Big Data Analytics

Assignment 3

Stream Processing

Course: CSE-557
Course Name: Big Data Analytics
Instructor: Dr. Vikram Goyal

Team 4

Deep Sharma	- 2020370
Mrishika Nair	- 2020389

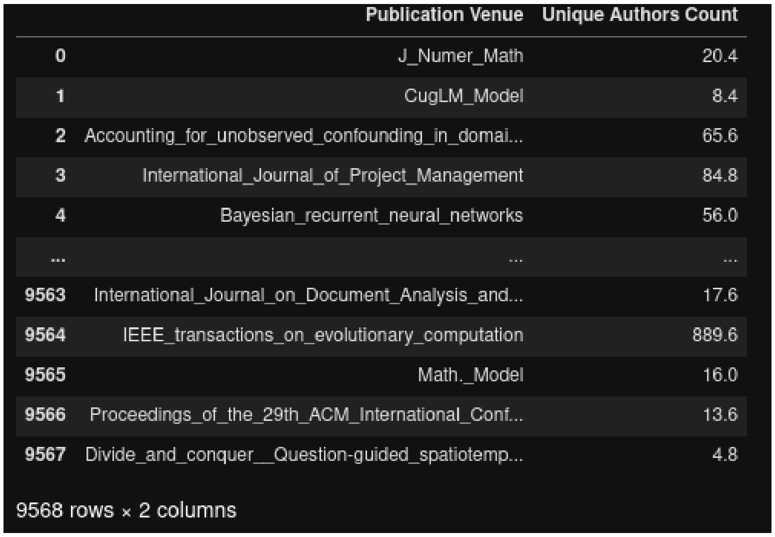
CONTENTS

1. Methodology

Generated streams using consumer and producer in Apache Kafka. Used Flajolet Martin Algorithm and DGIM Algorithm as asked.

2. Results

2.1. Unique Authors

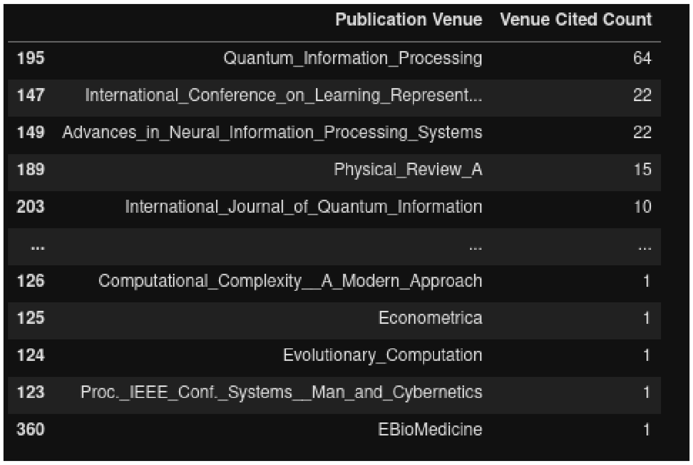


	Publication Venue	Unique Authors Count
0	J_Numer_Math	20.4
1	CugLM_Model	8.4
2	Accounting_for_unobserved_confounding_in_domai...	65.6
3	International_Journal_of_Project_Management	84.8
4	Bayesian_recurrent_neural_networks	56.0
...	...	...
9563	International_Journal_on_Document_Analysis_and...	17.6
9564	IEEE_transactions_on_evolutionary_computation	889.6
9565	Math._Model	16.0
9566	Proceedings_of_the_29th_ACM_International_Conf...	13.6
9567	Divide_and_conquer__Question-guided_spatiotemp...	4.8

9568 rows × 2 columns

Fig 2.1.1: Author counts for each publication venue ("topic") using Flajolet Martin Algorithm

2.2. Number of times a publication venue is cited in the last 500 items



	Publication Venue	Venue Cited Count
195	Quantum_Information_Processing	64
147	International_Conference_on_Learning_Represent...	22
149	Advances_in_Neural_Information_Processing_Systems	22
189	Physical_Review_A	15
203	International_Journal_of_Quantum_Information	10
...	...	...
126	Computational_Complexity__A_Modern_Approach	1
125	Econometrica	1
124	Evolutionary_Computation	1
123	Proc._IEEE_Conf._Systems__Man_and_Cybernetics	1
360	EBioMedicine	1

Fig 2.1.2: Number of times a publication venue is cited using DGIM Algorithm for the last 500 items in stream

3. References and Learning Resources

3.1. Libraries

1.	Matplotlib 3.6.0	- https://matplotlib.org
2.	Numpy 1.23.3	- https://numpy.org
3.	Pandas 1.4.4	- https://pandas.pydata.org
4.	Scikit Learn 1.1.3	- https://scikit-learn.org
5.	Apache Spark 3.3.2	- https://spark.apache.org
6.	Scipy 1.9.1	- https://scipy.org
7.	Murmur Hash 3	- https://pypi.org/project/mmh3
8.	Apache Kafka 3.4.0	- https://kafka.apache.org

3.2. Books

1. Mining of Massive Datasets, Anand Rajaraman, Jeffrey Ullman