

# ESSENTIALS OF DATA SCIENCE

## THEORY ACTIVITY NO.1

Name: Amayara Gani

Division: CS2

Roll No: CS2-61

PRN: 202401040273

---

Dataset Link: <https://archive.ics.uci.edu/dataset/410/paper+reviews>

### Import Libraries

```
✓ [93] import numpy as np
      import pandas as pd
```

### 1. Problem Statement-1

Problem Statement: Load the dataset and explore the structure

```
import json
with open('/content/reviews.json') as f:
    data = json.load(f)

reviews = []
for paper in data["paper"]:
    for review in paper["review"]:
        review["paper_id"] = paper["id"] # Adding paper_id to each review
        reviews.append(review)

# We are converting it into DataFrame
df = pd.DataFrame(reviews)
print(df.head())
```


```
confidence evaluation id lan orientation remarks \
0      4      1  1 es      0
1      4      1  2 es      1
2      5      1  3 es      1
3      4      2  1 es      1
4      4      2  2 es      0

      text      timespan  paper_id
0  - El artículo aborda un problema contingente y...  2010-07-05      1
1  El artículo presenta recomendaciones prácticas...  2010-07-05      1
2  - El tema es muy interesante y puede ser de mu...  2010-07-05      1
3  Se explica en forma ordenada y didáctica una e...  2010-07-05      2
4                                     2010-07-05      2
```

## 2. Problem Statement-2

Problem Statement: Display Basic Information About the Dataset

```
[71] df.describe()
```




	id	paper_id
count	405.000000	405.000000
mean	1.824691	84.945679
std	0.821362	49.854958
min	1.000000	1.000000
25%	1.000000	38.000000
50%	2.000000	92.000000
75%	2.000000	126.000000
max	4.000000	172.000000

## 3. Problem Statement-3

Identify Missing Values in the *Dataset*

```
[72] df.isnull().sum()
```



	0
confidence	2
evaluation	0
id	0
lan	0
orientation	0
remarks	0
text	0
timespan	0
paper_id	0

dtype: int64

#### 4. Problem Statement-4

Problem Statement: Find shape of the given dataset

```
[73] df.shape
```

➡ (405, 9)

#### 5. Problem Statement-5

Problem Statement: Display the first 5 rows of the dataset to get an overview of the structure of the data.

```
df.head()
```

	confidence	evaluation	id	lan	orientation	remarks	text	timespan	paper_id
0	4	1	1	es	0	- El artículo aborda un problema contingente y...	2010-07-05	1	
1	4	1	2	es	1	El artículo presenta recomendaciones prácticas...	2010-07-05	1	
2	5	1	3	es	1	- El tema es muy interesante y puede ser de mu...	2010-07-05	1	
3	4	2	1	es	1	Se explica en forma ordenada y didáctica una e...	2010-07-05	2	
4	4	2	2	es	0		2010-07-05	2	

Next steps: [Generate code with df](#) [View recommended plots](#) [New interactive sheet](#)

#### 6. Problem Statement-6

Problem Statement: Display the last 5 rows of the dataset to check for any recent entries.

```
[75] print(df.tail())
```

	confidence	evaluation	id	lan	orientation	remarks	text	timespan	paper_id
400	5	-1	1	es	-2		El trabajo pretende ofrecer una visión del uso...	2015-07-05	170
401	4	-1	2	es	-1		El paper está bien escrito y de fácil lectura...	2015-07-05	170
402	4	-1	1	es	0		Observación de fondo: No se presenta un ejemp...	2015-07-05	171
403	3	1	2	es	-1		Se propone un procedimiento para elaborar máqu...	2015-07-05	171
404	4	1	1	es	-1		El artículo describe básicamente los component...	2015-07-05	172

## 7. Problem Statement-7

Problem Statement: Display the first 5 reviews for paper ID 1.

```
[76] print(df[df['paper_id'] == 1].head())
```

```
↗ confidence evaluation id lan orientation remarks \
0      4      1  1 es      0
1      4      1  2 es      1
2      5      1  3 es      1

      text      timespan  paper_id
0 - El artículo aborda un problema contingente y...  2010-07-05      1
1 El artículo presenta recomendaciones prácticas...  2010-07-05      1
2 - El tema es muy interesante y puede ser de mu...  2010-07-05      1
```

## 8. Problem Statement-8

Problem Statement: Find the papers that received the most reviews.

```
[77] paper_reviews_count = df['paper_id'].value_counts()
      most_reviewed_papers = paper_reviews_count.head(10)
      print(most_reviewed_papers)
```

```
↗ paper_id
128      4
112      4
130      4
129      4
102      4
80       4
92       4
50       4
55       4
40       3
Name: count, dtype: int64
```

## 9. Problem Statement-9

Problem Statement: Count the Number of Reviews for Each Paper

```
[78] paper_reviews_count = df['paper_id'].value_counts()
      paper_reviews_count
```

paper_id	count
128	4
112	4
130	4
129	4
102	4
...	...
63	1
71	1
70	1
68	1
172	1

169 rows × 1 columns

dtype: int64

## 10. Problem Statement-10

Problem Statement: Filter reviews where the evaluation is very positive (evaluation = 2).

```
[79] positive_reviews = df[df['evaluation'] == 2]
      print(positive_reviews[['paper_id', 'text', 'evaluation']].head())
```

Empty DataFrame  
Columns: [paper\_id, text, evaluation]  
Index: []

## 11. Problem Statement-11

Problem Statement: Count the number of reviews written in Spanish (lan == 'es') and English (lan == 'en')

```
[80] language_distribution = df['lan'].value_counts()
      print(language_distribution)
```

```
lan
es   388
en    17
Name: count, dtype: int64
```

## 12. Problem Statement-12

Problem Statement: Filter out reviews that have fewer than 10 words.

```
[81] df['word_count'] = df['text'].apply(lambda x: len(x.split()))
      filtered_reviews = df[df['word_count'] >= 10]
      print(filtered_reviews[['paper_id', 'text', 'word_count']].head())
```

```
paper_id      text  word_count
0         1  - El artículo aborda un problema contingente y...      93
1         1  El artículo presenta recomendaciones prácticas...      94
2         1  - El tema es muy interesante y puede ser de mu...     211
3         2  Se explica en forma ordenada y didáctica una e...     200
5         2  Los autores describen una metodología para des...     299
```

## 13. Problem Statement-13

Problem Statement: Calculate the minimum, maximum, and average length (in words) of the reviews.

```
[82] review_length_stats = {
      'min_length': np.min(df['word_count']),
      'max_length': np.max(df['word_count']),
      'avg_length': np.mean(df['word_count'])
    }
      print(review_length_stats)
```

```
{'min_length': 0, 'max_length': 1007, 'avg_length': np.float64(160.51604938271606)}
```

## 14. Problem Statement-14

Problem Statement: Count the number of reviews in each language (es for Spanish and en for English)

```
[83] language_distribution = df['lan'].value_counts()
      language_distribution
```

```
count
lan
es    388
en     17
dtype: int64
```

## 15. Problem Statement-15

Problem Statement: Exclude reviews in English (lan == 'en') and focus only on the Spanish reviews.

```
[84] df_es = df[df['lan'] == 'es']
      df_es
```

```
confidence  evaluation  id  lan  orientation  remarks  text  timespan  paper_id  word_count
0           4           1   1   es           0  - El artículo aborda un problema contingente y...  2010-07-05      1         93
1           4           1   2   es           1  El artículo presenta recomendaciones prácticas...  2010-07-05      1         94
2           5           1   3   es           1  - El tema es muy interesante y puede ser de mu...  2010-07-05      1        211
3           4           2   1   es           1  Se explica en forma ordenada y didáctica una e...  2010-07-05      2        200
4           4           2   2   es           0  2010-07-05      2          0
...         ...         ...   ...   ...         ...  ...         ...         ...
400          5          -1   1   es          -2  El trabajo pretende ofrecer una visión del uso...  2015-07-05     170        110
401           4          -1   2   es          -1  El paper está bien escrito y de fácil lectura...  2015-07-05     170         89
402           4          -1   1   es           0  Observación de fondo: No se presenta un ejemp...  2015-07-05     171        134
403           3           1   2   es          -1  Se propone un procedimiento para elaborar máqu...  2015-07-05     171         69
404           4           1   1   es          -1  El artículo describe básicamente los component...  2015-07-05     172        107

388 rows x 10 columns
```

## 16. Problem Statement-16

Problem Statement: Display the last 5 reviews where the evaluation score is -2 (very negative).

```
[85] print(df[df['evaluation'] == -2].tail())
```

```
Empty DataFrame
Columns: [confidence, evaluation, id, lan, orientation, remarks, text, timespan, paper_id, word_count]
Index: []
```

## 17. Problem Statement-17

Problem Statement: Check if there are any duplicate rows in the dataset.

```
[86] print(df.duplicated().sum())
```

0

## 18. Problem Statement-18

Problem Statement: Sort the dataset by the 'timespan' column in descending order.

```
[87] print(df.sort_values(by='timespan', ascending=False))
```

```
confidence evaluation id lan orientation \
404      4          1   1  es         -1
364      3          1   2  es          1
346      4         -1   2  es         -1
347      4         -2   1  es         -2
348      4          1   2  es         -1
..      ...      ...  ..  ..         ...
82       4          2   2  es          2
83       2          1   3  es          1
84       4          2   1  es          2
85       5          2   2  es          2
0        4          1   1  es          0

remarks \
404
364 El artículo aborda una temática de gran releva...
346
347 Si se considera un track para mostrar proyecto...
348 Desde mi punto de vista, es un trabajo regular...
..      ...
82
83
84
85
0

text      timespan  paper_id \
404 El artículo describe básicamente los component... 2015-07-05      172
364 Los datos abiertos de Chile Este trabajo anali... 2015-07-05      152
346 el trabajo presenta una posible solución a una... 2015-07-05      143
347 El artículo presenta el desarrollo de una apli... 2015-07-05      144
```



## 19. Problem Statement-19

```

348 Es una buena aplicación, que usa los recursos ... 2015-07-05 144
.. ...
82 Es un excelente trabajo de investigación de ap... 2010-07-05 31
83 Este es un trabajo que muestra la experiencia ... 2010-07-05 31
84 Me ha gustado mucho Lo dejaría tal cual. 2010-07-05 32
85 La ponencia es muy completa es un tema importa... 2010-07-05 32
0 - El artículo aborda un problema contingente y... 2010-07-05 1

word_count
404 107
364 476
346 217
347 182
348 203
.. ...
82 66
83 106
84 8
85 21
0 93

[405 rows x 10 columns]
```

## 20. Problem Statement-20

Problem Statement: Get the count of unique values in the 'lan' column (language column).

```
[88] print(df['lan'].value_counts())
```

```

lan
es    388
en     17
Name: count, dtype: int64
```