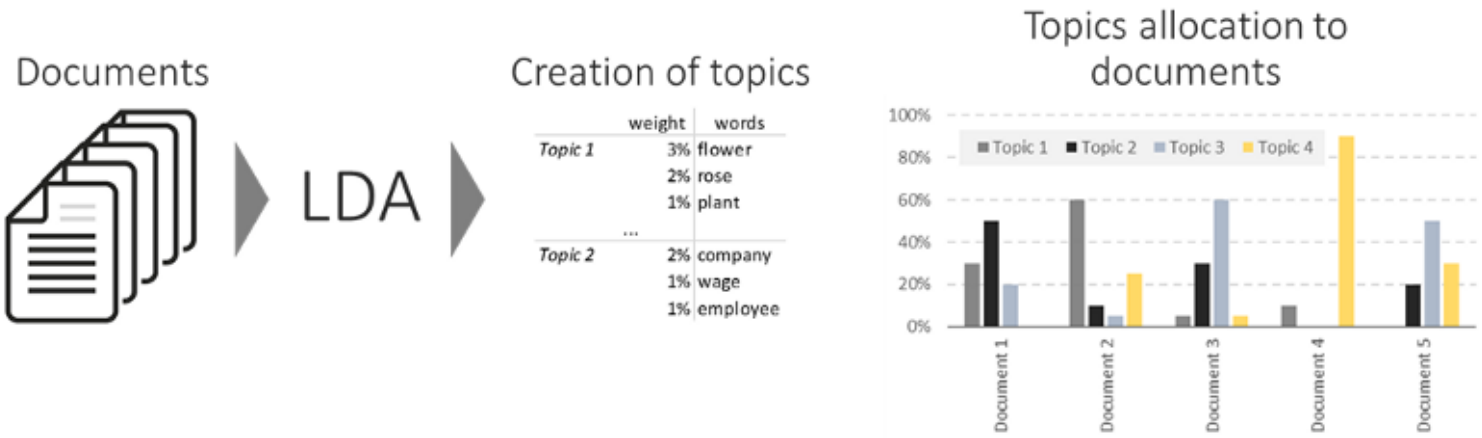
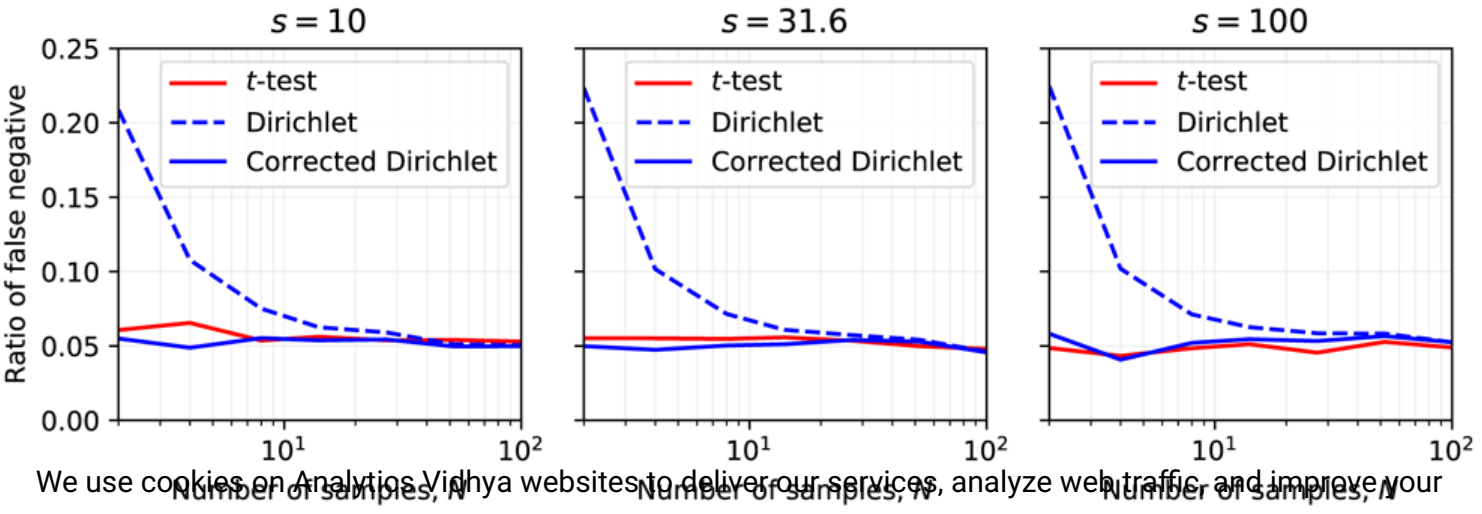


Diagram credits to: https://www.kdnuggets.com/2019/09/overview-topics-extraction-python-latent-dirichlet-allocation.html

Visualization can be done using various methods present in different libraries so the visualization graph might differ then the insight it gives is the same. It tells about the mixture of topics and their distribution in the data or different documents. While preparing even dimensionality reduction techniques like t-SNE can also be used for predicting with good frequent terms from the various documents. Some libraries used for displaying the topic modelling are sklearn, gensim…etc.

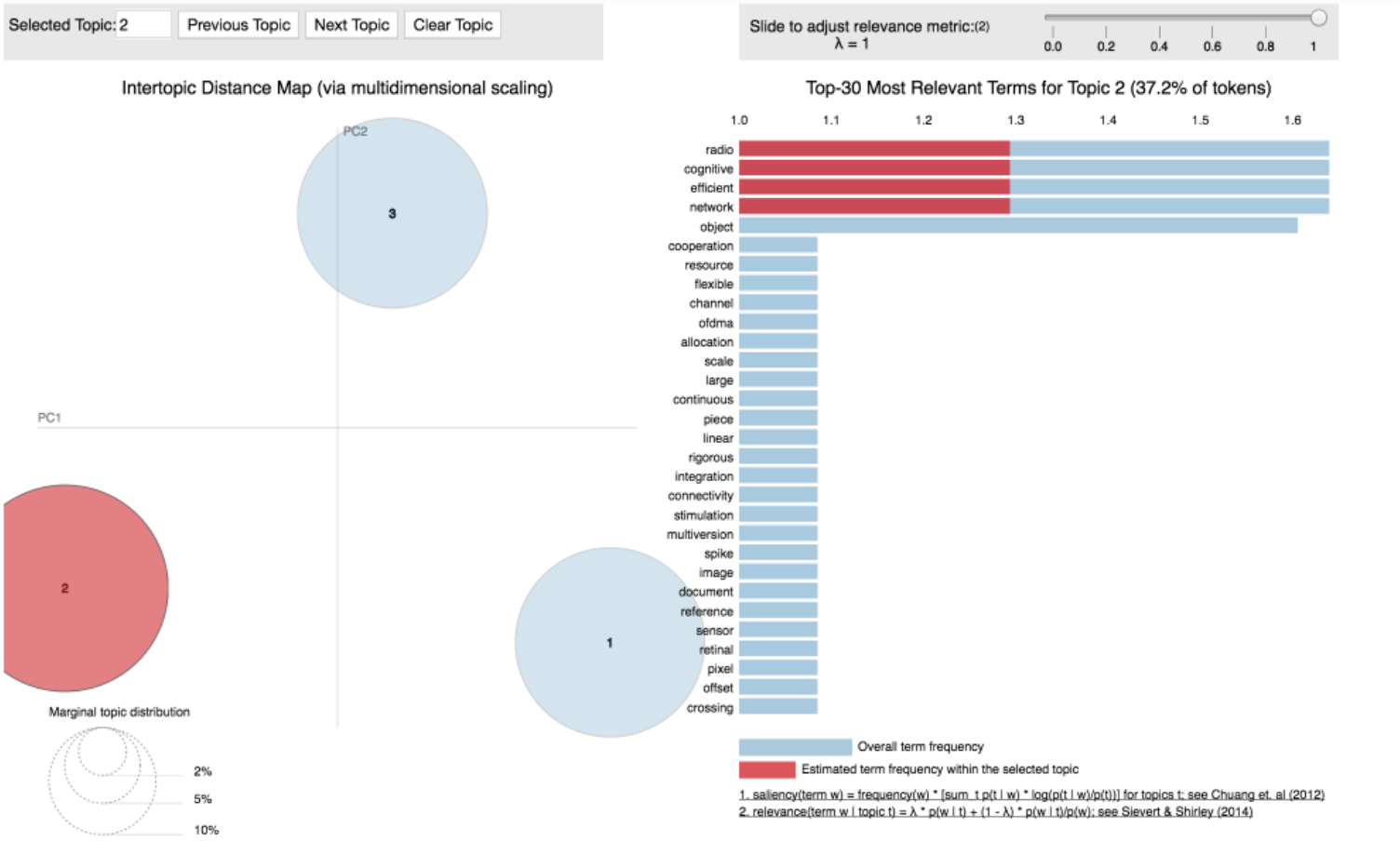


Data Visualization for Topic modelling:

Code for displaying or visualizing the topic modelling performed through LDA is:

Code for displaying or visualizing the topic modelling performed through LDA is:

```
import pyLDAvis.sklearn
pyLDAvis.enable_notebook()
pyLDAvis.sklearn.prepare(best_lda_model, small_document_term_matrix,small_count_vectorizer,md
s='tsne')
```



Applications of Topic Modelling:

1. Medical industry
2. Scientific research understanding
3. Investigation reports
4. Recommender System

5. Blockchain

6. Sentiment analysis

7. Text summarisation

8. Query expansion which can be used in search engines

And many more…

This is a short description of the use, working, and interpretation of results using Topic modeling in NLP with various benefits. Let me know if you have any queries. Thanks for reading.👦👩👩‍🎓🧚🙌 Stay safe and Have a nice day.😊

The media shown in this article are not owned by Analytics Vidhya and is used at the Author's discretion.

TAGS : BLOGATHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BLOGATHON/), TOPIC MODELLING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/TOPIC-MODELLING/)

NEXT ARTICLE

SweetViz Library – EDA in Seconds

(https://www.analyticsvidhya.com/blog/2021/05/sweetviz-library-eda-in-seconds/)

•••

PREVIOUS ARTICLE

Top 8 Python Libraries For Natural Language Processing (NLP) in 2021

(https://www.analyticsvidhya.com/blog/2021/05/top-8-python-libraries-for-natural-language-processing-nlp-in-2021/)

(https://www.analyticsvidhya.com/blog/author/yamini5/)

Yamini5 (Https://Www.Analyticsvidhya.Com/Blog/Author/Yamini5/)

LEAVE A REPLY

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

SUBMIT COMMENT

POPULAR POSTS

40 Questions to test a Data Scientist on Clustering Techniques (Skill test Solution) (https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/)

45 Questions to test a data scientist on basics of Deep Learning (along with solution) (https://www.analyticsvidhya.com/blog/2017/01/must-know-questions-deep-learning/)

Top 30 MCQs to Ace Your Data Science Interviews (https://www.analyticsvidhya.com/blog/2021/04/top-30-mcqs-to-ace-your-data-science-questions-interviews/)

Commonly used Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

Effective Data Visualization Techniques in Data Science Using Python (https://www.analyticsvidhya.com/blog/2021/08/effective-data-visualization-techniques-in-data-science-using-python/)

3 Interesting Python Projects With Code for Beginners! (https://www.analyticsvidhya.com/blog/2021/07/3-interesting-python-projects-with-code-for-beginners/)

COVID-19: A Medical diagnosis using Deep Learning (https://www.analyticsvidhya.com/blog/2021/08/covid-19-a-medical-diagnosis-using-deep-learning/)

30 Questions to test a data scientist on Tree Based Models (https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-tree-based-models/)

CAREER RESOURCES

16 Key Questions You Should Answer Before Transitioning into Data Science (https://www.analyticsvidhya.com/16-key-questions-data-science-career-transition/?&utm_source=Blog&utm_medium=CareerResourceWidget)

NOVEMBER 23, 2020

What is A Business Analyst and What is the Role of a business analyst in a Company? (https://www.analyticsvidhya.com/blog/2021/06/what-

is-a-business-analyst-and-what-is-the-role-of-a-business-analyst-in-a-company/?&utm_source=Blog&utm_medium=CareerResourceWidget)

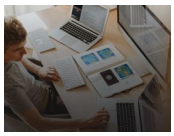
JUNE 28, 2021

Data Engineering – Concepts and Importance (https://www.analyticsvidhya.com/blog/2021/06/data-engineering-concepts-and-importance/?&utm_source=Blog&utm_medium=CareerResourceWidget)

JUNE 14, 2021

Here's What You Need to Know to Become a Data Scientist! (https://www.analyticsvidhya.com/blog/2021/01/heres-what-you-need-to-know-to-become-a-data-scientist/?&utm_source=Blog&utm_medium=CareerResourceWidget)

JANUARY 22, 2021



These 7 Signs Show you have Data Scientist Potential! (https://www.analyticsvidhya.com/blog/2020/12/these-7-signs-show-you-have-data-scientist-potential/?&utm_source=Blog&utm_medium=CareerResourceWidget)

DECEMBER 3, 2020

RECENT POSTS

Introducing Panda Gym -Understanding The Reinforcement Learning (https://www.analyticsvidhya.com/blog/2021/08/introducing-panda-gym-understanding-the-reinforcement-learning/)

AUGUST 14, 2021

Creating Continuous Action Bot using Deep Reinforcement Learning (https://www.analyticsvidhya.com/blog/2021/08/creating-continuous-action-bot-using-deep-reinforcement-learning/)

AUGUST 14, 2021

Better EDA with 3 Easy Python Libraries for Any Beginner (https://www.analyticsvidhya.com/blog/2021/08/better-eda-with-3-easy-python-libraries-for-any-beginner/)

AUGUST 14, 2021

What is relational about Relational Databases? (https://www.analyticsvidhya.com/blog/2021/08/what-is-relational-about-relational-databases/)

AUGUST 14, 2021

id=com.analyticsvidhya.android) vidhya/id1470025572)

Analytics Vidhya

About Us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-vidhya/) (https://www.facebook.com/AnalyticsVidhya) (https://www.youtube.com/channel/UCH6gDteHtH4hg3o2343iObA)

Privacy Policy Terms of Use Refund Policy