

ساختارهای زبانی و شبکه‌های هم‌رخدادی

علی اکرامیان، فاطمه رمضان‌زاده و علیرضا آستانه

پروژه‌ی درس علم شبکه، نیمسال اول ۱۴۰۲-۰۳

زبان، جنبه‌ای بنیادی در ارتباط و برهم‌کنش بین انسان‌هاست که نقش مهمی در انتقال افکار، ایده‌ها و احساسات دارد. پژوهشگران، در طول سال‌ها در رشته‌های مختلف، سعی کرده‌اند از منظرهای متفاوتی، برهم‌کنش‌های پیچیده‌ی زبانی را تحلیل و مطالعه کنند. یکی از پنجره‌هایی که برای نگاه کردن و مطالعه‌ی زبان‌ها می‌توان پیدا کرد، بررسی زبان‌ها به عنوان شبکه است. یکی از انواع این شبکه‌ها که می‌توان به‌وسیله‌ی آن زبان را مورد بررسی قرار داد، شبکه‌های هم‌رخدادی است به طوری که در یک متن، کلمات هم‌جوار را همسایه گرفته و بین آن‌ها می‌توانیم ارتباطی فرض کنیم. در مدل سازی این شبکه، در ساده‌ترین تقریب، کلمات مانند راس‌های گراف هستند و بین هر دو کلمه‌ی همسایه در یک جمله، یک یال رسم می‌شود. این گونه می‌توانیم این شبکه‌ی هم‌رخدادی را رسم کرده و خصوصیات آن را برای متن‌ها و شعرهای مختلف بدست آوریم.

از کاربردهای بررسی این گونه شبکه‌ی زبانی می‌توان به تحلیل‌های زبانی و تشخیص مدل و سبک متن اشاره کرد. همچنین ویژگی‌های متن‌های نوشته شده توسط افراد مختلف و نویسنده‌گان بزرگ مختلف، با هم فرق دارد و می‌توان با ویژگی‌های بدست آمده از شبکه‌هایی برای متن‌های نوشته شده این را تولید کرده است. حتی می‌توان تشخیص داد که آیا این متن توسط انسان نوشته شده است یا یک مدل زبانی بزرگ^۱ آن را تولید کرده است. حتی می‌توان با بدست آوردن ویژگی‌های یک شبکه‌ی زبانی، کلمات کلیدی را استخراج و متن را خلاصه نویسی کرد.

در این پروژه ما سعی می‌کنیم گام اول این هدف بزرگ، که بدست آوردن ویژگی‌های شبکه‌های زبانی است را انجام دهیم و برای چندین متن و شعر این ویژگی‌ها را بدست آوریم و با هم مقایسه کرده و تحلیل نماییم.

برای تولید شبکه‌ی هم‌رخدادی، باید چندین کار را انجام دهیم. در ابتدا باید متن را جمله‌بندی کرده و سپس جمله‌ها را نیز به کلمات جدا از هم تبدیل کرد. مثلاً متن پاراگراف اول همین صفحه را اگر جمله‌بندی کنیم چنین آرایه‌ای از جملات خواهیم داشت:

[زبان، جنبه‌ای بنیادی در ارتباط و برهم‌کنش بین انسان‌هاست که نقش مهمی در انتقال افکار، ایده‌ها و احساسات دارد.]، [پژوهشگران، در طول سال‌ها در رشته‌های مختلف، سعی کرده‌اند از منظرهای متفاوتی، برهم‌کنش‌های پیچیده‌ی زبانی را تحلیل و مطالعه کنند.]، [یکی از پنجره‌هایی که برای نگاه کردن و مطالعه‌ی زبان‌ها می‌توان پیدا کرد، بررسی زبان‌ها به عنوان شبکه است.]، [یکی از انواع این شبکه‌ها که می‌توان به‌وسیله‌ی آن زبان را مورد بررسی قرار داد، شبکه‌های هم‌رخدادی است به طوری که در یک متن، کلمات هم‌جوار را همسایه گرفته و بین آن‌ها می‌توانیم ارتباطی فرض کنیم.]، [در مدل سازی این شبکه، در ساده‌ترین تقریب، کلمات مانند راس‌های گراف هستند و بین هر دو کلمه‌ی همسایه در یک جمله، یک یال رسم می‌شود.]، [این گونه می‌توانیم این شبکه‌ی هم‌رخدادی را رسم کرده و خصوصیات آن را برای متن‌ها و شعرهای مختلف بدست آوریم.]

حال باید کلمه به کلمه نیز بکنیم، که در نهایت به این صورت خواهد شد:

[زبان, 'ا', 'جنبه‌ای', 'بنیادی', 'در', 'ارتباط', 'و', 'برهمکنش', 'بین', 'انسان‌هاست', 'که', 'نقش', 'مهمی', 'در', 'انتقال', 'افکار', 'ا', 'ایده‌ها', 'و', 'احساسات', 'دارد', 'پژوهشگران', 'در', 'طول', 'سال‌ها', 'در', 'رشته‌های', ' مختلف', 'ا', 'سعی', 'کرده‌اند', 'از', 'منظرهای', 'متفاوتی', 'برهمکنش‌های', 'پیچیده‌ی', 'زبانی', 'را', 'تحلیل', 'و', 'مطالعه', 'کنند', 'ا', 'یکی', 'از', 'پنجره‌هایی', 'که', 'برای', 'نگاه', 'کردن', 'و', 'مطالعه‌ی', 'زبان‌ها', 'می‌توان', 'پیدا', 'کرد', 'ا', 'بررسی', 'زبان‌ها', 'به', 'عنوان', 'شبکه', 'است', 'ا', 'یکی', 'از', 'انواع', 'این', 'شبکه‌ها', 'که', 'می‌توان', 'به‌وسیله‌ی', 'آن', 'زبان', 'را', 'مورد', 'بررسی', 'قرار', 'داد', 'شبکه‌های', 'هم‌رخدادی', 'است', 'به', 'طوری', 'که', 'در', 'یک', 'متن', 'ا', 'كلمات', 'هم‌جوار', 'را', 'هم‌سایه', 'گرفته', 'و', 'بین', 'آن‌ها', 'می‌توانیم', 'ارتباطی', 'فرض', 'کنیم', 'ا', 'در', 'مدل', 'سازی', 'این', 'شبکه', 'در', 'ساده‌ترین', 'تقریب', 'ا', 'كلمات', 'مانند', 'راس‌های', 'گراف', 'هستند', 'و', 'بین', 'هر', 'دو', 'كلمه‌ی', 'هم‌سایه', 'در', 'یک', 'جمله', 'ا', 'یک', 'یال', 'رسم', 'می‌شود', 'ا', 'این', 'گونه', 'می‌توانیم', 'این', 'شبکه‌ی', 'هم‌رخدادی', 'را', 'رسم', 'کرده', 'و', 'خصوصیات', 'آن', 'را', 'برای', 'متن‌ها', 'و', 'شعرهای', ' مختلف', 'بدست', 'آوریم']

حال باید علائم نگارشی را در این متن حذف کنیم. خروجی آن چنین چیزی خواهد شد:

[زبان, 'جنبه‌ای', 'بنیادی', 'در', 'ارتباط', 'و', 'برهمکنش', 'بین', 'انسان‌هاست', 'که', 'نقش', 'مهمی', 'در', 'انتقال', 'افکار', 'ایده‌ها', 'و', 'احساسات', 'دارد', 'پژوهشگران', 'در', 'طول', 'سال‌ها', 'در', 'رشته‌های', ' مختلف', 'سعی', 'کرده‌اند', 'از', 'منظرهای', 'متفاوتی', 'برهمکنش‌های', 'پیچیده‌ی', 'زبانی', 'را', 'تحلیل', 'و', 'مطالعه', 'کنند', 'یکی', 'از', 'پنجره‌هایی', 'که', 'برای', 'نگاه', 'کردن', 'و', 'مطالعه‌ی', 'زبان‌ها', 'می‌توان', 'پیدا', 'کرد', 'بررسی', 'زبان‌ها', 'به', 'عنوان', 'شبکه', 'است', 'یکی', 'از', 'انواع', 'این', 'شبکه‌ها', 'که', 'می‌توان', 'به‌وسیله‌ی', 'آن', 'زبان', 'را', 'مورد', 'بررسی', 'قرار', 'داد', 'شبکه‌های', 'هم‌رخدادی', 'است', 'به', 'طوری', 'که', 'در', 'یک', 'متن', 'كلمات', 'هم‌جوار', 'را', 'هم‌سایه', 'گرفته', 'و', 'بین', 'آن‌ها', 'می‌توانیم', 'ارتباطی', 'فرض', 'کنیم', 'در', 'مدل', 'سازی', 'این', 'شبکه', 'در', 'ساده‌ترین', 'تقریب', 'كلمات', 'مانند', 'راس‌های', 'گراف', 'هستند', 'و', 'بین', 'هر', 'دو', 'كلمه‌ی', 'هم‌سایه', 'در', 'یک', 'جمله', 'یک', 'یال', 'رسم', 'می‌شود', 'این', 'گونه', 'می‌توانیم', 'این', 'شبکه‌ی', 'هم‌رخدادی', 'را', 'رسم', 'کرده', 'و', 'خصوصیات', 'آن', 'را', 'برای', 'متن‌ها', 'و', 'شعرهای', ' مختلف', 'بدست', 'آوریم']

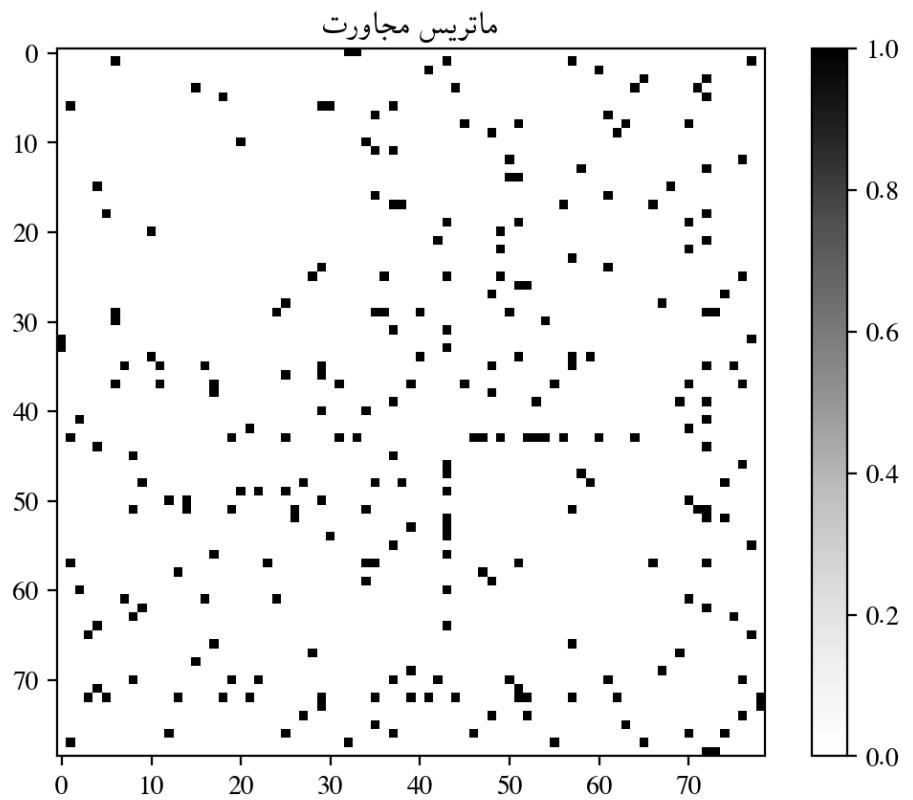
حال باید همه‌ی شکل‌های واژگان و افعال را به ریشه اصلی و مفردشان برد تا بتوانیم شبکه‌ی کامل و دقیقی داشته باشیم. این کار را با استفاده از تابع Lemmatizer که در کتابخانه‌ی پردازش زبان Hazm است انجام می‌دهیم. دو نمونه از کار این تابع را می‌بینیم:

hazm.Lemmatizer().lemmatize('متن') = ('متن‌ها')
hazm.Lemmatizer().lemmatize('آورده‌آور') = ('آوریم')

حال پس از این که این آرایه‌ها را به ریشه‌ها و مفردہایشان بردمیم تا یکسان شوند، گراف را تشکیل می‌دهیم و بین هر دو درایه‌ی مجاور در این آرایه یک یال می‌کشیم. حال می‌توان ماتریس مجاورت این گراف را بدست آورد.

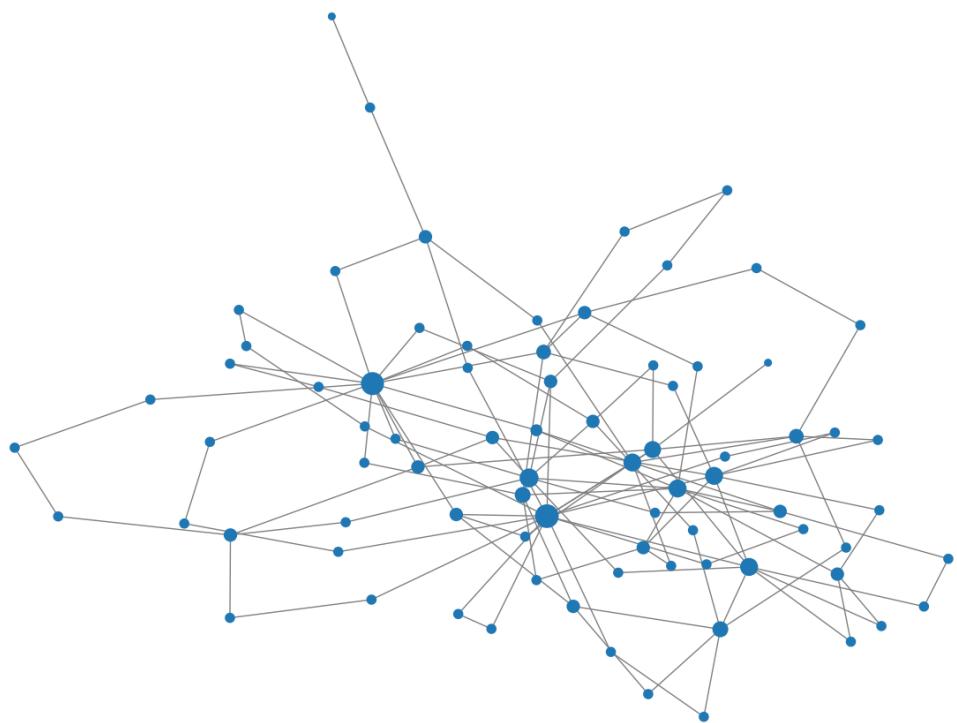
پس از بدست آوردن بدست آوردن گراف و ماتریس مجاورت آن می‌توان این گراف را به طور گرافیکی نمایش داد (که البته برای گراف‌ها با راس‌های خیلی بزرگ، خیلی کارا نیست).

در صفحه‌ی بعد، شکل ماتریس مجاورت را با رنگ روی شکل نشان داده‌ایم. سیاه‌ها نشان دهنده‌ی وجود یال بین دو راس بوده و سفیدها به معنی عدم وجود یال.



حال شکل گرافیکی گراف را نیز می‌بینیم. اندازه‌ی هر راس متناسب با درجه‌ی راس آن می‌باشد.

گراف کلمات برای متن



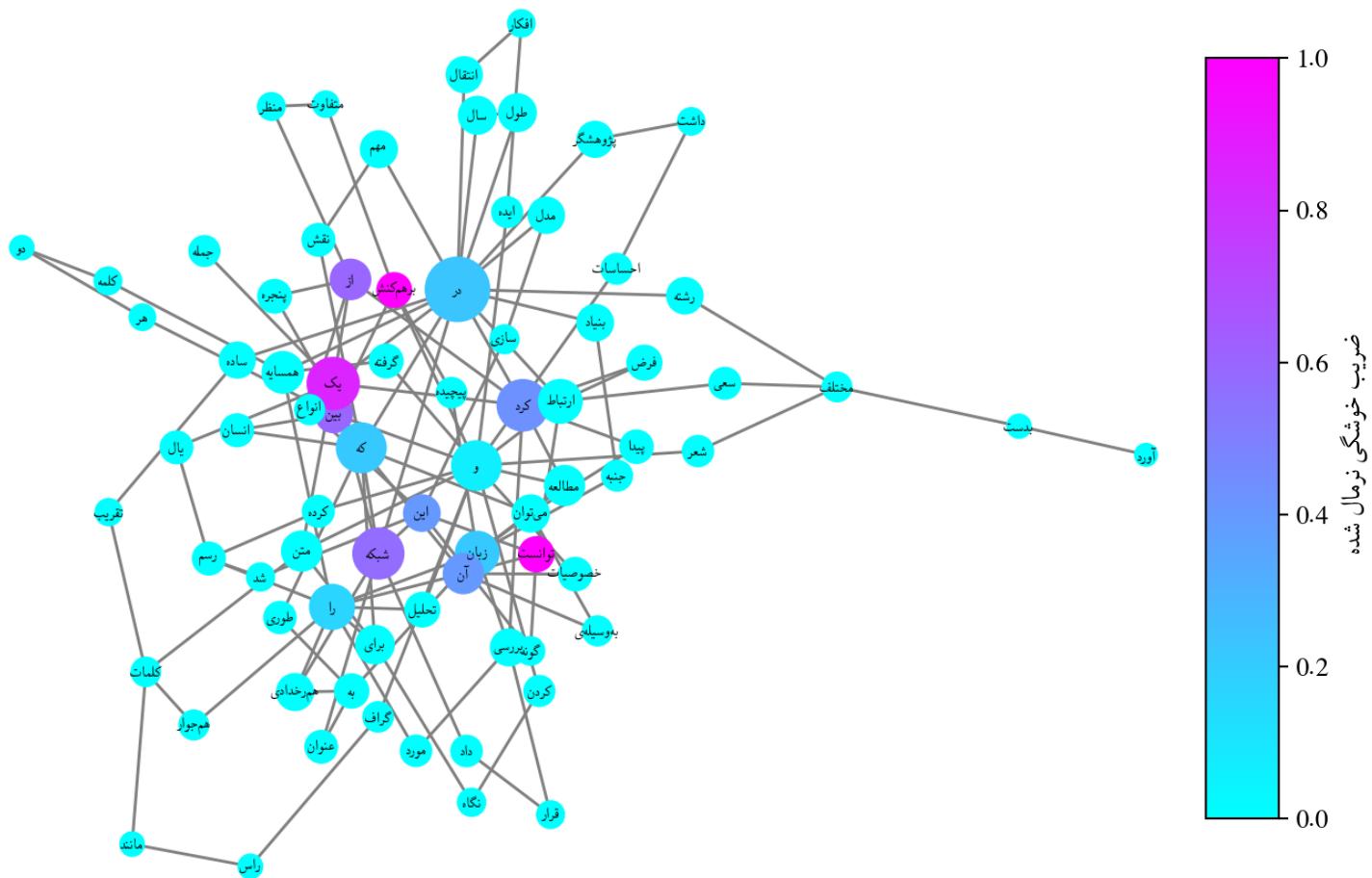
همین طور که در شکل نیز مشخص است، چندین راس با درجه‌ی نسبتاً بزرگ داریم ولی اکثر راس‌ها کوچک هستند.

حال می‌توانیم ویژگی‌هایی از قبل ضریب خوشگی، نزدیکی، میانگی و میزان اهمیت با تعریف ویژه برداری را نیز برای هر راس بدست آوریم و به نحوی نشان دهیم. نزدیکی و میانگی را در ارائه‌ی کلاس نشان دادیم پس در اینجا، ضریب خوشگی و اهمیت را با تعریف ویژه برداری نشان می‌دهیم.

ضریب خوشگی را برای هر راس حساب کرده و با رنگ آن را روی راس نشان می‌دهیم. برای نمایش بهتر نیز دو کار را انجام دادیم. اولاً، ضرایب خوشگی برای راس‌هایی با درجه راس پایین (مثلاً زیر چهار) به راحتی ممکن است یک شود ولی این برای ما خیلی ارزشمند نیست. پس درجه راس‌های زیر چهار را اگر ضریب خوشگی یک داشتند، حذف می‌کنیم و مثل بقیه نشان می‌دهیم. دوماً، برای تفکیک بهتر رنگ‌ها در نمایش، همه‌ی ضرایب خوشگی راس‌ها را به حداقل ضریب خوشگی بدست آمده در کل گراف تقسیم کرده هر ضریب خوشگی، عددی بین صفر و یک می‌شود که یک همان حداقل ضریب خوشگی یافت شده برای یک راس در گراف می‌باشد.

اندازه‌ی هر راس را نیز می‌توانیم متناسب با اهمیت با تعریف ویژه‌بردای آن رسم کنیم. پس در شکل حاصل، اندازه‌ی راس‌ها متناسب با اهمیت آن به صورت ویژه‌برداری است و رنگ آن‌ها نشان‌دهنده‌ی ضریب خوشگی آن‌ها می‌باشد.

گراف کلمات برای متن



با توجه به شکل چند نکته مورد توجه است. می‌توان دید که بیشترین درجه‌ی راس‌ها متعلق به حروف ربط، اضافه و پس از آن افعال است. اگر این کلمات را از شبکه حذف کنیم، پس از آن می‌توانیم کلمات کلیدی متن را استخراج کنیم. مثلاً در این متن، پس از حذف، کلمات «شبکه»، «زبان»، «ارتباط» و غیره ضریب خوشنگی و درجه‌ی راس بالایی دارند. البته نحوه‌ی بررسی و مقایسه‌ی درجه‌ی راس‌ها با ضریب خوشنگی فرق دارد. همچنین این پارامترها در متن‌های کوچک و کم تکرار مانند این پاراگراف خیلی خوب مشخص نیست ولی مثلاً پس از حذف حروف و افعال می‌بینیم که کلمه‌ی «شبکه» هم درجه‌ی راس بالایی دارد و هم اهمیت آن با تعریف ویژه مقداری بالا است و همچنین ضریب خوشنگی بالایی دارد. بالا بودن درجه راس که نشان دهنده‌ی تکرار آن در متن است و به نوعی با اهمیت آن نیز می‌تواند در ارتباط باشد. ولی ضریب خوشنگی یک کلمه می‌تواند نشان دهنده‌ی ارتباط آن کلمه و جایگاه آن در متن باشد و رابطه‌ی سطح بالاتری از شبکه را به ما گوشزد می‌کند. مثلاً کلمه‌ی «برهم‌کنش» اهمیت ویژه برداری خیلی بالایی ندارد ولی این جایگاه ویژه را در متن داشته که ضریب خوشنگی آن بالا رفته است.

لیست درجه‌ی راس‌ها به صورت زیر است:

('کلمه', درجه‌ی راس)

[('در', 15), ('و', 14), ('را', 9), ('که', 8), ('کرد', 8), ('شبکه', 7), ('یک', 7), ('این', 6), ('آن', 6), ('بین', 5), ('از', 5), ('متن', 4), ('مخالف', 4), ('برای', 4), ('بررسی', 4), ('رسم', 4), ('همسایه', 4), ('برهم‌کنش', 4), ('می‌توان', 4), ('ارتباط', 4), ('به', 4), ('توانست', 4), ('کلمات', 4), ('هم‌خدادی', 3), ('مطالعه', 3), ('داشت', 2), ('ساده', 2), ('طول', 2), ('عنوان', 2), ('سازی', 2), ('منظر', 2), ('به‌وسیله‌ی', 2), ('انتقال', 2), ('پیدا', 2), ('بدست', 2), ('سال', 2), ('متفاوت', 2), ('بنیاد', 2), ('پیچیده', 2), ('طوری', 2), ('فرض', 2), ('گونه', 2), ('نگاه', 2), ('تحلیل', 2), ('مانند', 2), ('گراف', 2), ('شد', 2), ('پنجره', 2), ('پژوهشگر', 2), ('جنبه', 2), ('رشته', 2), ('مورد', 2), ('خصوصیات', 2), ('انسان', 2), ('ایده', 2), ('گرفته', 2), ('راس', 2), ('کردن', 2), ('هم‌جوار', 2), ('کرده', 2), ('افکار', 2), ('انواع', 2), ('احساسات', 2), ('مدل', 2), ('قرار', 2), ('شعر', 2), ('تقریب', 2), ('یال', 2), ('دو', 2), ('کلمه', 2), ('سعی', 2), ('نقش', 2), ('داد', 2), ('مهم', 2), ('جمله', 1), ('آورده', 1)]

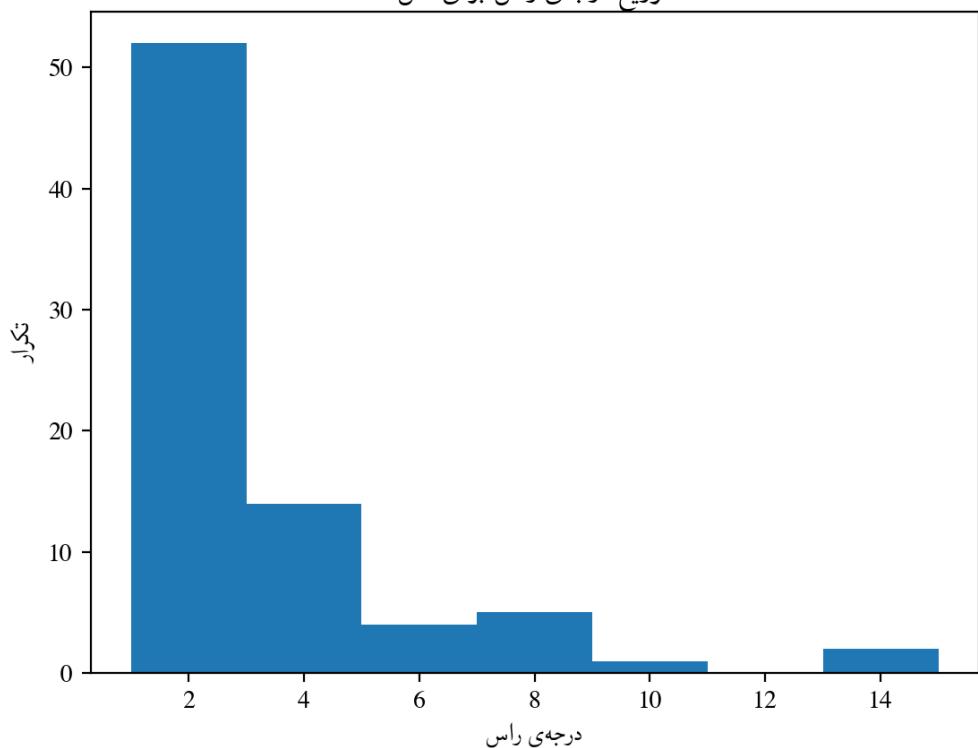
مطابق پاراگراف قبل، حروف ربط، اضافه و افعال به ترتیب بیشترین درجه‌ی راس‌ها را دارد هستند. پس از آن نیز کلمات کلیدی متن را می‌توان پیدا کرد که حروف ربط، اضافه و اشاره را با رنگ سبز، افعال را با رنگ قرمز و کلمات کلیدی را با آبی مشخص کردیم (فقط درجه‌ی راس‌های بالا). مثلاً طبق این لیست می‌توان متوجه شد که کلمات «زبان» و «شبکه» اهمیت بالایی در متن دارند.

متنی که بررسی کردیم شامل 78 راس و 127 یال است. که درجه‌ی راس میانگین 3.26 را به ما خواهد داد. همچنین میانگین اتصال راس‌ها² 2.11 است. ضریب خوشنگی میانگین نیز برای این گراف، 0.0529 است. در ضمن ضریب هم‌سنخ‌جویی نیز 0.215 - است. که نشان می‌دهد رئوس با درجه‌ی راس زیاد بیشتر تمایل دارند به درجه‌ی پایین‌ترها وصل شوند. و این نشان از ساختار ادبیاتی است که حروف ربط و اضافه متصل کننده‌ی کلمات به هم دیگر هستند و این حروف، خود را به هم وصل نخواهند کرد و ساختارهای همسایگی به صورت اتصال‌های کلمه به کلمه و حرف به کلمه بیشترین مقدار است و اتصال‌های حرف به کلمه نیز به قدری زیاد شده که این ضریب هم‌سنخ‌جویی منفی بدست آمده است. در قسمت بعدی این پارامترها را با اشعار مقایسه کرده و تفاوت‌های ساختاری در جمله بندی و استفاده از حروف و کلمات و افعال را خواهیم دید.

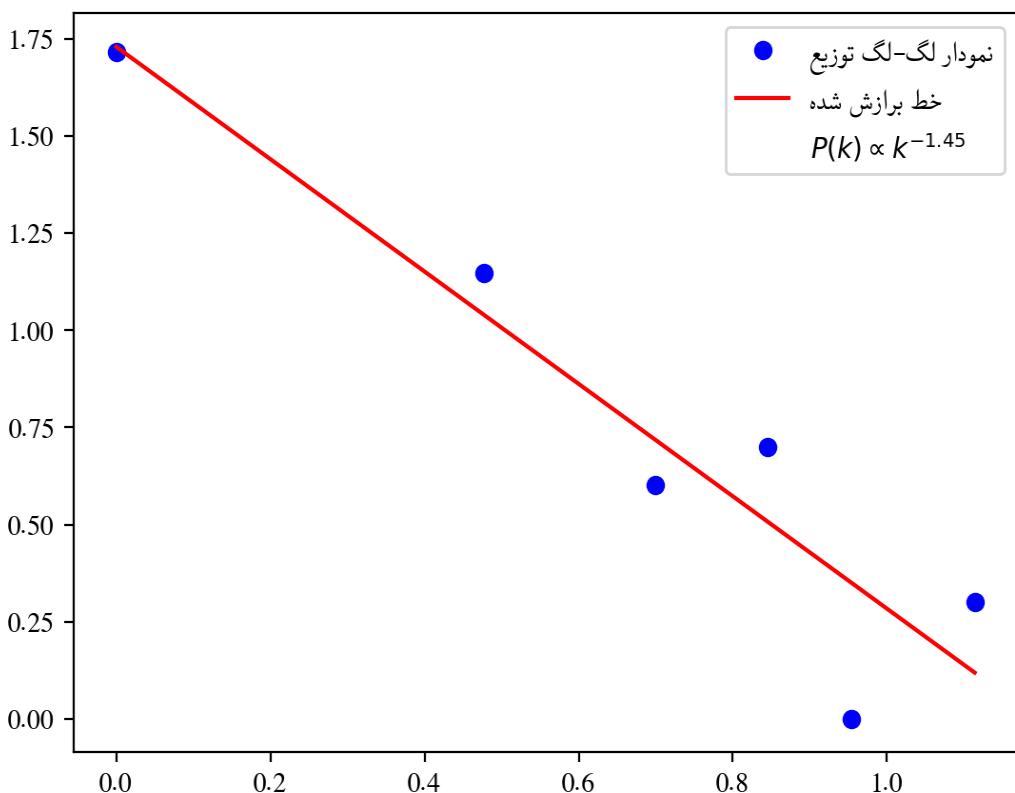
توزیع درجه راس

حال به سراغ توزیع درجه راس رفته و آن را رسم می‌کنیم. در بخش‌های بعدی این توزیع معنا پیدا می‌کند و اینجا با 78 راس، خیلی توزیع تمیز و خوبی نخواهیم رسید. ولی با این حال می‌آوریم.

توزیع درجه‌ی راس برای متن



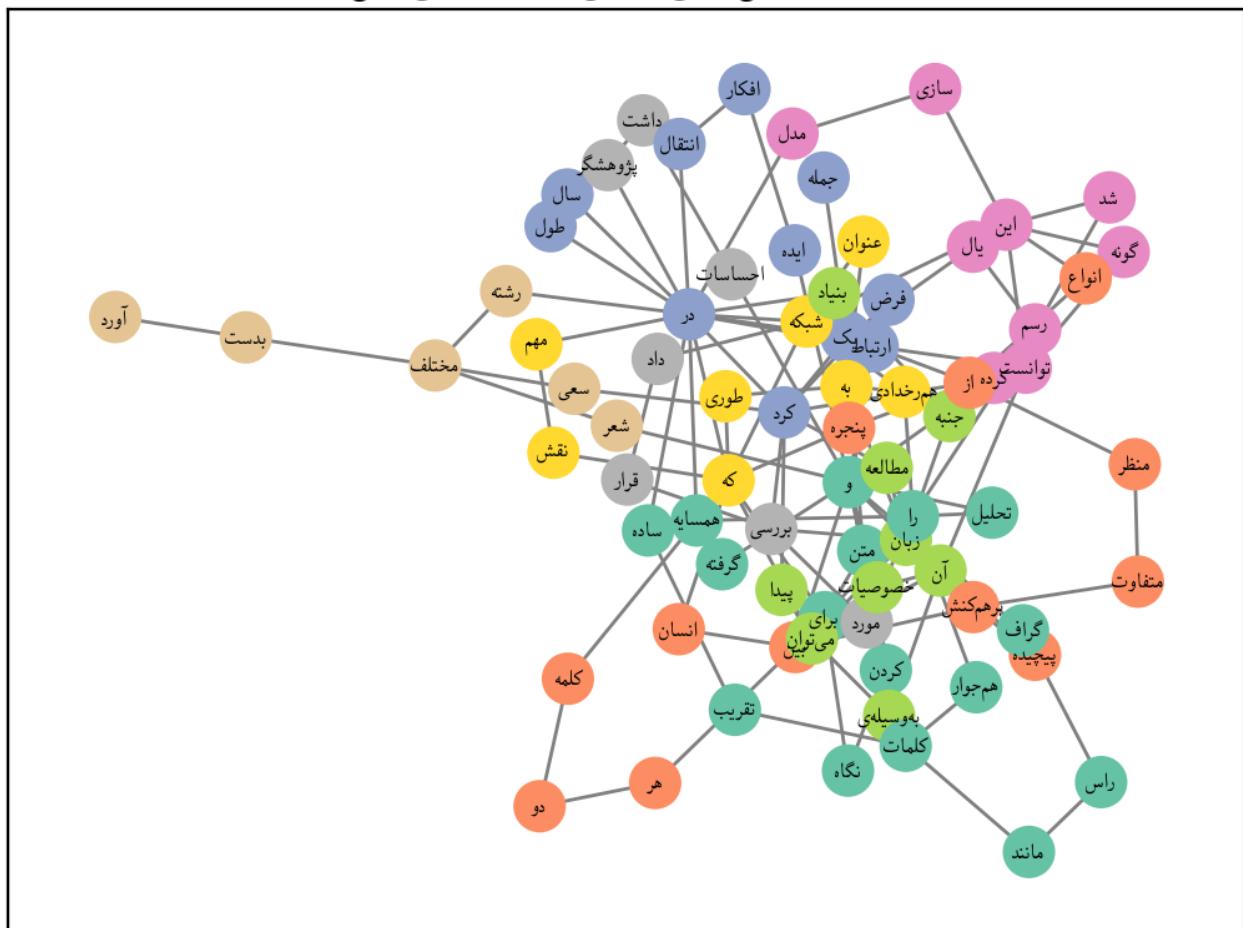
حال به توجه به شکل‌ها حدس می‌زنیم که توزیع توانی باید داشته باشیم. حال باید دانه بندی را نیز لگاریتمی کنیم تا بتوانیم خطی به این‌ها برازش کنیم. این کار را انجام می‌دهیم و نتیجه را می‌بینیم:



پس نمای تابع توانی برازش شده، $1.45 - \text{می باشد.}$

حال می توانیم این شبکه را نیز انجمان بندی کنیم:

گراف انجمان بندی شده کلمات برای متن



مقایسه‌ی شعر با متن

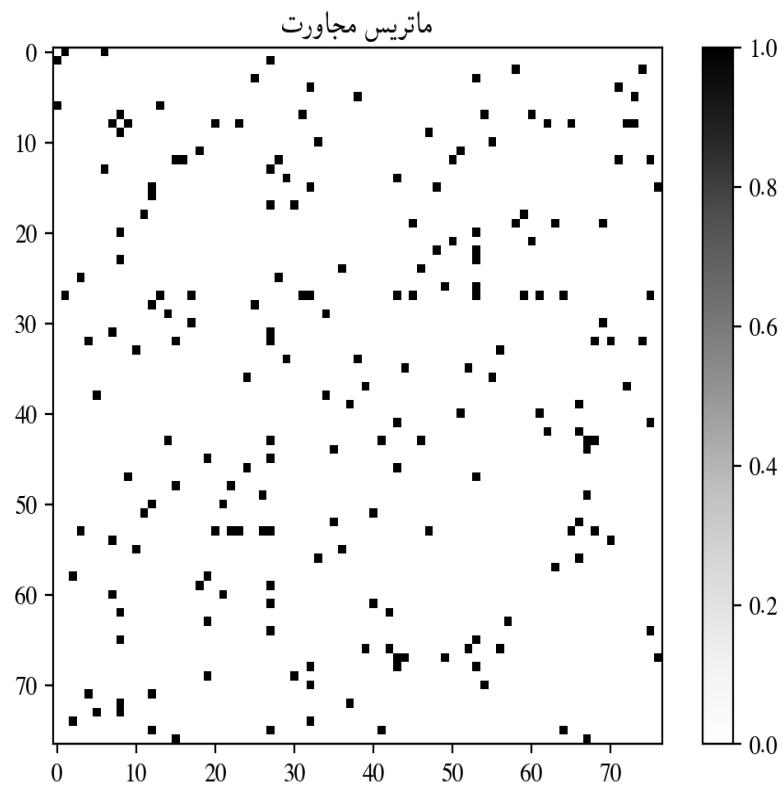
حال برای مقایسه، تکه‌ای کوچک از اشعار پروین اعتضامی را نیز همین کار را می‌کنیم (سعی می‌کنیم تعداد بیتی را انتخاب کنیم تا تعداد راس‌ها برابر شود).

بجرئت کرد روزی بال و پر باز
گذشت از بامکی بر جو کناری
شدش گیتی به پیش چشم تاریک
ز رنج خستگی درماند در راه
گه از تشویش سر در زیر پر کرد
نه اش نیروی زان ره بازگشتن
نه راه لانه دانستی کدامست
نه از خواب خوشی نام و نشانی

کیوتو بچه‌ای با شوق پرواز
پرید از شاخکی بر شاخساری
نمودش بسکه دور آن راه نزدیک
ز وحشت سست شد بر جای ناگاه
گه از اندیشه بر هر سو نظر کرد
نه فکرش با قضا دمساز گشتن
نه گفتی کان حوادث را چه نامست
نه چون هر شب حدیث آب و دانی

حال همین کارهایی که کردیم را با این بیت شعر نیز انجام داده و مقایسه می‌کنیم با متن.

در این صفحه، ماتریس مجاورت به شکل گرافیکی و همچنین شکل شبکه که اندازه‌ها متناسب با اهمیت ویژه برداری بودند و رنگ‌ها مشخصه‌ی ضریب خوشگی بودند رسم شده‌اند.



گراف کلمات برای شعر



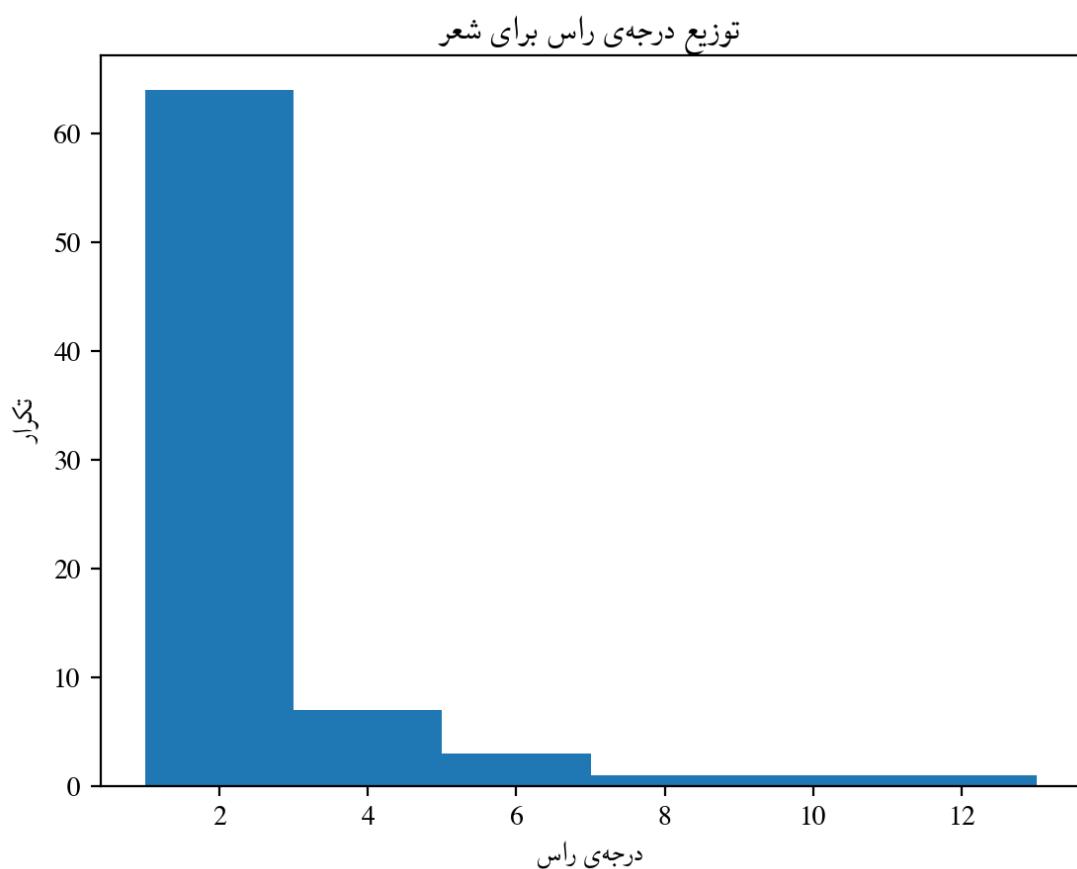
همین طور که در شکل‌ها نیز مشخص است، ماتریس مجاورت خیلی تُنگ بوده (نسبت به متن مورد بررسی) و همچنین راس‌هایمان از ضریب خوشگی بالایی برخوردار نیستند. باز هم از شکل می‌توان دید که شاخه‌هایی بلند وجود دارد که نشان از بودن تعداد خیلی بیشتری راس با درجه راس پایین (مثلاً دو) می‌باشد و بنظر اختلاف در درجه‌ی راس‌ها شدیدتر از متن است.

حال خصوصیات را بررسی می‌کنیم:

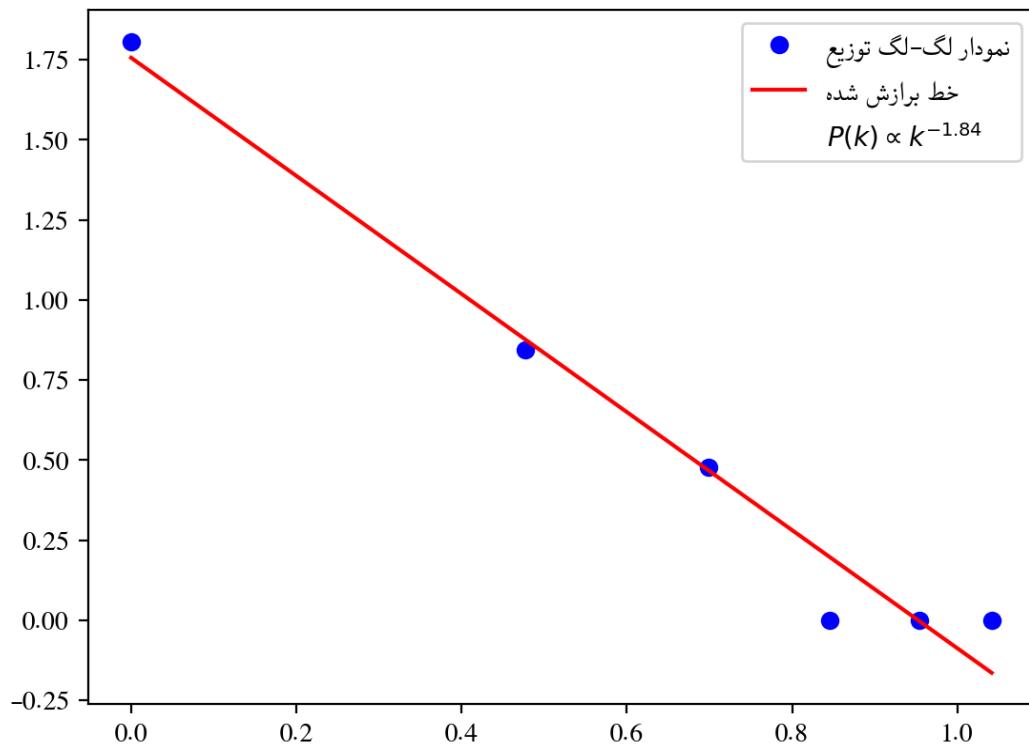
در گراف این شعر ما 77 راس و 100 سال داریم که میانگین درجه راس 2.59 را به ما خواهد داد. همچنین ضریب خوشگی میانگین نیز برای این گراف، 0.0153 است. در ضمن ضریب هم‌سنخ‌جويي نيز 0.099 - است.

حال می‌توان دید که با تعداد راس تقریباً برابر (78 در مقابل 77)، یال‌های کمتری در گراف است و میانگین درجه راس پایین‌تر از متن است. همچنین ضریب خوشگی میانگین تقریباً یک‌چهارم متن است که خیلی کمتر است! تفاوت جالب نیز بالاتر بودن میزان هم‌سنخ‌جويي است که یعنی در اشعار، بر خلاف متن‌ها، راس‌های هم‌درجه گویی بیشتر می‌خواهند به هم متصل شوند که این نشان‌دهنده‌ی این است که گویی پیوند بین کلمات و درجه‌ی پایین‌ترها بیشتر شده و به مانند متن نیست که کلمات خیلی دور و بر حروف اضافه و ربط باشند و ساختار شعری متفاوت است و نحوه‌ی برخورد با این حروف نیز در شعر خیلی با متن فرق دارد!

حال می‌توانیم دیگر خصوصیات آماری را نیز ببینیم:

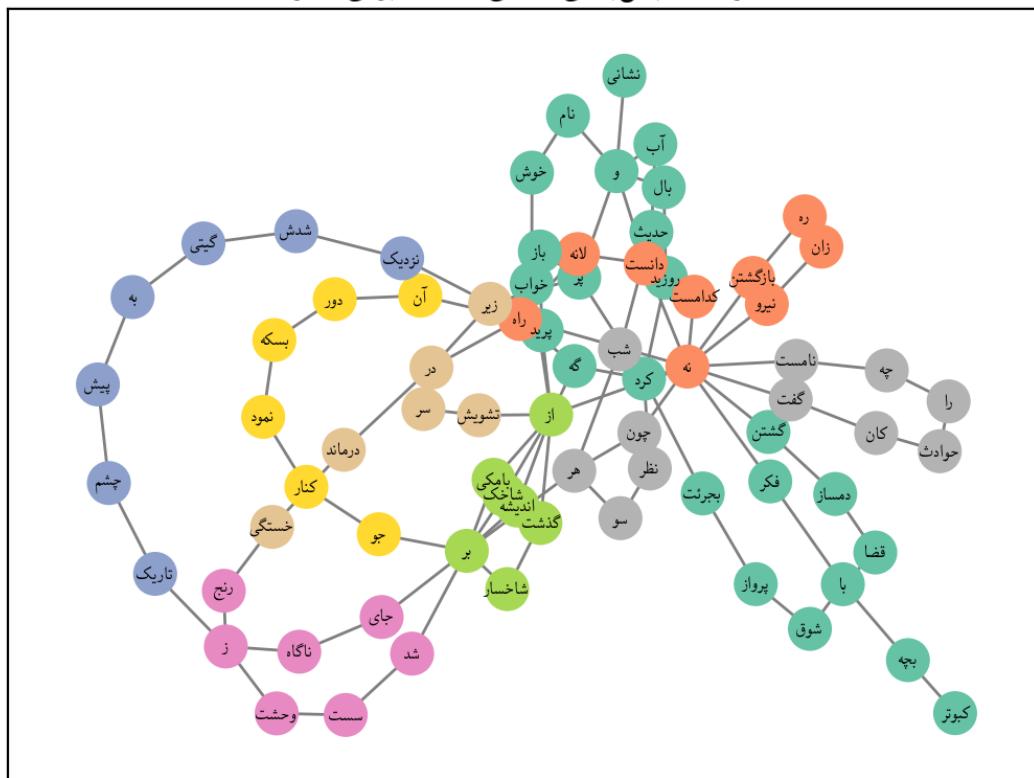


که همان‌طور که در شکل خود گراف نیز دیدیم، اختلاف بسیار شدیدتری در توزیع درجه‌ی راس‌ها با متن وجود دارد و شبیه آن تندتر می‌باشد.



که این شیب بیشتر یعنی نمای $1.84 -$ را می‌دهد که همان‌چیزی است که انتظار داشتیم.
این نیز انجمن‌بندی برای این شعر است:

گراف انجمن‌بندی شده کلمات برای شعر



خاصیت جهان کوچکی

یکی دیگر از خواصی که می‌توان در این گراف‌ها چک کرد خاصیت جهان کوچکی برای این متون است. برای این خاصیت باید ضریب خوشگی میانگین و طول مشخصه‌ی گراف را حساب کرده و با گراف رندوم با تعداد راس و یال برابر چک کنیم. برای این کار خوب است اندازه‌ی گراف بزرگ باشد ولی با این حال ما برای همین دو مثال خودمان نیز این را بررسی می‌کنیم. در بخش‌های بعدی این کار را با گراف‌های بزرگ‌تر انجام خواهیم داد.

حال طول مشخصه و ضریب خوشگی میانگین متن و شعر را پیدا کرده و با گراف رندومی که همین تعداد راس و یال را دارد (اینقدر این گراف رندوم را می‌سازیم تا همبند باشد و سپس مقایسه را انجام می‌دهیم چون احتمال خوبی وجود دارد که همبند نباشد) مقایسه می‌کنیم. همچنین برای مقایسه با گراف‌های رندوم تولید شده، یک آنسامبل 50 تایی از گراف‌های همبند رندوم تولید شده از روی گراف‌های اصلی متن و شعرمان می‌سازیم و میانگین طول مشخصه و میانگین ضریب خوشگی میانگین آن‌ها را گزارش می‌کنیم.

برای متن:

طول مشخصه‌ی گراف: 3.3496

ضریب خوشگی میانگین: 0.0528

طول مشخصه‌ی گراف رندوم ساخته شده از گراف متن: 3.8046

ضریب خوشگی میانگین گراف رندوم ساخته شده از گراف متن: 0.0346

برای شعر:

طول مشخصه‌ی گراف: 4.6490

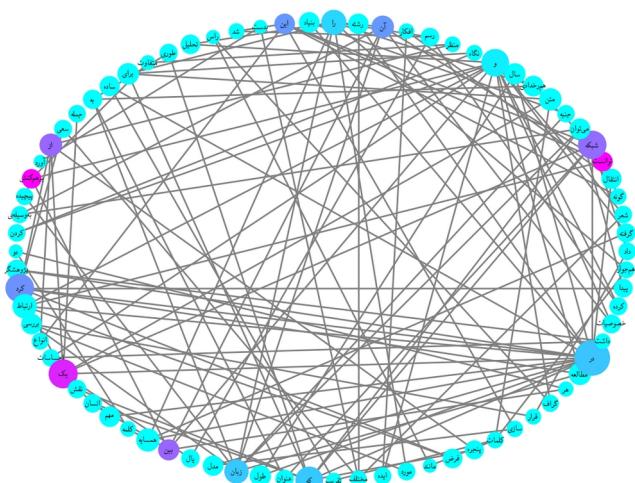
ضریب خوشگی میانگین: 0.0153

طول مشخصه‌ی گراف رندوم ساخته شده از گراف متن: 4.827

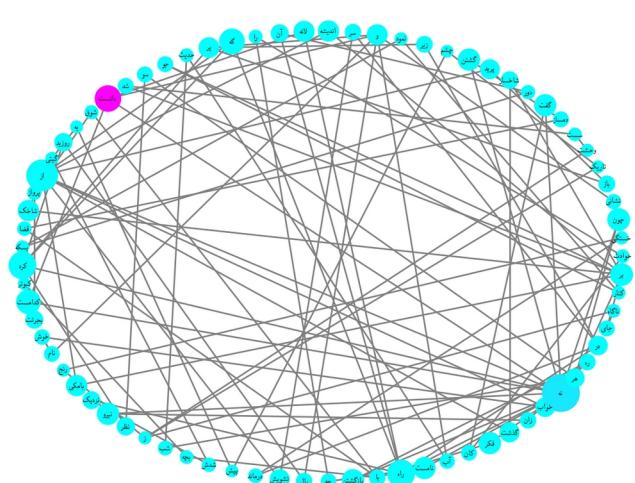
ضریب خوشگی میانگین گراف رندوم ساخته شده از گراف متن: 0.0121

حال می‌توان دید که برای متن، طول مشخصه بیشتر شده ولی ضریب خوشگی افت کرده است. این ما را به این که این شبکه‌ها می‌توانند جهان کوچک باشند رهنمون می‌کند. البته در شعر درست است که طول مشخصه بیشتر شده است ولی ضریب خوشگی بدست آمده خیلی نزدیک به ضریب خوشگی خود گراف شعر ما بوده باید بررسی روی گراف‌های بزرگ‌تری از متن احتمالاً صورت گیرد.

گراف کلمات برای متن



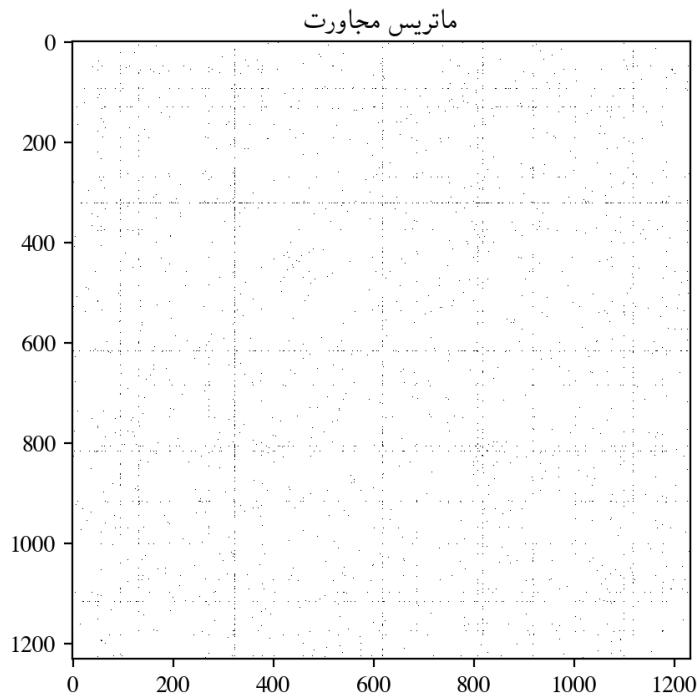
گراف کلمات برای شعر



بررسی متن‌ها و اشعار بزرگ‌تر

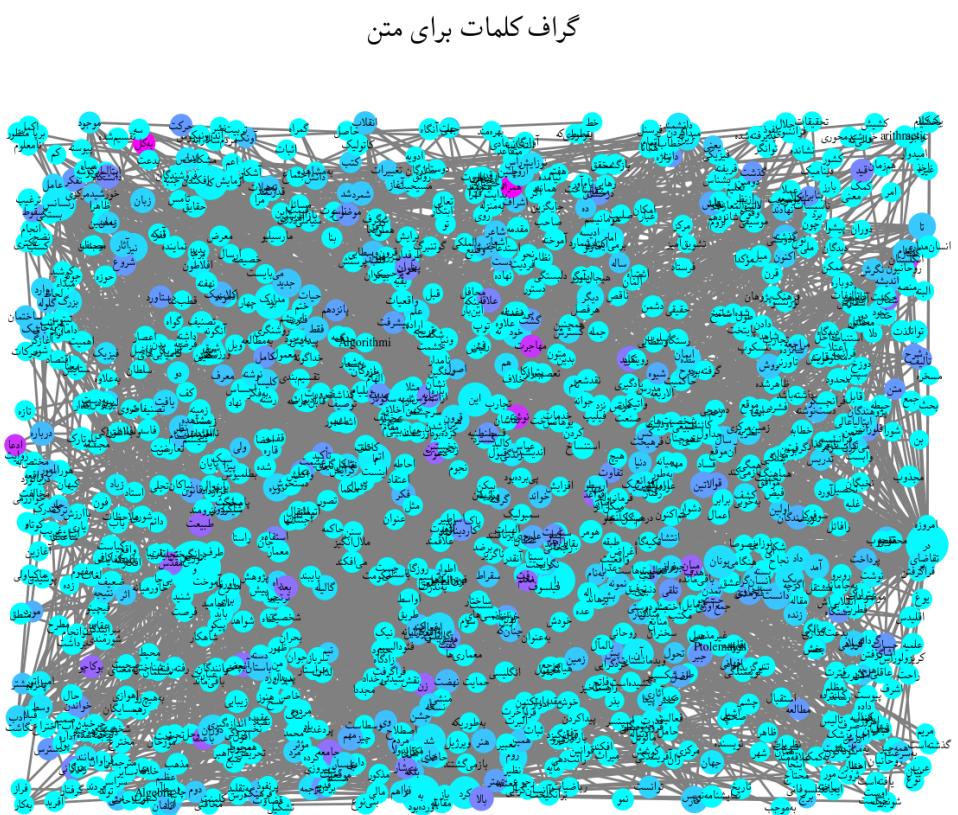
حال که با فضای کار آشنا شدیم، متن‌ها و اشعار بزرگ‌تر را مورد بررسی قرار می‌دهیم. متن‌های مورد بررسی مقاله‌ی رنسانس در ویکی‌پدیا و داستان شازده کوچولو است. شعرهای مورد بررسی نیز، بخش ضحاک از شاهنامه‌ی فردوسی، قصاید پروین اعتصامی و دفتر اول مثنوی معنوی است. که همه چیزی از حدود 1300 راس دارند.

متن اول: مقاله‌ی رنسانس ویکی‌پدیا

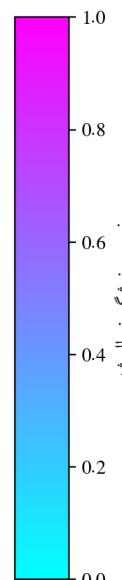
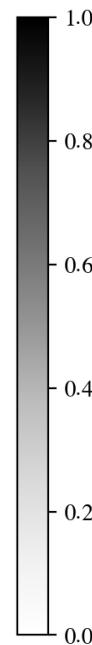


همان کمیت‌هایی که برای متن‌های کوچک محاسبه کردیم را اینجا نیز حساب می‌کنیم.

ماتریس مجاورت را به طور گرافیکی در رو به رو می‌بینیم:



این نیز گراف کلمات می‌باشد که البته خیلی ارزشی برای نمایش ندارد.



آمار تجمیعی برای این گراف:

تعداد راس‌ها: 1232

تعداد یال‌ها: 3082

میانگین درجه راس: 5.004

ضریب خوشگی میانگین: 0.1745

ضریب همسنخ‌جويي: - 0.1873

بیشترین درجه راس‌ها (درجات بالاتر از 30 را می‌آوریم):

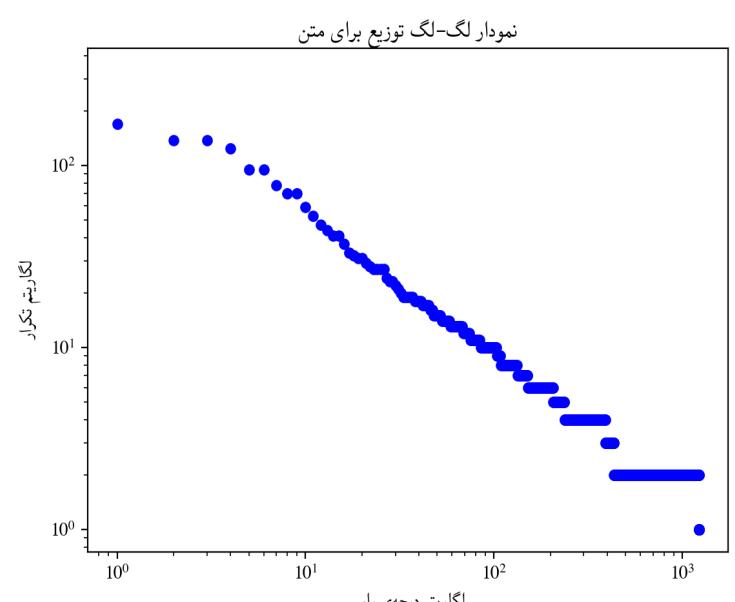
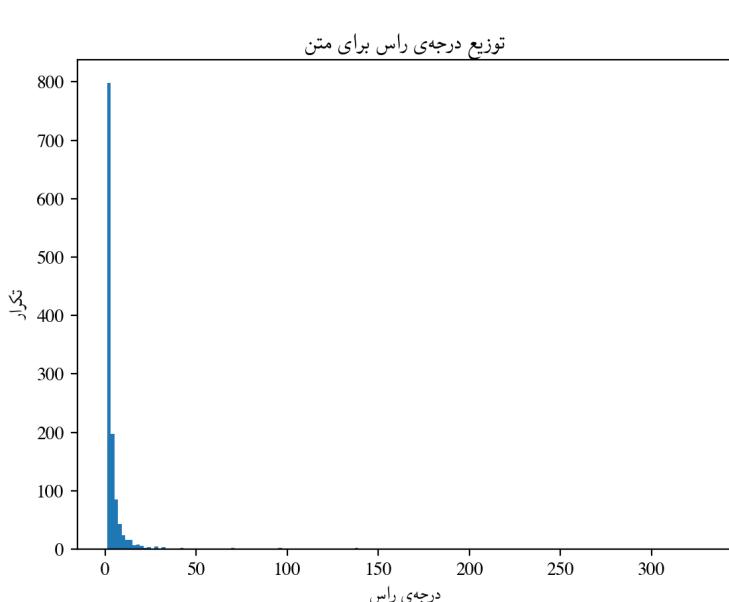
('و', 329), ('در', 169), ('به', 137), ('از', 124), ('را', 95), ('بود', 95), ('این', 78), ('کرد', 70)

('رنسانس', 70), ('با', 59), ('شد', 53), ('آن', 47), ('یونان', 44), ('برای', 41), ('کتاب', 41), ('خود', 37), ('او', 33)

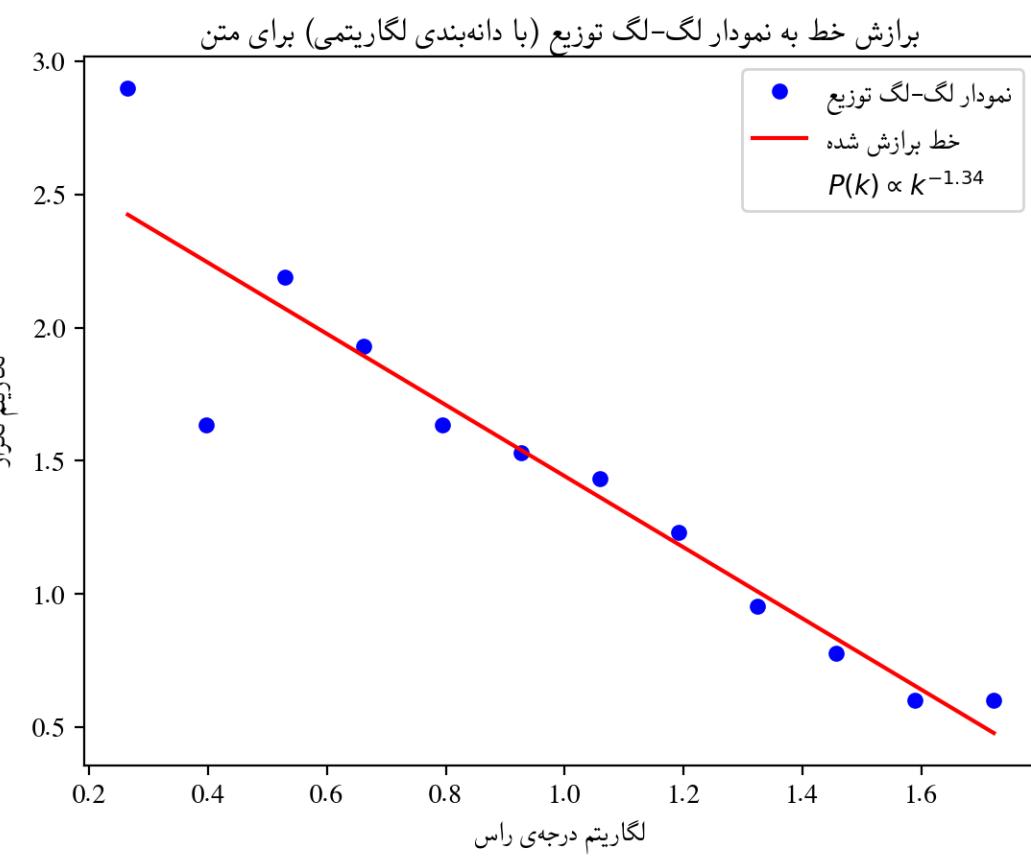
استخراج کلمات کلیدی پس از حذف افعال، حروف و ضمایر:

رنسانس - یونان - کتاب - انسان - ایتالیا - دوران

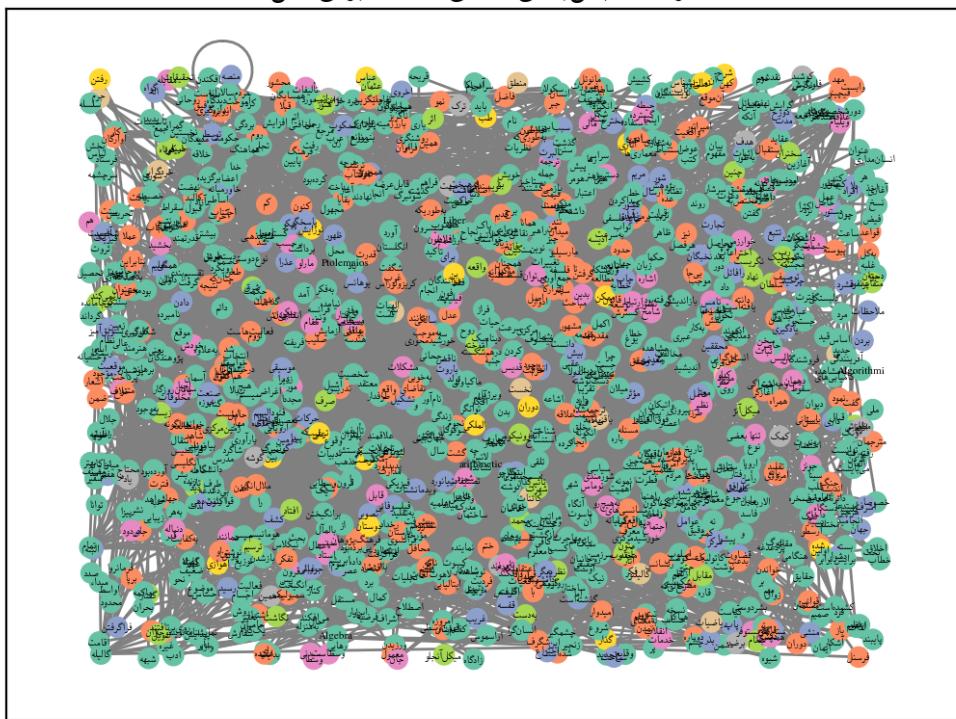
بررسی توزیع درجه راس‌ها:



حال در صفحه‌ی بعد شکل نمودار لگاریتمی به همراه دانه‌بندی لگاریتمی را خواهید دید که نمای آن را پس از برازش خط، 1.34 - بدست آوردیم.



حال انجمن‌بندی را نیز انجام می‌دهیم:

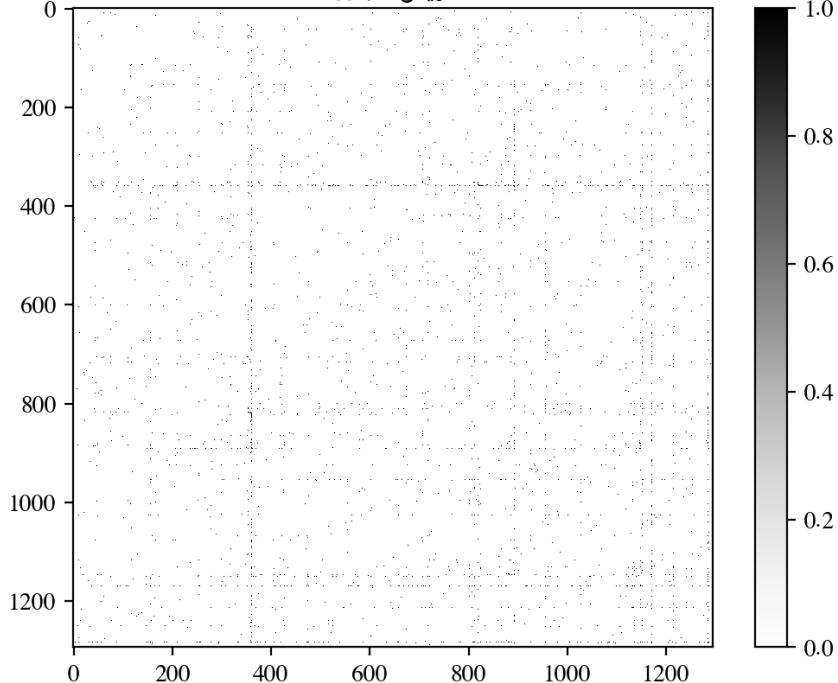


که به طور گرافیکی خیلی مشخص نیست چه اتفاقی دارد می‌افتد و کارهای پیشرفته‌تری برای تحلیل لازم است.

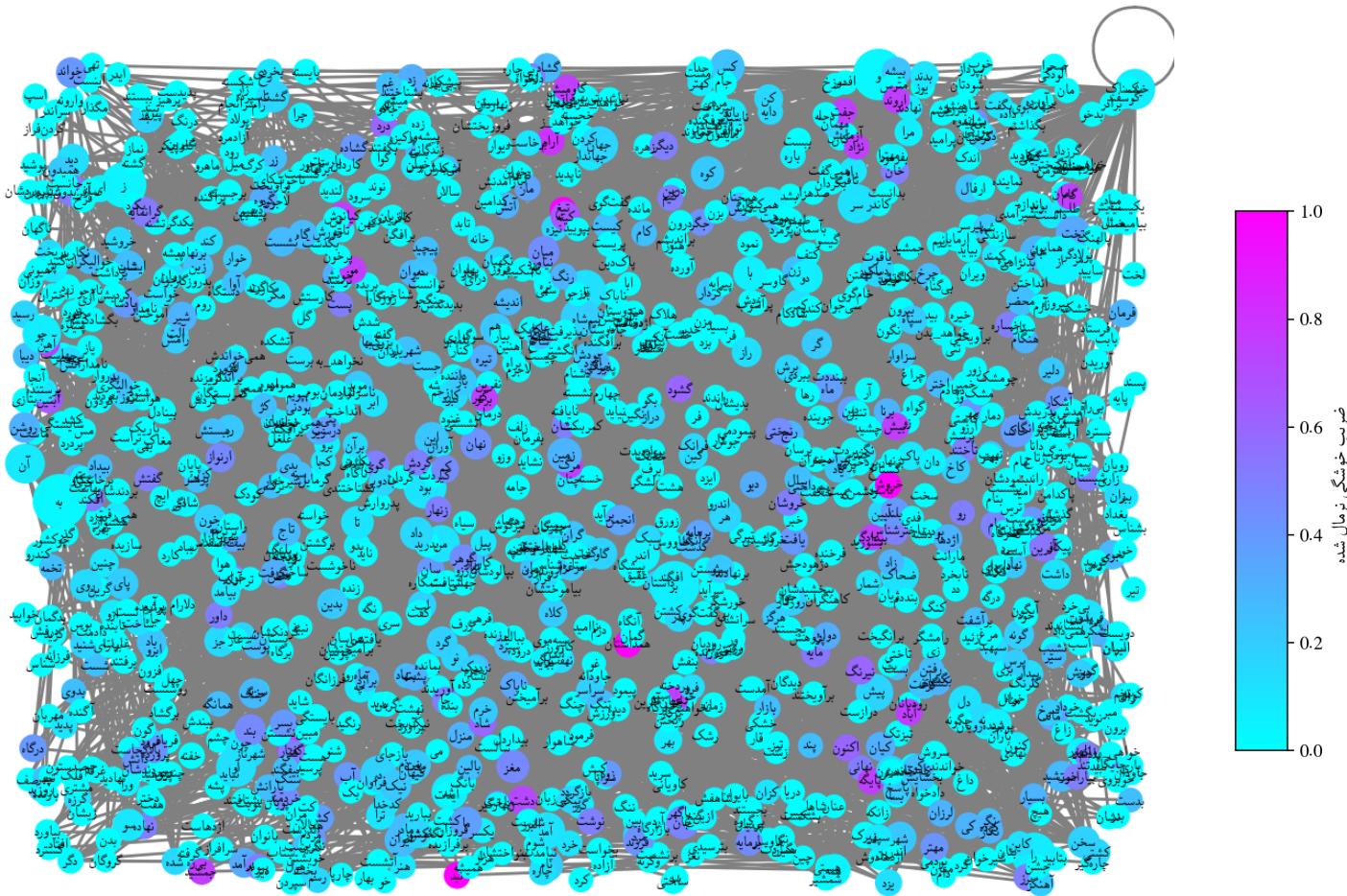
شعر اول: بخش ضحاک از شاهنامه‌ی فردوسی

ماتریس مجاورت و شکل گراف را می‌بینیم:

ماتریس مجاورت



گراف کلمات برای شعر



آمار تجمیعی برای این گراف:

تعداد راس‌ها: 1294

تعداد یال‌ها: 4393

میانگین درجه راس: 6.7897

ضریب خوشگی میانگین: 0.1571

ضریب همسنخ‌جويي: -0.1901

بیشترین درجه راس‌ها (درجات بالاتر از 50 را می‌آوریم):

('به', 254), ('و', 225), ('که', 159), ('از', 152), ('بر', 159), ('ز', 141), ('را', 111), ('یک', 87), ('بود', 86),

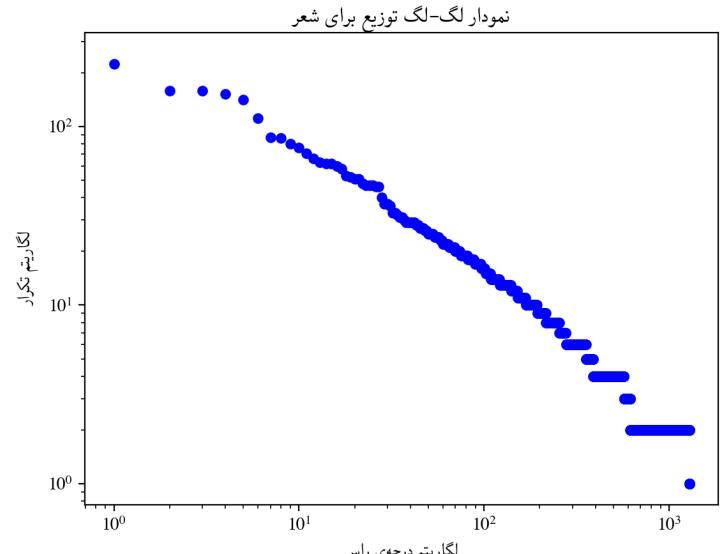
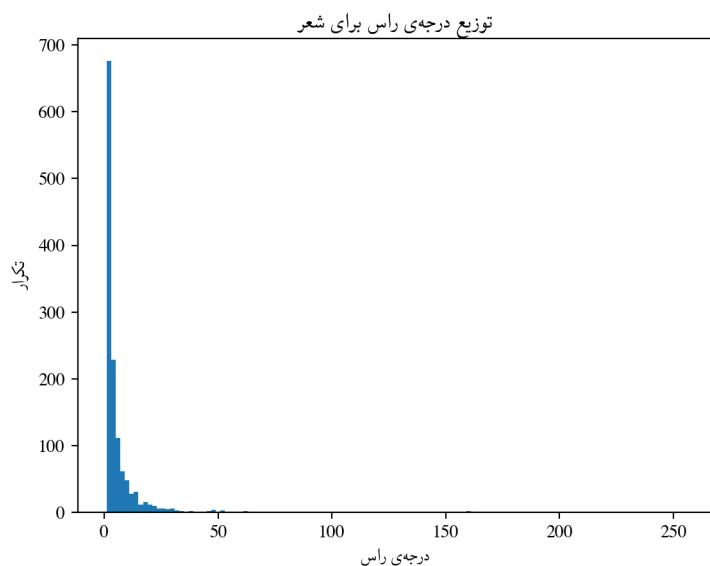
('آن', 80), ('چو', 76), ('تو', 71), ('او', 66), ('شد', 63), ('سر', 62), ('اندر', 62), ('ضحاک', 60), ('همی', 58),

('من', 53), ('با', 52), ('شاه', 51), ('آمد', 51)

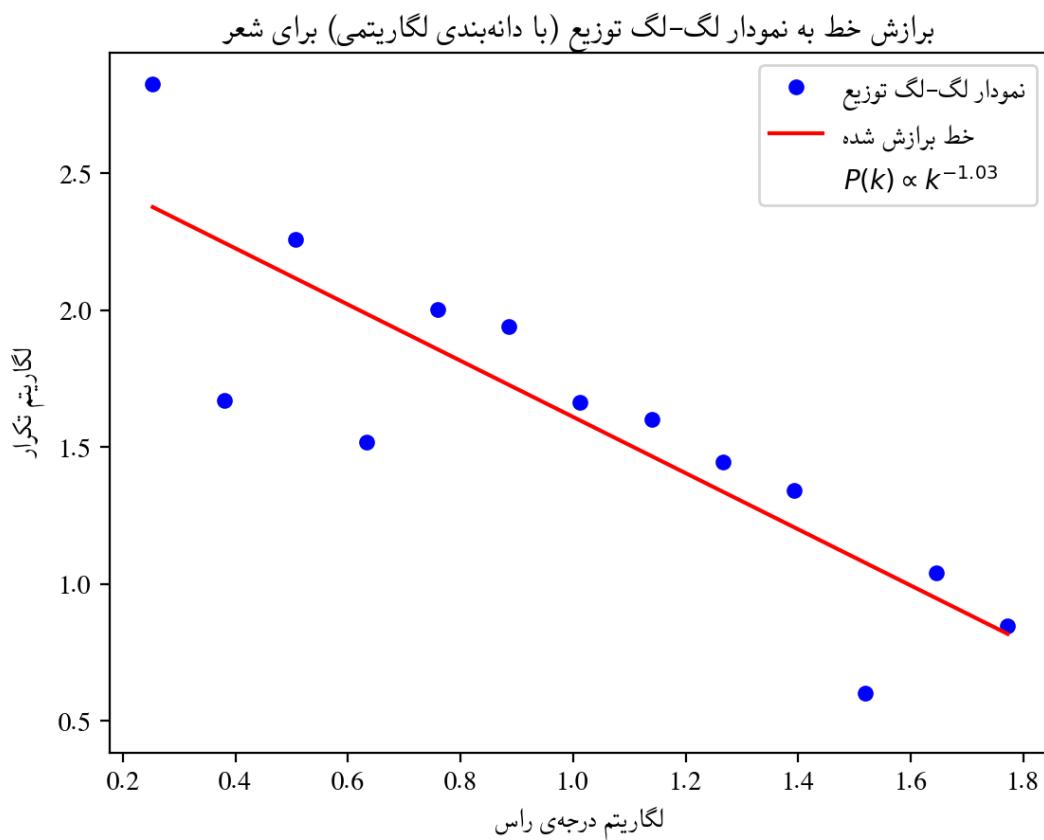
استخراج کلمات کلیدی پس از حذف افعال، حروف و ضمایر:

ضحاک - شاه

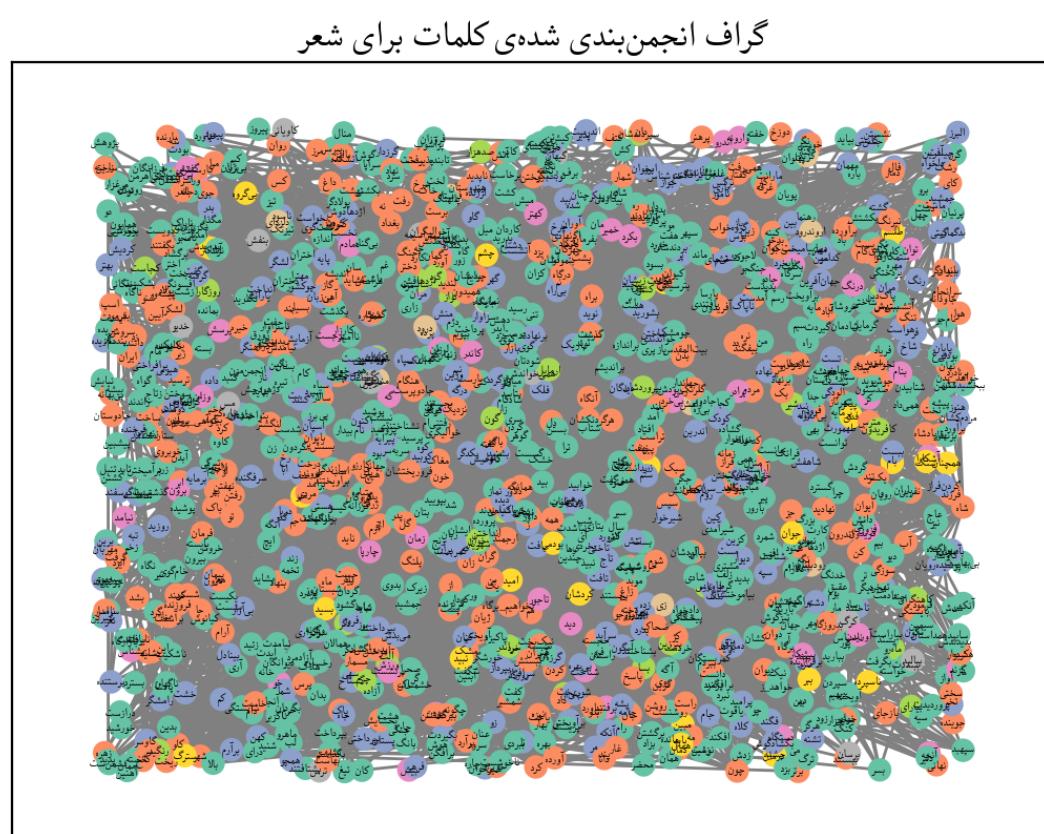
بررسی توزیع درجه راس‌ها:



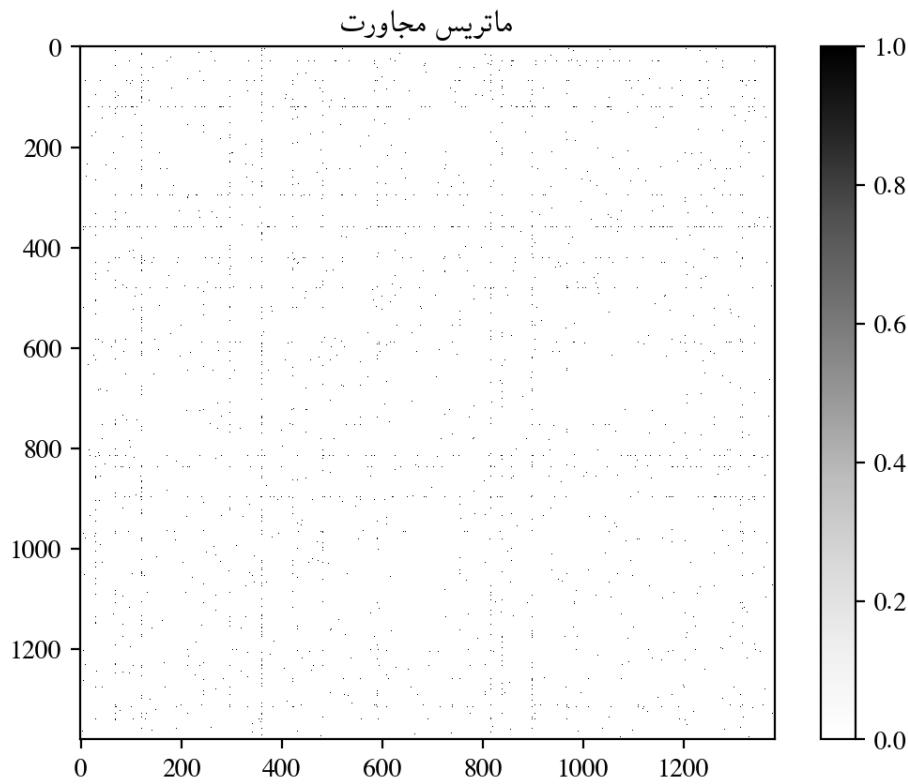
حال در صفحه‌ی بعد شکل نمودار لگاریتمی به همراه دانه‌بندی لگاریتمی را خواهید دید که نمای آن را پس از برازش خط، 1.03 - بدست آوردیم.



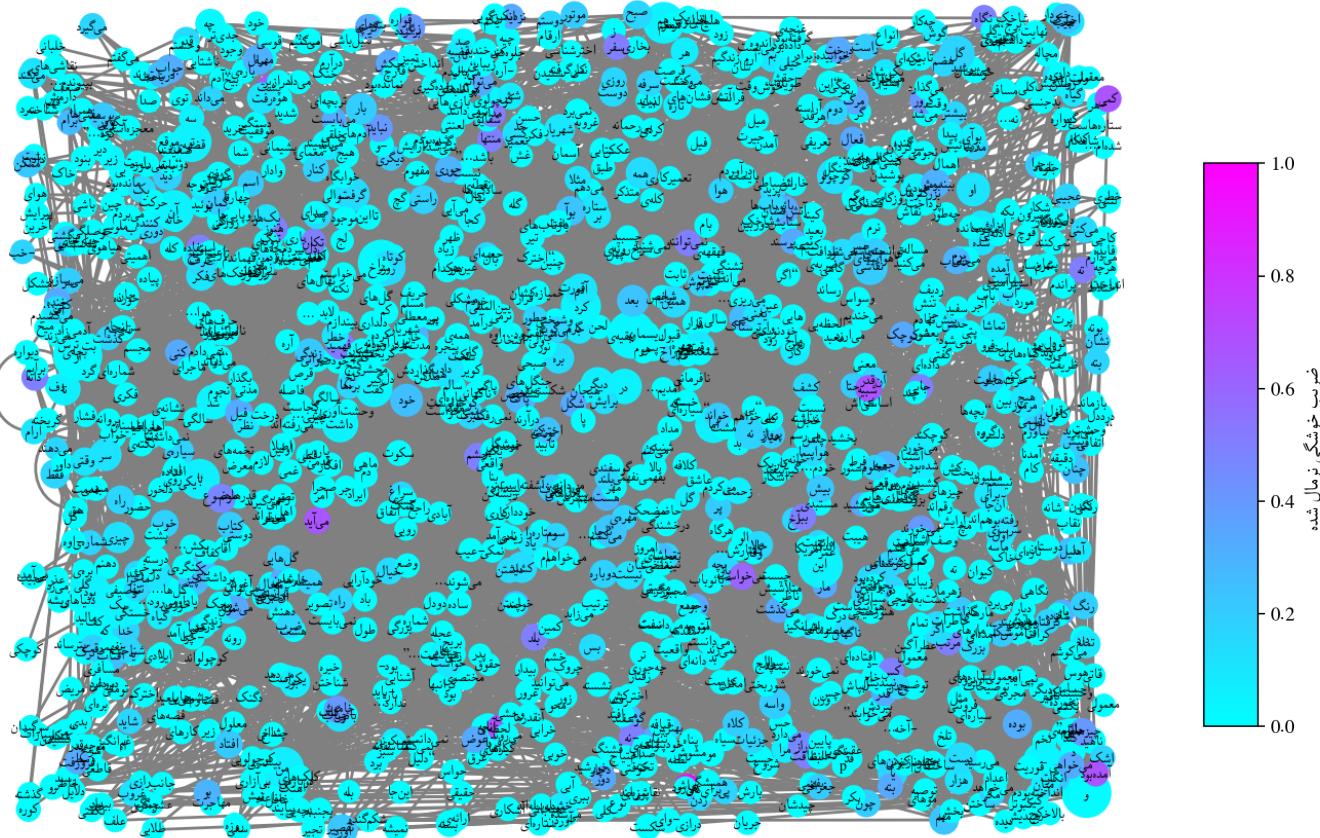
حال انجمن بندی را نیز انجام می‌دهیم:



متن دوم: داستان شازده کوچولو بررسی ماتریس مجاورت و شبکه گراف:



گراف کلمات برای متن



آمار تجمیعی برای این گراف:

تعداد راس‌ها: 1380

تعداد یال‌ها: 3738

میانگین درجه راس: 5.417

ضریب خوشگی میانگین: 0.1481

ضریب هم‌سنجی: -0.2159

بیشترین درجه راس‌ها (درجات بالاتر از 40 را می‌آوریم):

('که', 208), ('و', 205), ('از', 177), ('را', 150), ('به', 150), ('این', 116), ('یک', 104), ('کرد', 82), ('هم', 81)

('آن', 77), ('من', 76), ('بود', 73), ('تو', 73), ('با', 72), ('گفت', 57), ('داشت', 57), ('اما', 55), ('در', 53), ('شده', 52)

('خیلی', 49), ('دیگر', 43), ('تا', 40)

استخراج کلمات کلیدی پس از حذف افعال، حروف و ضمایر:

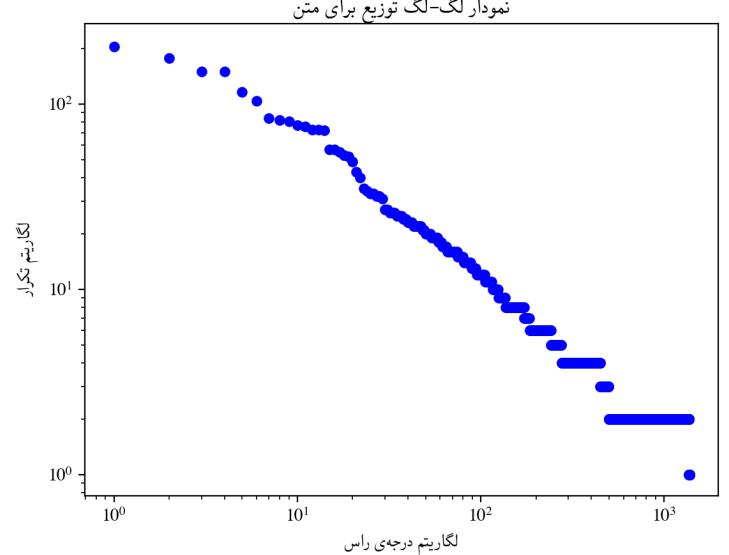
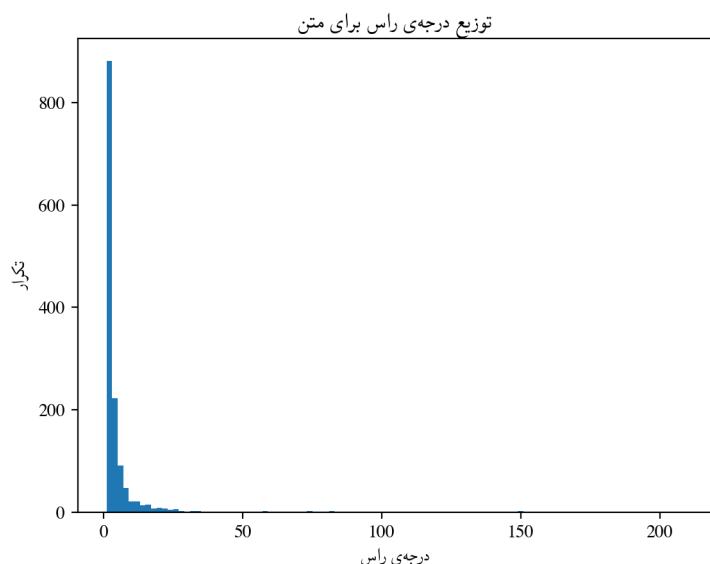
با توجه به این که داستان شازده کوچولو بیشتر مکالمه محور است و خط داستانی یا مرجعی را دنبال

نمی‌کند و بخش زیادی از داستان سوال و جواب و گفت‌وگو است، کلمات کلیدی‌ای خیلی نمی‌توان یافت.

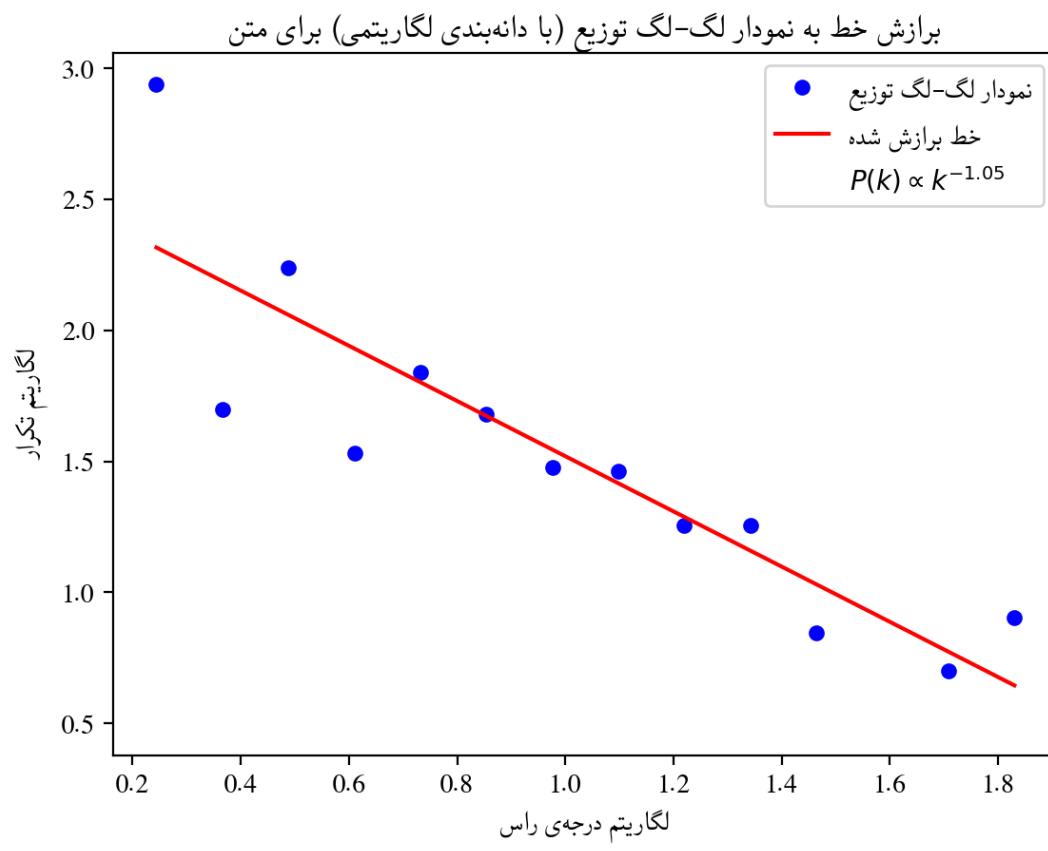
البته می‌توان دید که کلماتی که در متون مکالمه محور پر کاربردند، مثل ضمائر، کلمه‌ی گفت و ... در بین

درجه راس‌های بالا پیدا می‌شوند.

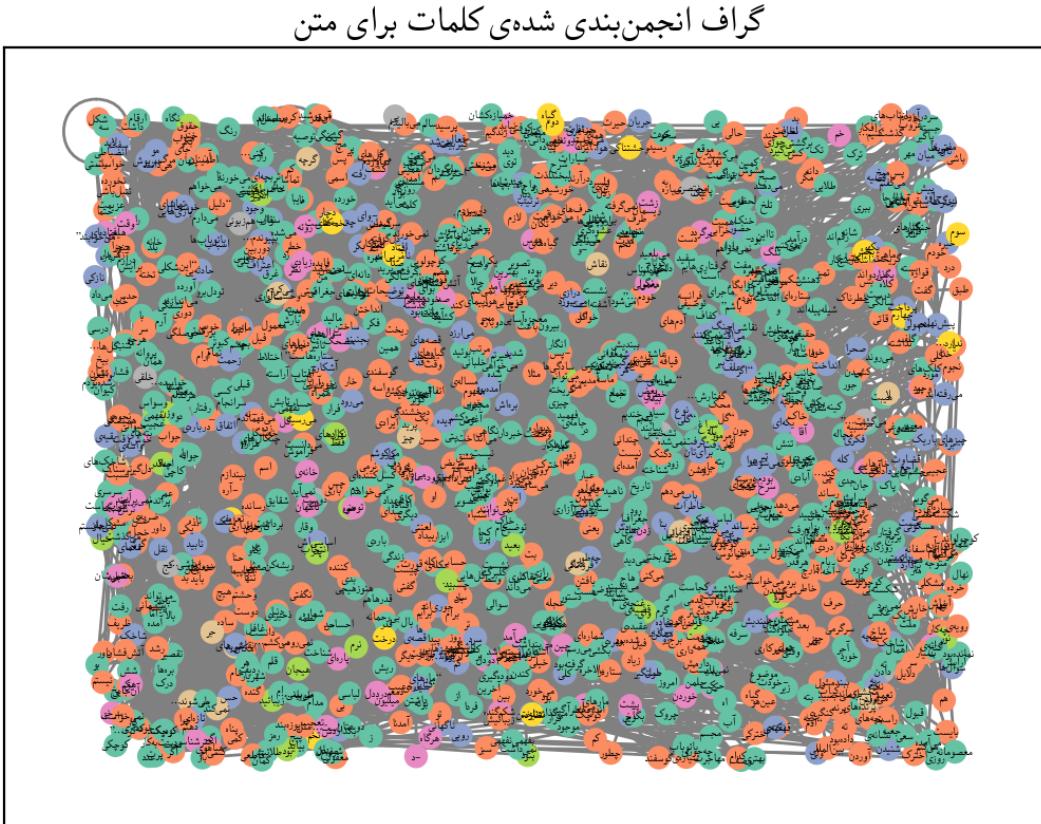
بررسی توزیع درجه‌های راس:



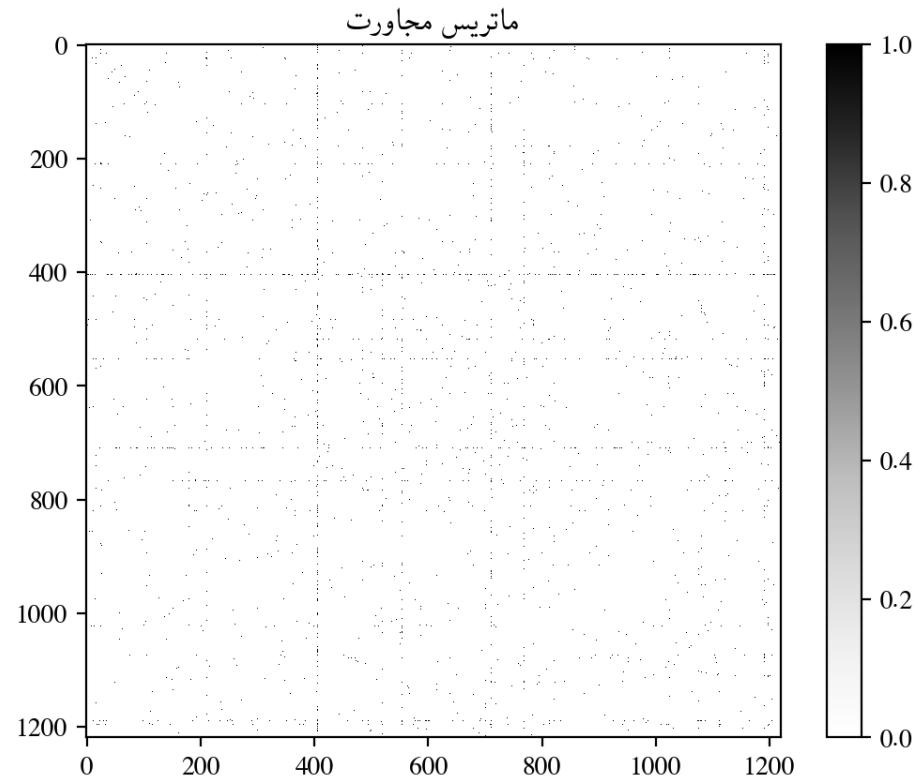
حال در صفحه‌ی بعد شکل نمودار لگاریتمی به همراه دانه‌بندی لگاریتمی را خواهید دید که نمای آن را پس از برازش خط، 1.05 - بدست آوردیم.



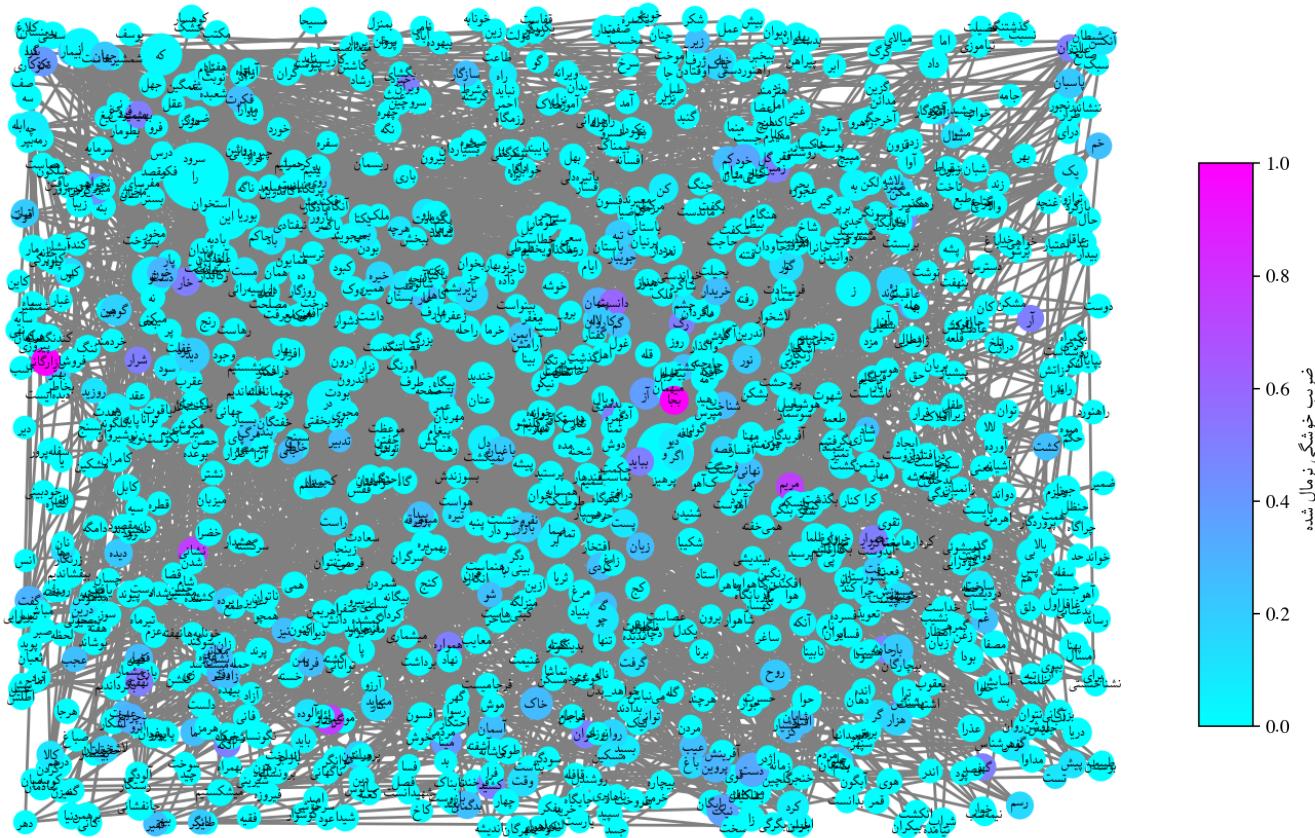
حال انجمن‌بندی را نیز انجام می‌دهیم:



شعر دوم: قصائد پروین اعتصامی



گراف کلمات برای شعر



آمار تجمیعی برای این گراف:

تعداد راس‌ها: 1219

تعداد یال‌ها: 2523

میانگین درجه راس: 4.139

ضریب خوشگی میانگین: 0.1121

ضریب هم‌سنخ‌جويي: - 0.1838

بیشترین درجه راس‌ها (درجات بالاتر از 20 را می‌آوریم):

('را', 257), ('و', 207), ('این', 110), ('که', 100), ('در', 87), ('از', 83), ('تو', 64), ('ز', 51), ('ای', 45), ('بر',

('چه', 34), ('با', 30), ('کرد', 29), ('چو', 26), ('به', 25), ('یک', 25), ('تا', 24), ('دل', 23), ('بی', 23),

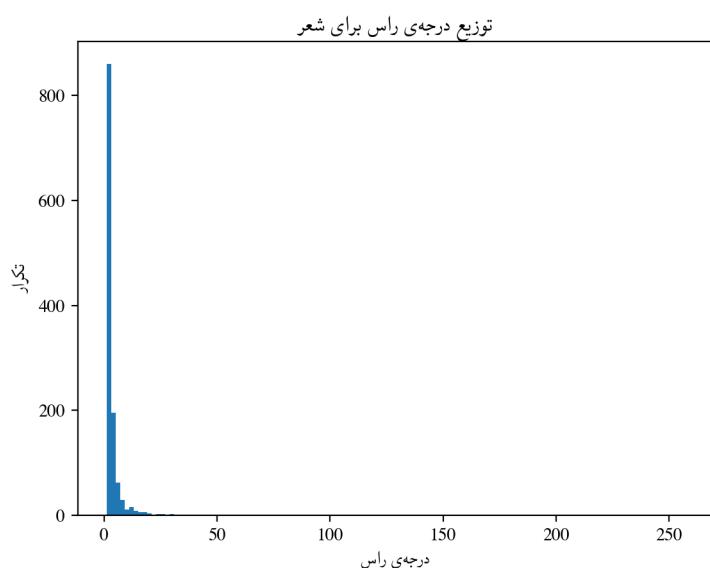
('بس', 21), ('شد', 20), ('هر', 20)

استخراج کلمات کلیدی پس از حذف افعال، حروف و ضمایر:

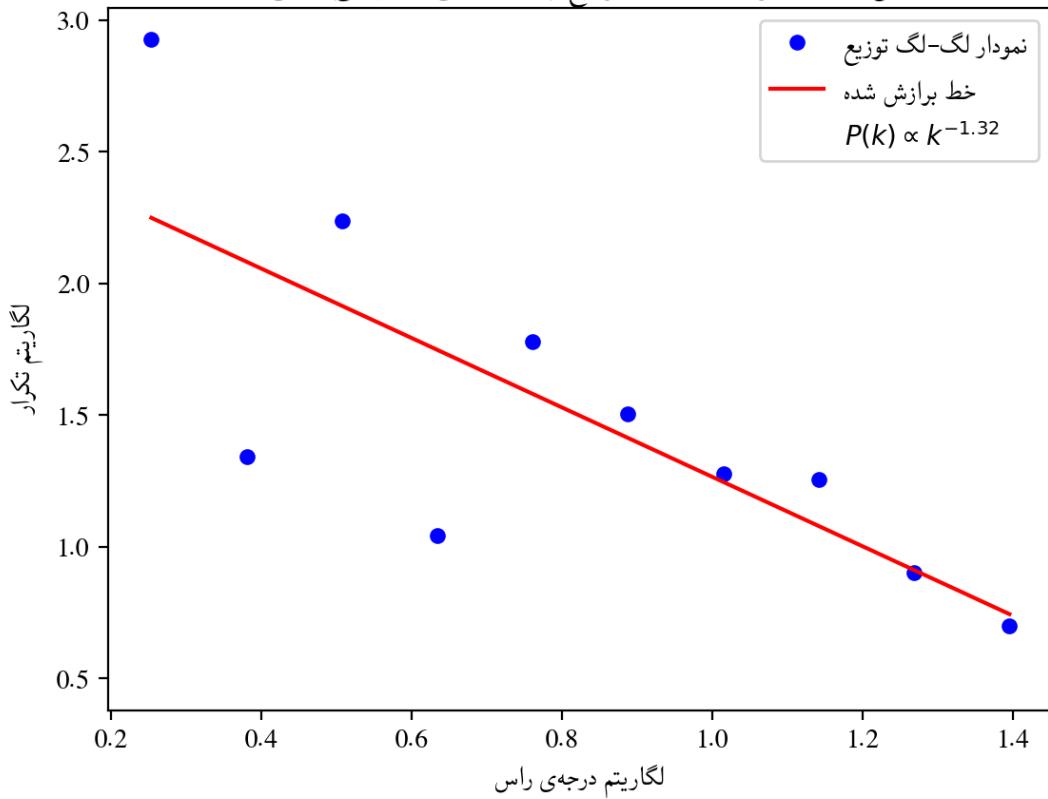
در این بین فقط شاهد ضمایر و افعال و حروف و غیره هستیم و خیلی کلمه‌ی کلیدی‌ای نمی‌توان استخراج

کرد. البته این که چند قصیده نیز هست در این بی تاثير نیست.

بررسی توزیع درجه‌های راس:

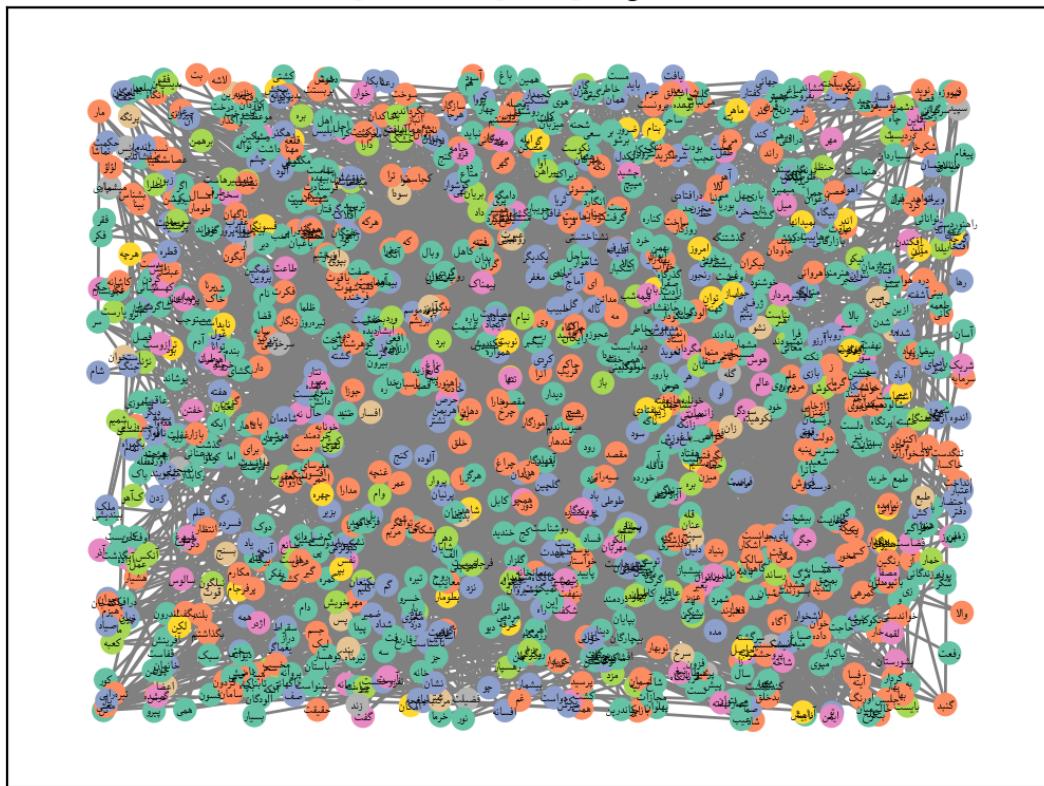


برازش خط به نمودار لگ-لگ توزیع (با دانه‌بندی لگاریتمی) برای شعر

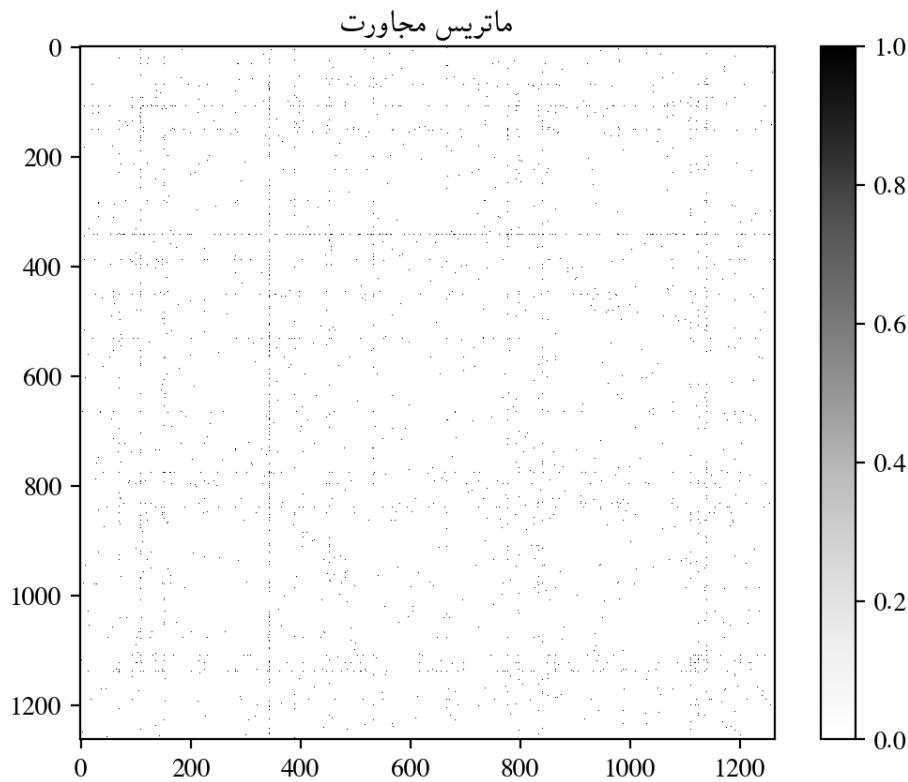


حال انجمن‌بندی را نیز انجام می‌دهیم:

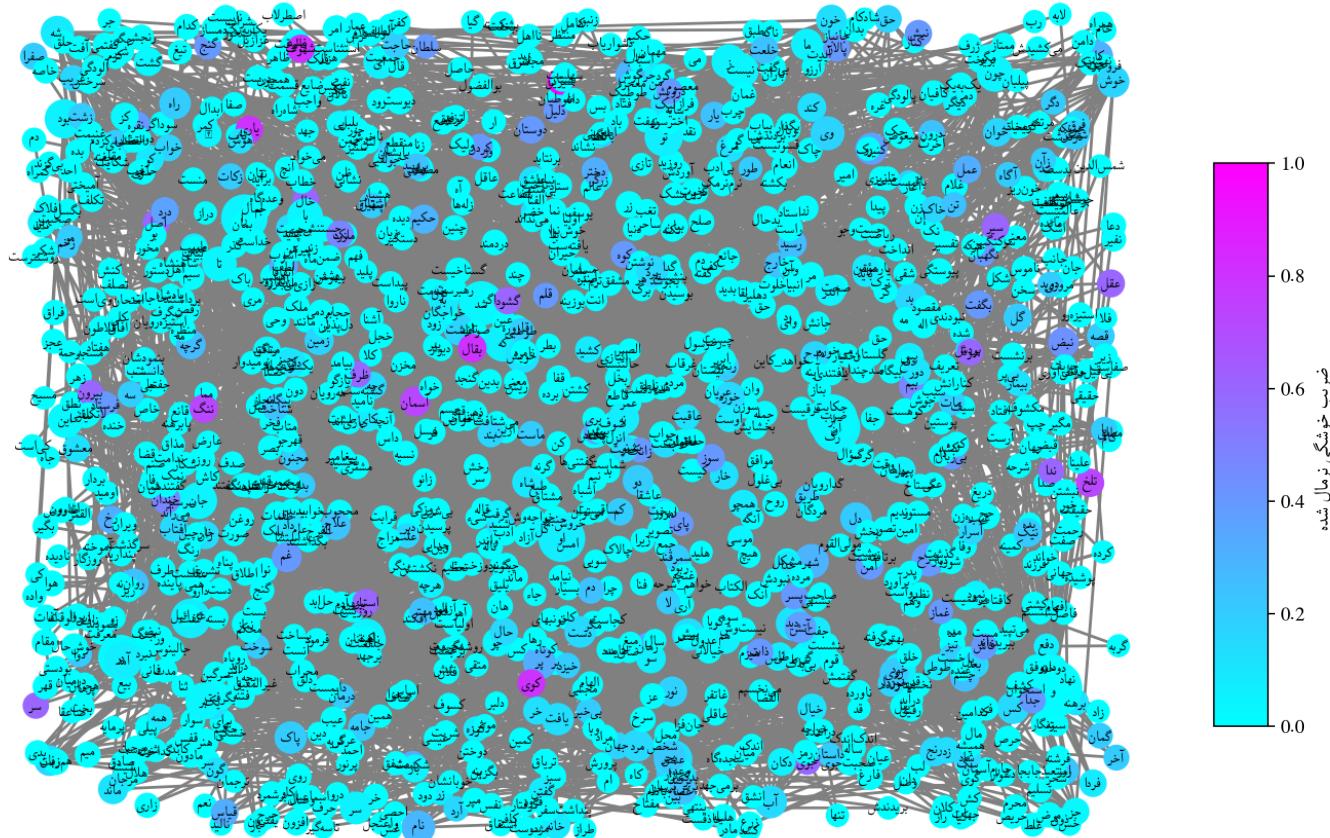
گراف انجمن‌بندی شده کلمات برای شعر



شعر سوم: بخشی از دفتر اول مثنوی معنوی
بررسی ماتریس مجاورت و شکل گراف:



گراف کلمات برای شعر



آمار تجمیعی برای این گراف:

تعداد راس‌ها: 1263

تعداد یال‌ها: 3219

میانگین درجه راس: 5.097

ضریب خوشگی میانگین: 0.1585

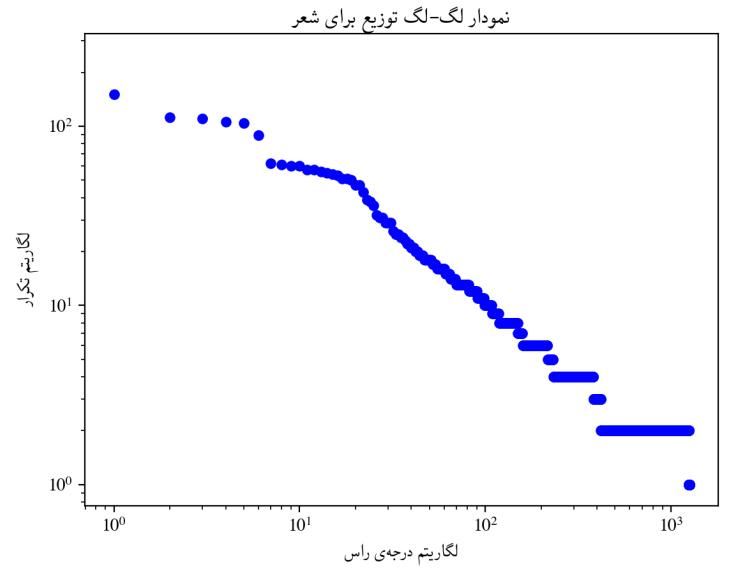
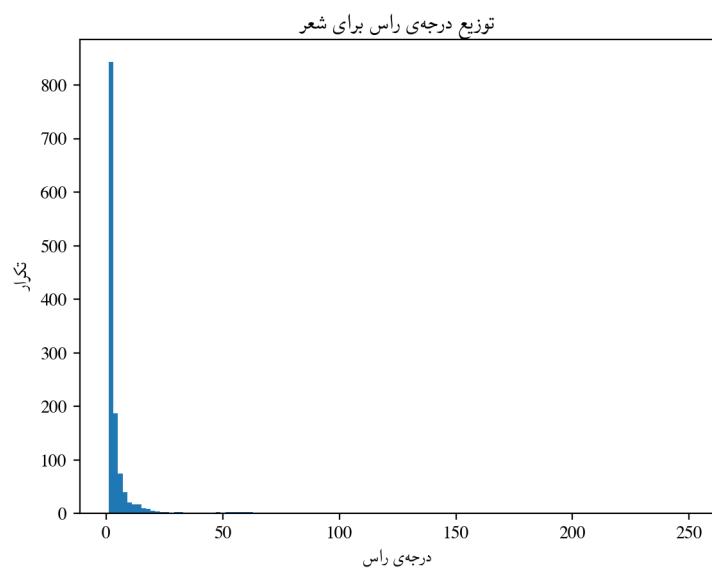
ضریب هم‌سنخ‌جويي: - 0.1874

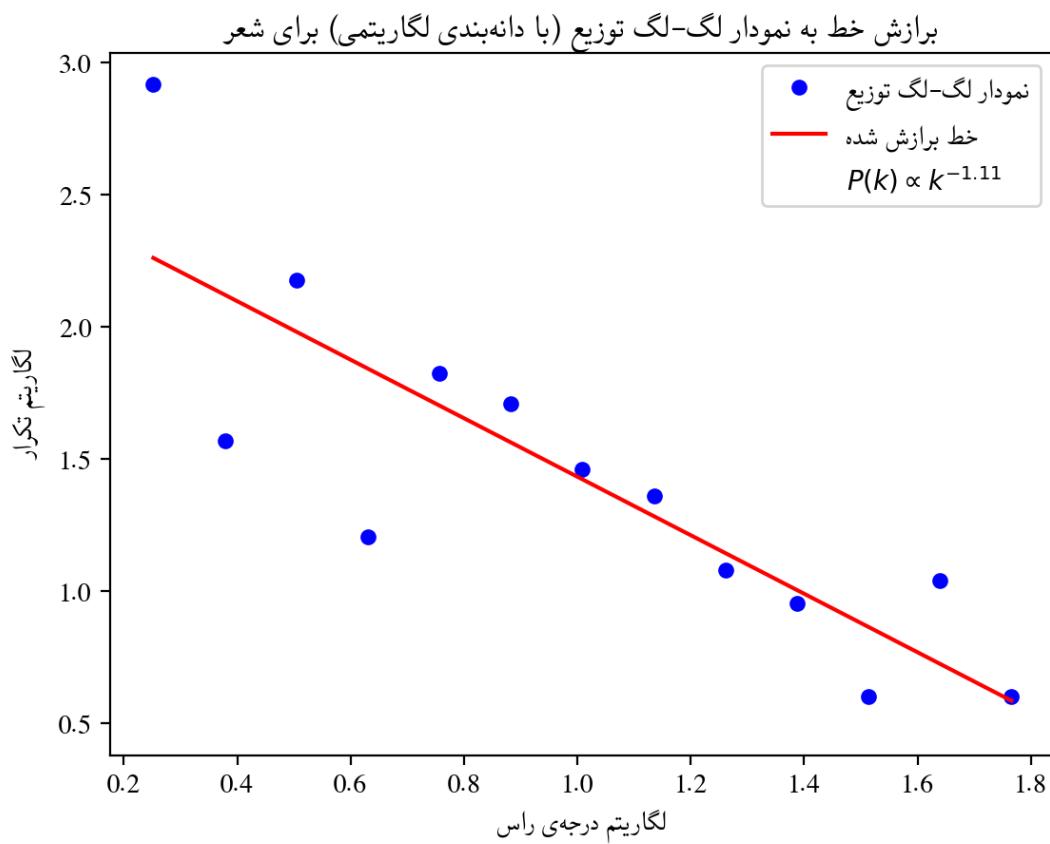
بیشترین درجه راس‌ها (درجات بالاتر از 40 را می‌آوریم):
 ('و', 249), ('از', 151), ('در', 112), ('آن', 110), ('را', 106), ('او', 89), ('شد', 104), ('که', 60), ('این', 60), ('گفت', 57), ('چون', 57), ('کرد', 56), ('تو', 55), ('نیست', 54), ('جان', 53), ('با', 43), ('هر', 51), ('تا', 50), ('من', 47), ('آمد', 47), ('ز', 51)

استخراج کلمات کلیدی پس از حذف افعال، حروف و ضمایر:

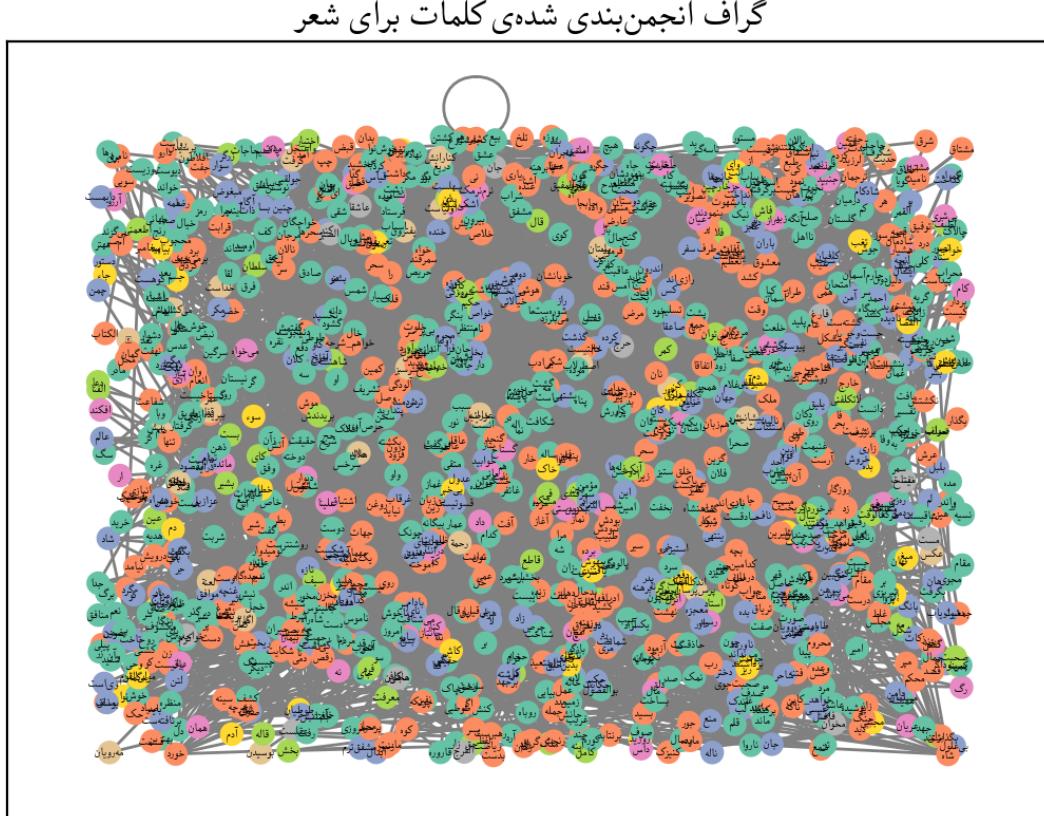
باز هم نمی‌توان خیلی کلمات کلیدی استخراج کرد! شاید چون سیر داستانی یکنواخت نداشته و تک موضوع نیست این اشعار.

بررسی توزیع درجه‌های راس:





حال انجمن بندی را نیز انجام می‌دهیم:



دیدیم که می‌توان با استفاده از روش‌ها و تکنیک‌های نظریه گراف و علم شبکه کارهای زیادی در زبان و تحلیل آن نجام داد. همچنین دیدیم که در متون منسجم‌تر و دارای خط فکری و داستانی یکنوا، می‌توان به راحتی کلمات کلیدی را استخراج و با آن‌ها خلاصه نویسی متن را انجام داد که این کار در اشعار کمتر است. می‌توان حتی با تحلیل کمی بیشتر، نوع اشعار (مثلاً قصیده، رباعی و غیره) را از هم تشخیص داد. حتی دیدیم که می‌توان شاعرها را نیز با این پارامترها مقداری از هم تمییز داد (البته با در نظر گرفتن نوع شعرشان).

تمام متن‌های استفاده شده و تحلیل شده به همراه کدهای این پروژه در [صفحه‌ی گیت‌هاب](#)³ موجود است.

مراجع:

- [1] ganjoor.net
- [2] <https://fa.wikipedia.org/wiki/رنسانس>

³ <https://github.com/Ali-Ekramian/Persian-Language-Processing-with-Network-Science-methods>