

# Text Processing: Handling Text Data

# Module Outline

1. Dataset Types: Structured vs Unstructured
2. Text Dataset
3. Reading Data
4. Regular Expressions
5. Pattern Matching in RegEx
6. Regular Expressions in Python
7. Regular Expressions on Real-World Dataset

# Dataset Types: Structured vs Unstructured Data

## Structured Dataset

- A dataset having fixed number of dimensions
- Tabular Data or Key Value Pairs

Emp_id	Emp_name	Designation	Salary	City
1001	RAHUL	Technical Leader	63k	LUCKNOW
1002	KARAN	Junior Developer	42k	JAIPUR
1003	SEEMA	Junior Developer	45k	AGRA
1004	SHREYA	Senior Developer	50k	KANPUR
1005	REET	Project Manager	86k	AMRITSAR
1006	GAURAV	Technical Leader	67k	DELHI

# Dataset Types: Structured vs Unstructured Data

## Structured Dataset

- A dataset having fixed number of dimensions
- Tabular Data or Key Value Pairs

Place	Maximum Temp in °C	Minimum Temp in °C
Pitampura	32.5	25.4
Akshardham	31.1	25
Yamuna Sports Complex	32.8	25.3
Delhi University	32.2	24.7
Safdarjung	29.7	23.3
Palam	31.9	24.9
Ridge	31.2	21
Lodhi Road	29.4	24
Noida	32.8	25

# Dataset Types: Structured vs Unstructured Data

## Structured Dataset

- A dataset having fixed number of dimensions
- Tabular Data or Key Value Pairs

```
{  
  "id": 123,  
  "name": "Sandeep",  
  "subject": "Computer",  
  "marks": 23  
},  
{  
  "id": 193,  
  "name": "Raja",  
  "subject": "Mathematics",  
  "marks": 25  
},  
{  
  "id": 223,  
  "name": "Smith",  
  "subject": "Geography",  
  "marks": 20  
},
```

# Dataset Types: Structured vs Unstructured Data

## Unstructured Dataset

- A dataset having no fixed dimensions
- Forms:
  - Audio



# Dataset Types: Structured vs Unstructured Data

## Unstructured Dataset

- A dataset having no fixed dimensions
- Forms:
  - Audio
  - Images

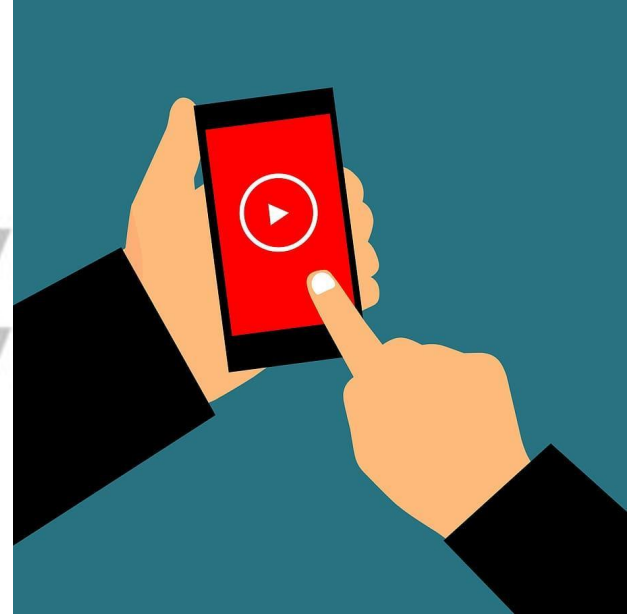
Cats



# Dataset Types: Structured vs Unstructured Data

## Unstructured Dataset

- A dataset having no fixed dimensions
- Forms:
  - Audio
  - Images
  - Videos





# Dataset Types: Structured vs Unstructured Data

## Unstructured Dataset

- A dataset having no fixed dimensions
- Forms:
  - Audio
  - Images
  - Video
  - Text



# Text Dataset

- Written form of any language
- Characters or words arranged together in a meaningful and formal manner
- Grammar rules and defined structures
- Examples:
  - Social Media: tweets, posts, comments



**Donald J. Trump** ✓

@realDonaldTrump

Following

The Fake News Media will not talk about the importance of the United Nations Security Council's 15-0 vote in favor of sanctions on N. Korea!

1:15 PM - 7 Aug 2017



**President Obama** @POTUS · Sep 16

Cool clock, Ahmed. Want to bring it to the White House? We should inspire more kids like you to like science. It's what makes America great.



440K

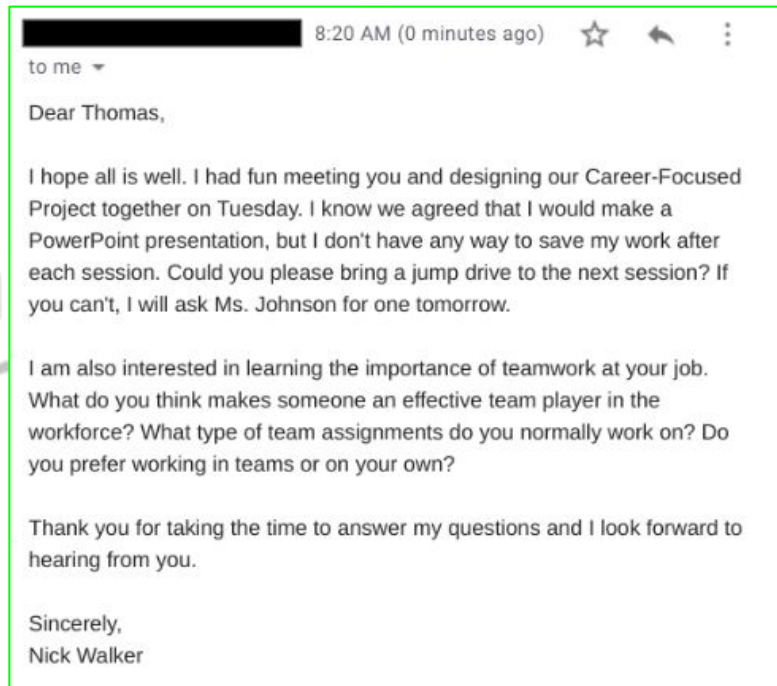


440K



# Text Dataset

- Written form of any language
- Characters or words arranged together in a meaningful and formal manner
- Grammar rules and defined structures
- Examples:
  - Social Media: tweets, posts, comments
  - Conversations: messages, emails, chats



# Text Dataset


- Written form of any language
- Characters or words arranged together in a meaningful and formal manner
- Grammar rules and defined structures
- Examples:
  - Social Media: tweets, posts, comments
  - Conversations: messages, emails, chats
  - Articles: news, blogs, transcripts

The **Moon** is Earth's only permanent natural satellite. It is the fifth-largest natural satellite in the Solar System, and the largest among planetary satellites relative to the size of the planet that it orbits (its primary). It is the second-densest satellite among those whose densities are known (after Jupiter's satellite Io).

The Moon is thought to have formed approximately 4.5 billion years ago, not long after Earth. There are several hypotheses for its origin; the most widely accepted explanation is that the Moon formed from the debris left over after a giant impact between Earth and a Mars-sized body called Theia.

The Moon is in synchronous rotation with Earth, always showing the same face, with its near side marked by dark volcanic maria that fill the spaces between the bright ancient crustal highlands and the

Moon ☾



Full Moon as seen from Earth's northern hemisphere

Designations

Adjectives



Thank You