

گزارش پروژه درس مبانی علم داده

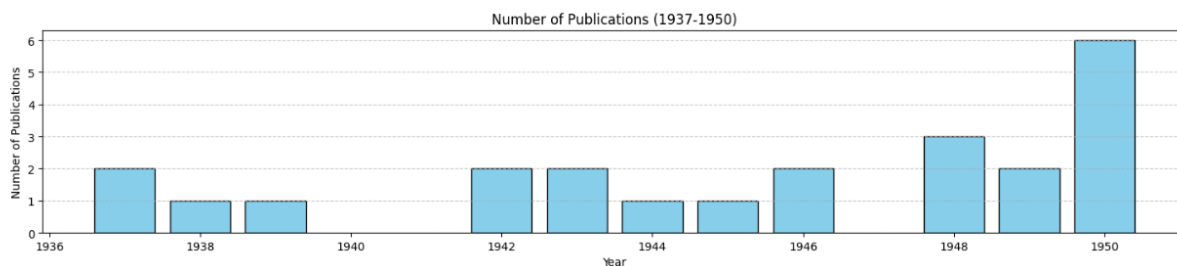
امیرحسین توکلی

۹۹۱۰۹۱۴۴

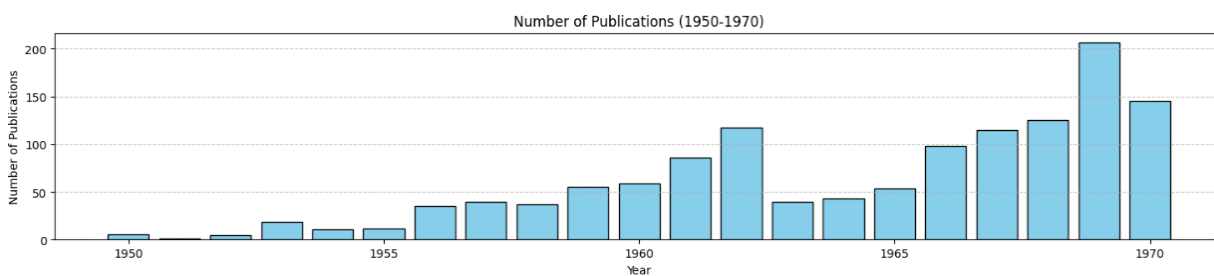
لینک Github:

https://github.com/AmirT000/FDS_project

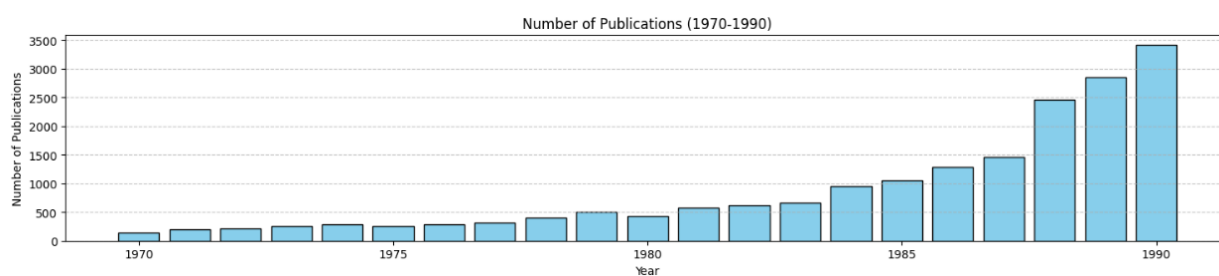
۱.۱.۱ میزان انتشار مقاله به طور چشمگیری افزایش پیدا کرده است. تعداد انتشار بین ۱۹۳۷ تا ۱۹۵۰:



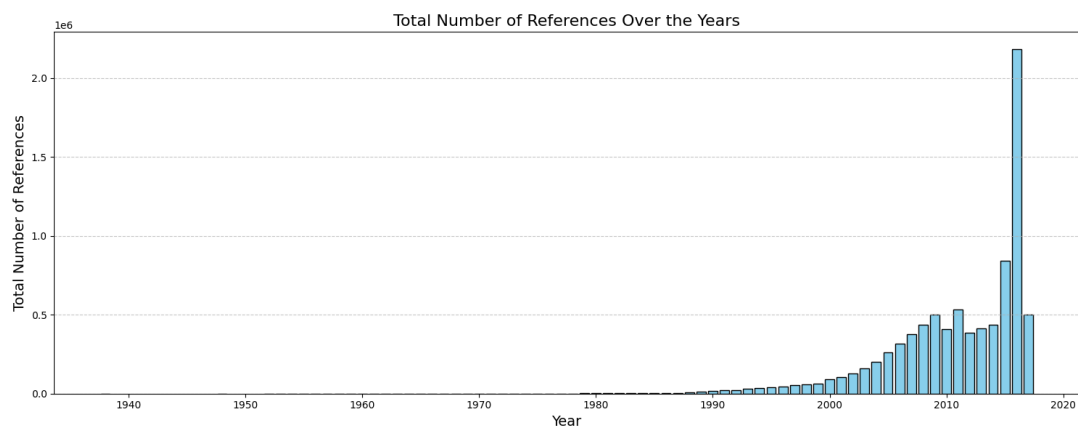
تعداد انتشار بین سال‌های ۱۹۷۰ تا ۱۹۹۰:



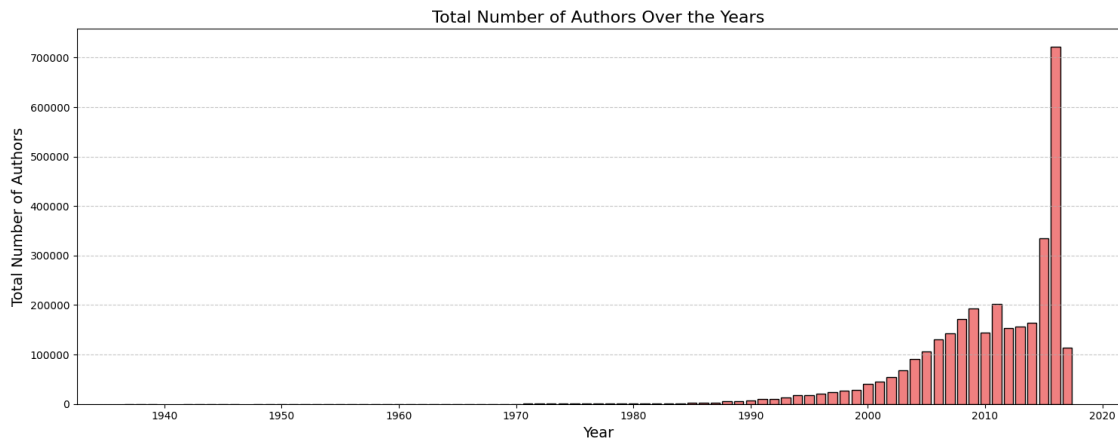
تعداد انتشار بین سال‌های ۱۹۷۰ تا ۱۹۹۰:



۲.۱.۱ تعداد رفرنس‌ها برحسب زمان:



۳.۱.۱ تعداد نویسندگان برحسب زمان شباهت زیادی به نمودار تعداد رفرنس‌ها در زمان دارد:



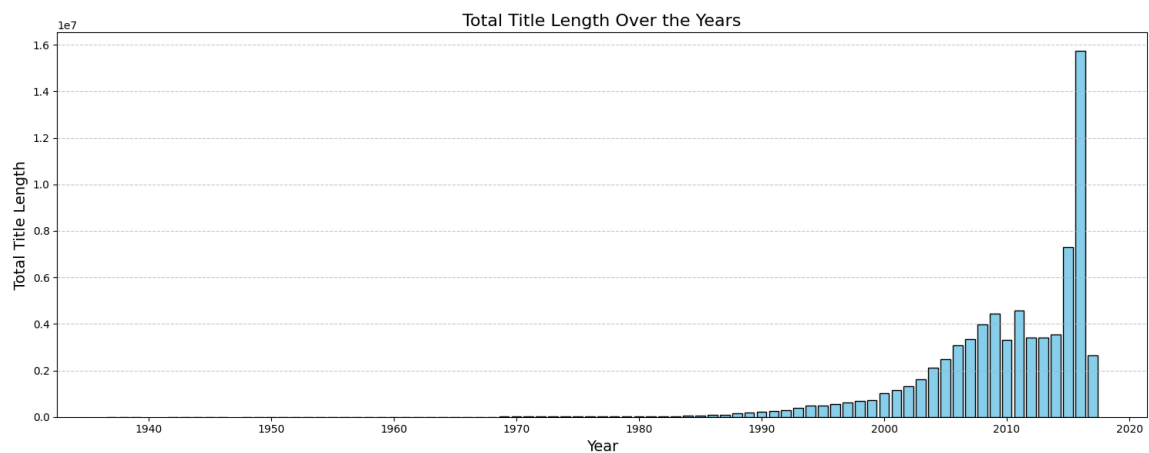
۴.۱.۱ Correlation نویسندگان و رفرنس‌ها:

Pearson Correlation Coefficient: 0.0560 (p-value: 0.0000)
Spearman Rank Correlation Coefficient: 0.0872 (p-value: 0.0000)

۵.۱.۱ Correlation نویسندگان و تعداد سایتیشن:

Pearson Correlation Coefficient: -0.0028 (p-value: 0.0052)
Spearman Rank Correlation Coefficient: -0.0166 (p-value: 0.0000)

۶.۱.۱



۱۱.۱.۱ ده مقاله برتر بر اساس تعداد رفرنس‌ها

title	num_references
Comprehensive frequency-dependent substrate no...	759
Time in Qualitative Simulation.	561
Bibliography on cyclostationarity	412
Fifty Years of MIMO Detection: The Road to Lar...	396
An Exploration of Enterprise Architecture Rese...	394
Structure and dynamics of molecular networks: ...	386
The NP-completeness column: An ongoing guide	363
Digital geometry	361
Deep Learning: Methods and Applications	343
Review: learning bayesian networks: Approaches...	326

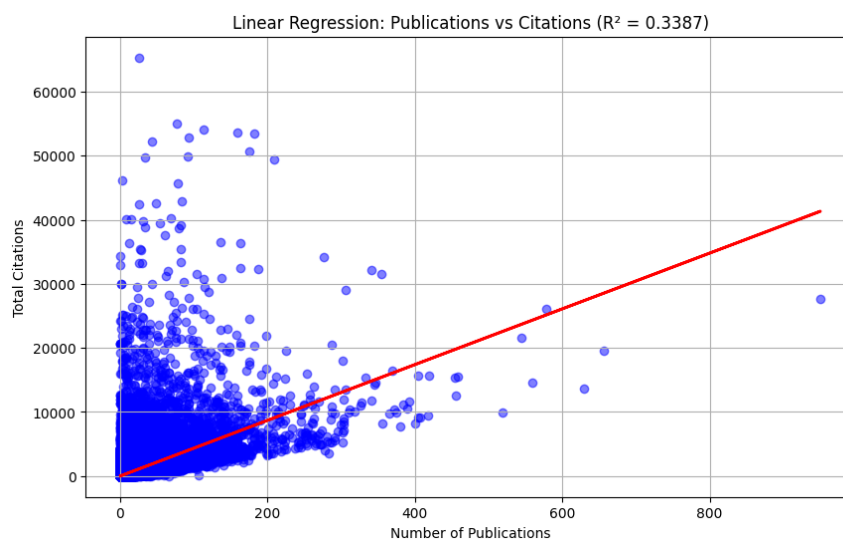
۱۲.۱.۱ ده مقاله برتر بر اساس تعداد سائیتیشن

title	n_citation
Distinctive Image Features from Scale-Invarian...	42508
Bowling alone: the collapse and revival of Ame...	34288
LIBSVM: A library for support vector machines	33016
Random Forests	28679
Support-Vector Networks	26114
MapReduce: simplified data processing on large...	24381
A fast and elitist multiobjective genetic algo...	24245
A theory for multiresolution signal decomposit...	24182
ImageNet Classification with Deep Convolutiona...	22884
Histograms of oriented gradients for human det...	22795

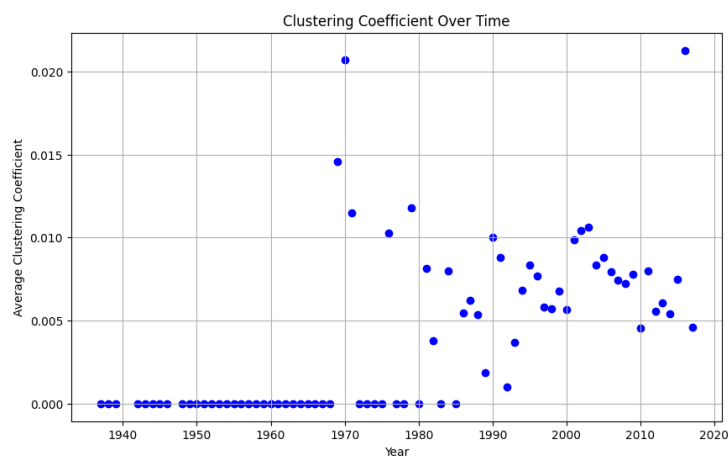
مشاهده می‌کنیم که این دو لیست هم اشتراکی با یکدیگر ندارند

۱۳.۱.۱ نمودار تعداد سائیتیشن و انتشار نویسنده‌ها را رسم کرده و Linear Regression انجام می‌دهیم.

مشخص است که خطای زیادی دارد و نمی‌توان پیشبینی کرد



۱.۲.۱ میانگین clustering coefficient شبکه سایتیشن برحسب سال:



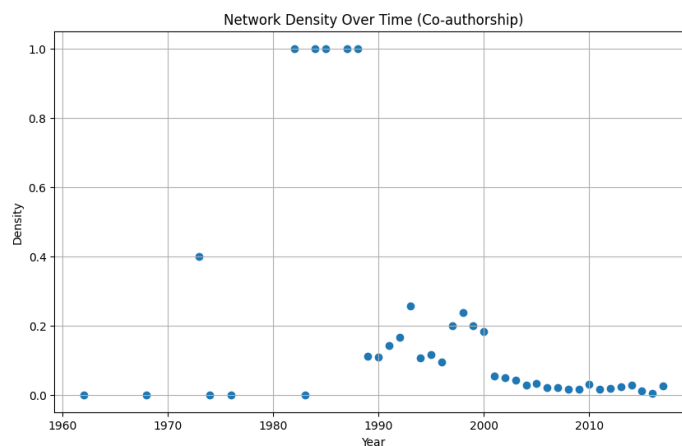
قطر گراف و ۱۰ مقاله موثر:

Average Path Length (SCC): 11.288044063779358
Diameter (SCC): 29

Top 10 Influential Papers (based on PageRank):

1. Paper ID: 6a6b9aa6-683f-4c7c-b06e-9c3018d10fd3, PageRank Score: 0.0002234734396393853
2. Paper ID: c1b6b493-01ef-420f-be44-7bacfe34e846, PageRank Score: 0.00019128161048528232
3. Paper ID: b944f77f-113b-4a02-ae5e-d4a124b8fd5b, PageRank Score: 0.00017979478285836238
4. Paper ID: f6bd8b64-684d-429a-aab5-8ff3a2c23cd6, PageRank Score: 0.00013839603811788994
5. Paper ID: 2659531e-eb9d-4dd5-b46f-10f66a4819c6, PageRank Score: 0.00011580422473495846
6. Paper ID: 748a2ab3-8b5f-4d0a-9e2d-af685089843a, PageRank Score: 0.00010567984642424661
7. Paper ID: e0f3a738-4ab2-40d1-ba44-506d81c1d230, PageRank Score: 9.731396921113041e-05
8. Paper ID: 8026f56a-a93e-4933-8ead-c9aa9e3f0498, PageRank Score: 9.394110306534417e-05
9. Paper ID: 7ccbdf09-a84e-4ad2-ab20-cb28b6c41155, PageRank Score: 9.327124792219694e-05
10. Paper ID: d3e00e7e-1c64-4d7a-b2b2-1ad98ba4c706, PageRank Score: 9.26653538193433e-05

۲.۲.۱ از هزار سمپل رندوم برای ساخت گراف استفاده کردیم. چگالی شبکه Co-authorship برحسب سال:



ده نویسنده تاثیرگذار:

Top 10 Influential Authors by Degree Centrality:				
	Author	Degree	Betweenness	Closeness
2691	Evgeni M. Zdobnov	0.008991	0.0	0.008991
2675	Laurent Falquet	0.008991	0.0	0.008991
2688	Marco Pagni	0.008991	0.0	0.008991
2687	Tom Oinn	0.008991	0.0	0.008991
2686	Nicola Mulder	0.008991	0.0	0.008991
2685	Beate Marx	0.008991	0.0	0.008991
2684	Rodrigo Lopez	0.008991	0.0	0.008991
2683	Youla Karavidopoulou	0.008991	0.0	0.008991
2682	Alexander Kanapin	0.008991	0.0	0.008991
2681	Daniel Kahn	0.008991	0.0	0.008991

۵ کامیونیتی به طور مثال:

Example of Author Communities (Showing 5 authors from each community):
Community 0: ['Maria G. Koziri', 'Panos Papadopoulos', 'Nikos Tziritas', 'Antonios N. Dadaliaris', 'Thanasis Loukopoulous'] ...
Community 1: ['Luís Fernando Orleans', 'Geraldo Zimbrão'] ...
Community 2: ['Artur Zawadzki', 'Marek Gorgon'] ...
Community 3: ['Yadong Wang', 'Jiankang Wu', 'Ashraf A. Kassim'] ...
Community 4: ['Arber Murturi', 'Burak Kantarci', 'Sema Oktug'] ...

۳.۲.۱

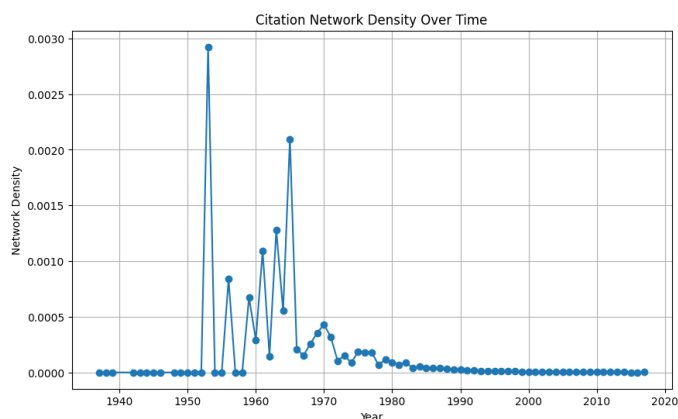
Top Interdisciplinary Clusters:
Community 1: 1070 venues → ['IEEE Computer Graphics and Applications', 'international conference in central europe on computer graphics and visualization', 'IE
Community 3: 987 venues → [nan, 'international conference on management of data', 'very large data bases', 'international conference on data engineering', 'Com
Community 0: 853 venues → ['international symposium on computers and communications', 'IEEE Communications Letters', 'IEEE Journal on Selected Areas in Communi
Community 2: 468 venues → ['IEEE Transactions on Information Theory', 'foundations of computer science', 'ACM Communications in Computer Algebra', 'ACM Sigsam
Community 4: 384 venues → ['programming language design and implementation', 'symposium on principles of programming languages', 'compiler construction', 'conf

Top Influential Venues (Degree Centrality):
nan: 0.8227
Communications of The ACM: 0.5078
Lecture Notes in Computer Science: 0.4786
IEEE Transactions on Pattern Analysis and Machine Intelligence: 0.438.
IEEE Transactions on Information Theory: 0.4289
systems man and cybernetics: 0.4246
IEEE Transactions on Knowledge and Data Engineering: 0.4241
ACM Computing Surveys: 0.4103
neural information processing systems: 0.4097
IEEE Computer: 0.4018

Top Influential Venues (PageRank):
nan: 0.0076
IEEE Transactions on Information Theory: 0.0030
Communications of The ACM: 0.0029
Lecture Notes in Computer Science: 0.0028
IEEE Transactions on Pattern Analysis and Machine Intelligence: 0.0027
systems man and cybernetics: 0.0025
neural information processing systems: 0.0025
IEEE Transactions on Knowledge and Data Engineering: 0.0024
IEEE Computer: 0.0023
ACM Computing Surveys: 0.0023

Examples of Interdisciplinary Connections: 2013: 27121 new venue connections
Examples: [(['computer software and applications conference', 'usenix security symposium'), ('International Journal of Network Security', 'international sym
2014: 27172 new venue connections
Examples: [(['Bulletin of The European Association for Theoretical Computer Science', 'principles of knowledge representation and reasoning'), ('Constraints
2015: 49880 new venue connections
Examples: [(['IEEE Wireless Communications', 'Journal of Computer Applications in Technology'), ('Computers & Graphics', 'international conference in centra
2016: 96649 new venue connections
Examples: [(['ACM Transactions on Mathematical Software', 'IEEE Transactions on Image Processing'), ('computer software and applications conference', 'usenix
2017: 16645 new venue connections
Examples: [(['ACM Journal on Emerging Technologies in Computing Systems', 'high performance computing and communications'), ('IEEE Communications Magazine',

۴.۲.۱ چگالی شبکه سائیتیشن بر حسب سال:



ده مقاله برتر:

Top 10 Bursting Papers (Highest Citation Growth):
Paper ID: b944f77f-113b-4a02-ae5e-d4a124b8fd5b, Citations: 5841
Paper ID: c1b6b493-01ef-420f-be44-7bacfe34e846, Citations: 5057
Paper ID: 6a6b9aa6-683f-4c7c-b06e-9c3018d10fd3, Citations: 3288
Paper ID: dd83785a-dd19-41e3-9b25-ebabbd48d336, Citations: 3279
Paper ID: e2f7a74a-8430-4463-94ce-fe85dfd309f9, Citations: 3242
Paper ID: f6bd8b64-684d-429a-aab5-8ff3a2c23cd6, Citations: 3235
Paper ID: 50dd56db-151d-4d62-8576-65f0ef6f381b, Citations: 2281
Paper ID: 8026f56a-a93e-4933-8ead-c9aa9e3f0498, Citations: 2279
Paper ID: 748a2ab3-8b5f-4d0a-9e2d-af685089843a, Citations: 2259
Paper ID: ebfca554-7a3c-4597-954b-07336a2e3030, Citations: 2238

۱.۲ ابتدا روی هزار سمپل رندوم، سه متود Louvain، spectral و hierarchical را تست کرده و clustering coefficient را برای هر کدام محاسبه می‌کنیم. Hierarchical بهترین نتیجه را می‌دهد:

```
Graph created with 3004 nodes and 4866 edges.  
Louvain Method: Communities: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99}  
Spectral Clustering: Communities: {0, 1, 2, 3, 4}  
<ipython-input-5-93e79364cae3>:69: ClusterWarning: The symmetric non-negative semi-definite matrix is not positive semi-definite.  
Z = sch.linkage(distance_matrix, method='ward')  
Hierarchical Clustering: Communities: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99}  
Louvain Method Clustering Coefficients: 0.6702794077223301  
Spectral Clustering Coefficients: 0.5785880424917773  
Hierarchical Clustering Coefficients: 0.9769357495881383  
Best clustering method: Hierarchical with coefficient 0.9769357495881383
```

سپس این متود را روی کل داده‌ها ترین می‌کنیم.