

تحلیل و بررسی مقالات علمی نوشتہ شده در ایران

پرنده علیزاده ۹۴۱۰۰۲۴

جمع آوری داده

داده‌های موجود برای مسئله به صورت چند فایل مجزا هستند که هر کدام قسمتی از داده را در برمی‌گیرد. در این قسمت توضیح کوتاهی از داده‌های موجود در هر فایل ارائه می‌کنیم:

- ۱ affiliation_auid : شناسه هر موسسه و دانشگاه را در scopus نشان می‌دهد.
- ۲ affiliation_status : اطلاعاتی راجع به موسسات و دانشگاه‌های مختلف در اختیار ما می‌گذارد. اطلاعاتی از قبیل تعداد مقالات در سایت‌های مختلف و اطلاعاتی راجع به تعداد نویسنده‌های آن موسسه.
- ۳ affiliation : نام و آدرس کامل دانشگاه‌ها و موسسات.
- ۴ article_author_affiliation : نویسنده‌ها و موسسات مقالات مختلف.
- ۵ articlesource : اطلاعاتی راجع به زورنال‌ها و ناشرهای مقالات.
- ۶ author_subaffiliation : اطلاعاتی راجع به دپارتمان و زیرموسسات نویسنده‌ها.
- ۷ authors : اطلاعاتی راجع به نویسنده‌ها و علایق تحقیقاتی آنها.
- ۸ subaffiliation : اطلاعاتی راجع به دپارتمان‌های موجود.
- ۹ article_details : اطلاعات کاملی راجع به هر مقاله، شامل نام، نویسنده‌ها، چکیده، کلمات کلیدی، منابع و ..
- ۱۰ articles : خلاصه‌ای از اطلاعات هر مقاله که کامل‌تر آن در مورد قبلی آمده است.

تحلیل اکتشافی داده

در داده ما 413,116 مقاله و 187,096 نویسنده مقاله وجود دارد.

تحلیل لغوی

با استفاده از عنوان و کلمات کلیدی مقالات و تجزیه آنها به کلمات و حذف لغات بی معنی و کم اهمیت تر به دو کنگکاوی زیر پاسخ میدهیم:

- ۱- لغات پر استفاده در عنوان مقالات. نمودار ۱ (ک در R P1.R)

women aqueous
temperature production optical
mechanical investigation assessment
ique disease parameters methods
nonlinear evaluation simulation
modified treatment optimization review
treatment quality
gas fuzzy design type distribution
in conditions carbon efficient Des
e water synthesis multi heat li
ty control networks
e characterization system power rats
s systems low acid neural
or algorithm oil network single
free performance induced flow nanogen
data nanoparticles process stress
determination liquid solution
field detection factors risk
oxide energy Study hybrid transfer
agement Application thermal
Analysis characteristics

۲- لغات پر استفاده در کلمات کلیدی مقالات. نمودار ۲ (کد در P2.R)

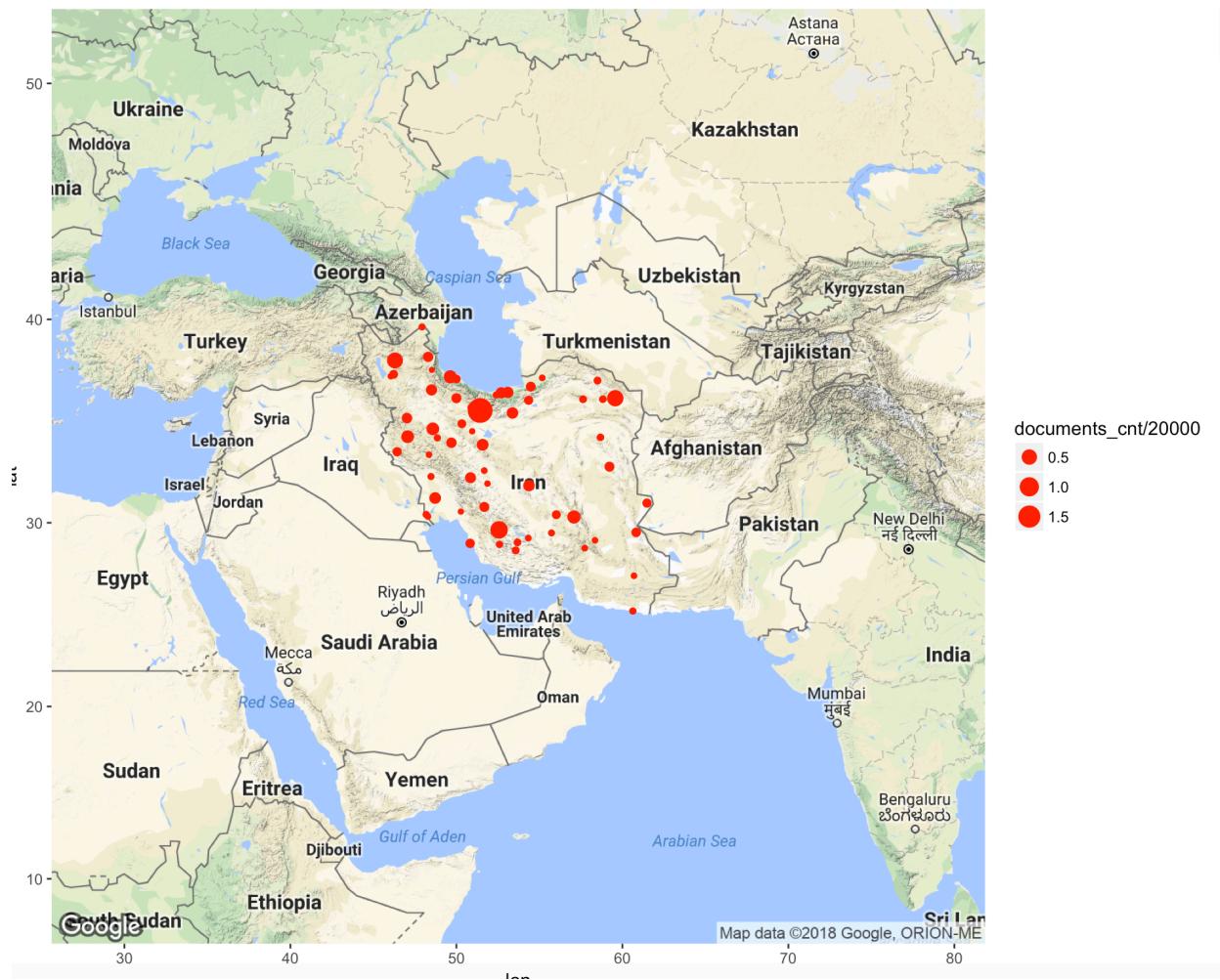


دو نمودار بالا آن قدری که انتظار میرفت شباهتی ندارند و نمودار اول که مربوط به عنوان است لغات کامپیوتري، زيستي، شيميايی و غيره در آن به چشم میخورد اما در نمودار دوم که کلمات کلیدی هستند عمدتاً لغات زيستي هستند. شاید بتوان نتيجه گرفت که مقالات زيستي بيشتر هستند اما در عنوان هاي آنها مباحث ديگري بررسی میشوند (مثل شيمي) و تاثيرات آنها روی زيست شناسی اندازه گرفته میشود.

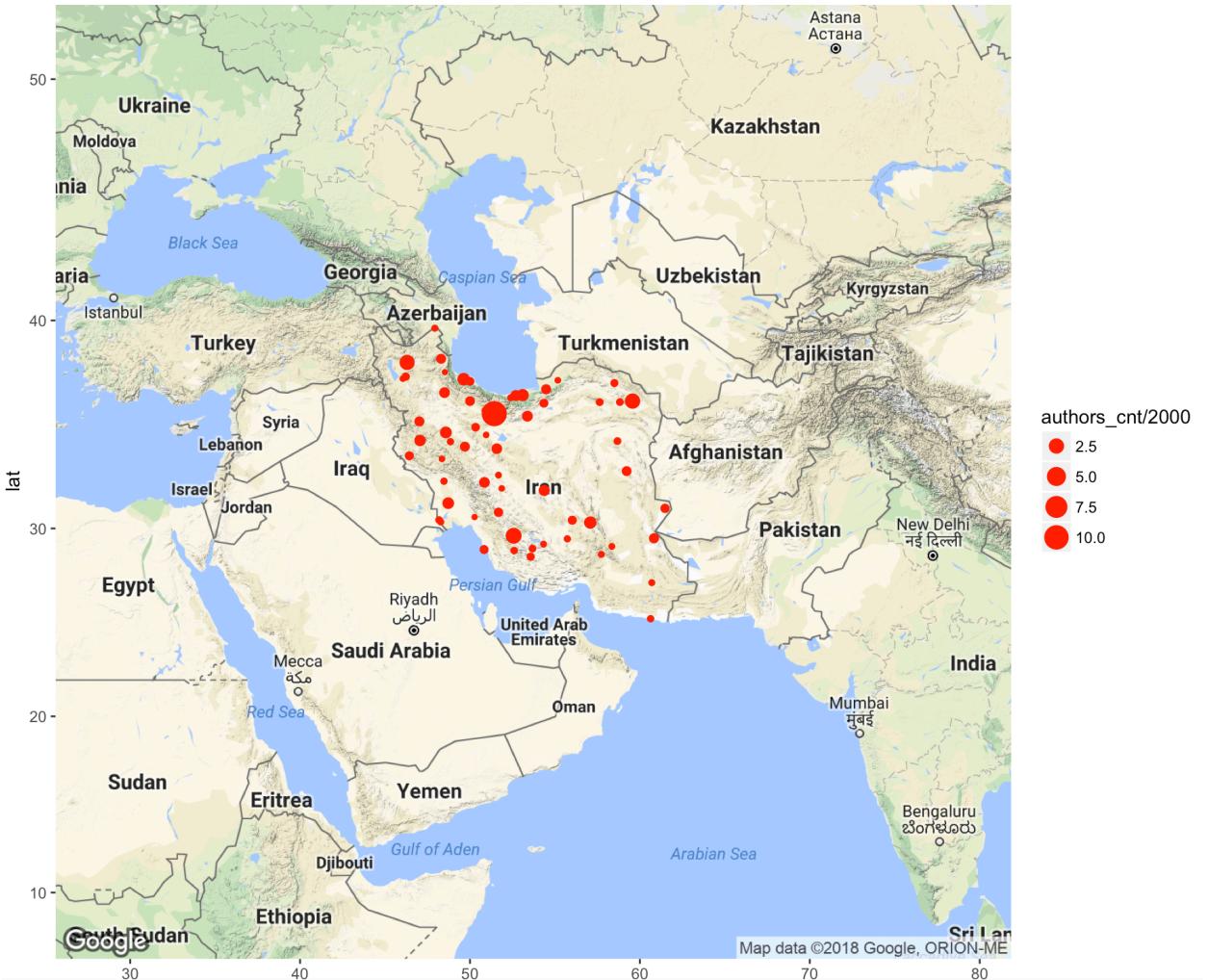
تحلیل مکانی

ابتدا از آدرس کامل موسسات شهر آنها استخراج شد و با استفاده از آن به دو کنگکاوی زیر پاسخ می‌دهیم:

۳- تحلیل تعداد مقالات بر حسب مکان. نمودار ۳. (کد در P3.R)



۴- تحلیل تعداد نویسنده ها بر حسب مکان. نمودار ۴. (کد در P3.R)

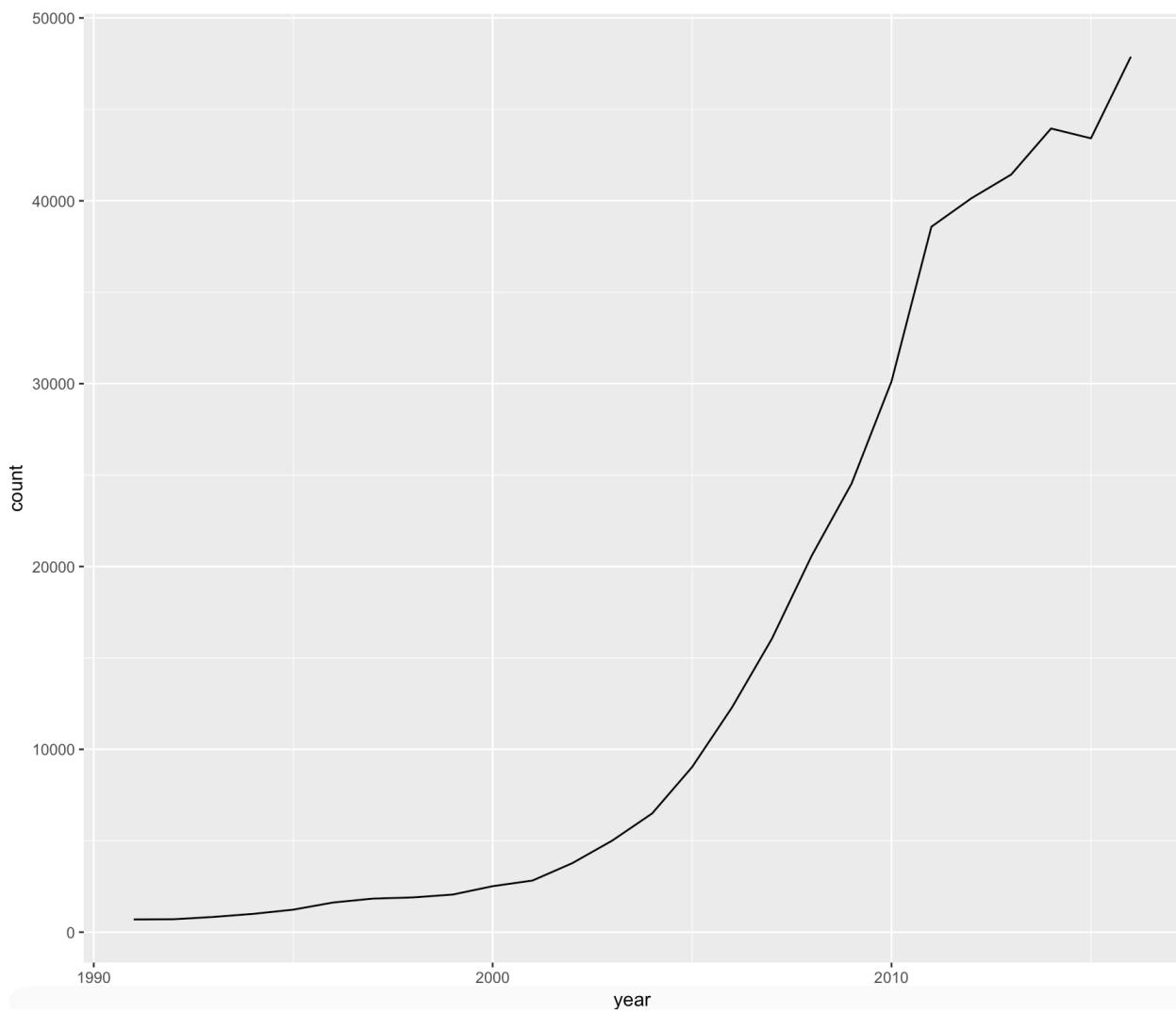


طبق مشاهده دو نمودار بالا به نظر میرسد که تعداد مقالات و نویسنده ها در مکان های مختلف متناسب است. با تست کوربیشن میتوان میزان این تناسب را اندازه گیری کرد که 0.9708857 به دست می آید که نزدیک به ۱ است و p -value نیز فرض \cdot بودن کوربیشن را رد میکند. بنابراین این دو کمیت باهم افزایش یا کاهش میابند که البته منطقی است.

تحلیل زمانی

۵- تحلیل تعداد مقالات بر حسب زمان. نمودار ۵. (کد در R) (P5.R)

تعداد مقالات از سال ۱۹۹۰ به بعد افزایشی با روند نمایی داشته که همینطور به رشد خود ادامه می دهد. البته این داده سال ۲۰۱۷ را به دلیل داشتن مقالات تا اواسط آن سال و کامل نبودن دارا نیست. همچنین در سال ۲۰۱۵ نسبت به سال قبل مقالات کاهش یافته که با روند همیشه صعودی آن متفاوت است.



تحلیل نویسندها

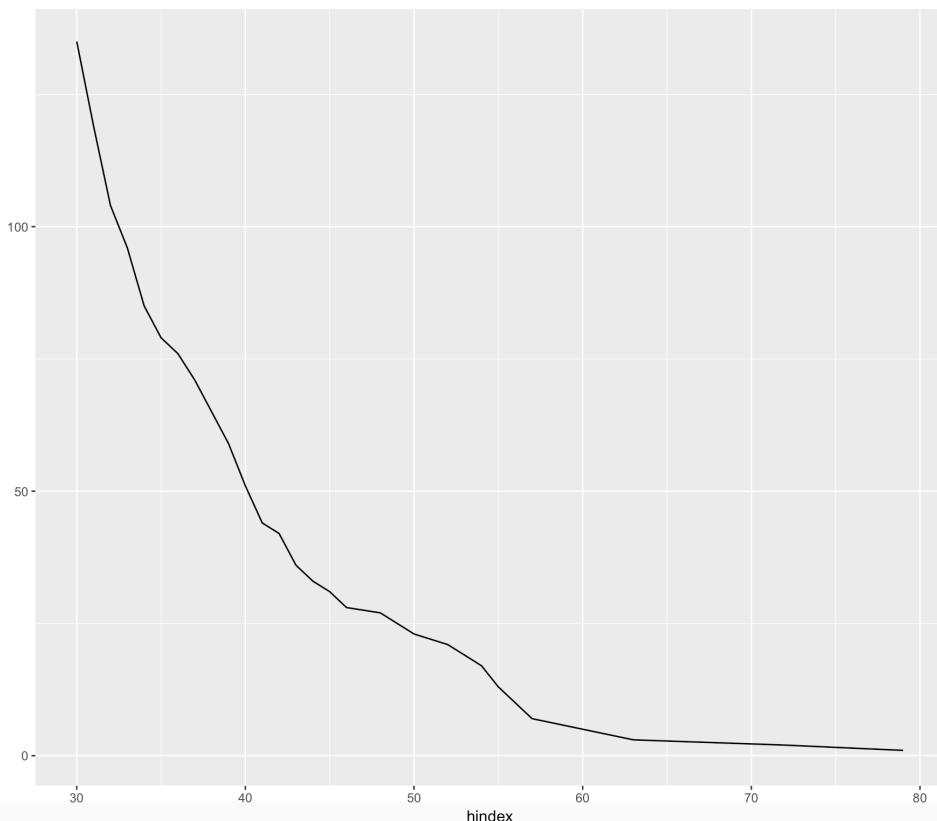
هر نویسنده به طور میانگین 5.589684 مقاله نوشته است و بیشترین تعداد مقاله برای یک فرد 871 مقاله است.

- ۶- برترین نویسنده از نظر h-index همراه با رشته های تحقیقاتی آنها. (کد در P6.R میانگین h-index نویسنده ایرانی ۱.673771 است.
- همچنین ۱۰ استاد برتر از این نظر در زیر آمده اند.
- با توجه به رشته های تحقیقاتی باز هم استادان برتر بیشتر در زمینه علوم تجربی و زیستی هستند.

Full name	H-index	Subject Areas
Ganjali, Mohammad Reza	79	Chemistry~Biochemistry~Genetics and Molecular Biology~Physics ~Materials ~Environment ~Pharmacology~ Toxicology ~Medicine ~Immunology and Microbiology~Energy
Khakzad, Mohsen	72	Physics~Biochemistry~ Genetics and Molecular Biology~Medicine~Multidisciplinary~Chemistry~Social Sciences~Psychology
Shamsipur, Mojtaba	63	Chemistry~Biochemistry~Genetics and Molecular Biology~Materials ~Physics ~Environment ~Pharmacology~ Toxicology ~Energy~Agricultural and Biological Sciences~Medicine~Immunology and Microbiology~Social Sciences~Multidisciplinary
Domiri Ganji, Davood	60	Engineering~Physics and Astronomy~Mathematics~Chemical Engineering~Materials Science~Chemistry~Energy~Computer Science~Multidisciplinary~Environmental Science~Biochemistry~ Genetics and Molecular Biology~Earth and Planetary Sciences~Medicine~Decision Sciences~Economics~ Econometrics and Finance

Salavati-Niassari, Masoud	57	Materials Science~Chemistry~Physics and Astronomy~Engineering~Chemical Engineering~Biochemistry~Genetics and Molecular Biology~Energy~Environmental Science~Medicine~Computer Science~ Earth and Planetary Sciences~Pharmacology~Toxicology and Pharmaceutics~Multidisciplinary~Business~ Management and Accounting~Agricultural and Biological Sciences~Mathematics
Abdollahi, Mohammad	57	Pharmacology~ Toxicology and Pharmaceutics~Medicine~Biochemistry~ Genetics and Molecular Biology~Environmental Science~Agricultural and Biological Sciences~Immunology and Microbiology~Veterinary~Chemistry~Neuroscience~Materials Science~Nursing~Social Sciences~Dentistry~Earth and Planetary Sciences~Chemical Engineering~Psychology~Health Professions~Arts and Humanities~Engineering~Computer Science~Multidisciplinary
Zeinali, Mohammad Hossein	56	Physics and Astronomy~Engineering~Mathematics~Computer Science~Medicine~Multidisciplinary~Social Sciences~Psychology
Fahim, A.	56	Physics and Astronomy~Engineering~Mathematics~Medicine~Multidisciplinary~Social Sciences~Psychology
Safarzadeh, Bita	56	Physics and Astronomy~Engineering~Mathematics~Medicine~Multidisciplinary~Social Sciences~Psychology~Agricultural and Biological Sciences~Computer Science~Biochemistry~ Genetics and Molecular Biology

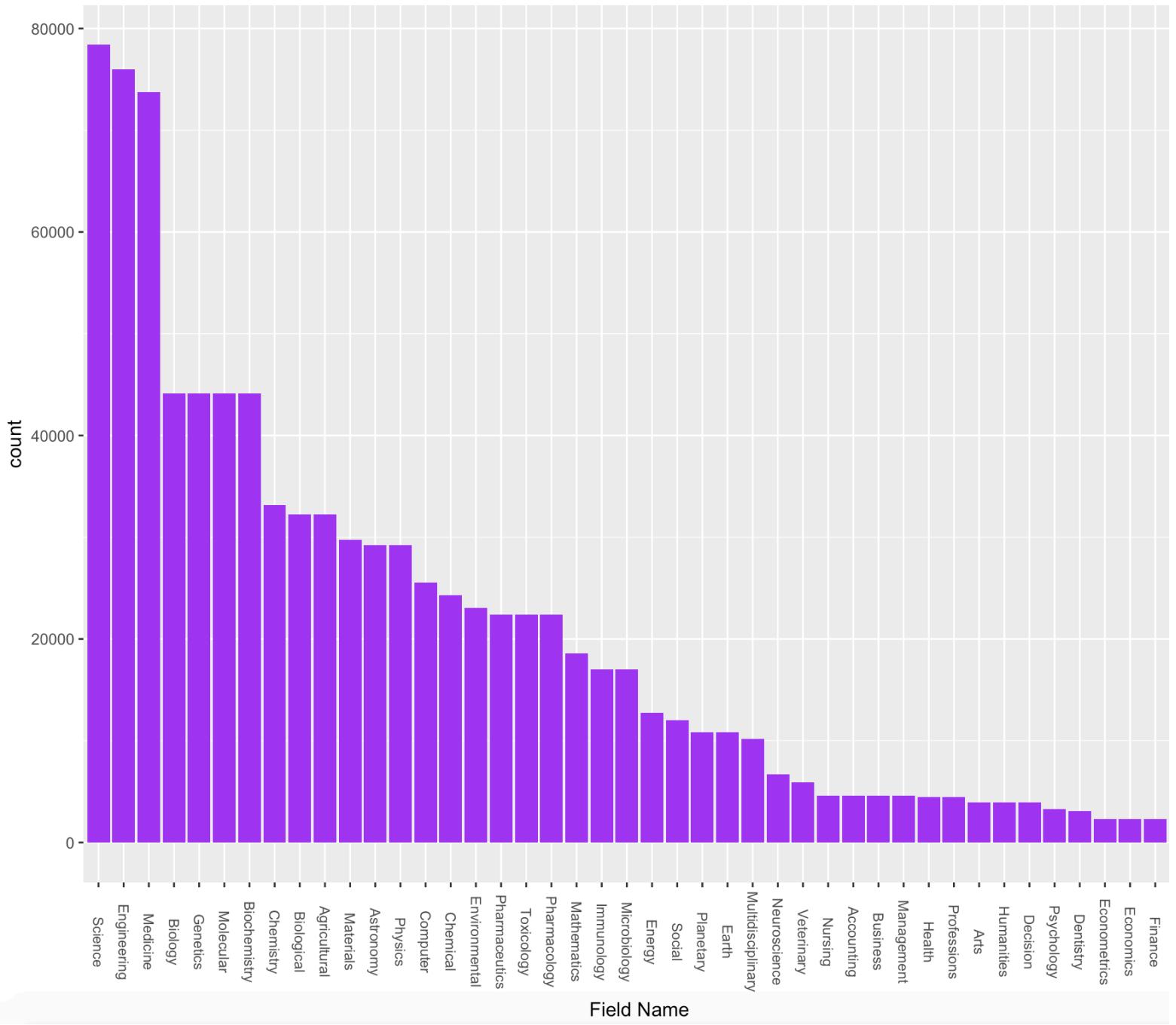
۷- نمودار تجمعی h-index بر حسب تعداد برای h-های بالای ۳۰ (کد در P7.R)



همانطور که در نمودار مشاهده میشود تعداد نویسنگان با h-index بالای ۵۰ حدود ۲۳ نفر است. که عددی بسیار کم است. همچنین h-index بالای ۳۰ نیز حدود ۱۲۰ نفر هستند.

۸ - بررسی تعداد افراد در رشته‌های تحقیقاتی مختلف (کد در R P8.R)

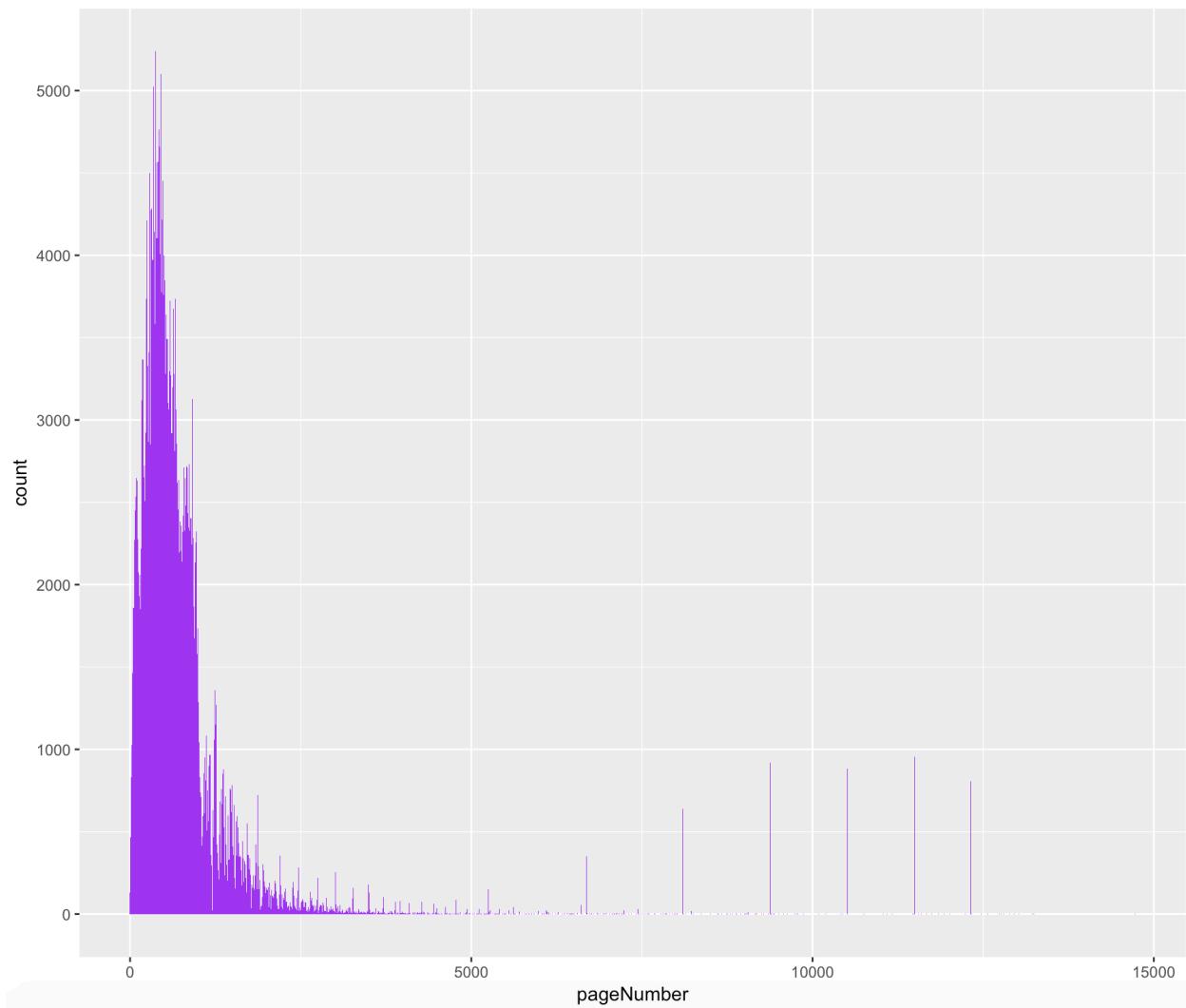
نمودار رشته‌های تحقیقاتی بر اساس تعداد افراد رسم شده است. علوم محض از مهندسی بیشتر رایج هستند و همچنین اکثر رشته‌های پژوهشی در شاخه‌های علوم تجربی قرار میگیرند.



تحلیل ویژگی های مقالات

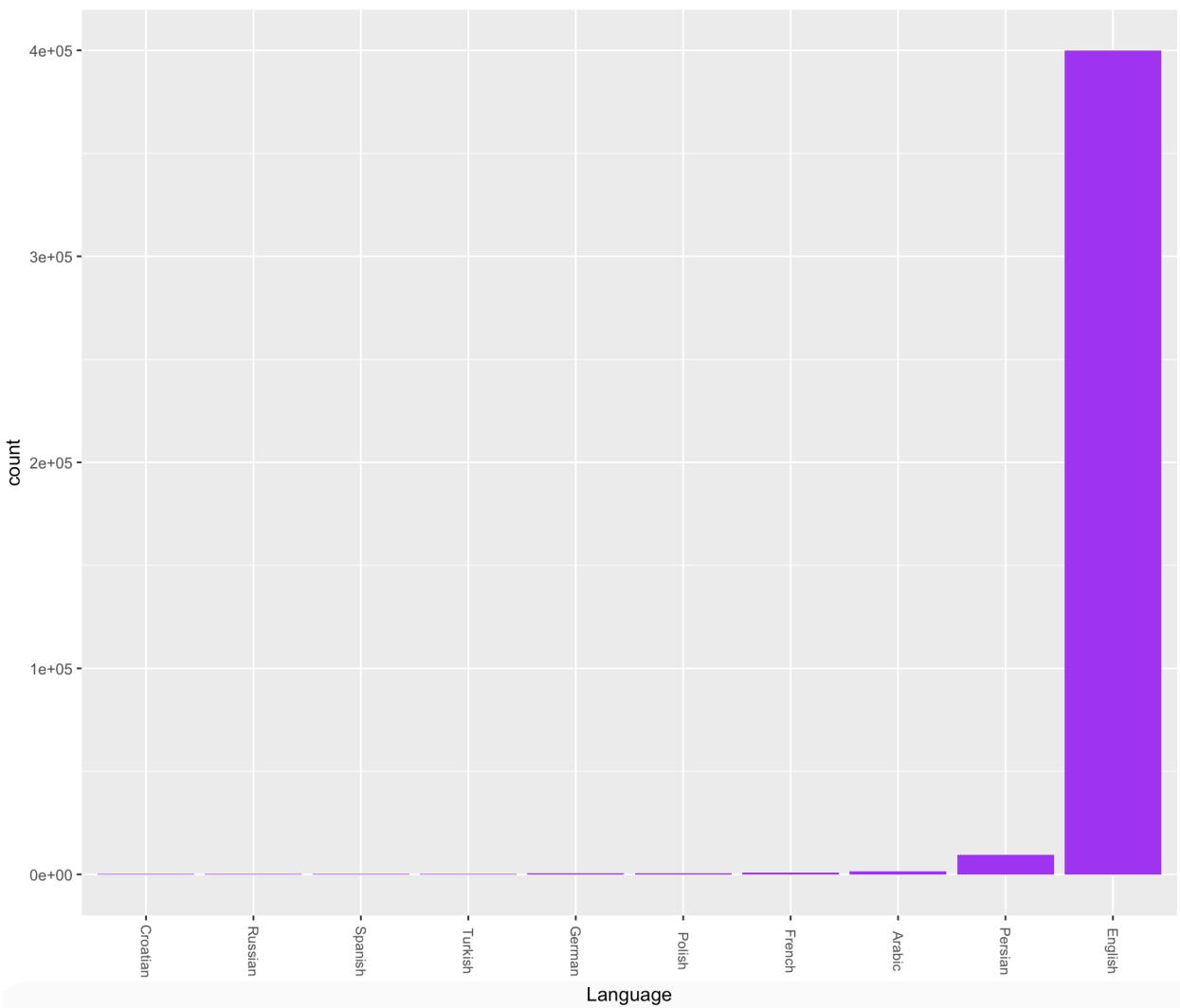
۹- بررسی تعداد صفحات مقالات (کد در P9.R)

میانگین تعداد صفحات ۸۰.۸ صفحه است. بنابراین در این داده با کتاب های نوشته شده توسط اساتید نیز سر و کار داریم. نمودار توزیع تعداد صفحات نیز در زیر رسم شده است.



۱۰- بررسی مقالات از نظر زبانی (کد در R (P10.R

زبان مورد استفاده اول انگلیسی و بعد با تفاوت فاحش فارسی قرار میگیرد. احتمالاً این پدیده به دلیل فرستادن مقالات برای منابع معتبر خارجی است که لازمه آن انگلیسی بودن زبان مقاله است.



بررسی شبکه نویسندها

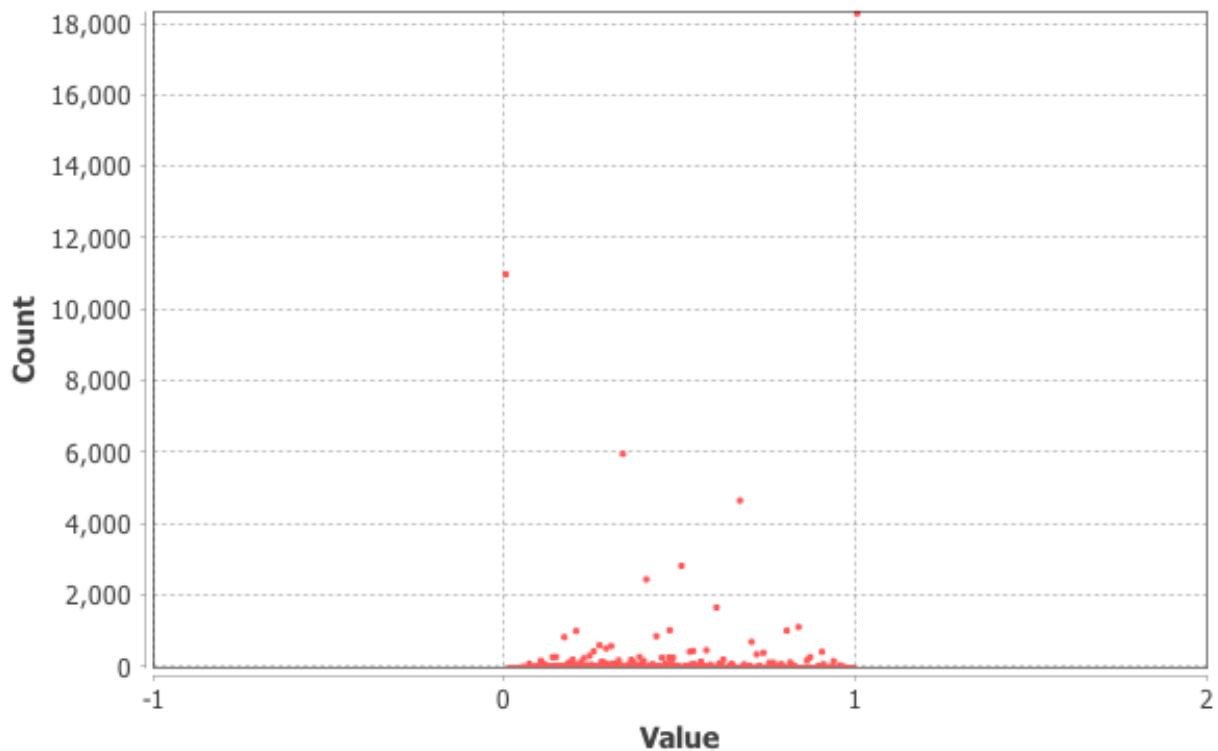
با توجه به نویسندهای مقالات مختلف، گرافی تشکیل شد که رئوس آن نویسندهای هستند و دو فرد به هم یال دارند اگر باهم مقاله مشترک داشته باشند. از این گراف رئوس درجه ۰ و ۱ برای راحتی و سریع تر شدن اجرای الگوریتم ها حذف شده اند. در این گراف که نوعی شبکه اجتماعی است می توان شاخص های متنوعی را بررسی کرد.

برای تحلیل شبکه اجتماعی علاوه بر زبان R از زبان C++ و نرم افزار Gephi نیز استفاده شده است. کدهای تشکیل گراف نویسندهای در R P14.R و authorsGraphGenerator.cpp موجود است. کمیت های زیر با نرم افزار Gephi اندازه گیری شده اند.

Clustering coefficient - ۱۱

معیاری برای میزان کامل بودن گراف است. در واقع به ازای هر راس نسبت همسایه های او که خود باهم همسایه هستند را نسبت به کل میسنجد و بین تمام راس ها میانگین گرفته میشود.

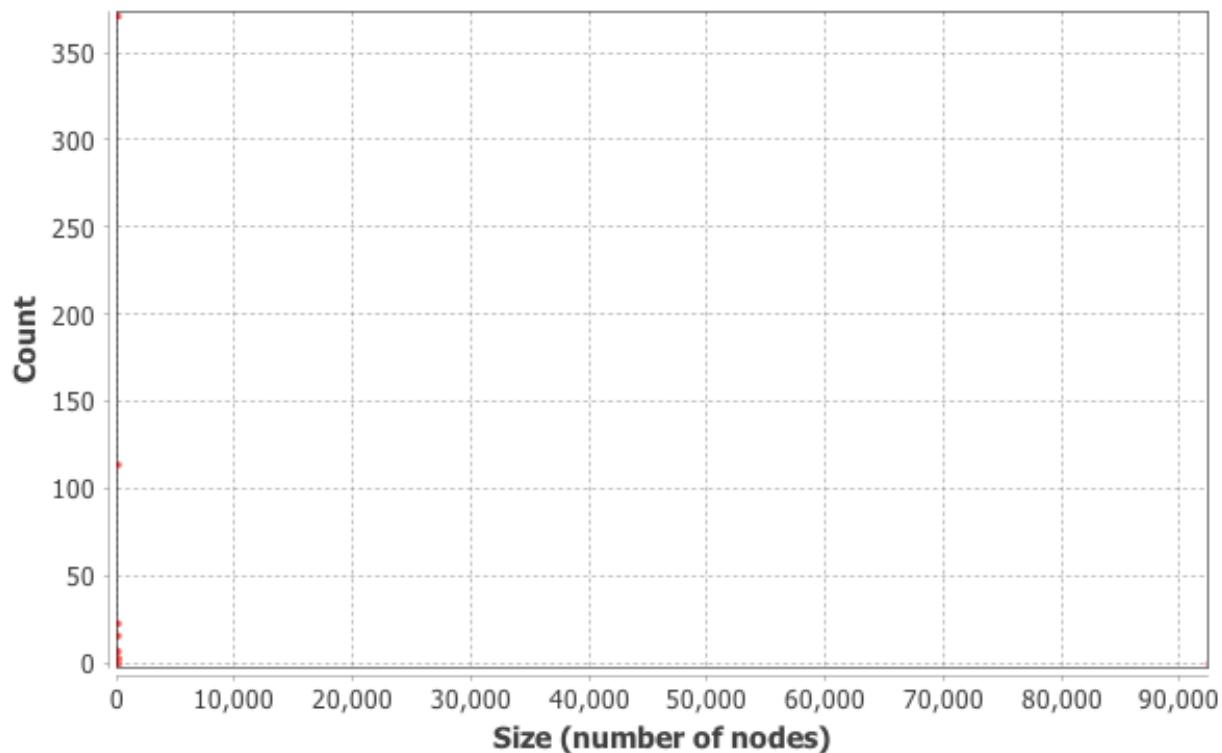
Clustering Coefficient Distribution



این مقدار در گراف نویسندهان 0.516 است و تعداد کل مثلث های گراف نیز 1057144 است.
نمودار توزیع Clustering coefficient نیز آورده شده است.

۱۲- تعداد مولفه های گراف
547

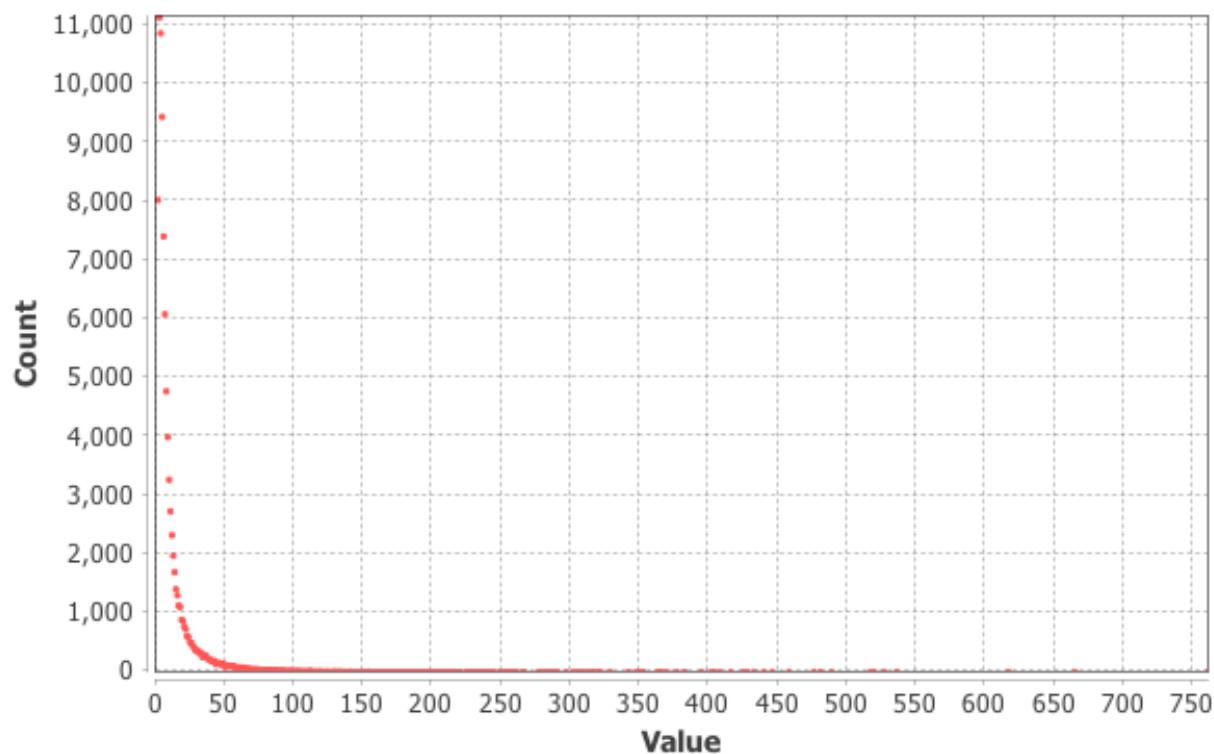
Size Distribution



نمودار توزیع اندازه مولفه ها

۱۳ - بررسی درجه ها
میانگین درجات : 11.426
نمودار توزیع درجات

Degree Distribution



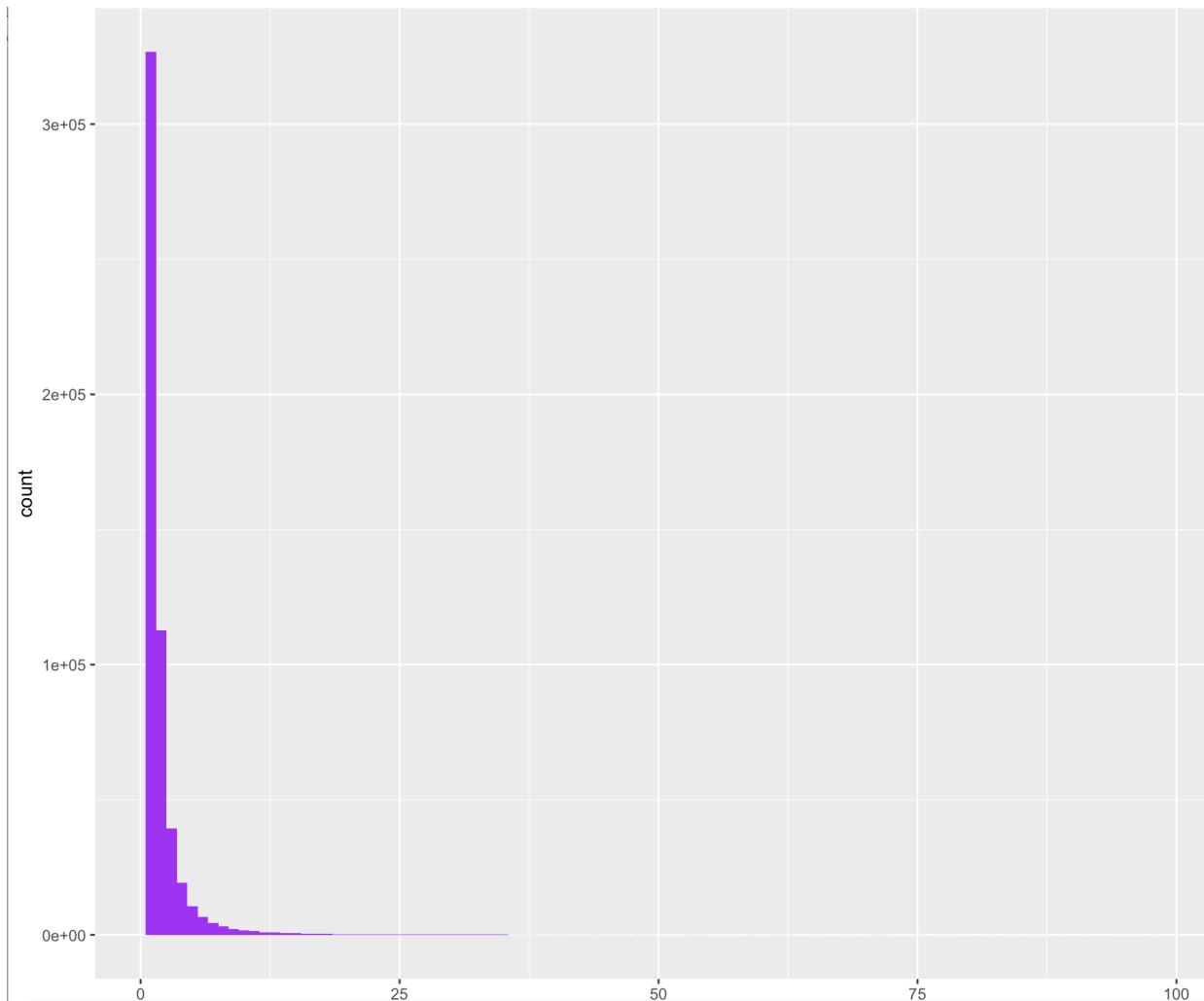
۱۴ - قطر گراف : ۱۷
۱۵ - شعاع گراف : ۱

۱۶ - تعداد مقالات بین نویسندهای (کد در R) (P16.R)

میانگین تعداد مقالات مشترک بین دو نویسنده : 2.128451

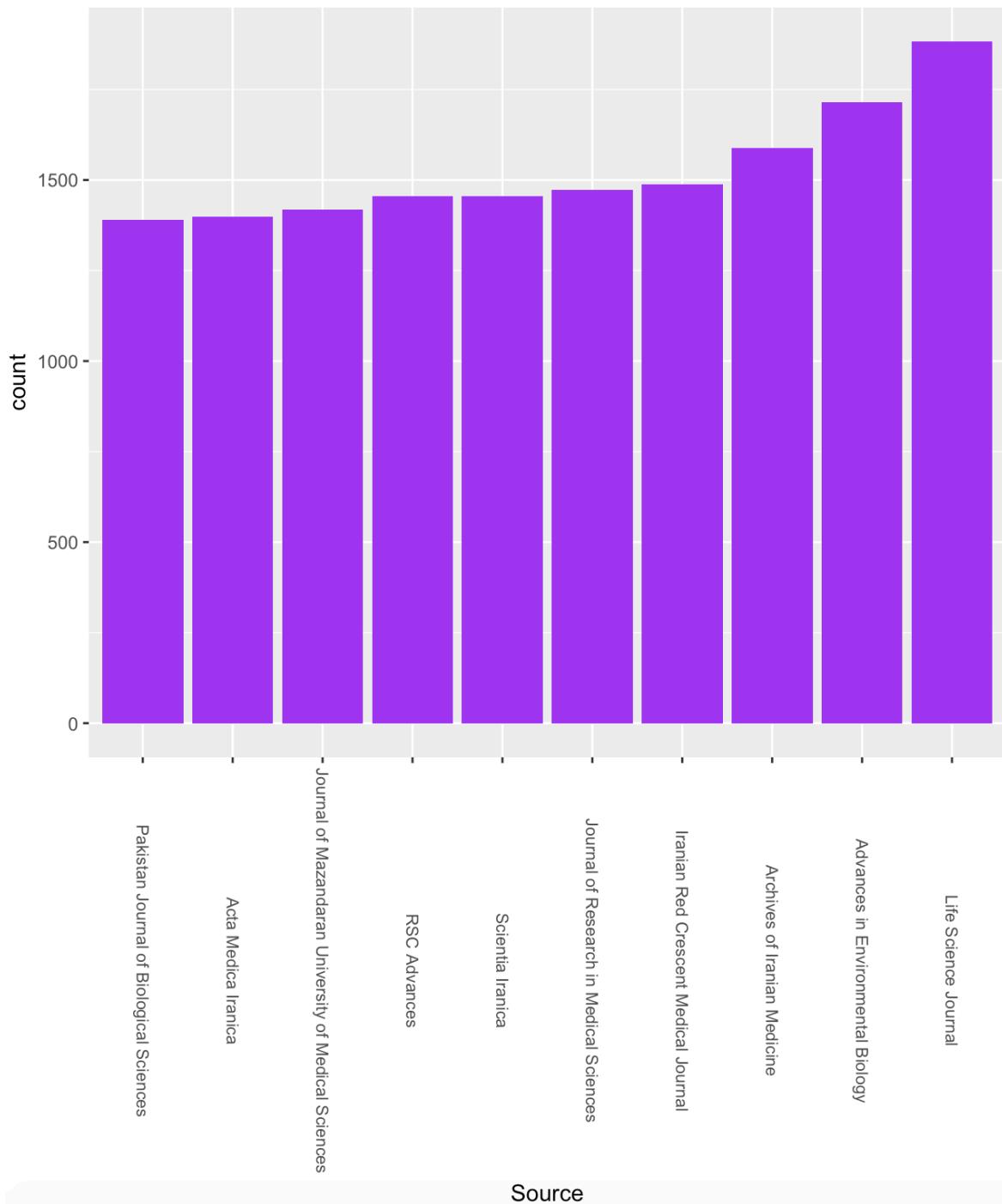
ماکریزم تعداد مقالات مشترک بین دو نویسنده: 604

نمودار توزیع :



بررسی منابع و ناشران مقالات

نمودار ۱۰ ناشر برتر به همراه تعداد مقالات آن رسم شده است. دوباره مشاهده میشود که اکثر آنها در رابطه با علوم های تجربی و زیستی هستند. (کد در R (P17.R)



بررسی پیشرفتی شبکه نویسندهان

با توجه به گرافی که از نویسندهان ساخته شد و در قسمت‌های قبل توضیح داده شد، قصد داریم این شبکه اجتماعی را به صورت دقیق‌تر بررسی کنیم.
از آنجایی که حجم داده بسیار زیاد بود نرم‌افزارهای تحلیل گرافی مثل Gephi پاسخگوی این سوالات نبودند و چون زبان R هم تا حد زیادی سرعت پایینی دارد، با استفاده از C++ بررسی‌های این قسمت را انجام داریم.

Average Path Length محاسبه

این کمیت به نوعی دوری راس‌ها از هم و در واقع نزدیکی نویسندها را بهم نشان میدهد. این کمیت به همراه قطر و شعاع برای بررسی شبکه‌های اجتماعی به کار می‌رود.
در واقع فاصله هر دو راس از هم محاسبه می‌شود و میانگین همه فواصل این کمیت را نشان میدهد.
که محاسبه گر این کمیت کد average-path-length.cpp است.
مقدار خروجی این کد برابر با ۱۰.۱۷۷۹ است. که نشان میدهد راس‌های گراف بهم تقریباً نزدیک هستند و شبکه به small world ها تا حد خوبی شبیه است.
اگر این مقدار با قطر که حدود ۱۷ است مقایسه شود متوجه می‌شویم که اکثر راس‌ها در فواصل نزدیک تری بهم هستند و نویسندها و اساتید (احتمالاً در یک دانشگاه باهم ارتباط‌های نزدیکی دارند) باهم ارتباط نزدیکی دارند.

محاسبه هموفیلی در شبکه نویسندهان کشور

هموفیلی در شبکه‌های اجتماعی، میزان تمایل افراد برای ارتباط با افراد شبیه به خود را نشان میدهد. این شباهت میتواند از جهات مختلف بررسی شود.
در اینجا به ازای هر نویسنده دانشگاه یا موسسه مربوط به او استخراج شده است و شباهت بین دو فرد یکسان بودن موسسه‌آنها تعریف می‌شود.
با این تعریف هموفیلی اندازه گیری شده در کل شبکه برابر با ۹۹٪ است که بسیار زیاد است.
این نشان میدهد که ارتباط بین دانشگاهی کمی در ایران صورت می‌گیرد و اکثر مقاله‌ها در یک دانشگاه انجام می‌شوند.
 بدیهی است که تعامل بین دانشگاه‌ها و اساتید آنها باعث پویایی بیشتر و پیشرفت علمی می‌شود و خوب است که از این نظر این موضوع مورد توجه قرار گیرد.
که این قسمت در P18.R و homophily.cpp است.

محاسبه multiplexity در شبکه نویسندهان کشور

multiplexity به معنای تعدد رابطه و شباهت بین دو راس در شبکه است. در واقع میزان قدرت یک رابطه را مشخص می‌کند.
در اینجا میتوانیم این تعریف را به ازای هر دو نفری که باهم یک مقاله داده‌اند تعداد مقاله‌های مشترک تعریف کنیم.
البته میتوانستیم روابط مختلف دیگری مثل دانشگاه مشترک، شهر مشترک و رشتہ مشترک را برای محاسبه این کمیت استفاده کنیم اما با محاسبه هموفیلی در قسمت قبل که ۹۹٪ شد نتیجه می‌گیریم دانشگاه و شهر نویسندهان مقاله یکسان است و این کمیت آنها را تمایز نمی‌کند. در مورد فیلدهای مورد علاقه، از آنجایی که هر فرد تعداد زیادی فیلد مورد علاقه دارد و این متغیر خیلی کلی در داده‌ها آورده شده خیلی کمیت مناسبی برای این موضوع نیست.

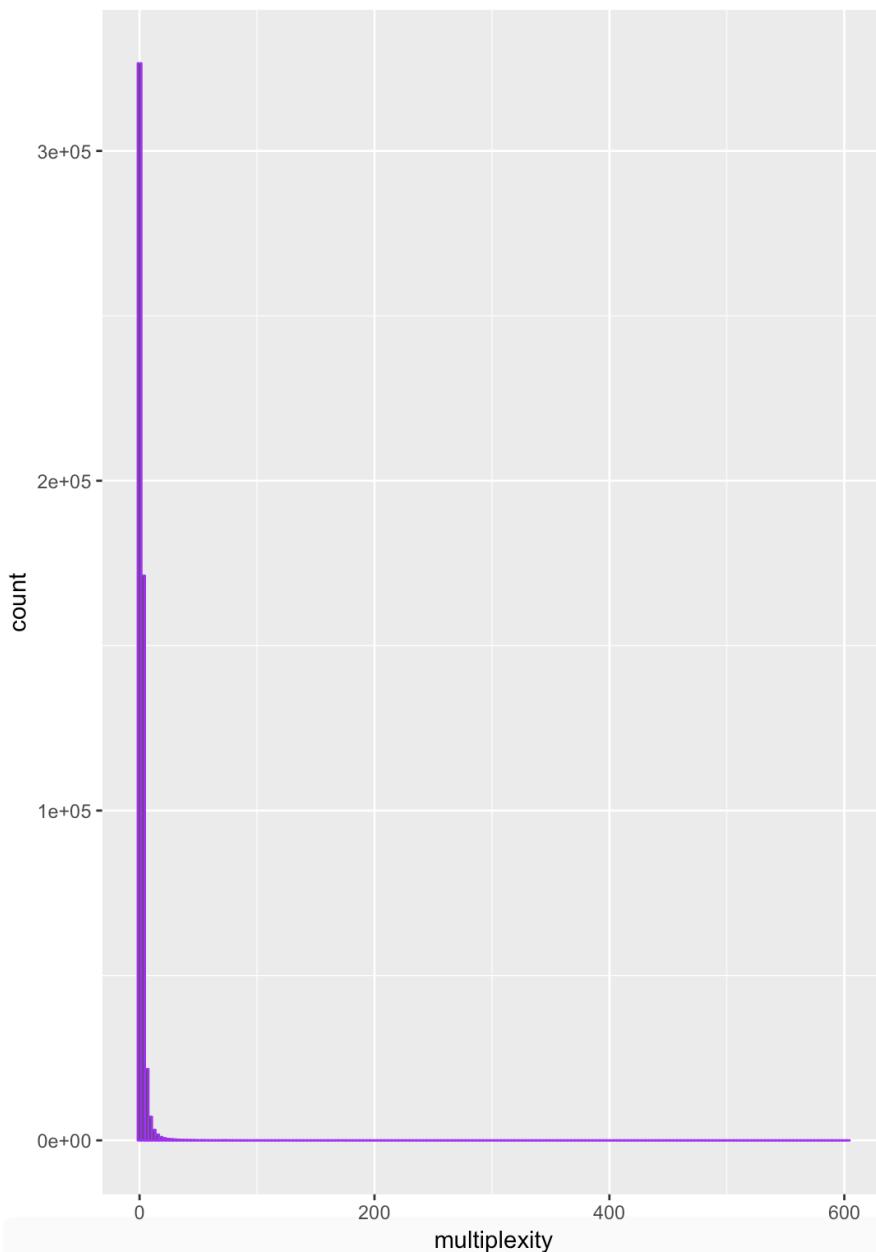
حال با استفاده از تعداد مقالات مشترک این مقدار را محاسبه کردیم.

میانگین: ۲.۱۲۸۸۰۵

ماکزیمم: ۶۰۴

مینیمم: ۱

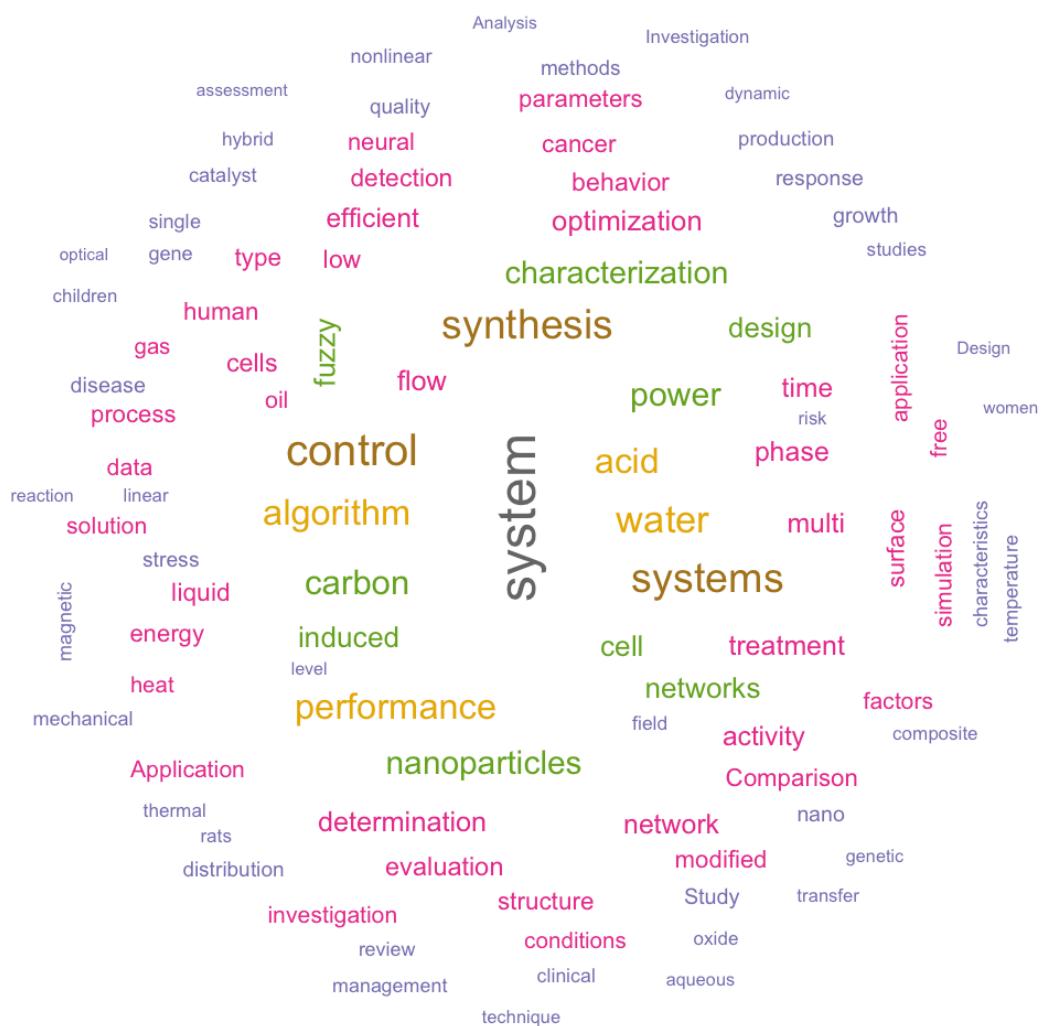
این به این معناست که اکثر نویسندهای یک مقاله برای مقاله بعدی نیز به هم رجوع کردند و هر دو نفر به طور میانگین بیشتر از ۲ مقاله باهم دارند.
نمودار توزیع این کمیت نیز در ادامه آمده است.
کد در P19.R



بررسی بعضی از موارد بالا در دانشگاه شریف

تحليل لغوی

P20.R لغات پر استفاده در عنوان مقالات. نمودار ۱ (کد در ۱-)



طبق این نمودار پر استفاده ترین لغت سیستم است که البته میتواند در همه رشته ها باشد اما شاید بین رشته برق و کامپیوتر تا حدی این کلمه پر استفاده تر باشد.

تحلیل نویسندها دانشگاه شریف

میانگین h-index نویسندها در دانشگاه شریف ۲.۱۹ است که نسبت به میانگین کل بیشتر است یعنی مقاله های با سایتیشن بیشتری نسبت به نویسندها کشور دارند.

برترین نویسندها دانشگاه شریف و رشته های آنها

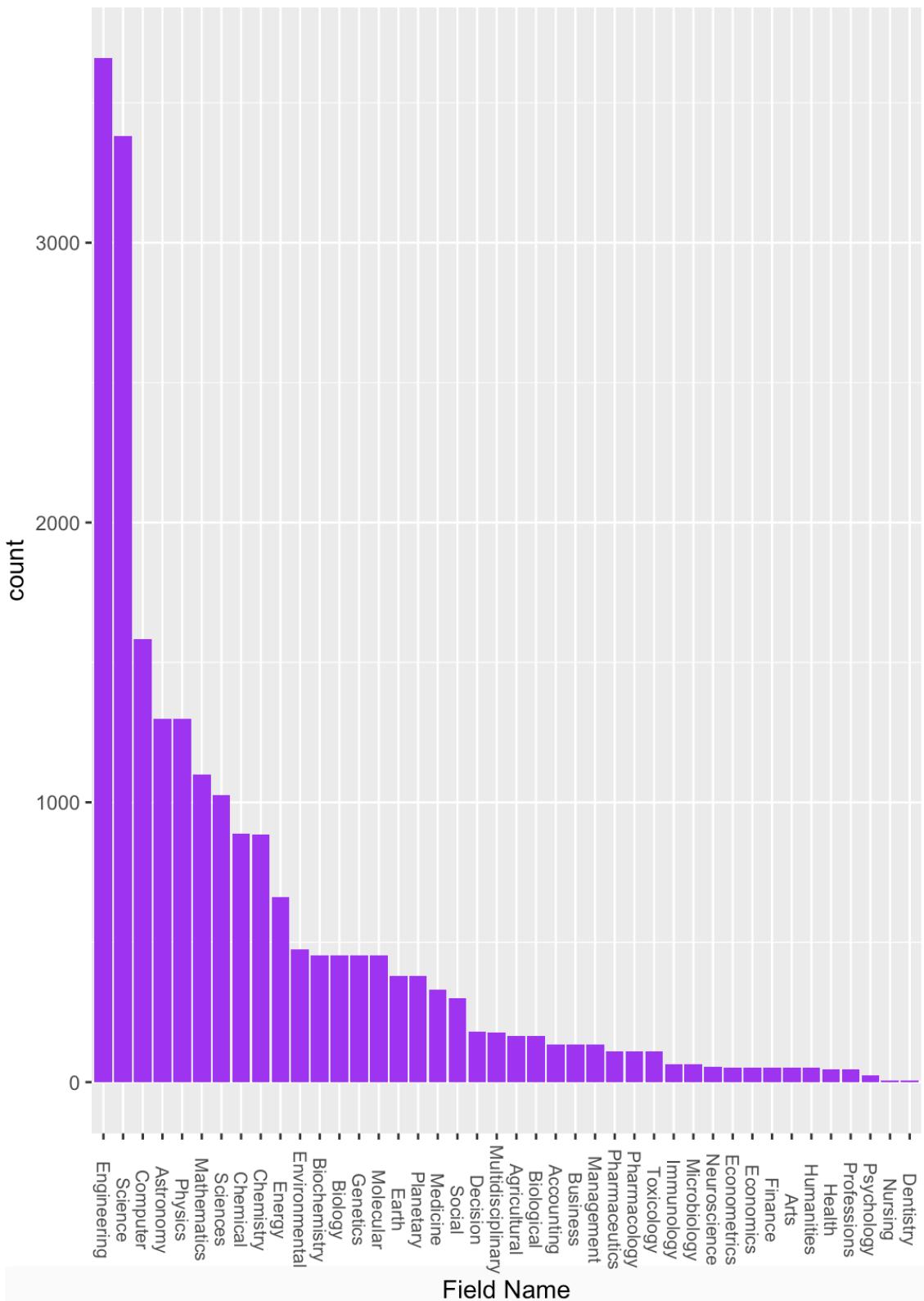
1	Mohammadi, Mohammad Reza	38	Physics and Astronomy~Materials Science~Engineering~Chemistry~Chemical Engineering~Computer Science~Environmental Science~Energy~Medicine~Biochemistry ~ Genetics and Molecular Biology~Mathematics~Arts and Humanities~Earth and Planetary Sciences
2	Parnianpour, Mohamad	37	Medicine~Engineering~Biochemistry~ Genetics and Molecular Biology~Computer Science~Chemical Engineering~Health Professions~Earth and Planetary Sciences~Social Sciences~Psychology~Mathematics~Neuroscience~Materials Science~Physics and Astronomy~Environmental Science~Nursing~Energy~Agricultural and Biological Sciences~Pharmacology~Toxicology and Pharmaceutics~Multidisciplinary
3	Fotuhi Firuzabad, Mahmud	31	Engineering~Energy~Computer Science~Mathematics~Environmental Science~Decision Sciences~Multidisciplinary~Materials Science~Physics and Astronomy~Chemical Engineering~Medicine~Economics~Econometrics and Finance~Biochemistry~Genetics and Molecular Biology

4	Pourjavadi, Ali	31	Materials Science~Chemistry~Chemical Engineering~Physics and Astronomy~Engineering~Biochemistry~Genetics and Molecular Biology~Agricultural and Biological Sciences~Environmental Science~Pharmacology~ Toxicology and Pharmaceutics~Mathematics~Medicine~Energy~Health Professions~Computer Science
5	Sadrnezhaad, Sayed Khatiboleslam	29	Materials Science~Engineering~Physics and Astronomy~Chemical Engineering~Chemistry~Environmental Science~Earth and Planetary Sciences~Biochemistry~ Genetics and Molecular Biology~Energy~Mathematics~Medicine~Computer Science~Pharmacology~ Toxicology and Pharmaceutics
6	Saidi, Mohammad Reza	29	Chemistry~Biochemistry~ Genetics and Molecular Biology~Pharmacology~ Toxicology and Pharmaceutics~Chemical Engineering~Materials Science~Environmental Science~Engineering~Medicine~Physics and Astronomy
7	Bagheri, Habib	28	Chemistry~Biochemistry~ Genetics and Molecular Biology~Environmental Science~Chemical Engineering~Engineering~Pharmacology ~ Toxicology and Pharmaceutics~Medicine~Physics and Astronomy~Agricultural and Biological Sciences
8	Khoei, Amir Reza	26	Engineering~Mathematics~Materials Science~Computer Science~Physics and Astronomy~Chemistry~Earth and Planetary Sciences~Chemical Engineering~Environmental Science

9	Taghavinia, Nima	26	Materials Science~Physics and Astronomy~Chemistry~Engineering~Chemical Engineering~Energy~Biochemistry~Genetics and Molecular Biology~Environmental Science~Medicine~Mathematics
10	Ghaderi, Elham	25	Materials Science~Chemistry~Physics and Astronomy~Engineering~Medicine~Biochemistry~ Genetics and Molecular Biology~Energy~Chemical Engineering
11	Haeri, Mohammad	25	Engineering~Computer Science~Mathematics~Physics and Astronomy~Chemical Engineering~Materials Science~Decision Sciences~Agricultural and Biological Sciences~Medicine~Chemistry~Economics~Econometrics and Finance~Multidisciplinary

بررسی رشته‌های نویسنده‌گان دانشگاه شریف

همان طور که در نمودار دیده می‌شود، بعد از دو فیلد عمدی که تقریباً همه را شامل می‌شود (مهندسی و علم محض) پر طرفدار ترین از نظر تعداد نویسنده‌گان رشته کامپیوتر و بعد نجوم است (که احتمالاً یا در فیزیک است یا هوافضا)



هدف این پژوهش، بررسی مقالات ایران و دریافت شهود نسبی نسبت به اطلاعات و آمارهای آن بود.
سعی شده بود که با آماره های نه چندان پیچیده و نمودارهای ساده این اطلاعات رسانده شود.

این داده، داده حجمی بود و بنابراین تحلیل آن با امکانات اولیه (یک لپتاپ) و با زبان آر خیلی راحت نبود.
همچنین علاوه بر حجم آن، داده خیلی تمیز و شاید خیلی هم با جزئیات و کامل نبود و بسیاری از تحلیلهایی که در
نظر داشتم روی آن قابل اجرا نبود.

سعی شده بود داده را از نظر شبکه های اجتماعی نیز بررسی کنم که این کار نیز عمدتاً با زبان سی پلاس پلاس و
با کمک آر و نرم افزار گفی انجام شد. انتخاب سی پلاس پلاس به دلیل سرعت بالای آن و راحتی کار در کار با
گراف بود. آر و گفی در گرافهای بزرگ بدون پردازش توزیع شده چندان پاسخگو نبودند و با زبان سی پلاس پلاس
هم هر آماره بعد از مدت زیادی جوابش مشخص میشد.
امیدوارم که تحلیلها و حقایقی که در پژوهش جمع‌آوری شد مفید و جالب باشد و امید است که آینده علمی کشور و به
خصوص دانشگاه شریف بسیار درخشنان باشد و پیشرفت زیادی داشته باشد.

با تشکر

پرند علیزاده
۹۴۱۰۰۲۴

https://github.com/praal/data_analysis_course