باسمه تعالی

نام و نام خانوادگی: مهران غفاریان

شماره دانشجویی: ۹۹۳۱۰۴۲

1) در این بخش می‌خواهیم نحوه استخراج توکن‌ها و نرمال‌سازی آنها در پروژه شما را بررسی کنیم. لطفا برای اولین خبر از مجموعه داده پردازش شده، متن اصلی خبر، توکن‌های استخراج شده در مرحله اول، نتیجه نرمال‌سازی توکن‌ها، حذف کلمات پرتکرار و ریشه‌یابی را به شکل مناسب نشان دهید. به عنوان مثال می‌توانید ابتدا متن اصلی خبر را خروجی دهید. سپس هر یک از توکن‌های استخراج شده که به گونه‌ای از هم جدا شده باشند را نمایش دهید و سپس هر یک از توکن‌ها که در فرآیند نرمال‌سازی تغییری می‌کنند را به دو شکل قبل از تغییر و بعد از تغییر نمایش دهید. در ادامه کلماتی از متن که حذف شده ‌اند را مشخص کنید و نهایتا نتیجه ریشه‌یابی روی فعل‌ها را مشخص کنید.

متن اصلی خبر:



توکن های استخراج شده در مرحله اول:

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'سردار', 'آزمون', 'ستاره', 'تیم', 'ملی', 'کشورمان', 'و', 'عضو', 'باشگاه', 'زنیت', 'به', 'دلیل', 'درخشش', 'بی', 'نظیرش', 'در', 'لیگ', 'روسیه', 'و', 'نزدیک', 'بودن', 'زمان', 'پایان', 'قراردادش', 'با', 'باشگاه', 'روسی', 'مشتریان', 'زیادی', 'پیدا', 'کرده\_است', '.', 'سیماک', 'سرمربی', 'زنیت', 'روز', 'گذشته', 'در', 'مصاحبه', 'با', 'رسانه', 'های', 'روسیه', 'اعلام', 'کرد', 'مسئله', 'آینده', 'آزمون', 'آسان', 'است', 'این', 'بازیکن', 'قصد', 'تمدید', 'قراردادش', 'را', 'ندارد', 'و', 'در', 'تابستان', 'به', 'باشگاه', 'جدیدی', 'می', 'رود', '.', 'در', 'همین', 'رابطه', 'سایت', '«', 'hitc', '»', 'به', 'تحلیل', 'حرف', 'های', 'سرمربی', 'زنیت', 'پرداخت', 'و', 'این', 'حرف', 'ها', 'را', 'فرصتی', 'استثنایی', 'برای', 'دو', 'باشگاه', 'اورتون', 'و', 'نیوکاسل', 'دانست', '.', 'این', 'رسانه', 'انگلیسی', '،', 'نوشت', ':', 'سردار', 'آزمون', 'ستاره', 'ایرانی', 'زنیت', 'در', 'لیگ', 'قهرمانان', 'اروپا', 'در', 'یک', 'بازی', 'جذاب', 'به', 'کابوس', 'هواداران', 'چلسی', 'تبدیل', 'شد', '.', 'این', 'ستاره', 'ایرانی', 'یک', 'گل', 'فوق', 'العاده', 'در', 'این', 'بازی', 'به', 'ثمر', 'رساند', 'و', 'اگر', 'واکنش', 'های', 'فوق', 'العاده', 'کپه', 'آ', 'نبود', 'می', 'توانست', 'چندین', 'بار', 'دیگر', 'دروزاه', 'شاگردان', 'توخل', 'را', 'باز', 'کند', '.', 'ستاره', 'ایرانی', 'از', 'دو', 'سال', 'گذشته', 'بنا', 'به', 'اعلام', 'مدیر', 'برنامه', 'اش', 'از', 'باشگاه', 'اورتون', 'با', 'ریاست', 'فرهاد', 'مشیری', 'ایرانی', 'پیشنهاد', 'داشته', 'و', 'یکی', 'از', 'گزینه', 'های', 'تقویت', 'خط', 'حمله', 'این', 'تیم', 'است', '.', 'از', 'طرف', 'دیگر', 'نیوکاسل', 'که', 'در', 'آستانه', 'سقوط', 'قرار', 'دارد', 'برای', 'رهایی', 'از', 'این', 'وضعیت', 'بد', 'می', 'خواهد\_با', 'مدیریت', 'جدید', 'سعودی', 'ها', 'سردار', 'آزمون', 'را', 'جذب', 'کند', 'تا', 'خط', 'حمله', 'اش', 'جانی', 'دوباره', 'بگیرد', '.', 'قرارداد', 'سردار', 'آزمون', 'تابستان', 'پیش', 'رو', 'با', 'زنیت', 'به', 'پایان', 'می', 'رسد', 'و', 'این', 'بازیکن', '17', 'میلیون', 'پوندی', 'به', 'صورت', 'رایگان', 'به', 'تیمی', 'دیگر', 'خواهد\_رفت', '.', 'نیوکاسل', 'و', 'اورتون', 'فرصتی', 'طلایی', 'برای', 'جذب', 'این', 'بازیکن', 'به', 'صورت', 'رایگان', 'را', 'دارند', '.', 'البته', 'مهاجم', 'ملی', 'پوش', 'ایرانی', 'از', 'سوی', 'باشگاه', 'های', 'یوونتوس', '،', 'لیون', 'و', 'بایر', 'لورکوزن', 'نیز', 'مورد', 'توجه', 'است', 'و', 'باید', 'دید', 'این', 'بازیکن', 'ایرانی', '26', 'ساله', 'چه', 'تصمیمی', 'برای', 'آینده', 'اش', 'در', 'تابستان', 'می', 'گیرد', '.', 'آزمون', 'از', 'سال', '2019', 'برای', 'زنیت', 'بازی', 'می', 'کند', '.', 'ستاره', 'ایرانی', 'در', 'مدت', 'کمتر', 'از', '3', 'سال', '3', 'بار', 'قهرمانی', 'لیگ', 'برتر', 'روسیه', 'را', 'تجربه', 'کرد', 'و', 'یک', 'بار', 'نیز', 'جایزه', 'آقای', 'گلی', 'لیگ', 'را', 'از', 'آن', 'کرد', '.', 'انتهای', 'پیام', '/', '\xa0']

تغیرات حاصل از نرمال‌سازی توکن‌ها:

[('17', '۱۷'), ('26', '۲۶'), ('2019', '۲۰۱۹'), ('3', '۳'), ('3', '۳'), ('\xa0', '')]

تغیرات حاصل از ریشه‌یابی:

]('گزارش', 'گزار'), ('خبرگزاری', 'خبرگزار'), ('تیم', 'تی'), ('ملی', 'مل'), ('کشورمان', 'کشور'), ('زنیت', 'زنی'), ('درخشش', 'درخش'), ('بی', 'ب'), ('نظیرش', 'نظیر'), ('زمان', 'زم'), ('پایان', 'پای'), ('قراردادش', 'قرارداد'), ('روسی', 'روس'), ('مشتریان', 'مشتری'), ('زیادی', 'زیاد'), ('کرده\_است', 'کرده\_اس'), ('سرمربی', 'سرمرب'), ('زنیت', 'زنی'), ('های', 'ها'), ('اعلام', 'اعلا'), ('کرد', 'کرد#کن'), ('آسان', 'آس'), ('است', 'اس'), ('قراردادش', 'قرارداد'), ('ندارد', 'داشت#دار'), ('تابستان', 'تابست'), ('جدیدی', 'جدید'), ('می', '#هست'), ('سایت', 'سای'), ('های', 'ها'), ('سرمربی', 'سرمرب'), ('زنیت', 'زنی'), ('پرداخت', 'پرداخ'), ('ها', '#هست'), ('فرصتی', 'فرصت'), ('استثنایی', 'استثنا'), ('برای', 'برا'), ('دانست', 'دانس'), ('انگلیسی', 'انگلیس'), ('نوشت', 'نوش'), ('ایرانی', 'ایران'), ('زنیت', 'زنی'), ('قهرمانان', 'قهرمان'), ('بازی', 'باز'), ('هواداران', 'هوادار'), ('چلسی', 'چلس'), ('شد', 'شد#شو'), ('ایرانی', 'ایران'), ('بازی', 'باز'), ('رساند', 'رساند#رسان'), ('واکنش', 'واکن'), ('های', 'ها'), ('نبود', 'بود#است'), ('می', '#هست'), ('توانست', 'توانس'), ('شاگردان', 'شاگرد'), ('ایرانی', 'ایران'), ('اعلام', 'اعلا'), ('اش', 'ا'), ('ریاست', 'ریاس'), ('مشیری', 'مشیر'), ('ایرانی', 'ایران'), ('یکی', 'یک'), ('های', 'ها'), ('تقویت', 'تقوی'), ('تیم', 'تی'), ('است', 'اس'), ('دارد', 'داشت#دار'), ('برای', 'برا'), ('رهایی', 'رها'), ('وضعیت', 'وضعی'), ('می', '#هست'), ('مدیریت', 'مدیری'), ('سعودی', 'سعود'), ('ها', '#هست'), ('اش', 'ا'), ('جانی', 'جان'), ('بگیرد', 'گرفت#گیر'), ('تابستان', 'تابست'), ('پیش', 'پی'), ('زنیت', 'زنی'), ('پایان', 'پای'), ('می', '#هست'), ('رسد', 'رسید#رس'), ('پوندی', 'پوند'), ('صورت', 'صور'), ('رایگان', 'رایگ'), ('تیمی', 'تیم'), ('خواهد\_رفت', 'خواهد\_رف'), ('فرصتی', 'فرصت'), ('طلایی', 'طلا'), ('برای', 'برا'), ('صورت', 'صور'), ('رایگان', 'رایگ'), ('دارند', 'داشت#دار'), ('مهاجم', 'مهاج'), ('ملی', 'مل'), ('پوش', 'پو'), ('ایرانی', 'ایران'), ('سوی', 'سو'), ('های', 'ها'), ('است', 'اس'), ('ایرانی', 'ایران'), ('تصمیمی', 'تصمیم'), ('برای', 'برا'), ('اش', 'ا'), ('تابستان', 'تابست'), ('می', '#هست'), ('گیرد', 'گرفت#گیر'), ('برای', 'برا'), ('زنیت', 'زنی'), ('بازی', 'باز'), ('می', '#هست'), ('ایرانی', 'ایران'), ('مدت', 'مد'), ('کمتر', 'کم'), ('قهرمانی', 'قهرمان'), ('برتر', 'بر'), ('کرد', 'کرد#کن'), ('آقای', 'آقا'), ('گلی', 'گل'), ('کرد', 'کرد#کن'), ('انتهای', 'انتها'), ('پیام', 'پیا'), ('', '#هست')[

توکن‌های حذف شده در مرحله حذف کلمات پرتکرار:

['و', 'در', 'به', 'از', 'این', 'که', 'با', 'را', 'کرد#کن', 'شد#شو', 'اس', 'داشت#دار', 'برا', 'آن', 'تی', 'کشور', 'یک', 'بود#است', 'بر', 'باز', 'تا', 'باید', 'فارس', 'گزار', 'پیا', 'دو', 'خبرگزار', 'انتها', 'مل', 'بازیکن', 'سال', 'قرار']

2) برای شاخص مکانی تولید شده موارد زیر را مشخص کنید:  
اندازه کل شاخص (اندازه فایل شاخص به بایت بیان کنید)

سه کلمه با طولانی‌ترین Postings list را بیان کنید (طول لیست بر اساس تعداد اسناد حاوی کلمه مد نظر است و نه تعداد دفعات تکرار کلمه در اسناد).

سه کلمه با کوتاه‌ترین Postings list را به همراه شماره اسناد و شماره مکان‌ها برای یکی از کلمات بیان کنید.

پس از حذف ۵۰ کلمه پرتکرار سایز آن برابر می‌شود با 1310808 بایت.

[(2595, 'خبرنگار'), (1957, 'امروز'), (1846, 'اظهار')]

که در کنار هر کلمه تعداد داکیومنت‌های آنها نیز ذکر شده است.

کلمات با کوتاه‌ترین‌ posting list ها به همراه تعداد داک‌هایشان:

[(1, '#justice'), (1, "وحد"), (1, '\*\*\*برخ')]

توکن: وحد

آیدی داک: 12050

مکان‌ها: [334]

3) در محاسبه بردار tf-idf اسناد مطلوب است:

طول بردار tf-idf هر سند (تعداد کلمات فرهنگ لغات) چند است؟

کلمه با بیشترین و کلمه با کمترین مقادیر idf به همراه مقدار idf آنها را گزارش کنید.

برای اولین سند از مجموعه اسناد، کلمه با بیشترین و کلمه با کمترین وزن tf-idf را بیان کنید.

شبیه‌ترین سند از میان کلیه اسناد موجود در مجموعه اسناد به اولین سند بر اساس شباهت کسینوسی را پیدا کرده و میزان شباهت کسینوسی دو سند و محتوای سند مورد نظر را ارائه کنید. تحلیل خود از نتیجه را بیان کنید.

تعداد کلمات فرهنگ لغت: 30287

کلمه ⁩صادر بیشترین مقدار 8.517193191416238 را دارد.

کلمه خبرنگار کمترین مقدار 0.6558513958162484 را دارد.

برای داکیومنت اول داریم:

token with most idf is کپه with the idf 8.517193191416238

token with least idf is کند with the idf 1.0578542961959418

شبیه‌ترین سند دارای لینک <https://www.farsnews.ir/news/14001101000129/> می‌باشد. که شباهت 0.6691515699578285 دارند.

محتوای سند:

به گزارش خبرگزاری فارس، نیکولو شیرا خبرنگار سرشناس ایتالیایی درباره ماجرای پیشنهاد باشگاه میلان به سردار آزمون مهاجم ایرانی زنیت سن پترزبورگ پرده برداشت. این خبرنگار ایتالیایی اعلام کرد مدیران باشگاه میلان برنامه ای برای جذب سردار آزمون مهاجم ایرانی زنیت روسیه در تابستان پیش رو ندارند. روسونری چندین بازیکن برای تقویت خط حمله اش درنظر دارد  ولی نام ستاره ایرانی زنیت در لیست میلان وجود ندارد. قرارداد ستاره ایرانی با تیم روسی در تابستان پیش رو به پایان می رسد و می تواند به هر باشگاهی به صورت رایگان برود. آزمون در این فصل 15 بازی در لیگ برتر روسیه انجام داده و 7 گل برای زنیت به ثمر رسانده است. البته ستاره ایرانی بنا به گزارش رسانه های اروپا از نیوکاسل، لیون، یوونتوس و لورکوزن پیشنهاد دارد. آزمون هیچ کدام از پیشنهادات غیر از باشگاه لیون را تایید نکرده و اصلی ترین مشتری مهاجم ملی پوش کشورمان بنا به گفته خودش لیون است. انتهای پیام/

تحلیل:

هر دو این خبر‌ها مربوط به امکان جابجایی سردار آزمون از تیم روسی زنیت به میلان هستند که بدلیل شباهت فراوان در موضوع خبر، شخصیت اصلی خبر و تیم‌های خاص نام برده شده دارای امتیاز بسیار بالایی هستند.

4) با یک مثال (یک کوئری دلخواه) تاثیر نرمال‌سازی طول بردار اسناد در نتایج بازیابی شده را بررسی کنید. یک کوئری انتخاب کنید، نتایج حاصل از بازیابی را برای دو حالتی که بردار اسناد نرمال‌سازی شود و یا نرمال‌سازی نشود مقایسه کنید و تحلیل خود از نتیجه را بیان کنید.

کوئری: تاثیر نوسانات قیمت دلار در بازار کالاهای اساسی

در صورت عدم نرمال سازی داکیومنت‌های طولانی‌تر امتیاز و در نتیجه رنک بالاتری می‌گیرند زیرا شانس وقوع ترم‌های مشترک بیشتری با کوئری را دارند. در حالیکه لزوما مرتبط‌تر نیست اند:



در حالیکه در صورت نرمال سازی داکیومنت‌ها بصورت عادلانه‌تری مقایسه می‌شوند و در نتیجه داکیومنت‌های طولانی و کوچک در برابر میزان شباهت مفاهیم خود با کوئری مقایسه می‌شوند:

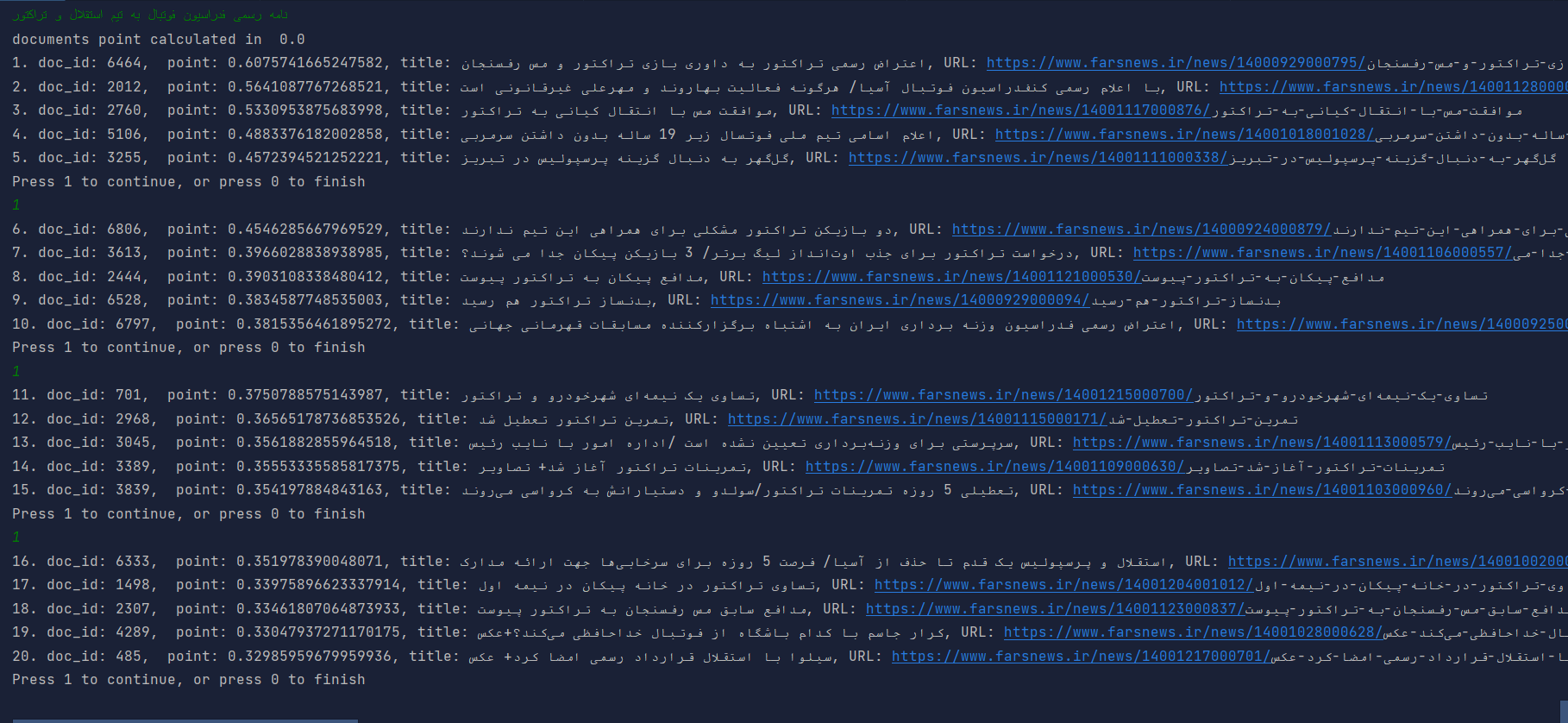


با برسی داکیومنت ها در دو حالت متوجه می‌شویم که نرمال سازی به افزایش شباهت داکیومنت‌های برگردانده شده کمک می‌کند.

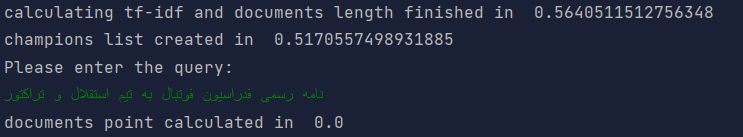
5) با یک مثال تاثیر استفاده از تکنیک لیست‌ قهرمانان بر سرعت و کیفیت نتایج جستجو را نشان دهید. برای یک کوئری دلخواه مقادیر زمان انجام فرآیند جستجو و نتایج حاصل از جستجو را برای دو حالت جستجوی عادی و جستجو با لیست قهرمانان گزارش کنید.

کوئری: نامه رسمی فدراسیون فوتبال به تیم استقلال و تراکتور

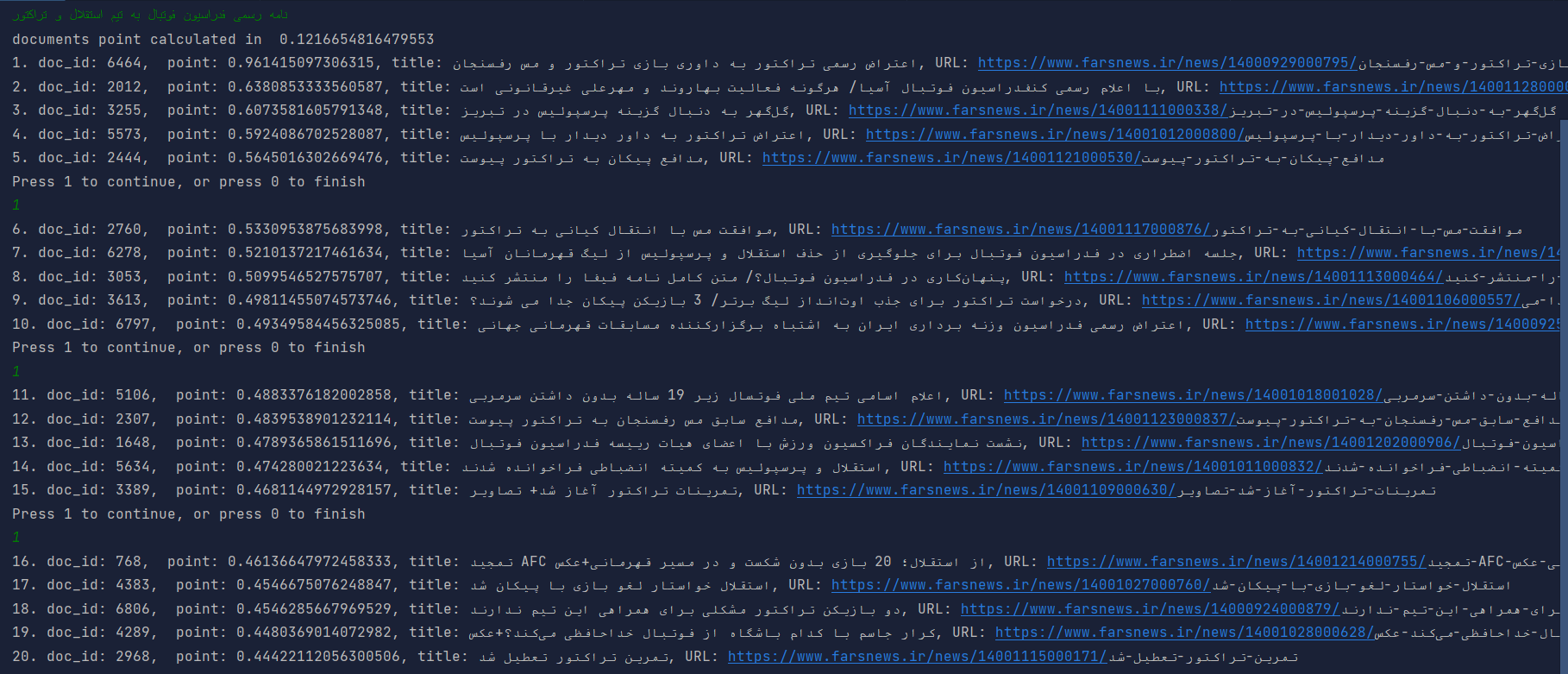
خروجی با استفاده از لیست قهرمانان



با صرف زمان نیم ثانیه‌ای برای ساخت لیست قهرمانان در هنگام پیش پردازش، می‌توان زمان محاسبه امتیاز داکیومنت‌ها را به نزدیک صفر رساند.



بدون استفاده از لیست قهرمانان:



در صورت عدم استفاده از لیست قهرمانان زمان محاسبه امتیاز داکیومنت‌ها بدلیل لزوم محاسبه و مقایسه تمامی داکیومنت‌ها زمان بیشتری می‌گیرد ولی بدلیل بررسی تمامی آنها در خروجی، داکیومنت‌ها امتیاز بالاتری دارند.

بطور کلی با صرف زمان کمی در پیش پردازش متون می‌توان زمان پاسخ گویی سیستم را به کوئری‌ها بسیار بالا برد هر چند امکان دارد کمی کیفیت شباهت‌ها کمتر باشد.