۱. در ابتدا فایل Json خوانده می‌شود.
۲. هر خبر شامل قسمت های مختلفی است که ما به همه بخش های آن نیازمند نیستیم. یک متغیر تعریف کرده و بخش های Content و Url و تیتر هر بخش را نگهداری می‌کنیم.
۳. متن هر خبر را Normalize می‌کنیم به این صورت که پیشوند و پسوند های کلمات را حذف می‌کند، سپس فاصله ها و نیم فاصله ها تصحیح می‌شوند. تبدیل های مربوط به نرمال سازی متن هایی که علائم نگارشی اینگلیسی دارند انجام شده و مانند فارسی می‌شوند. کسره، ضمه و فتحه از کلمات حذف می‌شود. کلماتی که به چند شکل نوشته می‌شود به یک فرم در می‌ایند برای مثال ریال و ریال یک چیز هستند و باید به یک فرم نوشته شوند.
۴. کلمات توکنایز می‌شوند به این شکل که اگر متن حاوی لینک بود با مقدار Link عوض می‌شود. اگر ایمیل بود به همین شکل با Email جا به جا می‌شود. فعل هایی که چند قسمتی اند بررسی می‌شود برای مثال "رفته بودم" و قسمت ها به یکدیگر می‌چسبند.
۵. کلمات توسط کتاب‌خانه Parsivar ریشه یابی می‌شوند. باید دقت داشت که کلماتی مانند "عنوان" ممکن است دچار تغییر شوند به این دلیل که ریشه یاب "ان" انتهای آن را حذف می‌کند.
۶. حال با استفاده از این کلمات یک دیکشنری حاوی Positional Index می‌سازیم.
۷. ۵۰ کلمه پرتکرار از این دیکشنری حذف می‌شوند به این دلیل که این کلمات حافظه زیادی را استفاده می‌کنند و در جست‌وجو کمکی به ما نمی‌کنند چون بیش‌تر شامل کلمات ربط و فعل هایی مانند "است" می‌باشند.
۸. حال به این Positional index که ساخته شده را در نظر می‌گیریم و به ازای هر کلمه در هر متن یک مقدار وزن tf-idf به این دیکشنری اضافه می‌کنیم.
۹. دیکشنری را برای استفاده آینده ذخیره می‌کنیم.
۱۰. مقدار وزن tf-idf را برای کوئری محاسبه می‌کنیم.
۱۱. حال یک تابع برای شباهت کسینوسی بین کوئری و داکیومنت ها می نویسیم. این تابع نرمال شده می‌باشد به این معنی که طول داکیومنت ها در پاسخ به کوئری در نظر گرفته شده اند.
۱۲. یک Champion List برای کلمات می‌سازیم به این شکل که به ازای هر کلمه k سندی که بیش‌ترین tf-idf را دارند در نظر می‌گیریم و ابتدا این لیست را بررسی می‌کنیم و اگر پاسخی نگرفتیم به سراغ بقیه کلمات می‌رویم. این کار باعث می‌شود که سرعت پردازی ما به طور چشمگیری افزایش پیدا کند.
۱۳. کوئری را از کاربر گرفته و پاسخ را بر می‌گردانیم.

۲.

الف) کوئری "فوتبال" جست‌وجو شد:

این خبر مربوط به کوئری می‌باشد و به دلیل عام بودن کوئری و سه بار ظاهر شدن این کلمه در متن با وجود کوتاه بودن متن این متن انتخاب شده است.

رونمایی از انتصاب‌های جدید در فدراسیون فوتبال

جلسه هیأت رئیسه فدراسیون فوتبال امروز با حضور اعضا برگزار شد.

در پایان این جلسه بیژن ذوالفقارنسب، پیشکسوت فوتبال ایران به عنوان رئیس کمیته فنی و توسعه فدراسیون انتخاب شد.

همچنین عباس صوفی، رئیس هیأت فوتبال همدان هم به عنوان مسئول کمیته موقت عمران و زیر ساخت فدراسیون منصوب شد.

ب) کوئری "اخبار سیاسی" جست‌وجو شد:

این خبر تا حد کمی به کوئری ارتباط دارد به این دلیل که طول متن خبر کم می‌باشد و جمله ای که در آن اطلاع می‌دهد که این خبر به گزارش گروه سیاسی تهیه شده است باعث شده است تا این خبر مرتبط با کوئری در نظر گرفته بشود.

ملک سازمانی وزیر کشور سابق در اختیار وزارت کشور است

به گزارش گروه سیاسی خبرگزاری فارس، پیرو انتشار مطالبی درباره ملک سازمانی وزیر کشور سابق، روابط عمومی این وزارتخانه اعلام کرد: به اطلاع می رساند که ملک مذکور در اختیار وزارت کشور قرار گرفته است.

در این خبر تاکید شده است: این ملک سازمانی به وزارت کشور تحویل داده شده است.

ج) کوئری "کپه" جست‌وجو شد:

در این خبر به دلیل این که کوئری ای که جست‌وجو شده خاص می‌باشد کاملا مربوط به کوئری جست‌وجو شده می‌باشد.

آزمون کابوس هواداران چلسی/رقابت مدیریت سعودی با ایرانی برای شکار مهاجم ۱۷ میلیون پوندی+عکس

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.

د) کوئری "درخت کریسمس" جست‌وجو شده است.

ستاره اسپانیایی؛ هدیه کریسمس گواردیولا به ژاوی+عکس

در این کوئری با وجود این که کلمه "درخت" موجود نمی‌باشد به دلیل بالا بودن idf کلمه کریسمس و کوتاه بودن متن. این متن به عنوان پاسخ به ما برگشته است با وجود این که تنها بخش کریسمس آن مرتبط است.

انتقال این بازیکن ۲۱ ساله اسپانیایی مورد توجه بلیچر ریپورت قرار گرفته و طرحی جالب در این باره را منتشر کرده است. در این طرح، تورس به هدیه کریسمس پپ گواردیولا به ژاوی تشبیه شده است.