

فاز دوم پروژه موتور جستجوی گلجفلی

۹۴۳۱۰۳۵

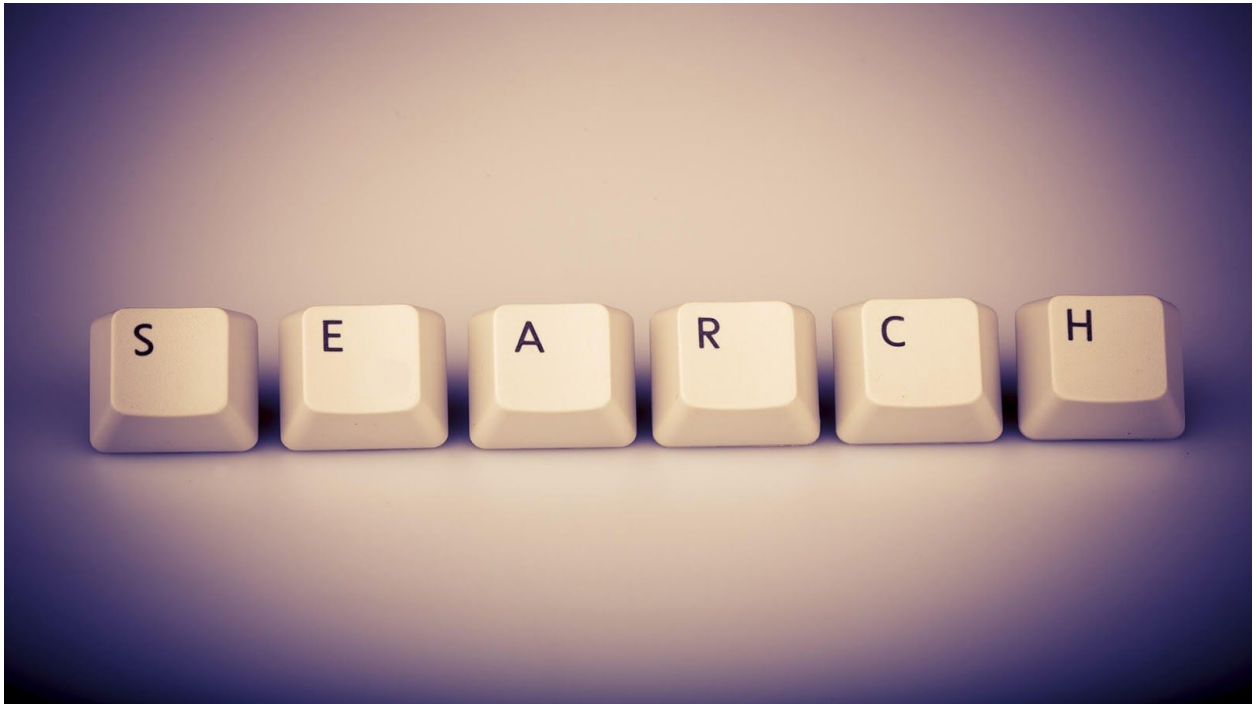
۹۴۳۱۰۳۶

۹۴۳۱۰۴۹

محمدرضا جبلی

علی جعفری

محمد مهدی گل محمدی



مقدمه

این پروژه در ۴ فاز تعریف شده است. در فاز اول به صورت صفر و یکی به نتایج سرچ‌ها نگاه میکردیم. در صورتی که در فاز دوم نتایج هر سرچ را رنک‌بندی میکنیم و ارزش هر داک را مشخص میکنیم.

توضیح مختصر

برای بخش نرمال سازی و همسان سازی و ... از کتابخانه ی آماده ی هضم استفاده کرده ایم. اما تمامی خطوط این کتابخانه را خوانده ایم و بلدیم. در فاز قبل با استفاده از داک های آماده یک دیکشنری ساخته میشد که هر سطر آن شامل یک postinglist بود که این postinglist مجموعه ای از مقاله ها بود و هر مقاله از مجموعه ای از position ها تشکیل میشد. پس زمانی که یک سرچ دریافت میکردیم با حرکت روی این دیکشنری کلمه ی مورد نظر را پیدا کرده و با استفاده از ۳ تابع not و neighbourhood و intersection عملیات مورد نظر کاربر را روی این پستینگلیست ها انجام داده و خروجی را به کاربر نشان میدادیم. (در قالب خاص خود)

در فاز دوم از این مرحله فراتر رفته و علاوه بر نگه داشتن تمام روند موجود در فاز قبل برای هر داک و همچنین کویری یک tfidf محاسبه میشود و سپس شباهت کویری با تمام داک هایی که خروجی فاز ۱ هستند گرفته میشود و هر چه شباهت بیشتر باشد در خروجی بالاتر به کاربر نمایش داده میشود. برای sort کردن نیز از یک max heap استفاده شده که زمان لازم برای اینکار را تا حد ممکن کوتاه تر کند.

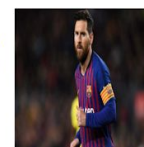
یک نمونه خروجی سیستم:

گل جفلی

لیونل

دو بازی مهمی که لیونل مسی از دست خواهد داد
khabaronline.ir

... خبرگزاری خبرآنلاین : لیونل مسی فوق ستاره ادعای رسانه های اسپانیایی لیونل مسی هر دوی ...
September 12th 2019, 00:10:00.000



افزایش چشمگیر قرارداد فن دایک با لیورپول
mashregnews.ir

... دایک در کنار لیونل مسی و کریستیانو ...
September 8th 2019, 14:55:00.000

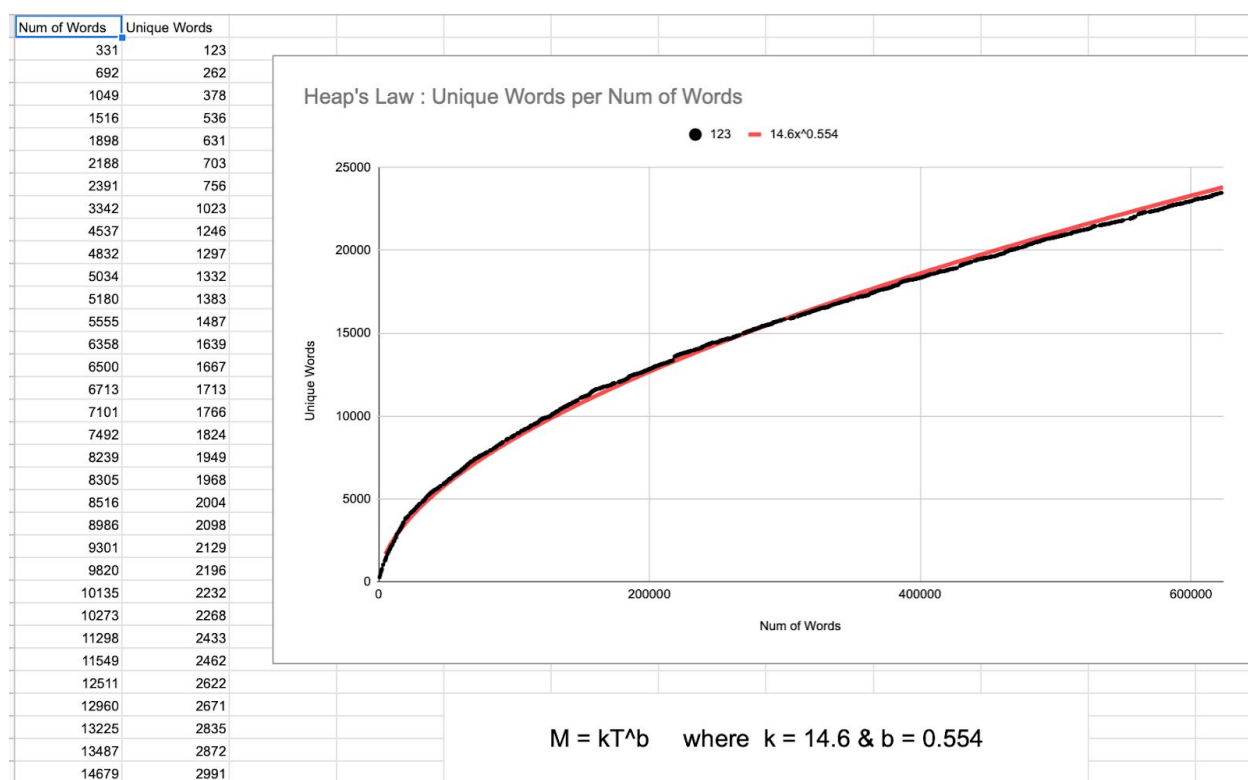


بخش امتیازی پروژه:

(۱)

بررسی Heap's Law:

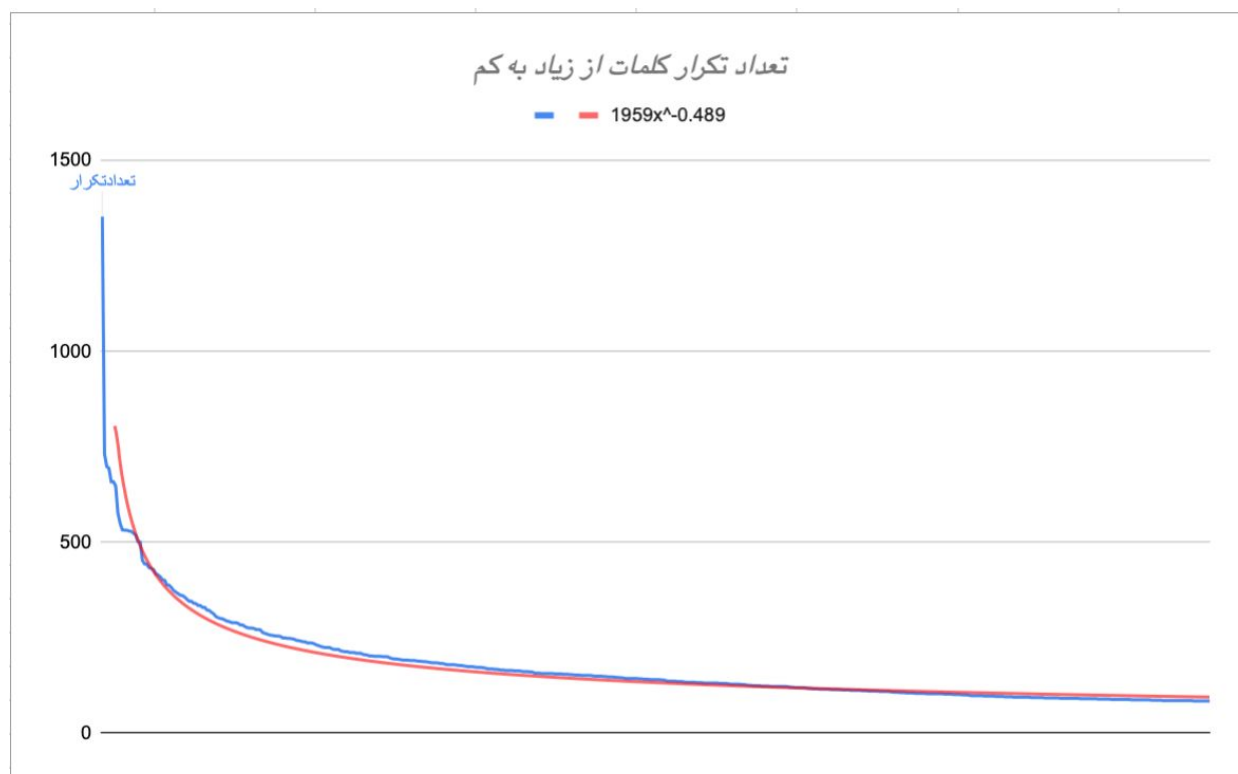
پس از اتمام حرکت بر روی هر doc یک نقطه‌ی (تعداد کلمات یونیک ، تعداد کل کلمات) تولید شده است و با استفاده از آن شکل زیر را داریم که کاملاً با یک خط که طبق فرمول heap است صدق میکند با $k=14.6$ و $b=0.554$



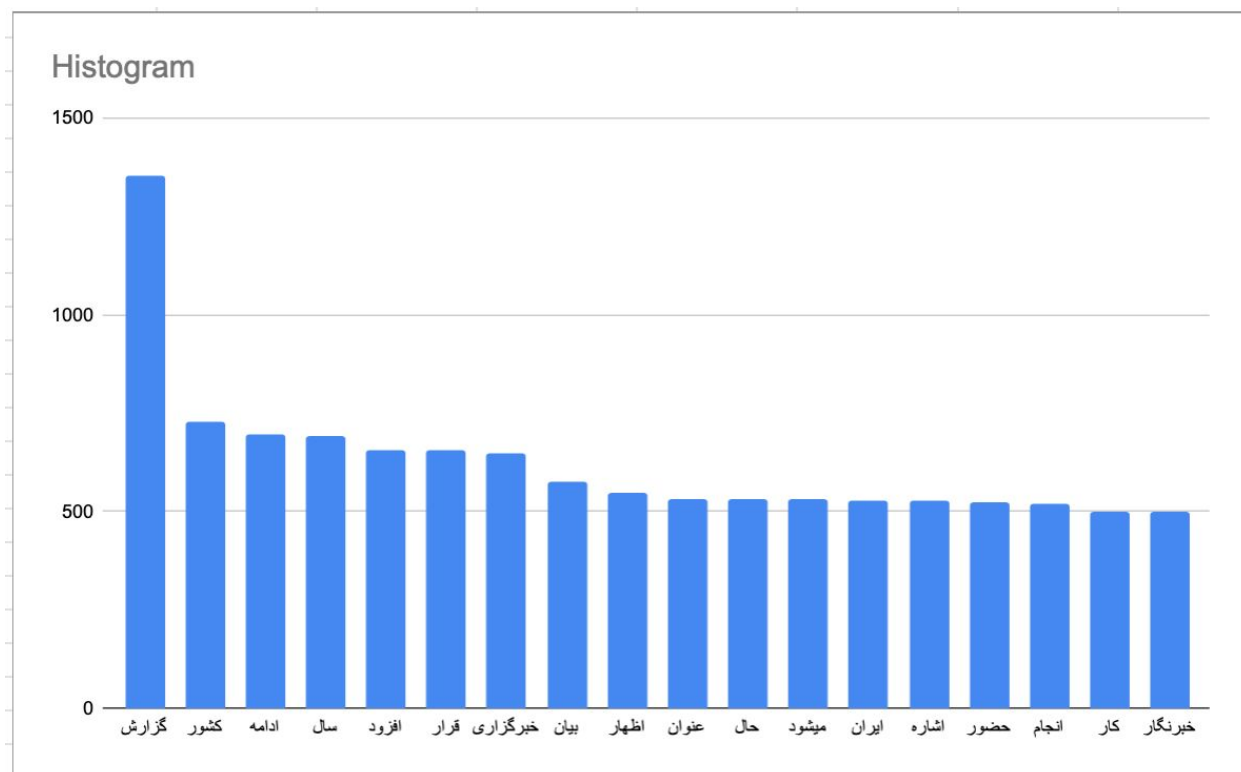
بررسی zipf's law:

بدلیل کم بودن تعداد مقالات موتور جستجو این قانون به طور کامل در موتور جستجویمان صدق نکرد. اما باز هم تا حدودی نحوه ی رابطه بین تعداد تکرار کلمات مختلف در مقالات را نشان داد.

در این شکل مشخص است که کلمات بیشتر نحوه توزیع به صورت $\frac{1}{x}$ دارند تا $\frac{1}{x^{0.489}}$



کلمات پرتکرار هم در تصویر زیر مشخص است:



وزن دهی های متفاوت: در کلاس Static یک متغیر تعریف کرده ام که مقادیر ۱ و ۲ و ۳ را میگیرد و وزن دهی موتور جستجو را تغییر میدهد. به طور کلی اگر بخواهم مقایسه ای بین این ۳ روش انجام دهم. به طور کلی در طول این پروژه متوجه شدم که ۲ نوع سبک نگاه کردن میتوان به شباهت کویری و داکيومنت داشت اولی این است که اگر تا جای ممکن داکيومنت کوتاه باشد اما کلمه ی ما را داشته باشد خیلی شبیه است. و مورد دوم اینکه خیلی به طول متن اهمیت ندهیم و فقط به تعداد حضور کلمات مورد نظرمان در خروجی اهمیت بدهیم.

روش وزن دهی اول واضحا تا حد زیادی تمایل به مورد دوم دارد. روش دوم تقریبا در وسط قرار میگیرد و روش سوم تمایل بیشتری به سبک اول دارد.

نتایج موتور جستجو توسط این ۳ روش مختلف برای یک کویری یکسان "استقلال" قابل مشاهده است که اطلاعاتی که در بالا دادم در آن کاملا قابل مشاهده خواهد بود:

الف) بهترین خروجی وزن دهی اول: ۱۶ استقلال - متن تقریبا بلند



مطمئنم استقلال با استراماچونی موفق می شود

علیرضا نیکبخت واحدی، بازیکن پیشین تیم استقلال اعتقاد دارد که این تیم با هدایت آندره استراماچونی موفق می شود.

mashregnews.ir

August 25th 2019, 16:02:00.000

به گزارش مشرق، علیرضا نیکبخت واحدی درباره شکست یک بر صفر استقلال مقابل ماشین سازی در هفته اول لیگ برتر اظهار داشت: بازی اول بود و زیاد نباید از تیمی که تعداد زیادی از بازیکنانش تغییر کرده اند، توقع داشت. باید به این تیم فرصت داد و نباید از الان نظر خاصی بدهیم. همیشه انتقادها از استقلال بوده است، اما به نظرم این تیم در دیدار مقابل ماشین سازی خوب بازی کرد، هر چند که توقع از استقلال بالاتر از این حرف هاست.

وی با تأکید بر اینکه هیچ تیمی به اندازه استقلال تغییرات نداشته است، افزود: استقلال در زمان وینفرد شفر به یکباره اوج گرفت. استقلال برای اینکه از ابتدای فصل بتواند اوج بگیرد، نباید به انتقادها توجه کند. اینکه بخواهیم از الان حاشیه درست کنیم، چیز جالبی نیست.

بیشتر بخوانید:

[استراماچونی نمی خواهد بماند](#)

بازیکن پیشین تیم فوتبال استقلال خاطرنشان کرد: بازیکنان فعلی استقلال متوجه هستند که در چه تیمی بازی می کنند. من همیشه با بهرامش دوست بوده ام و همیشه به او احترام می گذارم. مطمئنم این تیم با هدایت آندره استراماچونی موفق می شود. استقلال با

ب) بهترین خروجی وزن‌دهی دوم: ۵ استقلال - متن متوسط



هافبک استقلال، دربی را از دست داد؟

میزان نوشت: هافبک تیم فوتبال استقلال ممکن است دربی تهران را از دست بدهد.

khabaronline.ir

September 1st 2019, 15:52:00.000

فرشید باقری هافبک تیم فوتبال استقلال در جریان بازی این هفته تیمش مقابل فولاد خوزستان از ناحیه مچ پا دچار مصدومیت شد و از زمین بیرون رفت.

پس از بررسی وضعیت این بازیکن مشخص شد او باید عمل جراحی روی مچ پای خود انجام دهد. با این شرایط به نظر می‌رسد در صورتی که این هافبک عمل جراحی کند، دربی پایتخت بین تیم‌های استقلال و پرسپولیس را از دست خواهد داد.

دربی نیم‌فصل اول لیگ نوزدهم قرار است در هفته چهارم برگزار شود و باید دید در صورت غیبت این بازیکن برنامه‌های کادرفنی استقلال برای جایگزین او چه خواهد بود.

ج) بهترین خروجی وزن‌دهی سوم: ۴ استقلال - متن کوتاه



بودجه 7 میلیاردی در حساب باشگاه استقلال

با پیگیری هیات مدیره استقلال پول بازیکنان و طلبکاران سال‌های گذشته پرداخت خواهد شد.

ilna.ir

September 2nd 2019, 15:06:22.000



با پیگیری هیات مدیره استقلال پول بازیکنان و طلبکاران سال‌های گذشته پرداخت خواهد شد.

به گزارش خبرنگار ورزشی ایلنا طلب بازیکنان فصل گذشته این تیم تسویه خواهد شد.

با پیگیری‌های هیئت مدیره استقلال نزدیک به هفت میلیارد بودجه به حساب باشگاه واریزی شود تا طی امروز و فردا به حساب بازیکنانی که از فصل گذشته در تیم حضور دارند و همچنین بازیکنان فصل جدید واریز شود.

همچنین قرار است با تامین بودجه دیگر پول بازیکنان و طلبکاران سال‌های گذشته تیم نیز پرداخت شود.

اختتامیه

کد موجود به نحوی می‌باشد که به خوبی مفاهیم scalability رعایت شده و میتوان به راحتی فازهای آینده را به آن افزود.