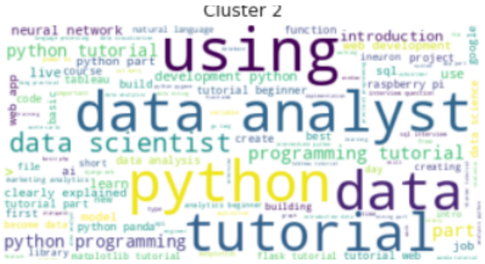
GIẢNG VIÊN: ĐẶNG NGỌC HOÀNG THÀNH

NHÓM SINH VIÊN:
Nguyễn Quỳnh Khánh Hà
Huỳnh Trần Anh Thy
Nguyễn Trịnh Thu Huyền
Nguyễn Văn Hoàng Dũng

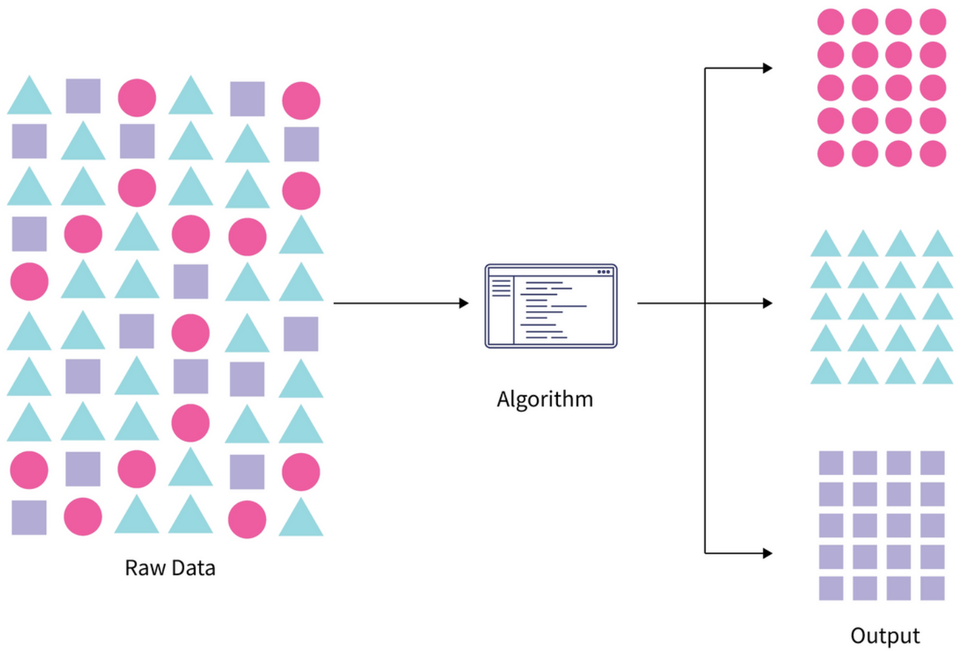
MACHINE LEARNING

PHÂN CỤM CHỦ ĐỀ **YOUTUBE VIDEO** BẰNG K-MEANS, MINIBATCH K-MEANS, AUTOENCODER VÀ DEEP EMBEDDED CLUSTERING



UEH – DATA SCIENCE

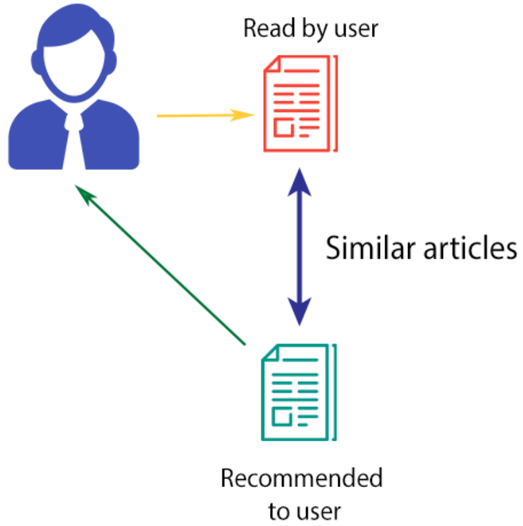
TOPIC MODELING



_ Là mô hình học máy không giám sát (unsupervised learning)

=> Nhóm các items có những đặc tính tương tự nhau

=> Giúp khai phá các chủ đề tiềm ẩn trong tập dữ liệu



Content-based Recommender Systems

DATA CRAWLING

CHANNEL DATA

	channelName	subscribers	views	totalVideos	playlistId
0	Joma Tech	2150000	173586160	113	UUV0qA-eDDICsRR9rPcnG7tw
1	Corey Schafer	1080000	83499114	231	UUCezIgC97PvUuR4_gbFUs5g
2	Tina Huang	478000	20045814	119	UU2UXDak6o7rBm23k3Vv5dww
3	sentdex	1220000	108485512	1237	UUfzlCWGWYyIQ0aLC5w48gBQ
4	365 Data Science	277000	11946346	215	UUEBpSZhI1X8WaP-kY_2LLcg
5	techTFQ	176000	8447865	85	UUnz-ZXXER4jOvuED5trXfEA
6	Data Professor	149000	4187498	309	UUV8e2g4IWQqK71bbzGDEI4Q
7	Luke Barousse	324000	14715055	126	UULLw7jmFsvfIVaUFsLs8mlQ
8	Krish Naik	729000	71597467	1640	UUNU_lfiiWBdtULKOw6X0Dig
9	Alex The Analyst	408000	15856671	197	UU7cs8q-gJRlGwj4A8OmCmXg
10	Data Science Dojo	93900	5309761	389	UUzL_0nIe8B4-7ShhVPfJkgw
11	StatQuest with Josh Starmer	893000	46463040	244	UUtYLUTtgS3k1Fg4y5tAhLbw
12	Ken Jee	237000	7708011	265	UUiT9RITQ9PW6BhXK0y2jaeg

YouTube DATA CRAWLING

VIDEO DATA

	video_id	channelTitle	title	description	tags	publishedAt	viewCount	likeCount	commentCount	duration
0	tmpXcc4pSPA	Data Professor	ChatGPT a threat for schools?	This video provides a brief overview on the th...	['ChatGPT', 'ChatGPT news', 'ChatGPT schools',...	2023-02-14T12:54:14Z	1158	28	6.0	PT52S
1	QlQPsvHSYfQ	Data Professor	How to use ChatGPT to Explain Code	In this video, I'll show you how to use ChatGP...	['ChatGPT', 'GPT3', 'GPT', 'GPT code', 'GPT co...	2023-01-29T13:05:38Z	5874	141	15.0	PT2M34S
2	ELJzUcYrAIQ	Data Professor	How to use ChatGPT to Generate Code in 90 seconds	In this video, I'll show you how to use ChatGP...	['ChatGPT', 'ChatGPT tutorial', 'ChatGPT code'...	2023-01-20T12:16:58Z	3538	81	10.0	PT1M39S
3	ty_lQUNTR0I	Data Professor	How to summarize text using ChatGPT	In this video, you'll learn how to summarize t...	['chatgpt', 'GPT', 'GPT3', 'AI', 'AI generated...	2023-01-13T11:06:56Z	28680	285	31.0	PT5M46S
4	Nj_zUMVuRUg	Data Professor	How to get started with ChatGPT for Beginners	In this video, you'll learn how to get started...	['ChatGPT', 'ChatGPT tutorial', 'ChatGPT tutor...	2023-01-08T05:55:39Z	10939	332	36.0	PT13M42S
...	...	...	...	...	...	...	...	...	...	...
5110	MTiaCUh1420	Krish Naik	Important libraries used in python Data Scienc...	Important libraries used in python Data Scienc...	['Machine Learning', 'Artificial Intelligence'...	2017-11-26T07:48:27Z	19508	173	15.0	PT8M31S
5111	DeT8mji0Jos	Krish Naik	Anaconda installation with Packages-Machine L...	Detailed explanation of anaconda python instal...	['Machine Learning', 'Artificial Intelligence'...	2017-11-26T06:16:34Z	74165	336	37.0	PT5M18S
5112	HrHJUc26Yxl	Krish Naik	What is Supervised Machine Learning-Machine L...	Detailed Explanation of Supervised Machine Lea...	['Machine Learning basics', 'Artificial Intell...	2017-11-26T04:50:08Z	33294	386	16.0	PT11M42S
5113	EqRsD3gqeCo	Krish Naik	What is Machine Learning in Data Science-Mach...	Detailed explanation of Machine Learning ,type...	['Machine learning basics and types', 'Data Sc...	2017-11-25T12:27:20Z	131181	818	46.0	PT10M
5114	qMLxWX49i8I	Krish Naik	Maeri unplugged by Krish and band	Maeri unplugged by krish and band	['Maeri', 'unplugged', 'version']	2014-06-01T07:20:59Z	5072	130	8.0	PT2M31S

5115 rows × 10 columns

PREPROCESSING

Missing value

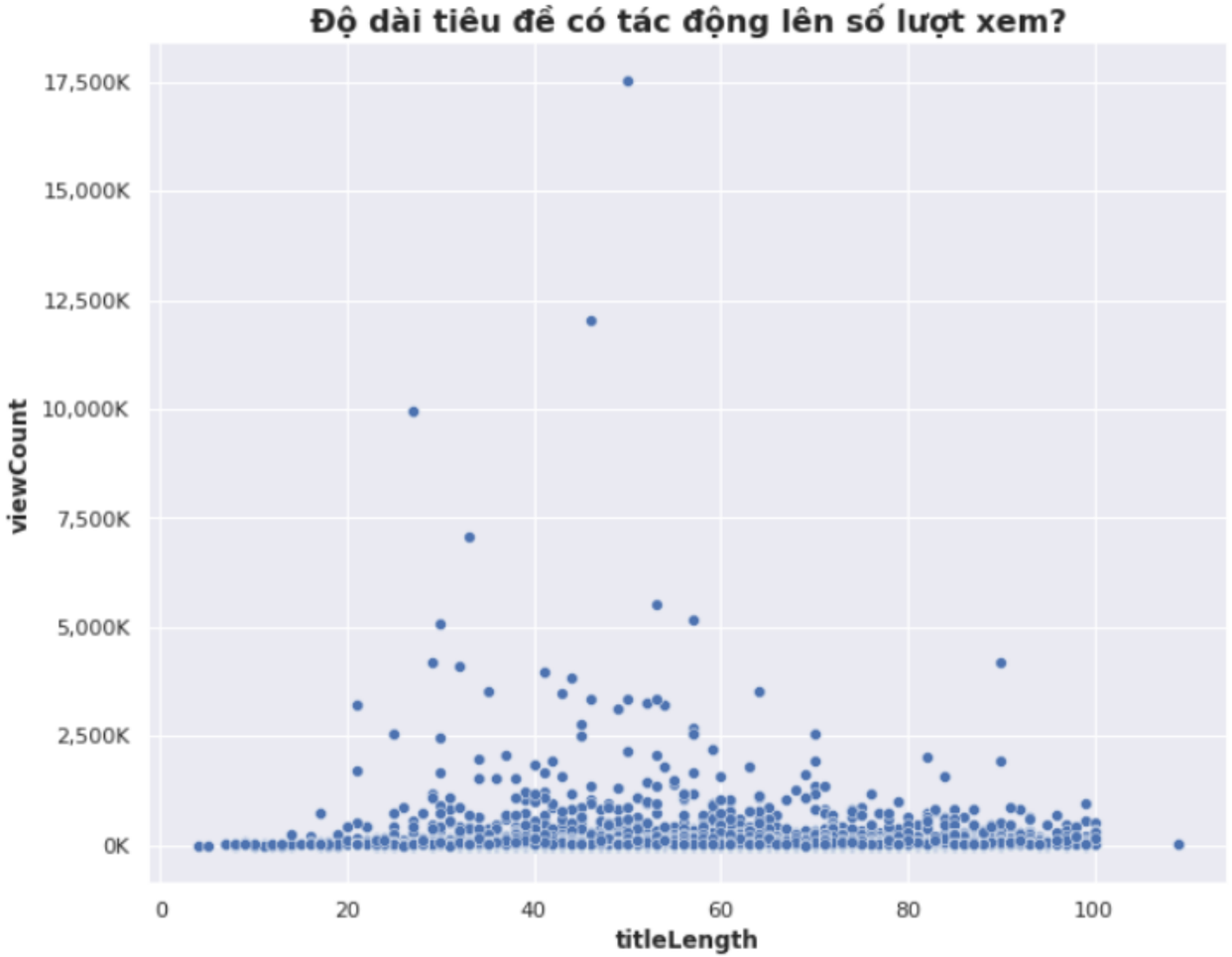
	video_id	channelTitle	title	description	tags	publishedAt	viewCount	likeCount	commentCount	duration
Total	0	0	0	17	377	0	0	0	2	0
Percent(%)	0.0	0.0	0.0	0.332356	7.370479	0.0	0.0	0.0	0.039101	0.0
Types	object	object	object	object	object	object	int64	int64	float64	object

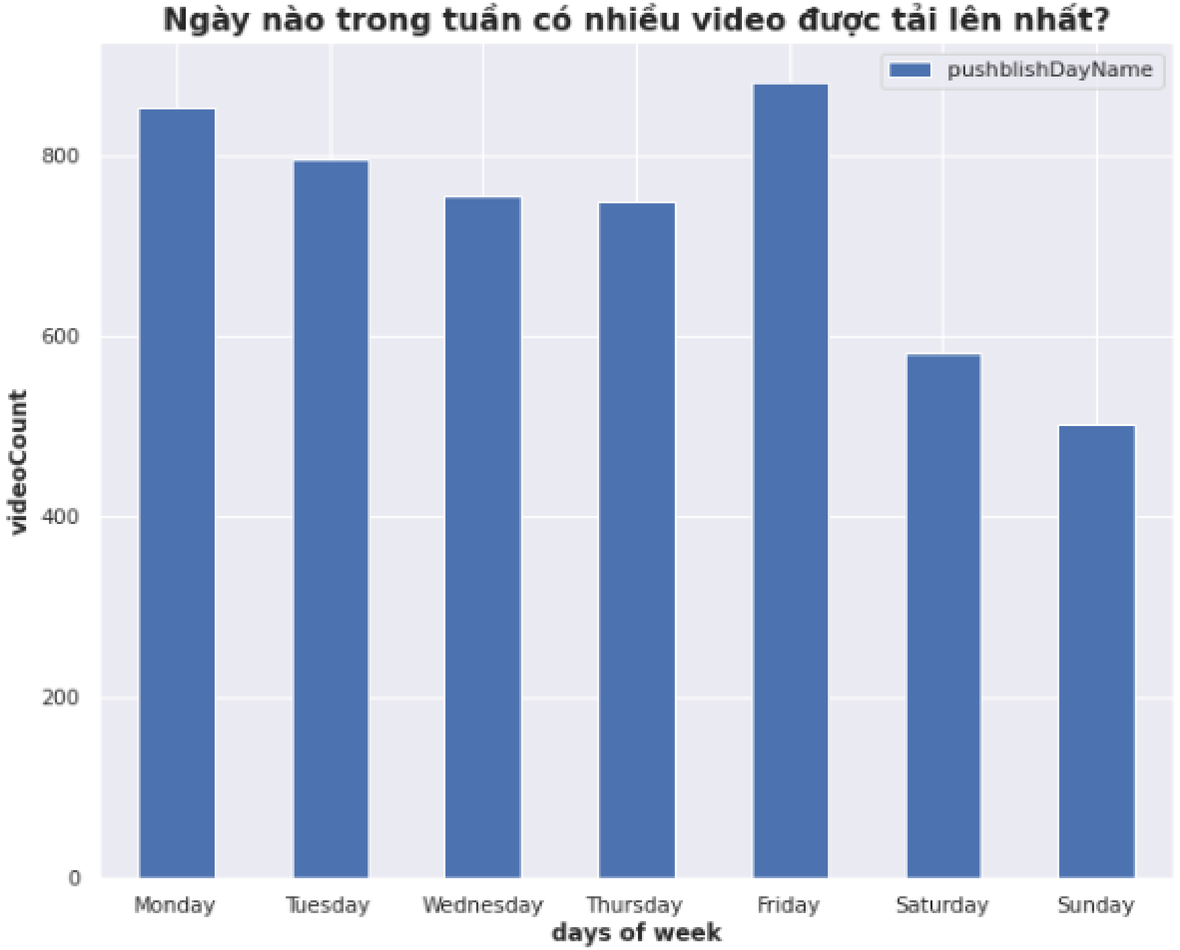
Feature Engineering

	title	likeCount	commentCount	duration	publishedAt
	ChatGPT a threat for schools?	28	6.0	PT52S	2023-02-14T12:54:14Z
)	use ChatGPT to Explain Code	141	15.0	PT2M34S	2023-01-29T13:05:38Z
	ChatGPT to Generate Code in 90 seconds	81	10.0	PT1M39S	2023-01-20T12:16:58Z
	summarize text using ChatGPT	285	31.0	PT5M46S	2023-01-13T11:06:56Z
	to get started with ChatGPT for Beginners	332	36.0	PT13M42S	2023-01-08T05:55:39Z

	pushblishDayName	durationSecs	likeRatio	commentRatio	titleLength
	Tuesday	52.0	24.179620	5.181347	29
	Sunday	154.0	24.004086	2.553626	34
	Friday	99.0	22.894291	2.826456	49
	Friday	346.0	9.937238	1.080893	35
	Sunday	822.0	30.350123	3.290977	45

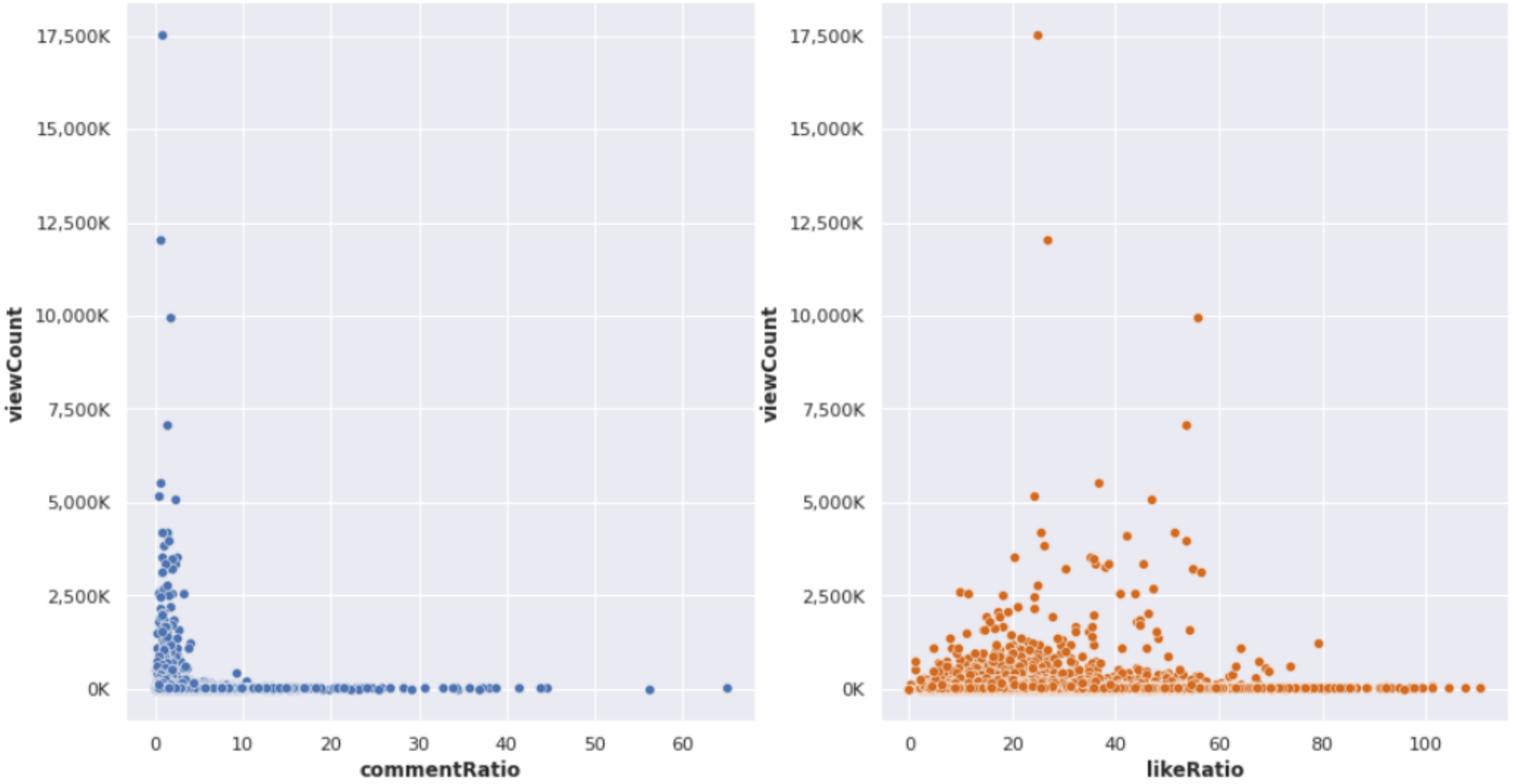
EDA


Độ dài tiêu đề có tác động lên số lượt xem?
viewCount
17,500K
15,000K
12,500K
10,000K
7,500K
5,000K
2,500K
0K
0
20
40
60
80
100
titleLength


Ngày nào trong tuần có nhiều video được tải lên nhất?
pushblishDayName
videoCount
800
600
400
200
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
days of week

EDA

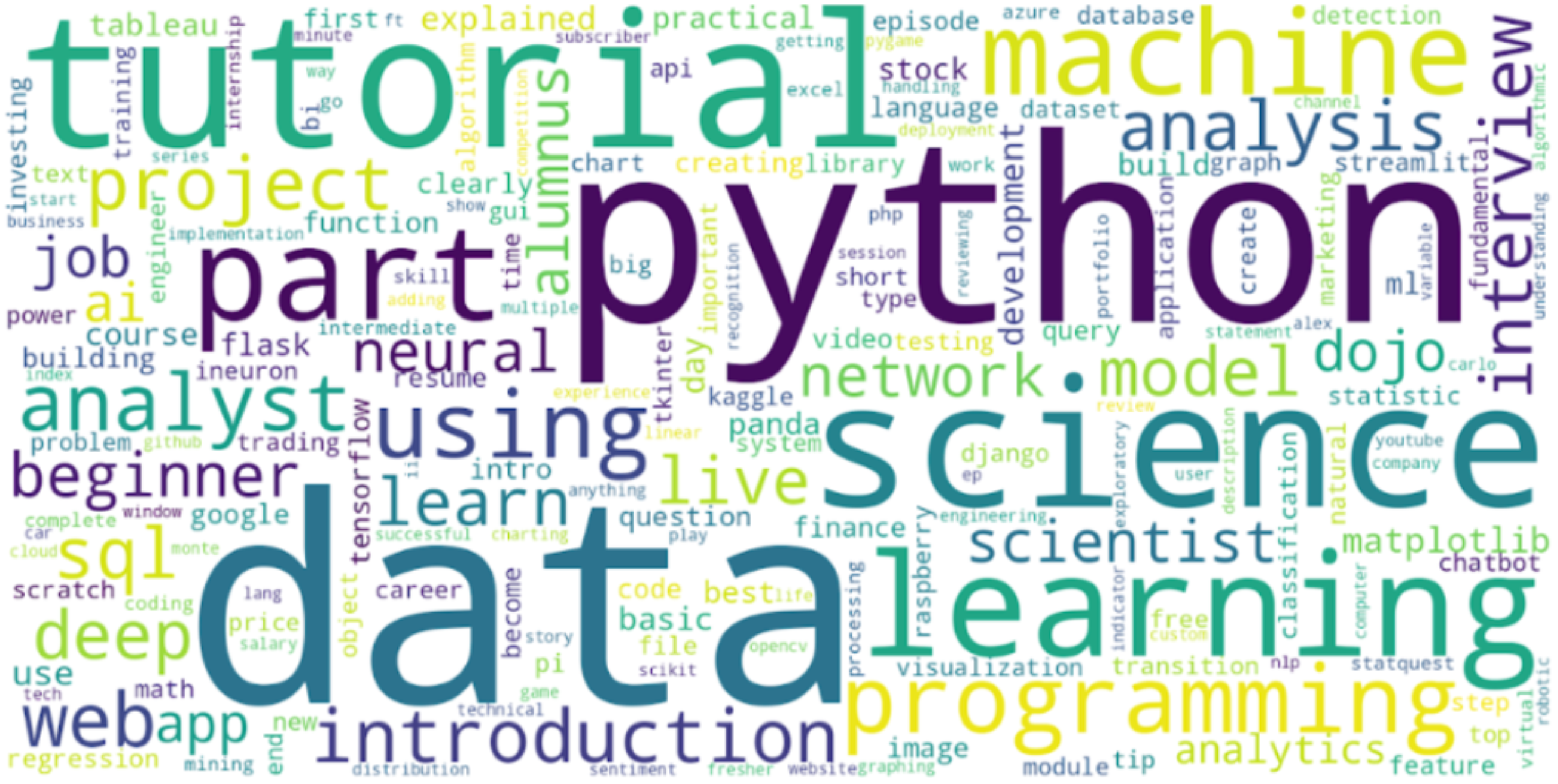
Bình luận và lượt thích có tương quan với lượt xem hay không?



EDA

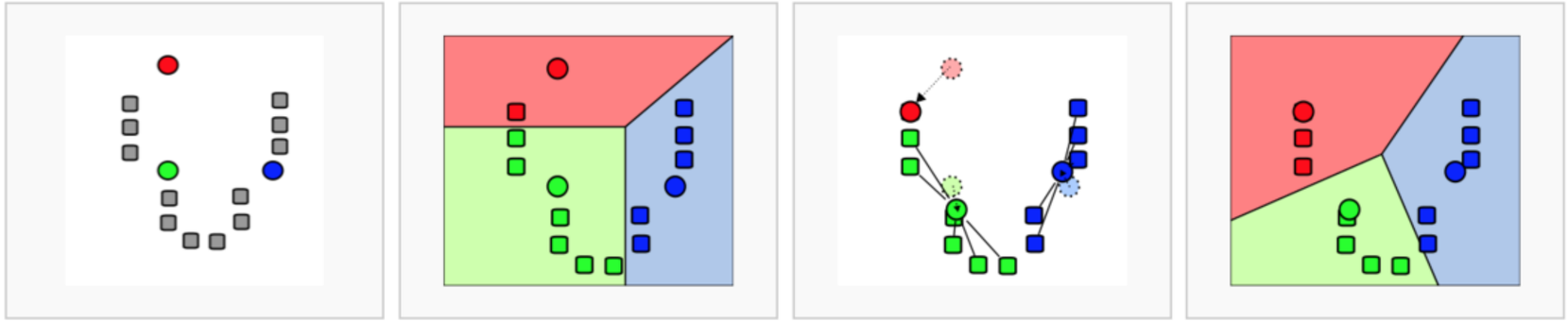
WORDCLOUD DESCRIPTION

WORDCLOUD TITLE



K –MEANS

Demonstration of the standard algorithm



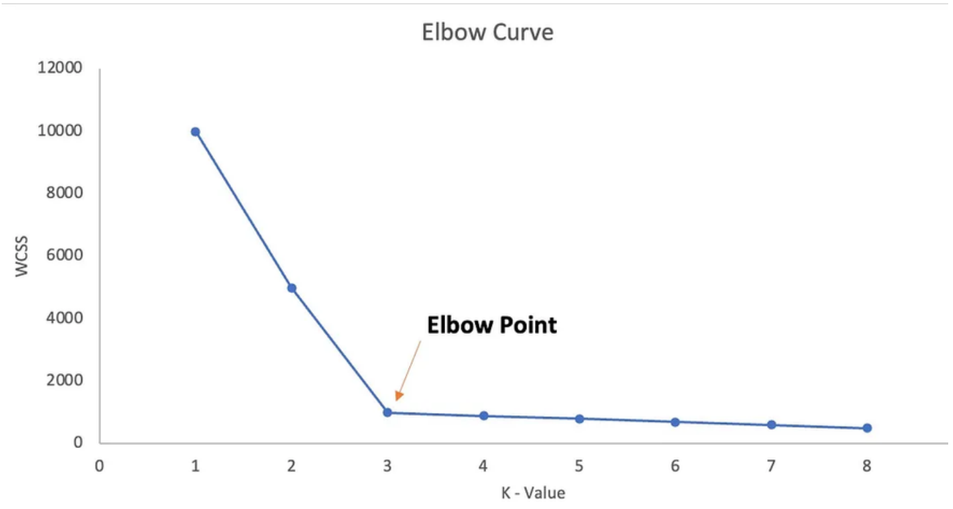
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

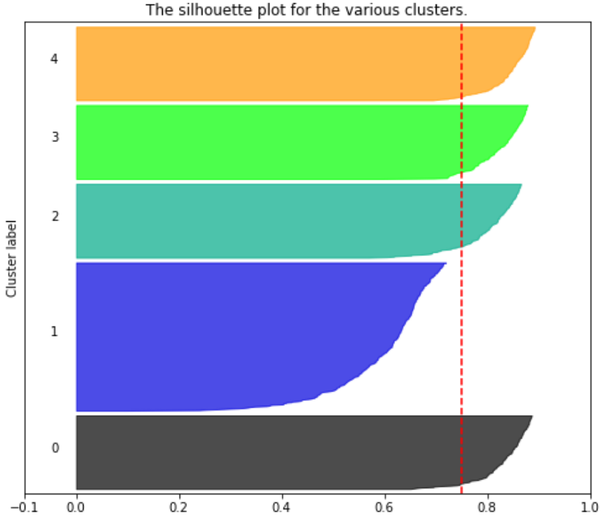
2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

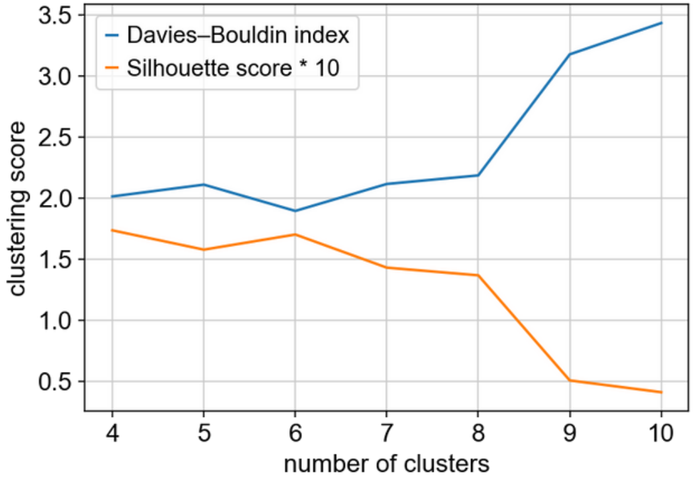
3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

ĐÁNH GIÁ







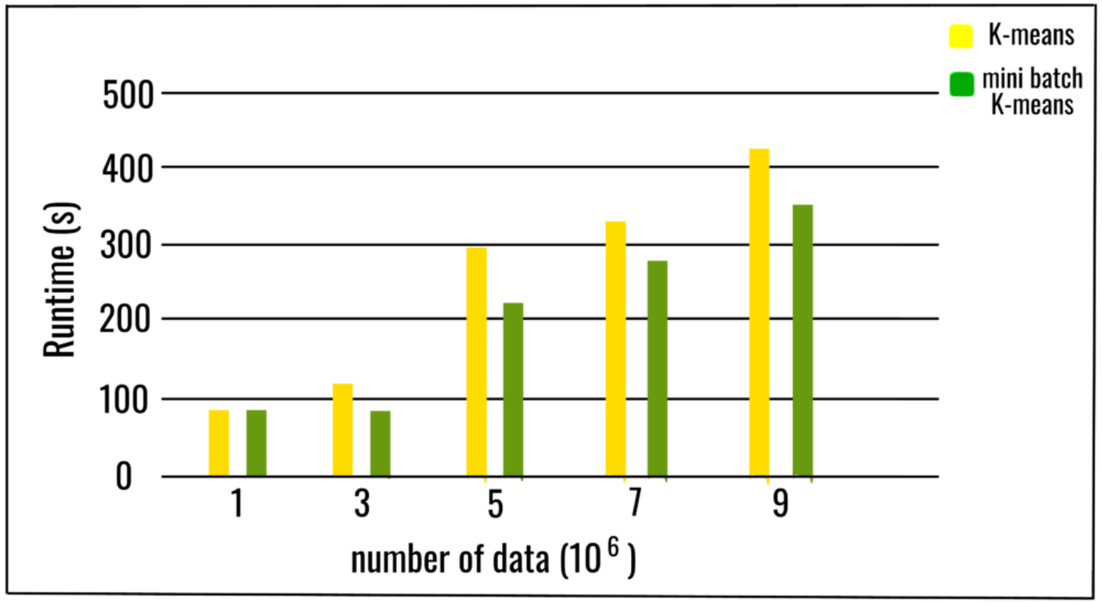
MINIBATCH K -MEANS

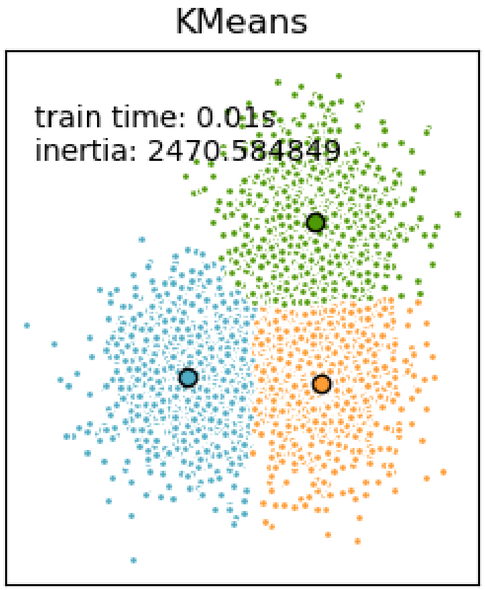
- **K-Means:**

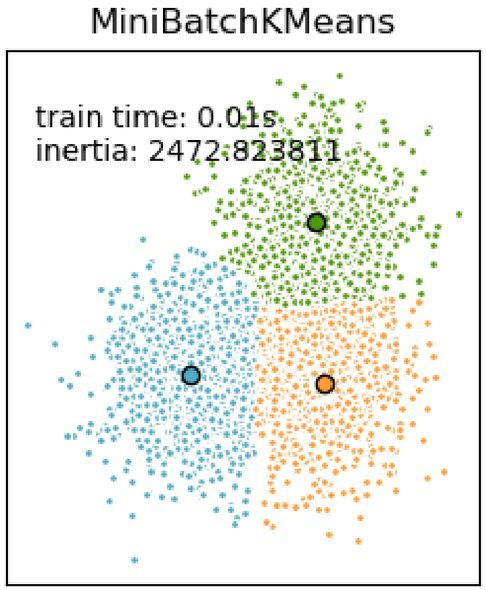
Khi kích thước tập dữ liệu tăng lên
-> Thời gian training tăng.
Vì cần toàn bộ tập dữ liệu trong bộ nhớ chính

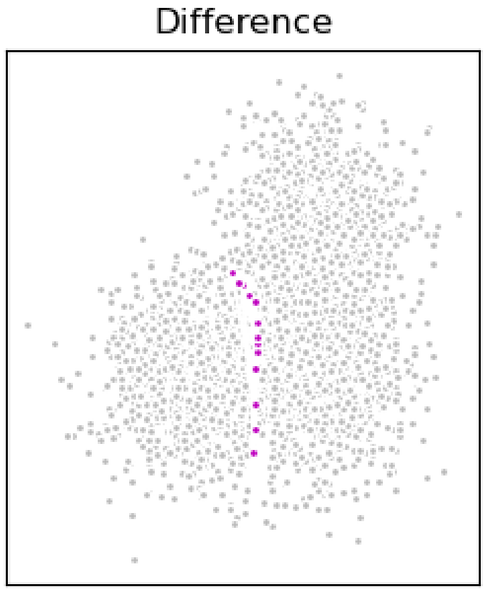
- **Minibatch K-Means:**

Giúp giảm chi phí thời gian & không gian của thuật toán

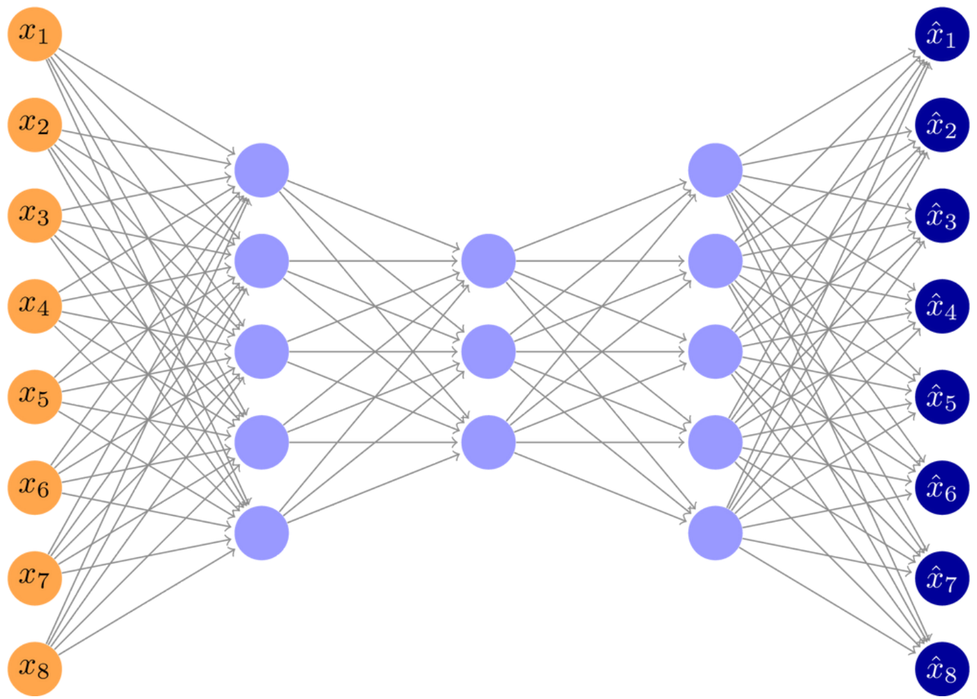








AUTOENCODER



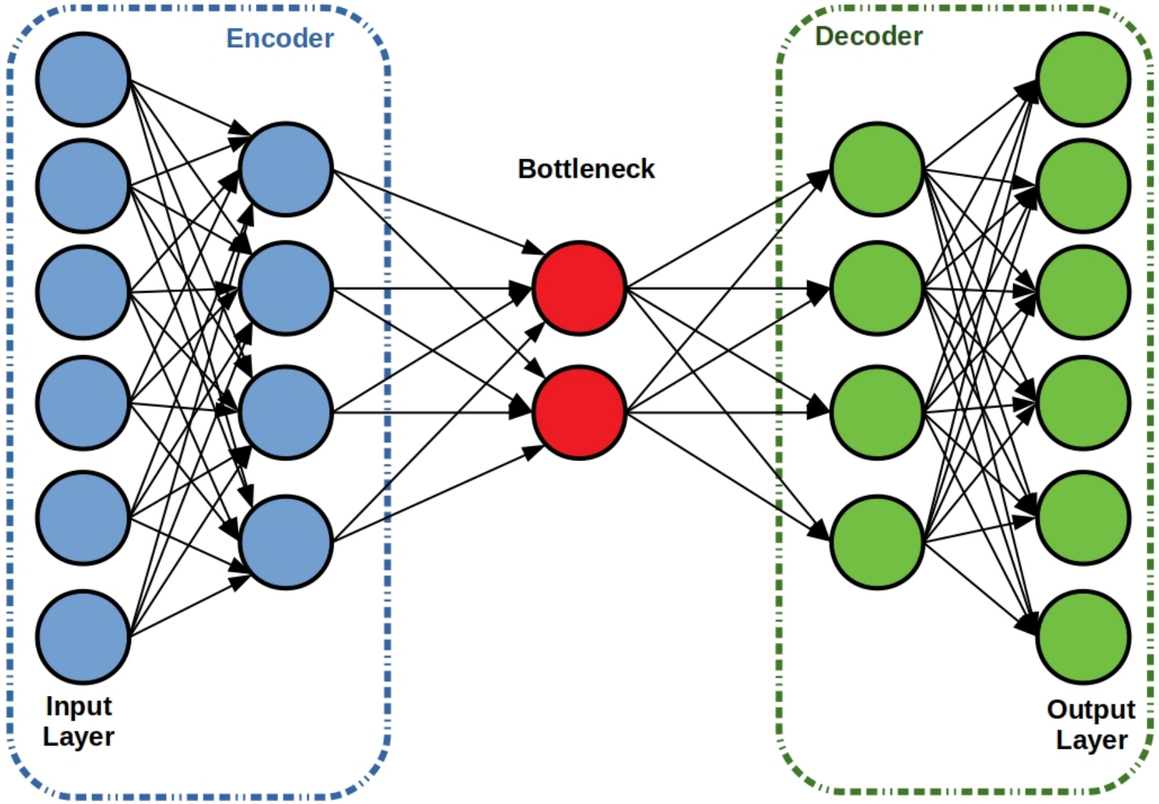
Original data

Reconstructed data

- Là mạng **ANN**, có khả năng học hiệu quả các <u>biểu diễn của dữ liệu đầu vào</u> mà không cần nhãn (*Unsupervised Learning*)

- Mục đích: Nén dữ liệu đầu vào -> Tái tạo lại output gần giống với input

AUTOENCODER



Cấu trúc của AutoEncoder

1. Encoder:

- Nén và ánh xạ dữ liệu đầu vào lên không gian tiềm ẩn (latent space)

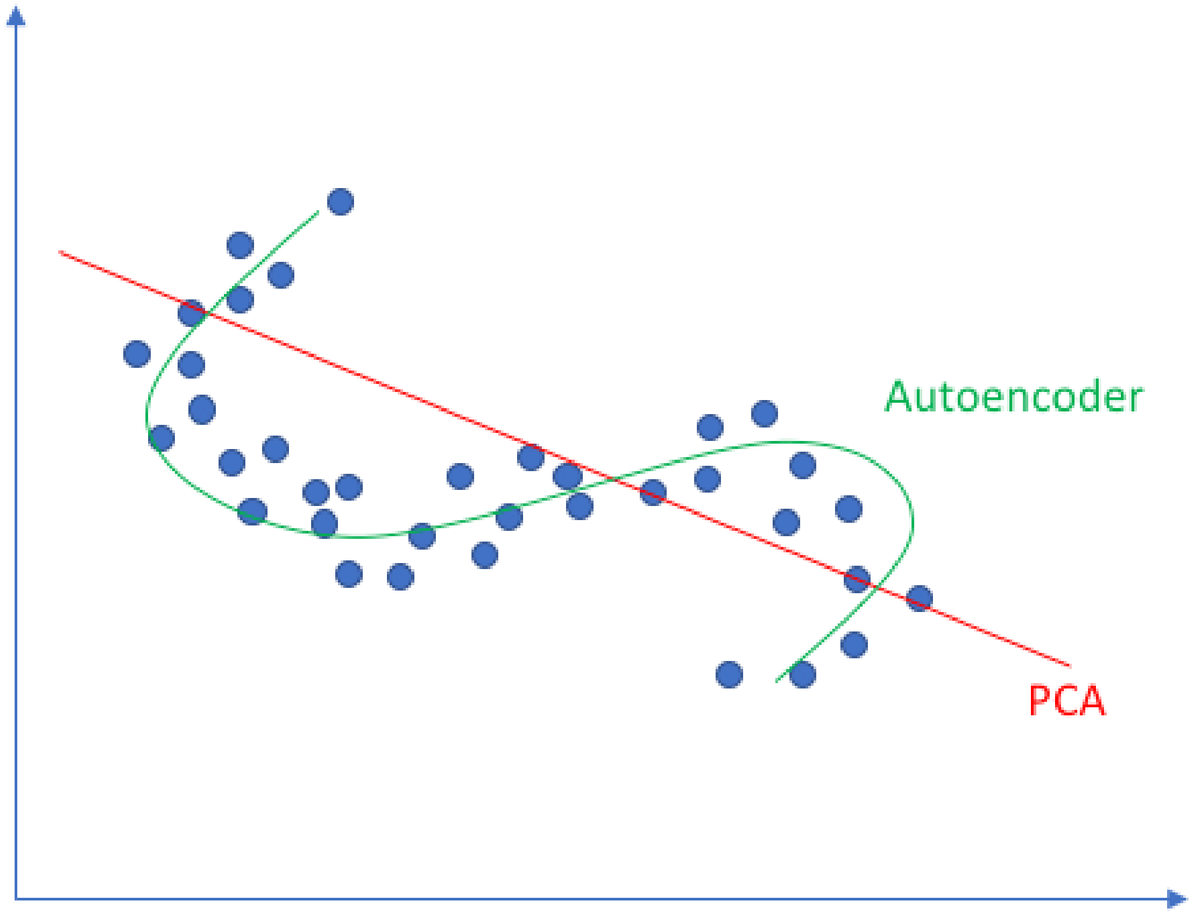
2. Bottleneck (**latent space**)**:**

- Là output của Encoder và là phần mang kích thước nhỏ nhất.
- Vì bottleneck mã hóa tối đa thông tin dữ liệu đầu vào, vì vậy nó chứa các đặc trưng, tri thức quan trọng nhất của dữ liệu.
- Từ đó có thể được dùng để trích xuất các đặc trưng tiềm ẩn của văn bản (feature extraction)

3. Decoder:

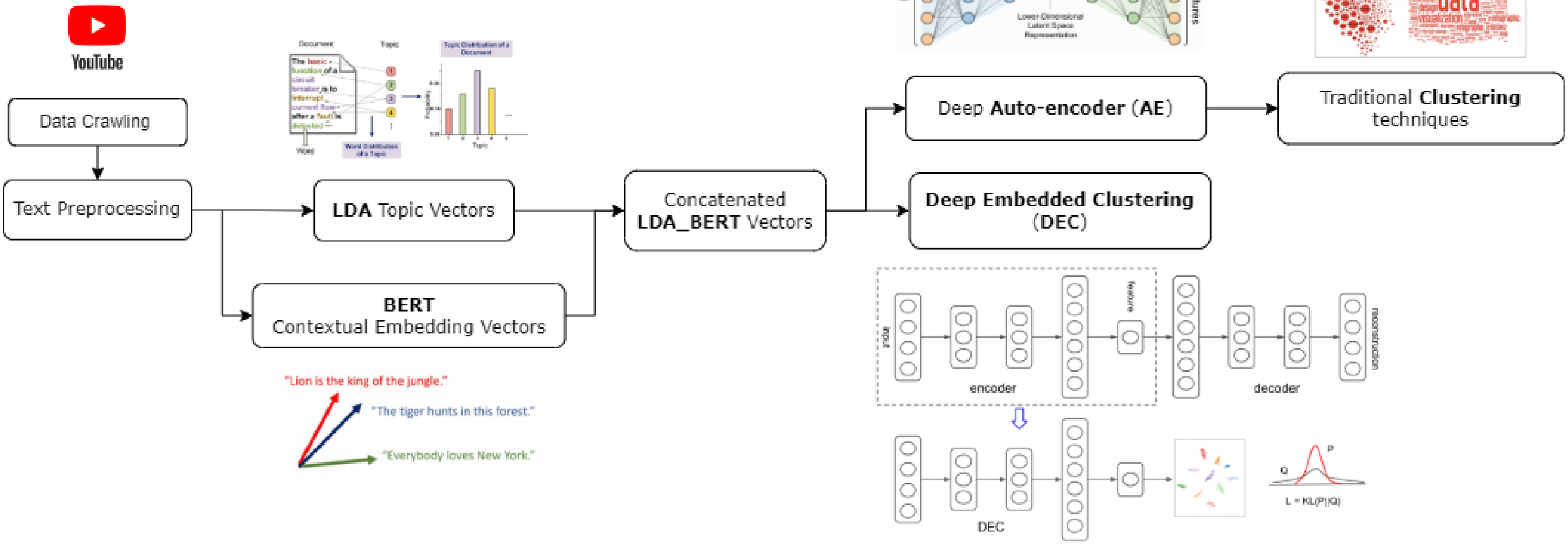
- Có nhiệm vụ giải mã data từ bottleneck để tái tạo lại dữ liệu đầu vào dựa trên các đặc trưng tiềm ẩn bên trong Bottleneck.

AUTOENCODER >< PCA



- **PCA** chỉ giảm chiều dữ liệu trên không gian tuyến tính đơn giản (*linear relationship*)
- **Autoencoder** có khả năng học các mối quan hệ phi tuyến tính phức tạp (*non-linear relationship*)

FLOWCHART


YouTube
Data Crawling
Text Preprocessing
Document
Topic
Word
Topic Distribution of a Document
Word Distribution of a Topic
Probability
Topic
LDA Topic Vectors
BERT
Contextual Embedding Vectors
"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."
Concatenated
LDA_BERT Vectors
Input Features
Encoding Network
Decoding Network
Reconstructed Features
Lower-Dimensional
Latent Space
Representation
Deep Auto-encoder (AE)
Traditional Clustering techniques
Deep Embedded Clustering (DEC)
input
feature
reconstruction
encoder
decoder
DEC
P
Q
L = KL(P||Q)

DEEP EMBEDDED CLUSTERING