به نام خدا

فاز دوم پروژه بازیابی اطلاعات

مدل سازی اسناد در فضای برداری

محمد جواد زندیه ۹۸۳۱۰۳۲

۱۳ خرداد ۱۴۰۱

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

سوال یک:

الف) یک پرسمان از کلمات ساده و متداول تک کلمه ای

پرسمان: جهان

نتایج: در فایل output1.txt قابل مشاهده است (برای تمام کوئری ها تنها پنج کوئری مرتبط اول را نمایش داده ایم)

تحلیل: برای پرسمان های تک تکلمه ای شاید این الگوریتم رتبه بندی ای که استفاده کردیم به خوبی عمل نکند زیرا میزان شباهت تعداد زیادی از داکیومنت ها با کوئری یکسان خواهد شد (برای همه داکیومنت هایی که این کلمه را دارند، میزان فاصله کسینوسی برابر یک میشود).

اثبات ریاضی:

$$\frac{\left(\left((1+\log(f_{t.d}))*log\left(\frac{N}{n_t}\right)\right)*\left((1+\log(1))*log\left(\frac{N}{n_t}\right)\right)\right)}{\sqrt{\left((1+\log(f_{t.d}))*log\left(\frac{N}{n_t}\right)\right)^2}*\sqrt{\left((1+\log(1))*log\left(\frac{N}{n_t}\right)\right)^2}}=$$

$$\frac{(1+\log(f_{t.d}))*\log^2(\frac{N}{n_t})}{\left((1+\log(f_{t.d}))*log\left(\frac{N}{n_t}\right)\right)*\left((1+\log(1))*log\left(\frac{N}{n_t}\right)\right)}=\frac{(1+\log(f_{t.d}))*\log^2(\frac{N}{n_t})}{(1+\log(f_{t.d}))*\log^2(\frac{N}{n_t})}=1$$

خروجی: واضح است که خروجی ها با پرسمان مرتبط اند (زیرا همگی شامل این کلمه هستند)، اما همانطور که در تحلیل گفته شد، رتبه بندی به خوبی عمل نمیکند، اما چون کلمه پرتکراری را داشتیم، تعداد سند بازیابی شده به پنج عدد میرسد.

:title

مدیر بی‌دستاورد، رئیس شد/ روزگار تلخ‌تر از تلخ برای والیبال ارومیه؟

:content

می‌کرد و به جام باشگاه‌های آسیا و جهان می‌رفت ، نه اینکه میزبانی جام

:title

پولادگر: درخشش‌های هندبال مایه آبرو و اعتبار بین‌المللی است/ پاکدل: زودتر از برنامه به اهداف خود رسیدیم

:content

عنوان نایب قهرمانی آسیا سهمیه مسابقات قهرمانی جهان را نیز بدست بیاورند . این

:title

واکنش اهالی تنیس به تبعیض نژادی/نادال: باید برای سختی‌ها آماده باشیم

:content

اتفاقی که افتاد بسیار متاسفم . در جهان واقعی این اتفاقات می‌افتد و من

من بسیار متاسفم . این تنیسور مطرح جهان گفت : ما مردم بسیار خوش‌شانسی

:title

مسابقات تنیس ایندین‌ولز| پایان ۳ هفته امپراطوری مدودوف/ جوکوویچ به صدر برمی‌گردد

:content

مقابل مانفیلس به ۳ هفته شماره یک جهان بودن خود پایان داد . مدودوف

جوکوویچ می‌تواند دوباره به صدر برترین تنیس‌بازان جهان برگردد . مدودوف بعد از این

من نیست . فکر می‌کردم شماره یک جهان بودن انگیزه بیشتری به من بدهد

:title

کار جالبی که ستاره فرمول یک جهان به افتخار مادرش انجام داد

:content

همیلتون دارنده ۷ عنوان قهرمانی فرمول یک جهان ، نام خود را به افتخار

شدن برای کسب رکورد هشتمین عنوان قهرمانی جهان است . این راننده ۳۷ ساله

همیلتون ادامه یابد . این راننده پرافتخار جهان درباره آمادگی‌اش برای کسب عنوان قهرمانی

جهان درباره آمادگی‌اش برای کسب عنوان قهرمانی جهان ، عنوان کرد : من برگشتم

ب) یک پرسمان از عبارات ساده و متداول چند کلمه ای

پرسمان: فوتبال قهرمانی آسیا

نتایج: در فایل output2.txt قابل مشاهده است (برای تمام کوئری ها تنها پنج کوئری مرتبط اول را نمایش داده ایم)

تحلیل: در این پرسمان، مقادیر محاسبه شده برای میزان شباهت، در کوئری ها متفاوت است و مشکل حالت الف را نخواهیم داشت. همانطور که در ادامه مشاهده میشود، داکیومنت هایی در رتبه بالاتر قرار دارند که تعداد بیشتری از کلمات کوئری را داشته باشند و ... (برای مقایسه، داکیومنت اول در رتبه بندی و داکیومنت چهارم در رتبه بندی علامت گذاری شده اند که میزان مرتبط بودن آنها با هم مقایسه گردد).

خروجی: هم رتبه بندی و هم مرتبط بودن را دارا می‌باشند. (از روی مقادیر شباهت میتوان متوجه شد که همگی نزدیک به ۱ هستند)

[('3177', 1.0), ('3715', 0.9999999999999999), ('6612', 0.9999999999999999), ('84', 0.9999999999999998), ('104', 0.9999999999999998)]

:title

بر خلاف ادعای صالحی امیری؛ یک فیفادی از دست رفت و خبری از برنامه نیست/ امیدها روی نوار بلاتکلیفی

:content

خبرنگار ورزشی خبرگزاری فارس ، تیم ملی فوتبال امید کشورمان در خرداد ماه باید

کشورمان در خرداد ماه باید در مسابقات قهرمانی آسیا در ازبکستان شرکت کند اما

در خرداد ماه باید در مسابقات قهرمانی آسیا در ازبکستان شرکت کند اما با

تیم‌ها همگروه نشود . در واقع تیم فوتبال امید باید در مسابقات قهرمانی آسیا

واقع تیم فوتبال امید باید در مسابقات قهرمانی آسیا با تیم‌های بسیار قوی چون

تیم فوتبال امید باید در مسابقات قهرمانی آسیا با تیم‌های بسیار قوی چون عربستان

شد و همین موضوع روند آماده‌سازی تیم فوتبال امید را کندتر می‌کند . آیا

مشکلات را کرده است ؟ طبعا مسابقات قهرمانی آسیا با مرحله مقدماتی متفاوت است

را کرده است ؟ طبعا مسابقات قهرمانی آسیا با مرحله مقدماتی متفاوت است و

کامل وارد مسابقات می‌شوند اما هنوز فدراسیون فوتبال زمانی برای آماده‌سازی و برنامه اردویی

فدراسیون داد که زمان کافی برای مسابقات قهرمانی آسیا وجود ندارد و تیم ملی

داد که زمان کافی برای مسابقات قهرمانی آسیا وجود ندارد و تیم ملی امید

:title

آغاز تمرینات تیم فوتبال امید در ابهام/ تکلیف قرارداد کادرفنی مشخص نیست

:content

ورزشی خبرگزاری فارس ، قراربود تیم ملی فوتبال امید کشورمان همزمان با برگزاری اردوی
آماده‌سازی خود را برای حضور در رقابت‌های قهرمانی زیر ۲۳ سال آسیا با حضور
حضور در رقابت‌های قهرمانی زیر ۲۳ سال آسیا با حضور نفرات لیگ برتری برگزار
تاجیکستان به عنوان صدرنشین مرحله مقدماتی رقابت‌های قهرمانی آسیا جواز حضور در مرحله نهایی
به عنوان صدرنشین مرحله مقدماتی رقابت‌های قهرمانی آسیا جواز حضور در مرحله نهایی مسابقات
دستیارانش باشد . باید از مدیران فدراسیون فوتبال هم که این روزها تمام تمرکز
سال آینده در بازی‌های آسیایی هانگژو و قهرمانی زیر ۲۳ سال آسیا حاضر شود
آسیایی هانگژو و قهرمانی زیر ۲۳ سال آسیا حاضر شود مشخص نمی‌کنند تا شاید
باشد ؟ مسئولان تیم امید و فدراسیون فوتبال در روزهای اخیر حتی مراحل اخذ

:title

پیشتازی قهرمان آسیا از چلسی در کسب جایزه بهترین باشگاه دنیا+عکس

:content

« گلوب ساکر » است تا بهترین‌های فوتبال در سال ۲۰۲۱ را معرفی کند
کمپانی اماراتی اهدا می‌کند جایزه بهترین باشگاه فوتبال است . در نظرسنجی بهترین باشگاه
بهترین باشگاه سال ۲۰۲۱ جهان الهلال قهرمان آسیا پیشتاز است . تیم سعودی با
سال ۲۰۲۱ قرار دارند . الهلال با قهرمانی در لیگ قهرمانان آسیا ۲۰۲۱ با
. الهلال با قهرمانی در لیگ قهرمانان آسیا ۲۰۲۱ با ۴ قهرمانی پرافتخارترین باشگاه
در لیگ قهرمانان آسیا ۲۰۲۱ با ۴ قهرمانی پرافتخارترین باشگاه قاره کهن شد .

:title

تمجید ویژه رئیس فدراسیون عراق از ناظم الشریعه

:content

حال آماده‌سازی برای حضور در مسابقات مقدماتی قهرمانی آسیا هستند . در همین رابطه

آماده‌سازی برای حضور در مسابقات مقدماتی قهرمانی آسیا هستند . در همین رابطه عدنان

در همین رابطه عدنان درجال رئیس فدراسیون فوتبال این کشور در اردوی تیم ملی

:title

محمدخانی: پرسپولیس باید مقابل استقلال انتحاری بازی کند

:content

بازی مساوی است با از دست رفتن قهرمانی لیگ ، چون اختلاف امتیازی با

دقایق و نتایج این مسابقه در تاریخ فوتبال ثبت می‌شود . هنوز در مورد

می‌آید . حامد لک در لیگ قهرمانان آسیا هم برای پرسپولیس خوب دروازه‌بانی کرد

پ) یک پرسمان دشوار و کم تکرار کلمه

پرسمان: نعل

نتایج: در فایل output3.txt قابل مشاهده است (برای تمام کوئری ها تنها پنج کوئری مرتبط اول را نمایش داده ایم)

تحلیل: مشابه حالت الف، رتبه بندی در اینجا به درستی عمل نمیکند، اما اسناد بازیابی شده مرتبط اند زیرا همگی، شامل این کلمه هستند.

خروجی: به علت اینکه این کلمه کم تکرار بوده است، تعداد اسناد بازیابی شده از حد در نظر رفته شده (پنج عدد) کمتر است.

:title

قرعه کشی لیگ قهرمانان آسیا ۲۰۲۲| فولاد مقابل استراماچونی و یاران قائدی و نوراللهی/ سپاهان در گروه D /غیبت غم انگیز سرخابی ها

:content

I : کوازکی ، گوانگژو ، جوهر نعل نهرو و برنده پلی آف ۳

:title

فتنه‌سازی نمی‌تواند ارابه در سنگلاخ‌مانده بیانیه‌نویسان زنجیره‌ای را نجات دهد

:content

ملت ، امروز در پی آنند با نعل وارونه و تحریف روایت انقلاب قهقرای

:title

بیانیه دفاتر بسیج دانشجویی شمال غرب کشور/ مردم آذربایجان اجازه حضور صهیونیست‌ها در کشورشان را ندهند

:content

و دست از حرکات « یکی به نعل و یکی به میخ » گونه

:title

سنگ‌اندازی اصلاح‌طلبانه در مسیر لغو تحریم‌ها / ایران از برجام خارج شد، یا آمریکا؟!

:content

سال ۱۴۰۰ در مقایسه با ۱۳۹۲ ! نعل وارونه اصلاح‌طلبان به وین حالا و

ت) یک پرسمان دشوار و کم تکرار چند کلمه ای

پرسمان: کنگره ضد تروریس

نتایج: در فایل output4.txt قابل مشاهده است (برای تمام کوئری ها تنها پنج کوئری مرتبط اول را نمایش داده ایم)

تحلیل: با توجه به کم تکرار بودن این کوئری، میزان شباهت اسناد بازیابی شده با کوئری، از سند دوم به بعد بسیار کم تر شده است. (از روش مقادیر شباهت هم میتوان متوجه شد)

مقادیر شباهت حاصل از این بازیابی:

[('6929', 0.9774002383257302), ('695', 0.4525091651152434), ('716', 0.4525091651152434), ('5107', 0.4525091651152434), ('6267', 0.4525091651152434)]

برای اسناد دوم به بعد، میزان شباهت کمتر از ۰.۵ شده است و ...

خروجی: رتبه بندی درست است، اما شباهت پایین است.

:title

توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

:content

تاریخ ۶ نوامبر سال ۱۹۸۷م . نمایندگان کنگره در واکنش به خریدهای نفت اداره‌ی

کردند . ریگان که نمی‌خواست کمتر از کنگره ضدتروریست جلوه کند ، ۳هفته بعد

. ریگان که نمی‌خواست کمتر از کنگره ضدتروریست جلوه کند ، ۳هفته بعد با

قانون مجازات‌ها برضد ایران و لیبی در کنگره تصویب شد . طبق این قانون

جانبه‌ی ایران در سال ۱۳۸۹ به تصویب کنگره‌ی این کشور رسید . این تحریم‌ها

:title

برگزاری مراسم روز درختکاری در فدراسیون ووشو/ ملی‌پوشان ۷۲ اصله نهال را غرس کردند

:content

عنوان کرد : در ستاد عالی دومین کنگره ملی شهدای ورزش مصوب شد که

:title

«پهلوانان ماندگار؛ ۵۱۳۵ شهید ورزشکار به نیت هر شهید یک درخت»/ کاشت نمادین درخت در فوتبال

:content

، سردار داود آذرنوش دبیر ستاد اجرایی کنگره شهدای ورزش کشور ، دکتر مهدی

:title

برگزاری سوپرجام فوتبال کشور به نام شهدای چوار

:content

قرار گرفت و ایشان در مراسم اولین کنگره شهدای استان ایلام ، بر جهانی

:title

مرادی: برخی رشته‌ها به جای شرکت در مسابقات جهانی انگار به پیک‌نیک رفته‌اند/ با این وضعیت باید اردوها را تعطیل کنم

:content

، علی مرادی در خصوص اینکه در کنگره جهانی چه اتفاقات رخ داد اظهار

سوال دو:

خروجی مربوط به پرسمان ها در فاز یک در فایل های output5.txt و output6.txt قرار دارد.

ب)

پرسمان: فوتبال قهرمانی آسیا

نتایج: در فایل output5.txt قابل مشاهده است.

تحلیل: در فاز یک، رتبه بندی به درستی انجام نمیشد و تنها هدف، پیدا کردن سند هایی بود که کلمات پرسمان را داشتند. در این پرسمان، سند اولی که بازیابی شده است مشابه فاز دوم است، اما مابقی اسناد میزان شباهت سند و پرسمان کم است و برخلاف فاز دو مشابهت زیادی ندارد.

خروجی:

Query: فوتبال قهرمانی آسیا

Title: بر خلاف ادعای صالحی امیری؛ یک فیفادی از دست رفت و خبری از برنامه نیست/ امیدها روی نوار بلاتکلیفی

URL: https://www.farsnews.ir/news/14001112000169/بر-خلاف-ادعای-صالحی-امیری-یک-فیفادی-از-دست-رفت-و-خبری-از-برنامه-نیست-

Content:

Title: جام ملت های فوتبال بانوان آسیا| کرونا بازی هم گروهی های ایران را لغو کرد

URL: https://www.farsnews.ir/news/14001103000897/جام-ملت-های-فوتبال-بانوان-آسیا|-کرونا-بازی-هم-گروهی-های-ایران-را-لغو

Content:

Title: اعلام زمان احتمالی پایان لیگ برتر و تغییر در برنامه‌ریزی مسابقات پس از قرعه کشی لیگ قهرمانان

URL: https://www.farsnews.ir/news/14001027000320/اعلام-زمان-احتمالی-پایان-لیگ-برتر-و-تغییر-در-برنامه‌ریزی-مسابقات-پس-از

Content:

فصل ۲۰۲۲ لیگ قهرمان آسیا اظهار جزو معدود سال هوادار فوتبال ایران هیجان مراسم قرعه‌کشی دنبال کرد&کن بابت حضور نماینده

Title: گفت‌وگوی ویژه فارس با مردی که با جادوگر فوتبال جنگید/ داستان حیرت‌انگیز از اعمال نفوذ در مستطیل سبز

URL: https://www.farsnews.ir/news/14000917000716/گفت‌وگوی-ویژه-فارس-با-مردی-که-با-جادوگر-فوتبال-جنگید-داستان

Content:

Title: فوت بازیکن پیشین تیم فوتسال ارژن شیراز

URL: https://www.farsnews.ir/news/14000926000473/فوت-بازیکن-پیشین-تیم-فوتسال-ارژن-شیراز

Content:

علی عرب زاده پایه تیم ملی فوتسال کشوربود فاجعه جامعه فوتبال فوتسال افغانستان .» انتهای پیام / م

ت)

پرسمان: کنگره ضدتروریست

نتایج: در فایل output6.txt قابل مشاهده است.

تحلیل: تنها سند اولی که در فاز دوم هم استخراج شده بود را نشان میدهد، زیرا تنها آن سند بود که هر دو عبارت را داشت، در حالیکه در فاز دوم اسنادی که فقط کنگره را داشتند هم بازیابی شده و رتبه بندی شده بود، در واقع همانطور که گفتیم، فاز یک فقط اسنادی که تمام عبارات را داشتند بر میگرداند و رتبه بندی درستی هم نداشت اما در این فاز هر دو این مشکل بر طرف شده است.

خروجی:

Query: کنگره ضدتروریست

Title: توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

URL: https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین

Content:

کاربرد نظامی نیزبودند گسترش تاریخ ۶ نوامبر سال ۱۹۸۷م نماینده کنگره واکنش خرید&خر نفت اداره انرژی آمریکا ایران لایحه مجلس
