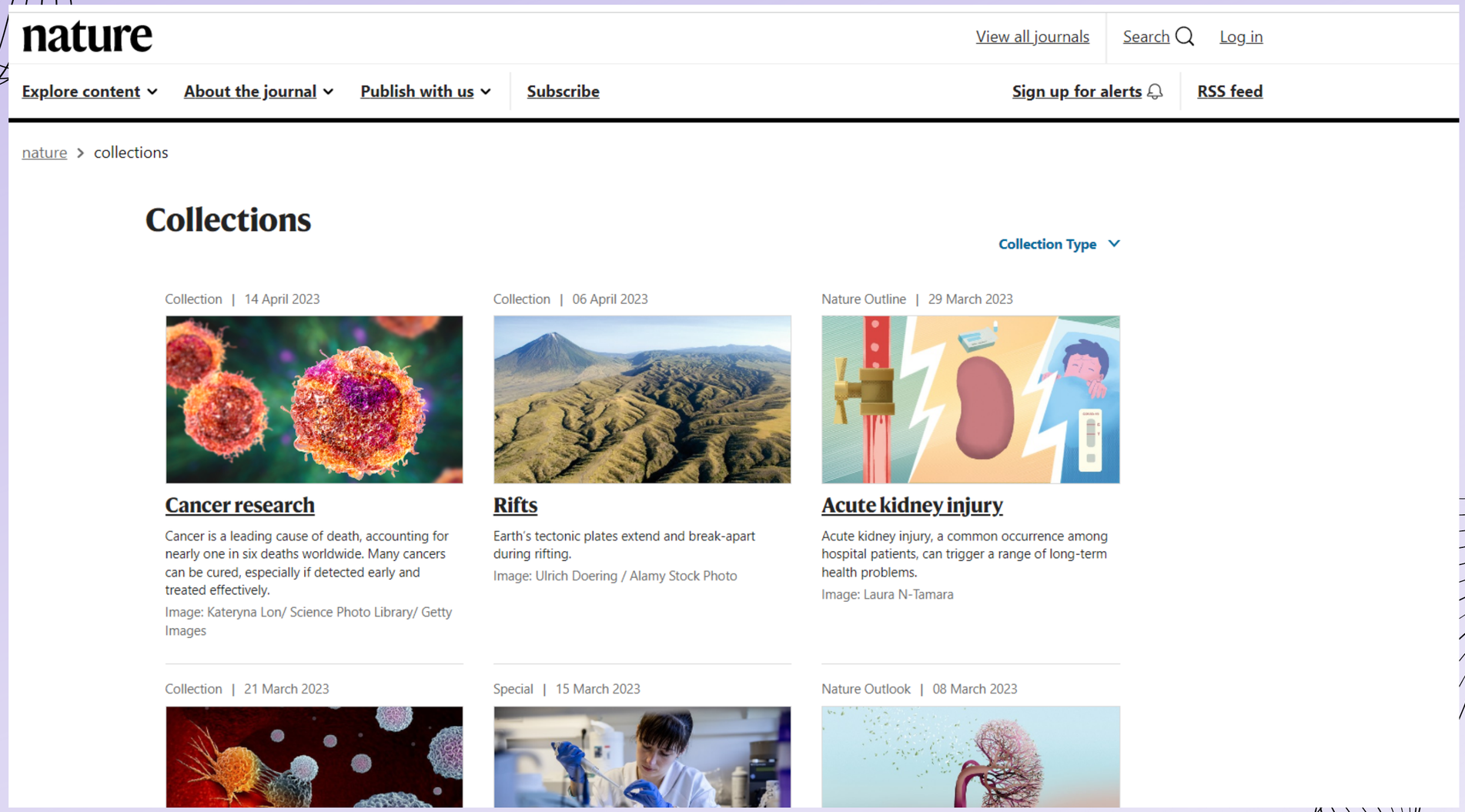
DATA SCIENCE
PROJECT

RESEARCH PAPER PUBLICATION ANALYSIS

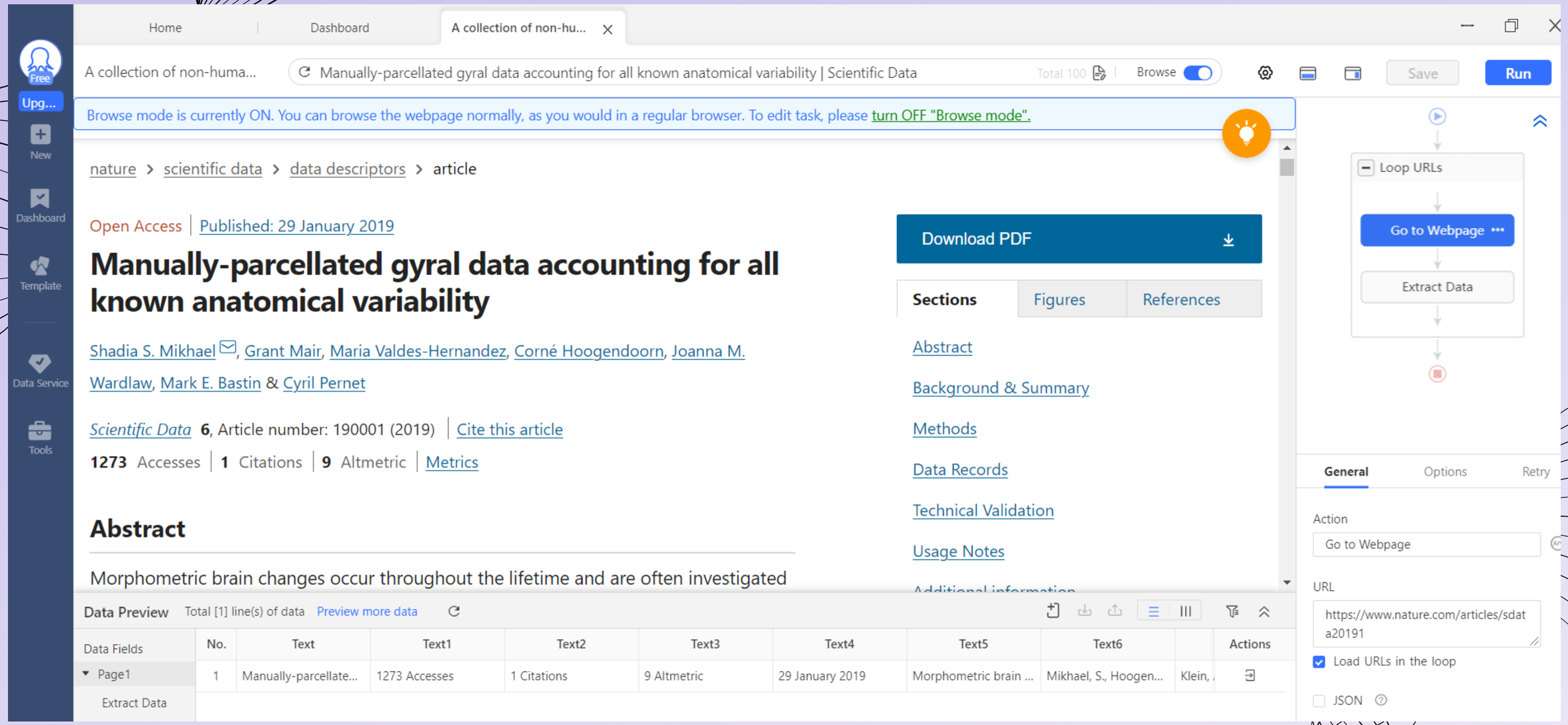
112003110 - DHANASHREE PAWAR
112003111 - NEHA PAWAR
112003130 - ABHISHEK SHELAR

DATA SCRAPING - WEBSITE

https://www.nature.com/nature/collections



DATA SCRAPING - OCTOPARSE



	A	B	C	D	E	F	G
1	Text	Text1	Text2	Text3	Text4	Text5	Text9
2							
3	Detecting and correcting systematic variation in large-scale RNA sequencing data	30k Accesses	113 Citations	55 Altmetric	24 August 2014	High-throughput RNA sequenc	Irizarry, R.A. et al. Multiple
4	The concordance between RNA-seq and microarray data depends on chemical tre	28k Accesses	311 Citations	74 Altmetric	24 August 2014	The concordance of RNA-sequ	Hamburg, M.A. Advancing r
5	Generation of the epicardial lineage from human pluripotent stem cells	10k Accesses	120 Citations	30 Altmetric	21 September 2014	The epicardium supports card	DeRuiter, M.C., Poelmann,
6	The special case of gene therapy pricing	7535 Accesses	50 Citations	77 Altmetric	09 September 2014		Wilson, J.M. Genet. Eng. Ne
7	Scientific rigor and the art of motorcycle maintenance	4420 Accesses	34 Citations	74 Altmetric	09 September 2014	Open Access articles citing thi	Johnson, G. New truths tha
8	Patents or patients: who loses?	1904 Accesses	5 Citations	14 Altmetric	09 September 2014		Grabowski, H. Nat. Rev. Dru
9	OpenBiome remains open to serve the medical community	2084 Accesses	35 Citations	4 Altmetric	09 September 2014		Ratner, M. Nat. Biotechnol.
10	Field trial of Xanthomonas wilt disease-resistant bananas in East Africa	3708 Accesses	54 Citations	57 Altmetric	09 September 2014	Open Access articles citing thi	Tripathi, L. et al. Plant Dis.
11	Stepping into the sunshine	692 Accesses	6 Citations	1 Altmetric	09 September 2014		Bhandari, M. et al. CMAJ 17
12	Startups on the menu: Alnylam	2306 Accesses	3 Altmetric	Metricsdetail:	09 September 2014		
13	New plant species through grafting	3226 Accesses	3 Citations	1 Altmetric	09 September 2014		Fuentes, I., Stegeman, S., G
14	Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleas	31k Accesses	435 Citations	269 Altmetric	21 September 2014	Current antibiotics tend to be	Centers for Disease Contro
15	Building a curriculum for bioentrepreneurs	6424 Accesses	6 Citations	12 Altmetric	09 September 2014		Berret, D. Adjuncts are bett
16	Wheat rescued from fungal disease	3295 Accesses	10 Citations	8 Altmetric	09 September 2014	Open Access articles citing thi	Voytas, D.F. Annu. Rev. Plar
17	Research Highlights	825 Accesses	Metricsdetails		09 September 2014		
18	The devil in the details of RNA-seq	15k Accesses	25 Citations	31 Altmetric	09 September 2014	Open Access articles citing thi	SEQC/MAQC-III Consortium.
19	Mobility, retention and productivity of genomics scientists in the United States	2038 Accesses	9 Citations	5 Altmetric	09 September 2014		Kenney, M. Biotechnology:
20	Bringing RNA-seq closer to the clinic	6949 Accesses	16 Citations	20 Altmetric	09 September 2014	Open Access articles citing thi	MAQC Consortium. Nat. Bio
21	Enzyme clustering accelerates processing of intermediates through metabolic cha	18k Accesses	263 Citations	35 Altmetric	28 September 2014	We present a quantitative mc	James, C.L. & Viola, R.E. Pro
22	Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize	16k Accesses	166 Citations	89 Altmetric	12 October 2014	C4 and C3 photosynthesis diff	Sage, R.F., Sage, T.L. & Koca
23	A bioinspired omniphobic surface coating on medical devices prevents thrombos	29k Accesses	470 Citations	170 Altmetric	12 October 2014	Thrombosis and biofouling of	McCarthy, P.M. & Smith, W.A
24	Honing our reading skills	3217 Accesses	4 Altmetric	Metricsdetail:	09 September 2014		
25	Recent patent applications in differential gene expression	1166 Accesses	1 Altmetric	Metricsdetail:	09 September 2014		
26	People	588 Accesses	Metricsdetails		09 September 2014		
27	Focus on RNA sequencing quality control (SEQC)	7035 Accesses	75 Altmetric	Metricsdetail:	09 September 2014		
28	Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivat	70k Accesses	1060 Citations	88 Altmetric	03 September 2014	Components of the prokaryoti	Barrangou, R. et al. CRISPR
29	Comprehensive characterization of complex structural variations in cancer by dire	21k Accesses	60 Citations	50 Altmetric	26 October 2014	The development of high-thro	Kandoth, C. et al. Mutation
30	Allosteric targeting of receptor tyrosine kinases	10k Accesses	59 Citations	23 Altmetric	07 November 2014	The drug discovery landscape	Hopkins, A.L. & Groom, C.R.
31	Corporate venture capital and Cambridge	4426 Accesses	2 Citations	25 Altmetric	09 October 2014		Pammolli, F., Magazzini, L.
32	Development of the clinical next-generation sequencing industry in a shifting po	5937 Accesses	19 Citations	32 Altmetric	09 October 2014		Crawford, J.M. & Aspinali, M
33	The scope of patent protection for gene technology in China	1987 Accesses	8 Citations	7 Altmetric	09 October 2014	Open Access articles citing thi	Zhu, L. & Guo, J. Guangdong
34	First Rounders Podcast: Mary Tanner	1272 Accesses	1 Altmetric	Metricsdetail:	09 October 2014		
35	Reversal of diabetes with insulin-producing cells derived in vitro from human plu	57k Accesses	968 Citations	327 Altmetric	11 September 2014	Transplantation of pancreatic	Barton, F.B. et al. Improven
36	Nature's contribution to today's pharmacopeia	2835 Accesses	27 Citations	1 Altmetric	09 October 2014	Open Access articles citing thi	Lipinski, C. & Hopkins, A. N
37	A blueprint of cell identity	2654 Accesses	2 Citations	1 Altmetric	09 October 2014		Cahan, P. et al. Cell 158, 90
38	Restoring the pharmaceutical industry's reputation	53k Accesses	42 Citations	99 Altmetric	09 October 2014		Harris Interactive. The Harr
39	Engineered liposomes sequester bacterial exotoxins and protect from severe inv	17k Accesses	146 Citations	253 Altmetric	02 November 2014	Gram-positive bacterial patho	Stefani, S. & Goglio, A. Met
40	Miniaturizing wireless implants	2336 Accesses	26 Citations	2 Altmetric	09 October 2014	Open Access articles citing thi	Beck, H. et al. Am. J. Cardio
41	How deep is enough in single-cell RNA-seq?	10k Accesses	22 Citations	17 Altmetric	09 October 2014	Open Access articles citing thi	Pollen, A.A. et al. Nat. Biote
42	Biopharmaceutical benchmarks 2014	25k Accesses	711 Citations	61 Altmetric	09 October 2014	Open Access articles citing thi	Walsh, G. Biopharmaceutic
43	Bridging the gap between invention and commercialization in medical devices	3456 Accesses	9 Citations	9 Altmetric	09 October 2014		Yock, P. et al. Sci. Transl. M
44	Research Highlights	672 Accesses	1 Altmetric	Metricsdetail:	09 October 2014		

OBJECTIVES

- Domain Extraction
- Finding Popular Author Domainwise
- Predicting Citation of Domain
- Trend analysis of a domain
- H-index for each author

DATA PREPROCESSING -

30k Accesses 113 Citations 55 Altmetric 24 August 2014

Convert dates into a canonical format CF:D:2014-08-24

```
import pandas as pd

# Read the file and specify which column is the date
convert_dates = pd.read_excel("publn_dataset.xlsx")

# Output with dates converted to YYYY-MM-DD
convert_dates["Text4"] = pd.to_datetime(convert_dates["Text4"]).dt.strftime("%Y-%m-%d")
convert_dates["Text4"] = "CF:D:" + convert_dates["Text4"]
convert_dates.to_excel("publn_dataset1.xlsx")
```

Taking only number from Accesses,Citation 30k 113 55

Removing k and adding 1000 for Accesses

30k ---> 30000

```
import pandas as pd
fileName = 'publn_dataset5.xlsx'
df = pd.read_excel(fileName)
#df.dropna(inplace = True)
df = df[df["Text1"].notna()]
df["Text1"] = df["Text1"].replace({"k":"*1e3"},
df.to_excel("removedk.xlsx")
```

DATA PREPROCESSING -

Input -Detecting and correcting systematic variation in large-scale RNA sequencing data

Output - detect correct systematic variation large scale rna sequence data

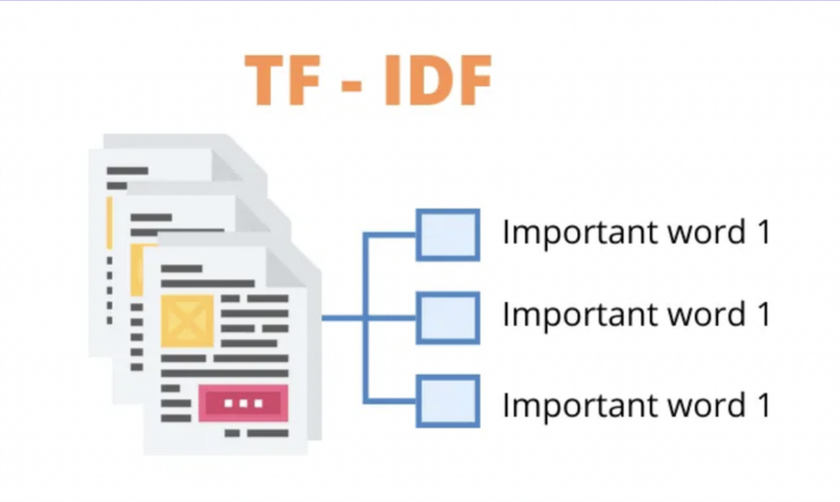
```
import pandas as pd
import string

# Define a function to preprocess the text
def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()

    # Remove punctuation
    text = text.translate(str.maketrans('', '', string.punctuation))

    # Remove stop words
    stop_words = ['the', 'a', 'an', 'and', 'or', 'in', 'of', 'to', 'for', 'with', 'on', 'at', 'by', 'from']
    tokens = text.split()
    tokens = [word for word in tokens if word not in stop_words]

    # Join the tokens back into a string
    text = ' '.join(tokens)

    return text

# Load the Excel file
df = pd.read_excel('file.xlsx', sheet_name='Sheet1')

# Apply the preprocessing function to the text column
df['preprocessed_text'] = df['Text3'].apply(preprocess_text)

# Save the preprocessed data to a new Excel file
df.to_excel('preprocessed_file.xlsx', sheet_name='Sheet1', index=False)
```

Preprocessed Data

1	Unnamed:	Text	preprocessed_text
2	0	Detecting and correcting systematic variation in large-scale	detect correct systemat variat large-scal rna sequenc data
3	1	The concordance between RNA-seq and microarray data de	concord rna-seq microarray data depend chemic treatment tr
4	2	Generation of the epicardial lineage from human pluripoten	gener epicardi lineag human pluripot stem cell
5	3	The special case of gene therapy pricing	special case gene therapi price
6	4	Scientific rigor and the art of motorcycle maintenance	scientif rigor art motorcycl mainten
7	5	Patents or patients: who loses?	patent patient lose
8	6	OpenBiome remains open to serve the medical community	openbiom remain open serv medic commun
9	7	Field trial of Xanthomonas wilt disease-resistant bananas in	field trial xanthomona wilt disease-resist banana east africa
10	8	Stepping into the sunshine	step sunshin
11	9	Startups on the menu: Alnylam	startup menu alnylam
12	10	New plant species through grafting	new plant speci graft
13	11	Sequence-specific antimicrobials using efficiently delivered F	sequence-specif antimicrobi use effici deliv rna-guid nucleas
14	12	Building a curriculum for bioentrepreneurs	build curriculum bioentrepreneur
15	13	Wheat rescued from fungal disease	wheat rescu fungal diseas
16	14	The devil in the details of RNA-seq	devil detail rna-seq
17	15	Mobility, retention and productivity of genomics scientists ir	mobil retent product genom scientist unit state
18	16	Bringing RNA-seq closer to the clinic	bring rna-seq closer clinic
19	17	Enzyme clustering accelerates processing of intermediates t	enzym cluster acceler process intermedi metabol channel
20	18	Comparative analyses of C4 and C3 photosynthesis in develc	compar analys c4 c3 photosynthesi develop leav maiz rice
21	19	A bioinspired omniphobic surface coating on medical device	bioinspir omniphob surfac coat medic devic prevent thrombos
22	20	Honing our reading skills	hone read skill
23	21	Recent patent applications in differential gene expression	recent patent applic differenti gene express
24	22	Focus on RNA sequencing quality control (SEQC)	focu rna sequenc qualiti control seqc
25	23	Rational design of highly active sgRNAs for CRISPR-Cas9–me	ration design highli activ sgrna crispr-cas9–medi gene inactiv
26	24	Comprehensive characterization of complex structural varia	comprehens character complex structur variat cancer directli
27	25	Allosteric targeting of receptor tyrosine kinases	alloster target receptor tyrosin kinas

Sheet1

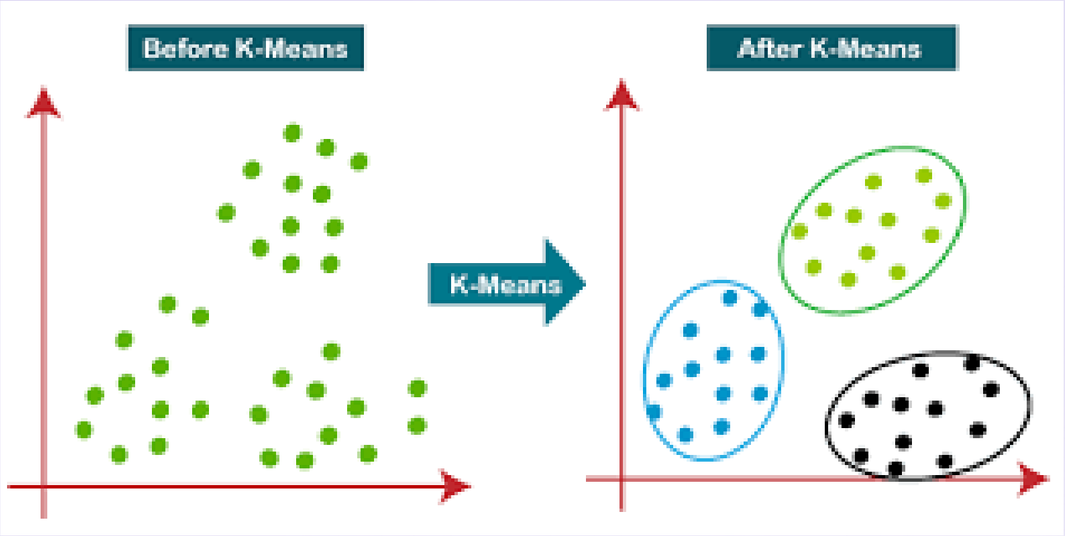
DOMAIN EXTRACTION



The TfidfVectorizer is used to transform the list of research paper titles into a matrix of TF-IDF features.

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

KMeans algorithm is used to predict the domain of the title based on the similarity of its features to the features of the research paper titles in each domain.



```python
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans

# load research paper titles from Excel sheet
df = pd.read_excel('name.xlsx')

df = df.replace(np.nan, '', regex=True)
# extract research paper titles
titles = df['Text'].tolist()

# initialize TfidfVectorizer
vectorizer = TfidfVectorizer(stop_words='english')

# vectorize research paper titles
X = vectorizer.fit_transform(titles)

# initialize KMeans clustering algorithm
kmeans = KMeans(n_clusters=16, random_state=100)

# train KMeans clustering algorithm
kmeans.fit(X)

# load new research paper titles from Excel sheet
new_titles_df = pd.read_excel('file_keywords.xlsx')

# create empty list to store predicted domains
domains = []

# loop through each new research paper title and predict its domain
for i, row in new_titles_df.iterrows():
    new_title = row['keywords']
    if isinstance(new_title, str):
        # vectorize new research paper title
```

```python
# loop through each new research paper title and predict its domain
for i, row in new_titles_df.iterrows():
    new_title = row['keywords']
    if isinstance(new_title, str):
        # vectorize new research paper title
        new_X = vectorizer.transform([new_title])

        # predict domain label of new research paper title
        label = kmeans.predict(new_X)[0]

        # map domain label to corresponding domain
        domain_mapping = {
            0: 'Oncology',
            1: 'Immunology',
            2: 'Cardiology',
            3: 'Neurology',
            4: 'Pediatrics',
            5: 'Epidemiology',
            6: 'Pharmacology',
            7: 'Genetics',
            8: 'Biomedical engineering',
            9: 'Anatomy',
            10: 'Natural Language Processing',
            11: 'Physics',
            12: 'Biology',
            13: 'Finance',
            14: 'Environmental Science',
            15: 'Computer Science'
        }
        domain = domain_mapping[label]
    else:
        domain = np.nan
    domains.append(domain)
```

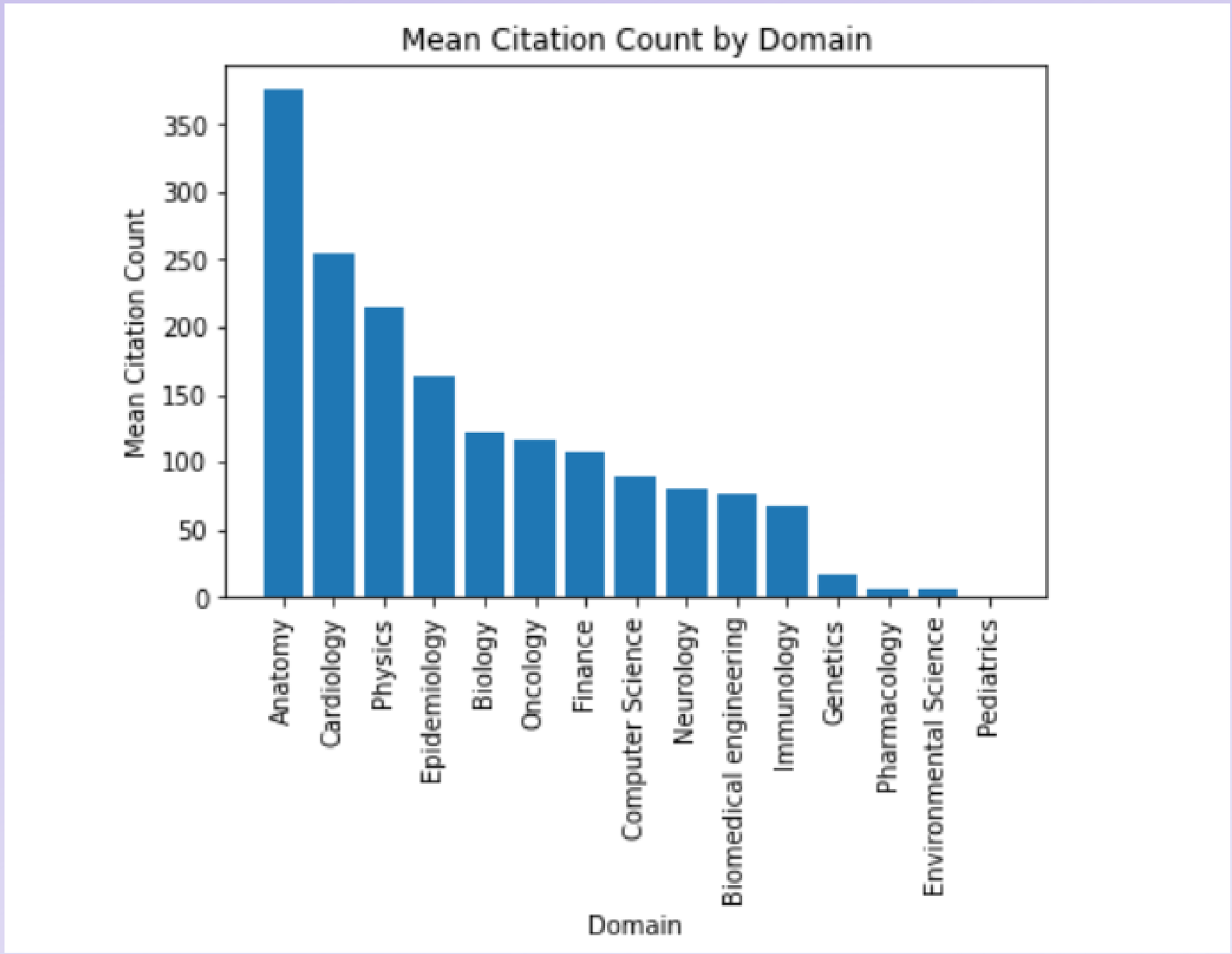
Unnamed:	Text	preprocessed_text	keywords
0	Detecting and correcting systematic variation in large-scale	detect correct systemat variat large-scal rna sequenc data	['detecting', 'systematic variation', 'rna']
1	The concordance between RNA-seq and microarray data de	concord rna-seq microarray data depend chemic treatment tr	['rna-seq', 'microarray data', 'chemical treatment', 'transcript ab
2	Generation of the epicardial lineage from human pluripoten	gener epicardi lineag human pluripot stem cell	['generation', 'epicardial lineage', 'human pluripotent stem cells'
3	The special case of gene therapy pricing	special case gene therapi price	['special case', 'gene therapy']
4	Scientific rigor and the art of motorcycle maintenance	scientif rigor art motorcycl mainten	['scientific', 'motorcycle maintenance']
5	Patents or patients: who loses?	patent patient lose	['patents']
6	OpenBiome remains open to serve the medical community	openbiom remain open serv medic commun	['openbiome', 'medical community']
7	Field trial of Xanthomonas wilt disease-resistant bananas in	field trial xanthomona wilt disease-resist banana east africa	['field', 'xanthomonas', 'wilt disease-resistant bananas', 'africa']
8	Stepping into the sunshine	step sunshin	['stepping']
9	Startups on the menu: Alnylam	startup menu alnylam	['startups', 'alnylam']
10	New plant species through grafting	new plant speci graft	['new plant species']
11	Sequence-specific antimicrobials using efficiently delivered F	sequence-specif antimicrobi use effici deliv rna-guid nucleas	['sequence-specific', 'rna-guided']
12	Building a curriculum for bioentrepreneurs	build curriculum bioentrepreneur	[]
13	Wheat rescued from fungal disease	wheat rescu fungal diseas	['wheat', 'fungal disease']
14	The devil in the details of RNA-seq	devil detail rna-seq	['rna-seq']
15	Mobility, retention and productivity of genomics scientists ir	mobil retent product genom scientist unit state	['mobility', 'genomics scientists']
16	Bringing RNA-seq closer to the clinic	bring rna-seq closer clinic	['bringing rna-seq']
17	Enzyme clustering accelerates processing of intermediates t	enzym cluster acceler process intermedi metabol channel	['enzyme', 'accelerates processing']
18	Comparative analyses of C4 and C3 photosynthesis in develc	compar analys c4 c3 photosynthesi develop leav maiz rice	['comparative', 'c4', 'c3']
19	A bioinspired omniphobic surface coating on medical device	bioinspir omniphob surfac coat medic devic prevent thrombos	['omniphobic surface', 'medical devices prevents thrombosis']
20	Honing our reading skills	hone read skill	['honing', 'reading skills']
21	Recent patent applications in differential gene expression	recent patent applic differenti gene express	['recent', 'patent applications', 'differential gene expression']
22	Focus on RNA sequencing quality control (SEQC)	focu rna sequenc qualiti control seqc	['focus', 'rna', 'quality control', 'seqc']
23	Rational design of highly active sgRNAs for CRISPR-Cas9–me	ration design highli activ sgrna crispr-cas9–medi gene inactiv	['rational', 'active sgrnas', 'crispr-cas9–mediated', 'gene inactiva
24	Comprehensive characterization of complex structural varia	comprehens character complex structur variat cancer directli	['comprehensive', 'complex structural variations', 'genome sequ
25	Allosteric targeting of receptor tyrosine kinases	alloster target receptor tyrosin kinas	['allosteric', 'receptor tyrosine kinases']

DOMAIN EXTRACTION

0: 'Oncology',
1: 'Immunology',
2: 'Cardiology',
3: 'Neurology',
4: 'Pediatrics',
5: 'Epidemiology',
6: 'Pharmacology',
7: 'Genetics',
8: 'Biomedical engineering',
9: 'Anatomy',
10: 'Natural Language Processing',
11: 'Physics',
12: 'Biology',
13: 'Finance',
14: 'Environmental Science',
15: 'Computer Science'

The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance

Oncology


Mean Citation Count by Domain
Mean Citation Count
350
300
250
200
150
100
50
0
Anatomy
Cardiology
Physics
Epidemiology
Biology
Oncology
Finance
Computer Science
Neurology
Biomedical engineering
Immunology
Genetics
Pharmacology
Environmental Science
Pediatrics
Domain

POPULAR AUTHOR DOMAINWISE

```python
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import NMF
from sklearn.preprocessing import normalize

# read Excel sheet into a pandas DataFrame
df = pd.read_excel('example_updated3.xlsx')

# fill in missing values with an empty string
df = df.fillna('')

# extract TF-IDF features from the article text
vectorizer = TfidfVectorizer()
tfidf = vectorizer.fit_transform(df['Text'])
tfidf_norm = normalize(tfidf)

# apply non-negative matrix factorization (NMF) to cluster the articles into topics
nmf = NMF(n_components=10)
W = nmf.fit_transform(tfidf_norm)
H = nmf.components_

# identify the most important topics for each domain
domainwise_top_topics = {}
for domain in df['Domain'].unique():
    # extract articles for the current domain
    domain_articles = df[df['Domain'] == domain]

    # extract TF-IDF features from the article text for the current domain
    domain_tfidf = vectorizer.transform(domain_articles['Text'])
    domain_tfidf_norm = normalize(domain_tfidf)

    # apply NMF to the TF-IDF features for the current domain
    domain_W = nmf.transform(domain_tfidf_norm)

    # identify the top topics for the current domain
    domain_top_topics = domain_W.sum(axis=0).argsort()[::-1][:5].tolist()
    domainwise_top_topics[domain] = domain_top_topics
```

```python
# identify the authors who have written the most articles in the top topics for each domain
domainwise_popular_authors = {}
for domain, top_topics in domainwise_top_topics.items():
    # extract articles that belong to the top topics for the current domain
    topic_articles = df[df['Text'].apply(lambda x: any(str(topic) in x for topic in top_topics))]

    # group the articles by author and count the articles
    author_counts = topic_articles.groupby('Names1')['Text'].count().reset_index()

    # sort the authors by the number of articles they have written
    popular_authors = author_counts.sort_values('Text', ascending=False)['Names1'].tolist()

    # exclude the name 'united' from the list of popular authors
    if 'United' in popular_authors:
        popular_authors.remove('United')

    if 'OReilly' in popular_authors:
        popular_authors.remove('OReilly')

    if 'CONVERGE' in popular_authors:
        popular_authors.remove('CONVERGE')

    # select the top 5 authors (excluding 'united') for the current domain
    domainwise_popular_authors[domain] = popular_authors[:5]

# display the top authors for each domain
for domain, authors in domainwise_popular_authors.items():
    print(f'Top authors in {domain}: {", ".join(authors)}')
```

POPULAR AUTHOR DOMAINWISE

```
Top authors in Neurology: Baker, James, Schrenk, Watson, Arregoces
Top authors in Oncology: Baker, James, Chen, Lowe, Arregoces
Top authors in Immunology: Zetsche, Chen, Zhu, Arregoces, Cho
Top authors in Environmental Science: Zetsche, Chen, Zhu, Lowe, Thomas
Top authors in Anatomy: Zetsche, Chen, Zhu, Lowe, Thomas
Top authors in Epidemiology: Zetsche, Zhu, Arregoces, Lowe, Chen
Top authors in Biology: Watson, James, Chen, Lowe, Abugessaisa
Top authors in Pediatrics: Hsu, Cong, James, Baker, Doudna
Top authors in Genetics: Watson, James, Chen, Lowe, Abugessaisa
Top authors in Pharmacology: Hsu, Cong, James, Baker, Watson
Top authors in Physics: Hsu, Cong, Baker, James, De
Top authors in Biomedical engineering: Hsu, Cong, Zetsche, Doudna, Watson
Top authors in Computer Science: Watson, James, Chen, Lowe, Abugessaisa
Top authors in Cardiology: Zetsche, Zhu, Arregoces, Lowe, Chen
Top authors in Finance: Zetsche, Chen, Zhu, Arregoces, Cho
```

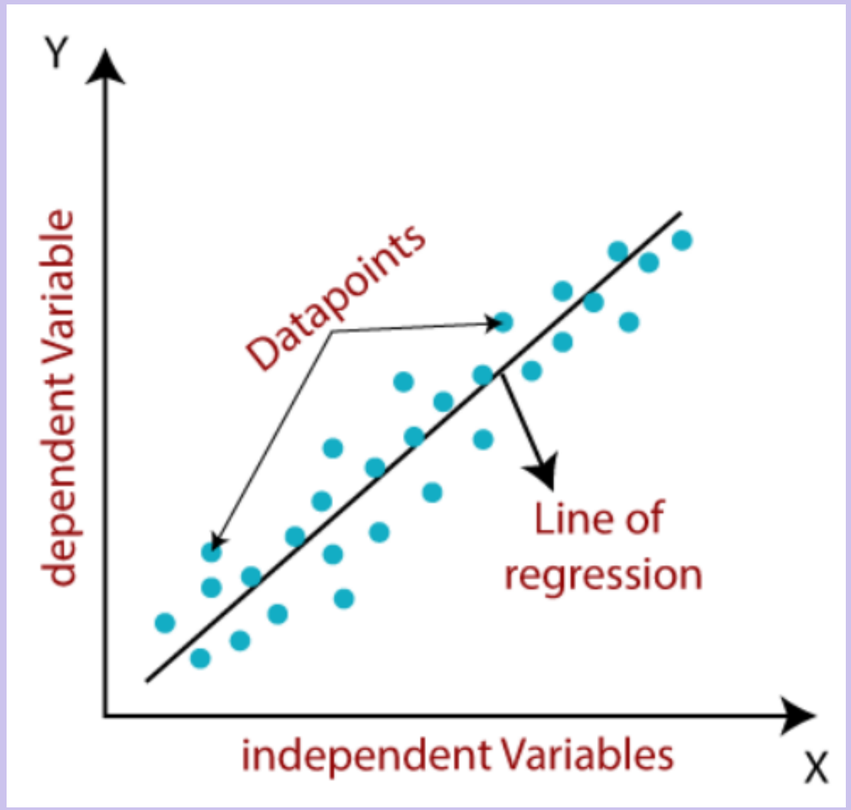
PREDICTING CITATION OF DOMAIN

Model Used :- Linear Regression

Linear regression is a supervised machine learning algorithm that models the linear relationship between the input features and the output variable.

Parameters used :-

- Different domains represented as binary variables in the input dataset
- The corresponding citation counts for each paper



Reference image :-
https://www.javatpoint.com/linear-regression-in-machine-learning

PREDICTING CITATION OF DOMAIN

```python
import pandas as pd
from sklearn.linear_model import LinearRegression

# Load dataset into a pandas DataFrame
df = pd.read_csv("domainnew.csv")

# Prepare X and y variables for linear regression
X = pd.get_dummies(df["Domain"], drop_first=True)
y = df["Citation"]

# Train linear regression model
model = LinearRegression()
model.fit(X, y)
```

```python
# Make predictions on new observations
new_data = pd.DataFrame({"Domain": ["covid", "economics", "mechanics"],
                         "Oncology" : [0, 0, 0],
                         "Immunology" : [1, 0 ,0],
                         "Cardiology": [0, 0, 0],
                         "Neurology":[0, 0, 0],
                         "Pediatrics":[0, 0, 0],
                         "Epidemiology":[0, 0, 0],
                         "Pharmacology":[1, 0, 0],
                         "Genetics":[1, 0, 0],
                         "Biomedical engineering":[1, 0, 0],
                         "Anatomy":[1, 0, 0],
                         "Natural Language Processing":[0, 0, 0],
                         "Physics":[0, 0, 1],
                         "Biology":[1, 0, 0],
                         "Finance":[0, 1, 0],
                         "Environmental Science":[0, 0, 0],
                         "Computer Science":[0, 0, 0]
                        })
#new_X = new_data.drop("Domain", axis=1)
new_X = new_data[X.columns]
predictions = model.predict(new_X)

# Print predicted citation counts for new observations
for i, pred in enumerate(predictions):
    print(f"Predicted citations for {new_data.iloc[i]['Domain']}: {pred}")
```

```
Predicted citations for covid: -1211.800129883292
Predicted citations for economics: 108.06060606060589
Predicted citations for mechanics: 215.14285714285668
```

H-INDEX OF AUTHORS

The h-index is based on the number of publications and the number of citations that those publications have received.

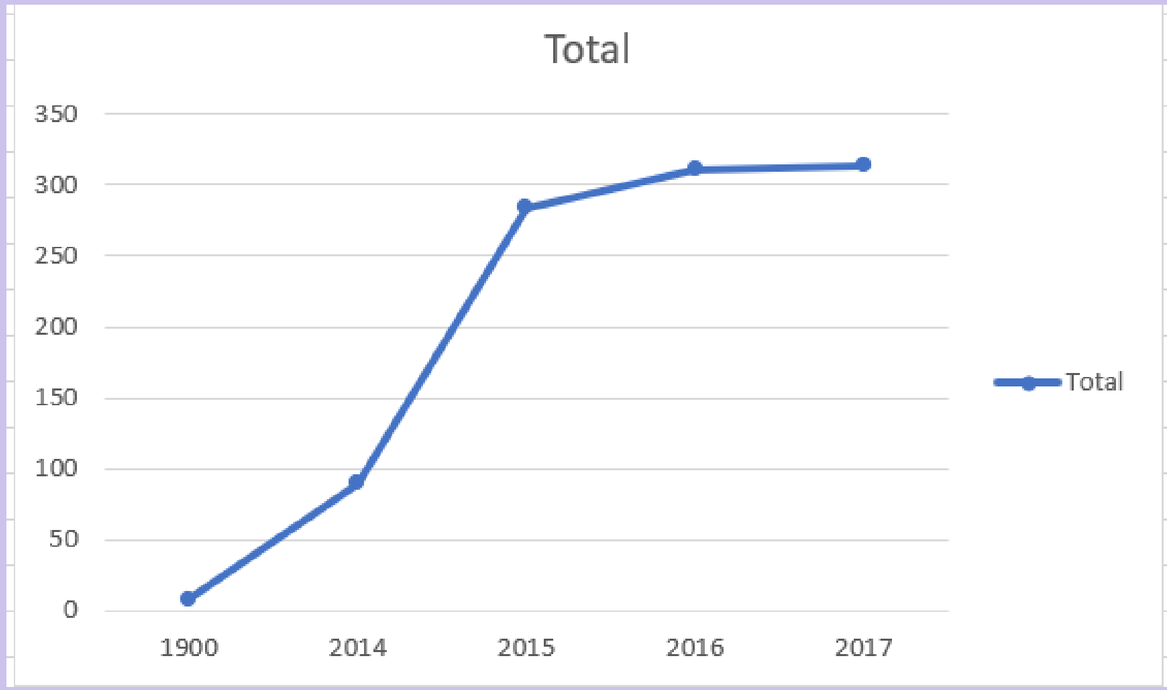
```python
import pandas as pd

# Load data from Excel file
#file_path = input("Enter the path to the Excel file: ")
df = pd.read_excel('name1.xlsx')

# Group papers by author
grouped_data = df.groupby('Author')

# Calculate h-index for each author
for author_name, group in grouped_data:
    # Sort papers by number of citations in descending order
    group_sorted = group.sort_values('citation', ascending=False)

    # Calculate h-index
    h_index = 0
    for i, row in group_sorted.iterrows():
        if row['citation'] >= (i + 1):
            h_index += 1
        else:
            break

    # Print the h-index for the current author
    print(f"The h-index of {author_name} is {h_index}.")
```

```
The h-index of Centers is 5.
The h-index of Cerletti is 1.
The h-index of Cermak is 1.
The h-index of Ceroni is 0.
The h-index of Cerrone is 0.
The h-index of Cevallos is 0.
The h-index of Chadwick is 2.
The h-index of Chaffer is 1.
The h-index of Chalhoub is 2.
The h-index of Chamberlain is 2.
The h-index of Chambers is 5.
The h-index of Chan is 1.
The h-index of Chandler is 0.
The h-index of Chang is 2.
The h-index of Chapin is 1.
The h-index of Chapman is 1.
The h-index of Chappell is 1.
The h-index of Chari is 1.
```

Trend analysis of a domain using citations

Years	Count of citations
1900	8
2014	90
2015	284
2016	311
2017	314
Grand Total	**1007**

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the citation data from a CSV file
df = pd.read_csv('publn_dataset.csv')

# Convert the publication year column to datetime format
df['Years'] = pd.to_datetime(df['Years'], format='%Y')

# Group the papers by year and count the number of citations
citation_counts = df.groupby(df['Years'].dt.year)['citations'].sum()

# Create a line chart of the citation counts over time
citation_counts.plot(kind='line', title='Citation Trends Over Time')

# Add axis labels
plt.xlabel('Years')
plt.ylabel('Count of citations')

# Show the plot
plt.show()
```



- By performing trend analysis of a domain using citations,researchers can gain valuable insights into the state of research field and use this information for predicting future trends of the domain.
- Trend analysis can help identify whether a particular research area is gaining or losing popularity.

THANK YOU