گزارش پروژه بازیابی اطلاعات

پارسا غفرانی ۴۰۱۳۱۰۲۸

فاز ۱:

نرمال سازی متون:

نام فایل: tokenization

فاصله گذاری صحیح:

در تابع transform_space_to_half_space، به دنبال فرمت‌های مختلف گفته شده در تعریف پروژه میگردیم و اگر با فاصله از حروف بعد از خود باشند، آن را با نیم فاصله جایگذین میکنیم. برای این کار از regex و پیدا کردن پترن‌ها استفاده میکنیم.

تعویض یونیکد:

در تابع change_unicode، زیر شاخه‌های ۲و۳و۴ در تعریف پروژه را برای نرمال سازی، انجام میدهیم. برای این کار هم چیزی شبیه به بالا انجام میدهیم، با این تفاوت که مجموعه‌ای از pattern ها را با یک حرف یا یک عبارت جایگزین میکنیم. در این تابع همچنین اعداد انگلیسی را با فارسی جایگذین میکنیم.

در تابع separate_numbers، بین عدد تا کلمه‌ی بعد از آن فاصله میگذاریم تا هنگام tokenize کردن به مشکلی بر نخوریم.

ریشه‌یابی:

برای این کار در تابع stemming با استفاده از کتابخانه‌ی parsivar و تعریف یک stemmer، ریشه‌ی هر توکن را پیدا میکنیم. تابع FindStem در این کتابخانه، به صورت مبتنی بر قانون و حذف پیشوندها و پسوندها، تلاش میکند تا ریشه‌ی کلمات را پیدا کند.

حذف علامت‌های نگارشی:

در تابع delete_punctuation با استفاده از regex، هرجا یکی از علامت‌ای نگارشی دیده شود، آن را با فاصله جایگذین میکنیم.

در تابع normalizer تمام توابع بالا را یکدور صدا میزنیم تا به صورت یک pipeline عمل کند و متن را نرمال سازی بکند.

توکنایز کردن:

در تابع tokenizer ابتدا متن را به صورت نرمال در می‌آوریم، سپس تمام حالت‌های فاصله را با یک فاصله جایگذین میکنیم و بر اساس space، split میکنیم.

همچنین در تابع count_word_frequencies، تعداد وقوع کلمات یک متن به خصوص را میشماریم.

ساخت شاخص معکوس

فایل postingslist:

در این فایل یک کلاس با همین نام تعریف شده که موجودیت postingslist را مدل سازی میکند.

تابع add:

با وقوع یک کلمه در یک سند، آن را به postingslist مربوطه اضافه میکند و تکرار آن را نیز افزایش میدهد.

تابع tf:

تکرار یک کلمه را در یک سند محاسبه میکند و ذخیره میکند.

تابع idf:

Idf مربوط به یک کلمه را محاسبه میکند.

تابع get_list_word:

لیستی شامل داکیومنت‌هایی که کلوه را دارند بر میگرداند (شبیه به non-positional-inverted-index)

تابع get_freq:

تعداد سندهای یکتایی که شامل کلمه هستند را برمیکرداند.

در تابع create_champion_list: با استفاده از یک heap، چمپیون لیست مربوط به یک کلمه را با استفاده از اسناد با وزن بالاتر میسازیم.

فایل postingslists:

در این فایل و کلاسی با همین نام، به مدیریت تمام postingslist های مربوط به تمام کلمات میپردازیم.

تابع set_valid_documents: در این تابع شماره سندهایی که به بررسی آن‌ها باید پرداخت مشخص میشود.

تابع add_term_to_index: در این تابع، یک کلمه را به شاخص خود اضافه میکنیم. (اگر وجود نداشت باید آن را initialize کنیم.)

تابع calculate_term_relevance: در این تابع، امتیاز tf-idf مربوط به یک کلمه را محاسبه میکنیم.

تابع calculate_term_frequency: در این تابع، میزان تکرار یک کلمه در یک سند را ذخیره میکنیم.

تابع calculate_inverse_document_frequency: در این تابع، idf مربوط به یک کلمه را ذخیره میکنیم.

تابع compute_document_weights: در این تابع ابتدا امتیاز tf-idf را با توابع قبلی و فایل قبلی محاسبه میکنیم. سپس با محاسبه‌ی طول آن، آن را نرمال سازی میکنیم.

تابع get_documents_for_term: لیستی از سندهایی که شامل یک کلمه هستند را دریافت میکنیم.

تابع get_document_weight: وزن مربوط به هر سند را میگیریم.

تابع generate_champion_lists: لیست چمپیون مربوط به کلمات را تولید میکند.

تابع get_champion_list_for_term: با دادن یک کلمه، لیست چمپیون نظیر را برمیگرداند.

تابع create_word_frequency_list: لیستی از کلمات، به همراه تکرارشان در کل سندها میسازد.

تابع remove_most_frequent_terms: ۵۰ کلمه‌ی پر تکرار را حذف میکند و برمیگرداند.

فایل run_file:

این فایل همان فایلی است که باید اجرا شد تا موتور جست و جوی ما کارش را آغاز کند.

تابع read_data: فایل داده‌ی داده شده را میخواند و در متغیر data ذخیره میکند.

تابع gather_valid_indices: شماره سندهایی که باید بازیابی شوند را پیدا میکند.

تابع preprocess: این تابع، هر سند را نرمال سازی و توکنایز میکند، مقادیر مروبطه را به شاخص ما اضافه میکنند و شاخص را ذخیره میکند.

تابع search_query: این تابع، با ورودی گرفتن یک کوئری از کاربر، آن را پردازش میکند و سندهای مرتبط به آن را بر اساس میزان ارتباط، رتبه بندی میکند.

۵۰ کلمه پر تکرار:

stop_words :

[('به', ۱۲۱۸۰), ('فارس', ۱۲۱۳۳), ('پیام', ۱۲۱۱۹), ('در', ۱۲۰۶۲), ('انتهای', ۱۲۰۵۹), ('خبرگزاری', ۱۱۹۹۹), ('و', ۱۱۹۹۳), ('از', ۱۱۴۳۸), ('این', ۱۱۳۶۱), ('با', ۱۱۱۴۷), ('گزارش', ۱۰۴۷۸), ('را', ۱۰۲۱۰), ('که', ۹۶۵۸), ('اس', ۹۱۵۵), ('کرد', ۸۳۹۳), ('برای', ۸۳۷۲), ('داشت&دار', ۷۹۱۶), ('شد', ۷۸۹۵), ('کرد&کن', ۷۸۸۴), ('شد&شو', ۷۶۴۱), ('بود&باش', ۶۵۲۶), ('ان', ۶۴۵۰), ('تا', ۶۲۸۹), ('خبرنگار', ۶۲۸۰), ('وی', ۶۲۱۸), ('یک', ۶۱۶۴), ('بر', ۶۱۴۵), ('خود', ۵۹۲۴), ('تیم', ۵۸۱۴), ('کشور', ۵۷۴۱), ('هم', ۵۵۸۹), ('گفت', ۵۵۸۲), ('گرفت&گیر', ۵۳۹۹), ('داد&ده', ۵۲۰۹), ('قرار', ۵۰۸۲), ('خواست&خواه', ۴۹۶۶), ('امروز', ۴۸۶۴), ('داشت&دارد', ۴۸۳۰), ('باید', ۴۸۱۰), ('داد', ۴۶۷۵), ('بازی', ۴۵۱۰), ('توانست&توان', ۴۴۷۶), ('اظهار', ۴۴۳۰), ('اما', ۴۴۲۳), ('اسلامی', ۴۳۵۸), ('کند', ۴۳۲۱), ('سال', ۴۳۰۱), ('ادامه', ۴۲۹۰), ('حضور', ۴۲۶۶), ('ما', ۴۰۸۶)]

پرسمان از کلمات ساده و متداول تک کلمه‌ای:

ایران:

Title: برخلاف شایعات مطرح شده ؛ بازی ایران و عراق با حضور ۱۰ هزار تماشاگر برگزار می شود

URL: https://www.farsnews.ir/news/14001106000296/برخلاف-شایعات-مطرح-شده-بازی-ایران-و-عراق-با-حضور-۰-هزار-تماشاگر

Content:

به گزارش خبرنگار ورزشی خبرگزاری فارس، بر خلاف شایعات در مورد برگزاری دیدار ایران و عراق دیدار دو تیم با حضور تماشاگران برگزار خواهد شد. طبق پیگیری های خبرنگار ورزشی خبرگزاری فارس، تماشاگران با رعایت پروتکل‌های بهداشتی در ورزشگاه دیدار دو تیم را نظاره‌گر خواهند شد. این دیدار با حضور ۱۰ هزار تماشاگر برگزار می‌شود. دیدار تیم‌های ایران و عراق ساعت ۱۸ فردا در ورزشگاه آزادی برگزار می‌شود. انتهای پیام/

Title: با حکم رئیس‌جمهور، مختارپور رئیس سازمان اسناد و کتابخانه ملی شد

URL: https://www.farsnews.ir/news/14000921000552/با-حکم-رئیس‌جمهور-مختارپور-رئیس-سازمان-اسناد-و-کتابخانه-ملی-شد

Content:

به گزارش خبرنگار حوزه دولت خبرگزاری فارس، با حکم آیت الله سید ابراهیم رئیسی، رئیس‌جمهور اسلامی ایران، علیرضا مختارپور به عنوان رئیس جدید سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران منصوب شد. انتهای پیام/

Title: اعلام ساعت دیدارهای تیم ملی مقابل عراق و امارات

URL: https://www.farsnews.ir/news/14001027000611/اعلام-ساعت-دیدارهای-تیم-ملی-مقابل-عراق-و-امارات

Content:

به گزارش خبرگزاری فارس و به نقل از سایت فدراسیون فوتبال، کنفدراسیون فوتبال آسیا با ارسال نامه‌ای ساعت دو بازی ایران مقابل عراق و امارات را اعلام کرد. بر اساس نامه ارسالی AFC دو بازی تیم ملی فوتبال ایران مقابل عراق (۷ بهمن) و امارات (۱۲ بهمن) راس ساعت ۱۸:۰۰ برگزار خواهد شد. انتهای پیام/

هر سه پرسمان بازگردانده شده توسط موتور بازیابی، شامل کلمه‌ی ایران هستند و به پرسمان مربوطه ارتباط دارند.

پرسمان از عبارت ساده و چند کلمه‌ای:

اخبار ورزشی:

Title: تست کرونای طارمی منفی شد/حضور هاشمیان در اردوی تیم ملی

URL: https://www.farsnews.ir/news/14001107000211/تست-کرونای-طارمی-منفی-شد-حضور-هاشمیان-در-اردوی-تیم-ملی

Content:

به گزارش خبرنگار ورزشی خبرگزاری فارس،با اعلام مجتبی خورشیدی سرپرست تیم ملی تست کرونای مهدی طارمی منفی شده و این بازیکن مشکلی برای حضور در دیدار مقابل عراق ندارد. سرپرست تیم ملی همچنین از حضور هاشمیان در اردوی تیم ملی خبر داد و اعلام کرد وی امروز تیم ملی را مقابل عراق همراهی خواهد کرد. انتهای پیام/

Title: بازیکنان استقلال به تیم ملی دعوت نمی‌شوند

URL: https://www.farsnews.ir/news/14001101000260/بازیکنان-استقلال-به-تیم-ملی-دعوت-نمی‌شوند

Content:

به گزارش خبرنگار ورزشی خبرگزاری فارس، سرمربی تیم فوتبال بزرگسالان کشورمان از عدم دعوت بازیکنان استقلال به تیم ملی خبر داد. با اعلام دراگان اسکوچیچ، به خاطر شرایط کرونایی حاکم بر اردوی آبی پوشان، تیم ملی به احتمال زیاد نمی‌تواند از بازیکنان استقلال استفاده کند. تیم ملی فوتبال کشورمان در روزهای هفتم و دوازدهم بهمن به ترتیب با تیم‌های عراق و امارات متحده عربی دیدار خواهد داشت. انتهای پیام/

Title: طاهری: ژستد و یامگا مشکلی برای بازی با نساجی ندارند/ به شایعات توجهی نداریم

URL: https://www.farsnews.ir/news/14001209000526/طاهری-ژستد-و-یامگا-مشکلی-برای-بازی-با-نساجی-ندارند-به-شایعات-توجهی

Content:

به گزارش خبرنگار ورزشی خبرگزاری فارس، از صبح امروز شایعاتی پیرامون بازی کردن کوین یامگا و رودی ژستد در صفحات مجازی منتشر شده و گفته می‌شود این ۲ بازیکن به دلیل پایان مجوز کار در بازی با نساجی نمی‌توانند به میدان بروند. خبرنگار ورزشی خبرگزاری فارس در این زمینه با بیژن طاهری سرپرست تیم فوتبال استقلال تماس گرفت که وی در این زمینه گفت: این خبر صحت ندارد و فکر نمی‌کنم کسی به این شایعات توجهی داشته باشد. از صبح امروز این خبر مطرح شده ولی به ما چیزی اعلام نشده و این ۲ بازیکن مشکلی برای همراهی استقلال ندارند. دیدار تیم‌های نساجی و استقلال از ساعت ۱۷ امروز در ورزشگاه امام رضا برگزار می‌شود. انتهای پیام/

تمام عبارات بازگردانده شده مرتبط به پرسمان هستند و دقت شود که به دلیل ریشه یابی، اخبار و خبرنگار هم ریشه و یکسان در نظر گرفته شده اند.

عبارت دشوار تک کلمه‌ای:

کریسمس:

Title: ستاره اسپانیایی؛ هدیه کریسمس گواردیولا به ژاوی+عکس

URL: https://www.farsnews.ir/news/14001007000739/ستاره-اسپانیایی-هدیه-کریسمس-گواردیولا-به-ژاوی-عکس

Content:

به گزارش خبرگزاری فارس، فران تورس ستاره اسپانیایی باشگاه منچستر سیتی با قراردادی تا سال ۲۰۲۷ به باشگاه بارسلونا پیوست. 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. انتهای پیام/

Title: کی‌روش «دیکتاتور» لقب گرفت/اختلاف مرد پرتغالی با مصری‌ها به خاطر کریسمس+عکس

URL: https://www.farsnews.ir/news/14001005000165/کی‌روش-دیکتاتور-لقب-گرفت-اختلاف-مرد-پرتغالی-با-مصری‌ها-به-خاطر-کریسمس

Content:

به گزارش خبرگزاری فارس، کارلوس کی روش سرمربی نام آشنا برای ایرانی‌ها این روزها در تیم ملی فوتبال مصر حاشیه های زیادی دارد و موافقان و مخالفانی پیدا کرده است. در جدیدترین خبر عبدالناصر زیدان یکی از خبرنگاران مطرح ورزشی مصر به شدت به کارلوس کی روش هجوم آورد او را «دیکتاتور» نامید. این خبرنگار در مصاحبه با سایت «صدی البلد» گفت: کارلوس کی روش یک دیکتاتور است. او نگاه بالا از پایین در تیم ملی مصر دارد. سرمربی پرتغالی قصد دارد محمد شریف(مهاجم الاهلی) را نابود کند. خط زدن این بازیکن دیوانه کننده است. البته این تنها خبرسازی کی روش در مصر نبود. سایت «المصری الیوم» نیز با تیتر «بحران میان کی روش با فدراسیون به خاطر کریسمس» به اختلافات این مربی با فدراسیون پرداخت. این رسانه مصری نوشت: کارلوس کی روش قرار بود اردوی تیم ملی برای آماده شدن در رقابت های مقدماتی جام جهانی ۲۰۲۲ از تاریخ ۲۸ دسامبر

آغاز کند ولی به دلیل تعطیلات کریسمس اردو را به تعویق انداخت که واکنش فدراسیون فوتبال را به همراه داشت. فدراسیون فوتبال مصر با این اقدام کی روش مخالفت کرد و اختلافات بین طرفین به خاطر تعطیلات کریسمس بالا گرفته است. انتهای پیام/

Title: کشتار در ورزشگاه فوتبال در آستانه سال جدید

URL: https://www.farsnews.ir/news/14001005000143/کشتار-در-ورزشگاه-فوتبال-در-آستانه-سال-جدید

Content:

به گزارش خبرگزاری فارس، در آستانه سال نو و روزهای کریسمس، مردم در ورزشگاه فوتالزای برزیل برای خوشحالی و شادمانی دور هم جمع شدند که یک حادثه مرگبار و خونین رخ داد. به گفته دبیرخانه امنیت عمومی ایالت سئارا برزیل، پنج نفر در هنگام جشن کریسمس در یک زمین فوتبال بر اثر شلیک گلوله کشته و ۶ نفر دیگر زخمی شدند. رسانه‌های محلی ذکر می‌کنند که این جنایت ناشی از نزاع بین ۲ جناح جنایتکار در برزیل بوده و بر همین اساس ۳ نفر بازداشت شده‌اند. مقامات ایالت سئارا تنها توانسته‌اند ۲ نفر از قربانیان را شناسایی کنند که یکی از آنها ۲۱ ساله و یکی دیگر ۲۶ ساله هستند. هر ۲ دو نیز سابقه جرم و جنایت داشته‌اند و به دلیل حمل سلاح گرم غیرقانونی، حضور در انجمن‌های جنایتکارانه و بر هم زدن آرامش شهر سابقه داشته‌اند. انتهای پیام/

تمام انساد برگشت داده شده کاملا مرتبط به عبارت پرسمان هستند!

عبارت دشوار چند کلمه‌ای:

وزارت ورزش:

Title: برگزاری مراسم تشییع دو شهید گمنام با حضور اهالی ورزش +فیلم

URL: https://www.farsnews.ir/news/14001028000330/برگزاری-مراسم-تشییع-دو-شهید-گمنام-با-حضور-اهالی-ورزش-فیلم

Content:

به گزارش خبرگزاری فارس، امروز (سه شنبه) مراسم تشییع و خاکسپاری ۲ شهید گمنام در محوطه وزارت ورزش و جوانان برگزار شد. در این مراسم جمعی از جوانان، ورزشکاران و قهرمانان و روسای فدراسیون ها و مسئولان وزارت ورزش نیز حضور داشتند. فیلم زیر را مشاهده کنید: انتهای پیام/

Title: مجوز ورود تماشاگران در بازی‌های تیم ملی صادر شد+عکس

URL: https://www.farsnews.ir/news/14001025000216/مجوز-ورود-تماشاگران-در-بازی‌های-تیم-ملی-صادر-شد-عکس

Content:

به گزارش خبرنگار ورزشی خبرگزاری فارس، علیرضا بهادری جهرمی اعلام کرد ستاد ملی مقابله با کرونا به ریاست رئیس جمهور با درخواست وزارت ورزش مبنی بر حضور تماشاگر در بازی‌های تیم ملی فوتبال ایران برای مسابقات انتخابی جام جهانی موافقت کرد. سقف ورود تماشاگر محدود و رعایت دستورالعل‌های بهداشتی ضروری است و وزارت ورزش مسئول نظارت بر رعایت دستورالعمل‌ها است. تیم ملی فوتبال ایران روز هفتم بهمن ماه به مصاف تیم ملی فوتبال عراق خواهد رفت. انتهای پیام/

Title: رئیس فدراسیون بسکتبال حکم ریاستش را از وزیر ورزش گرفت

URL: https://www.farsnews.ir/news/14001214000951/رئیس-فدراسیون-بسکتبال-حکم-ریاستش-را-از-وزیر-ورزش-گرفت

Content:

به گزارش خبرگزاری فارس، طی جلسه‌ای حکم ریاست جواد داوری، رییس فدراسیون بسکتبال توسط سیدحمید سجادی وزیر ورزش و جوانان اهدا شد. سید محمد پولادگر معاون توسعه ورزش قهرمانی و حرفه‌ای، اسماعیل احمدی مشاور وزیر و سرپرست حوزه وزارتی و کاوه احمدی مدیرکل روابط عمومی وزارت ورزش و جوانان نیز در این جلسه حضور داشتند. همچنین در این جلسه از زحمات عسگریان سرپرست سابق فدراسیون نیز تقدیر شد. انتهای پیام/

تمام اسناد بازیابی شده در این پرسمان نیز کاملا مرتبط با کوئری هستند و حاوی کلمات آن هستند.