بنام خدا

مهران غفاریان ۹۹۳۱۰۴۲

پرسمان اول: خبرنگار

تیتر داکیومنت اول: حفظ کرامت خبرنگاران از سوی دولت‌مردان ضروری است

جملات حاوی کلمه:

به گزارش حوزه دولت خبرگزاری فارس، علی بهادری جهرمی سخنگوی دولت به رفتار سرپرست فرمانداری رفسنجان با یک خبرنگار واکنش نشان داد و در صفحه شخصی خود در توییتر نوشت:

دوران بدزبانی با خبرنگاران گذشته است؛ دولت مردمی، پاسخگو و حفظ کرامت خبرنگاران از سوی دولت‌مردان ضروری است.»

این داکیومنت با توجه به اینکه تعداد رویدادهای کلمه مورد نظر زیاد بوده و از جهتی طول داکیومنت نیز کوتاه بود دارای امتیاز بالایی شده. همچنین به وضوح داکیومنت به کلمه سرچ شده بسیار نزدیک می‌باشد.

پرسمان دوم: لیگ برتر فوتبال ایران

تیتر داکیومنت اول: جابری: نامه ای درباره انصراف شهروند از حضور در لیگ برتر فوتسال ارسال نشده است

جملات حاوی کلمات:

آرش جابری در گفت وگو با خبرنگار ورزشی فارس در مورد انصراف تیم فوتسال شهروند ساری از حضور در نیم فصل دوم لیگ برتر گفت: تا این لحظه هنوز نامه ای به دست سازمان لیگ مبنی بر انصراف این تیم از لیگ برتر نرسیده است؛ اما تیم امید شهروند بعد از اینکه مدارک خود را برای حضور در لیگ امید ارسال نکرد با پیگیری های سازمان لیگ این تیم رسما از حضور در لیگ امیدهای کشور انصراف داده است.

رئیس و دبیر سازمان لیگ در ادامه گفت: دو روز است که لیگ امیدهای کشور با حضور ۱۱ تیم در دو گروه آغاز شده و تا ۲۸ دی ادامه دارد و در حال حاضر هم در رشت حضور دارم.

در این داکیومنت با توجه به تعدد کلمه لیگ و سایر کلمات و همچنین طول کمتر داکیومن تطابق بیشتری با پرسمان اصلی دارد و همچنین مرتبط تر می‌باشد.

پرسمان: فدراسیون

تیتر داکیومنت اول: واریزی جدید به فدراسیون‌ها/پرداختی کمیته ملی المپیک صددرصد شد

جملات حاوی کلمه:

به گزارش خبرنگار ورزشی خبرگزاری فارس، کمیته ملی المپیک باقیمانده اعتبارات خود را به حساب فدراسیون‌های ورزشی امروز واریز کرد تا ایت فدراسیون‌ها در سال ۱۴۰۰ صددرصد بودجه خود را دریافت کنند.

کیکاووس سعیدی دبیر کل کمیته ملی المپیک با اعلام این خبر به فارس، گفت: امروز مبلغ ۴۲ میلیارد ریال به حساب فدراسیون‌ها واریز می‌شود، با این حساب صددرصد بودجه فدراسیون‌ها از سوی کمیته کامل می‌شود.

وی افزود: مازاد بر سهمیه بودجه، کمیته ۱۰ تا ۱۵ میلیارد تومان هم به فدراسیون‌هایی که المپیک را پیش رو داشتند کمک کرد هر چند در این بین برخی فدراسیون‌های دیگر هم از مازاد اعتبارات بهره بردند.

کمیته ملی المپیک در طول ۴ سال گذشته هر سال ۱۰۰ درصد بودجه فدراسیون‌ها را پرداخت کرده است. این احتمال وجود دارد تا پایان سال در صورت داشتن اعتبار، باز هم به فدراسیون‌ها کمک شود.

تعدد کلمه کمیاب فدراسیون باعث شده این داکیومنت بیشترین امتیاز را بگیرد که ارتباط بیشتری با پرسمان ما نیز دارد.

پرسمان: انتقال سردار آزمون از زنیت به لورکوزن

تیتر داکیومنت اول: سردار آزمون به بایرلوکوزن آلمان پیوست+عکس

جملات حاوی کلمات:

به گزارش خبرگزاری فارس، سردار آزمون با باشگاه لورکوزن به توافق رسید تا در آغاز فصل ۲۰۲۳-۲۰۲۲ به این تیم ملحق شود. قرارداد آزمون با زنیت در تابستان امسال به پایان می رسد و این بازیکن به صورت بازیکن آزاد راهی لورکوزن خواهد شد.

قرارداد آزمون با لورکوزن ۵ ساله و تا پایان فصل ۲۰۲۷ خواهد بود.

با توجه به تعدد کلمات خاص پرسمان در این داکیومنت، داکیومنت انتخابی ارتباط و امتیاز بیشتری با پرسمان دارد.

تشریح عملکرد موتور جست و جو

در ابتدا تمامی داکیومن‌ها بررسی می‌شوند. در هر مرحله ابتدا داکیومنت نرمالایز می‌شود. در ادامه متن داکیومنت توکنایز می‌شود. در ادامه هر توکن بررسی می‌شود و پس از ریشه‌یابی و حذف کلمات غیر کاربردی نظیر علامات نگارشی و عبارات اضافه شده در انتهای کلمات نظیر علامت‌های جمع حذف می‌شوند. در نهایت توکن‌های داکیومنت به همراه اندیس مکانی آنها برگردانده می‌شوند.

بعد از تعیین توکن‌های هر داکیومنت شاخص مکانی کل مجموعه را بروز می‌کنیم. بصورتیکه برای کل شاخص مجموعه یک دیکشنری در نظر می‌گیریم که هر ترم یک کلید می‌باشد به شاخص مکانی خود که شامل فرکانس آن ترم در کل مجموعه، آی‌دی‌اف ترم، و یک دیکشنری که شاخص‌های مکانی آن ترم در آن قرار دارند. در این دیکشنری هر داک آیدی کلیدی هست به شاخص مکانی آن ترم در آن داکیومنت که شامل فرکانس آن کلمه در آن داکیومنت و لیستی از مکان‌های رویداد آن کلمه در آن داکیومنت. برای هر داکیومنت این شاخص بروز می‌شود تا در نهایت شاخص مجموعه بصورت کامل مشخص شود.

در ادامه لیستی از هر توکن و فرکانس ان در کل مجموعه ساخته می‌شود که در ادامه ۵۰ کلمه پرتکرار در بین آنها از شاخص کل مجموعه حذف می‌شوند.

قدم بعدی محاسبه tf-idf می‌باشد. در این مرحله با پیمایش بر روی تمامی ترم‌های مجموعه، دیکشنری tf-idf بروز می‌شود. در این دیکشنری هر ترم کلیدی هست به دیکشنری دیگر که در آن برای آن ترم، هر داک آیدی کلیدی می‌باشد به امتیاز tf-idf آن ترم در آن داکیومنت. در همین حین دیکشنری مربوط به طول داکیومنت ها نیز بروز می‌شوند که در آن هر داک آیدی کیلیدی می‌باشد به مجموع مربعات فرکانس‌های ترم‌های آن داکیومنت.

در مرحله بعد لیست قهرمانان برای هر ترم ساخته می‌شود که در آن برای هر ترم در مجموعه لیستی از داکیومنت‌ها به همراه tf-idf آن ترم در ان داکیومنت مرتب می‌شوند تا در ادامه پس از مرتب سازی با پیمایش بر روی این لیست تعداد مطلوبی از این داکیومنت‌ها انتخاب شوند که یا امتیاز مطلوبی دارند و یا اینکه تعداد داکیومنت‌های لیست قهرمانان آن ترم تعداد حداقلی داشته باشند. در این دیکشنری هر ترم کلیدی هست به دیکشنری دیگر که در آن هر داک آیدی کلیدی می‌باشد به tf-idf آن کلمه در آن داکیومنت.

در ادامه پرسمان کاربر دریافت می‌شود. پس از توکنایز کردن آن و محاسبه فرکانس کلمات در آن به روش مشابه گذشته، می‌توان امتیاز هر داکیومنت را محاسبه کرد.

در این مرحله برای هر توکن پرسمان به امتیاز داکیومنت‌های آن ترم افزوده می‌شود. که مقدار امتیاز افزوده شده برابر ضرب tf-idf آن ترم در آن داکیومنت در tf آن ترم در پرسمان می‌باشد.

در نهایت امتیاز داکیومنت‌ها بر طول آنها تقسیم می‌شود تا نرمال سازی انجام شود. سپس این امتیازها به شیوه مشابه قبل مرتب می‌شوند تا به تترتیب امتیاز آنها در اختیار کاربر قرار بگیرند.

در انتها در صورت تمایل کاربر، شاخص پیاده سازی شده ذخیره می‌شود تا در کاربردهای بعدی موتور جست‌وجو، یا از شاخص ذخیره شده استفاده شود و یا اینکه دوباره با پردازش داکیومنت ها شاخص جدید ساخته شود.