به نام خدا

گزارش پروژه بازیابی اطلاعات غزل آشفته دل 40031004

1.ساخت شاخص مکانی

1.1.مجموعه داده

برای دیتاست یک کلاس ساخته شده که دو فیلد path و data دارد. برای ایجاد دیتاست آدرس آن داده می‌شود و توسط متد read_dataset این فایل json خوانده می‌شود و مقدار آن به عنوان فیلد data آبجکت ساخته شده قرار می‌گیرد. در این کلاس متدهای دیگری مثل get_fields برای دریافت فیلدهای هر خبر و read_data_at_index برای خواندن یک خبر خاص با استفاده از ایندکس داده شده و get_titles برای دریافت لیستی از title خبرها و get_contents برای دریافت لیستی از content های خبرها و read_title_at_index برای خواندن title خبر موجود در ایندکس داده شده و read_content_at_index برای خوندن content خبر موجود در ایندکس داده شده و ... برای کار با دیتاست نوشته شده است.

2.1.پیش‌پردازش اسناد

ابتدا متن نرمالایز می‌شود و قوانین نرمال‌کردن که ذکر شده روی آن اجرا می‌شود. سپس متن نرمال‌شده را توکن توکن می‌کنیم. سپس 50 توکنی که بیشترین تکرار را در کل کالکشن داشتند پیدا و جزو stop word ها در نظر می‌گیریم. البته کاراکترهای غیر حروف الفبایی مثل . یا (و ... از قبل به عنوان stop word در نظر گرفتیم که اینها جدا از 50 کلمه خواسته شده در دستور کار هستند.

```
stopWords =
['.','?','/','،',':',')','(','»','«','[',']','{','}',',','؛',
'*','#','=','+','-','$','!','@','%',

'^','&','_',';','~','`','"','\'','|','\\','<','>']
```

50 کلمه پیداشده :

و 219248

در 164489

به 132859

از 92983

این 82244

که 75464

با 69021

را 67492

است 57644

برای 30431

کرد 24733

تیم 22088

هم 21703

ما 19764

شد 19471

یک 17671

آن 16599

بود 16566

باید 16204

تا 15857

کشور 15762

بر 15652

وی 15488

بازی 15310

خود 14552

مجلس 14519

اسلامی 14464

گفت 13874

فارس 13807

مردم 13094

گزارش 13055

پیام 12829

ایران 12620

خبرگزاری 12406

انتهای 12252

اما 12111

دولت 11595

شود 11287

داشت 10753

دارد 10644

سال 10432

اینکه 10060

ملی 9785

قرار 9479

دو 8943

می‌شود 8805

کند 8652

کار 8425

نیز 8339

امروز 8324

ریشه یابی توسط کتابخانه آماده parsivar انجام شده است.

3.1.ساخت شاخص مکانی

پس از پیش پردازش داکیومنت‌ها، به ازای هر کلمه در کالکشن، یک آبجکت Term ایجاد می‌کنیم. کلاس Term حاوی فیلدهایی مثل total_frequency برای نگهداری تعداد تکرار در کل کالکشن و frequency_in_docs برای نگهداری تعداد تکرار در هر داکیومنت و position_in_docs برای نگهداری پوزیشن(های) کلمه در هر داکیومنت و weight_in_docs برای نگهداری امتیاز tf-idf که بعدا به آن خواهیم پرداخت و champion_list برای نگهداری لیست قهرمانان. بدین ترتیب شاخص مکانی خود را ذخیره می‌کنیم.

```
class Term:
    def __init__(self):
        self.total_frequency = 0
        self.frequency_in_docs = {}
        self.position_in_docs = {}
        self.weight_in_docs = {}
        self.champion_list = {}
```

1.2.مدل‌سازی اسناد در فضای برداری

با توجه به فرمول‌های قرار گرفته دو تابع یکی برای محاسبه tf و دیگری برای محاسبه idf می‌نویسیم.

```
def tf(self,term, doc_id):
    t = self.pos_index[term]
    if t.frequency_in_docs[doc_id] > 0:
        return 1 + math.log10(t.frequency_in_docs[doc_id])
    return 0

new *
def idf(self,term):
    t = self.pos_index[term]
    nt = len(t.frequency_in_docs.keys())
    return math.log10(self.get_size() / nt)
```

سپس به ازای تک تک term ها در داکیومنت‌ها، وزن term در آن داکیومنت را با ضرب مقدار tf در idf به دست می‌آوریم و در weight_in_docs مربوط به آن داکیومنت، ذخیره می‌کنیم.

2.2.پاسخ‌دهی به پرسمان در فضای برداری

در تابع norm_docs نرم تک تک داکیومنت‌ها را به دست می‌آوریم و آن‌ها را در docs_norm دیتاست که یک لیست است، ذخیره می‌کنیم. ایندکس i ام در این لیست، نشان‌دهنده نرم داکیومنت i ام است. برای به دست آوردن نرم هر داکیومنت باید وزن هر کلمه در آن را به توان دو برسانیم و با هم جمع کنیم و در انتها رادیکال بگیریم. زمانی که یک کوئری وارد می‌شود ابتدا باید عملیات پیش پردازش را مشابه کاری که در قسمت پیش پردازش کل کالکشن داشتیم، انجام دهیم (نرمال‌سازی، توکنایز کردن، حذف stop word ها، ریشه‌یابی). سپس در بین توکن‌های پیش‌پردازش شده کوئری، به ازای هر کلمه، تعداد تکرار آن در کوئری را به دست می‌آوریم. سپس در یک حلقه دیگر روی کلمات کوئری پیمایش می‌کنیم. وزن هر کلمه در کوئری برابر است با term frequency که در مرحله قبل حساب کردیم ضربدر idf آن کلمه. در دیکشنری scores کلیدها داکیومنت‌ها و مقادیر، امتیاز آن‌ها هستند. در بین داکیومنت‌هایی که آن کلمه را دارند، به امتیاز آن داکیومنت در scores مقدار وزن کلمه در کوئری را در وزن آن کلمه در داکیومنت ضرب می‌کنیم و اضافه می‌کنیم. بعد اتمام حلقه، هر یک از مقادیر socres را بر نرم داکیومنتش تقسیم کنیم(چون نرم کوئری برای همه یکسان است، دیگر نیاز به تقسیم بر آن نیست) و در نهایت سورت می‌کنیم به صورت نزولی و k تای اول را برمی‌گردانیم.

3.2. افزایش سرعت پردازش پرسمان

برای ساخت لیست قهرمانان، قبل از دریافت کوئری باید به ازای هر کلمه وزنش را در داکیومنت‌هایی که آن را دارند به دست آوریم و نزولی سورت کنیم و k تای اول را در لیست قهرمانان آن کلمه که فیلدی از Term است ذخیره کنیم. در زمان دریافت کوئری تمام مشابه قسمت قبل عمل می‌کنیم با این تفاوت که به ازای هر کلمه در کوئری به جای بررسی کل داکیومنت‌هایی که آن کلمه را دارند و به دست آوردن وزن و ضرب وزن در وزن کلمه در کوئری و اضافه کردن امتیاز داکیومنت، باید روی لیست قهرمانان مربوط به داکیومنت پیمایش کنیم.

4.2. گزارش

در همه موارد k را برابر سه قرار دادیم و سه خبر اول بازگردانده می‌شود.

الف)یک پرسمان از کلمات ساده و متداول تک کلمه‌ای

کوئری : باشگاه

باشگاه استقلال مالیات و عملکرد مالی سال 98-99 را تسویه کرد

به گزارش خبرگزاری فارس و به نقل از باشگاه استقلال، با اعلام مصطفی احمدی مدیر مالی این باشگاه، مالیات‌های حقوق و عملکرد سال مالی ۹۸ - ۹۹ این باشگاه امروز پرداخت و تسویه مالی مذکور اخذ شد. به این ترتیب باشگاه استقلال هیچگونه بدهی از این بابت در سال مالی مذکور به اداره دارایی ندارد. انتهای پیام/

مجیدی دوباره مظاهری را در اختیار باشگاه قرار داد

به گزارش خبرگزاری فارس و به نقل از روابط عمومی باشگاه استقلال، با اعلام کادرفنی آبی پوشان، رشید مظاهری به دلیل «تخلف انضباطی در تمرین روز پنج‌شنبه ۹ دی» در اختیار باشگاه قرار گرفته و این دروازه‌بان به جلسه کمیته انضباطی باشگاه استقلال دعوت شده است. انتهای پیام/

مدافع پرسپولیس به فولاد پیوست

به گزارش خبرگزاری فارس و به نقل از سایت باشگاه پرسپولیس، مهدی شیری در پی توافق نهایی که صورت گرفت از باشگاه پرسپولیس جدا شد. باشگاه پرسپولیس در ازای دریافت مبلغی با صدور رضایت‌نامه این بازیکن موافقت کرد و به این ترتیب، این بازیکن به طور قطعی از پرسپولیس جدا شد. او قرار است برای ادامه فوتبال خود به فولاد خوزستان ملحق شود. باشگاه پرسپولیس برای مهدی شیری آرزوی موفقیت می‌کند. انتهای پیام/

همانطور که می‌بینیم، هر سه سند با کوئری زده شده مرتبط هستند و موارد مربوط در اسناد هایلایت شده است. در سند اول و دوم بازگردانده شده حتی در تایتل هم کلمه باشگاه قرار دارد.

الف) یک پرسمان از عبارات ساده و متداول چند کلمه‌اي

کوئری : نشست خبری

نشست خبری ربیعی تحریم شد

به گزارش خبرگزاری فارس از شیراز، خبرنگاران بعد از بازی فجر سپاسی و مس رفسنجان حاضر نشدند در نشست خبری محمد ربیعی حاضر شوند. ربیعی به دلیل اینکه در نشست خبری قبل از بازی حاضر نشده بود باعث ناراحتی خبرنگاران شیرازی شده بود و به همین دلیل حاضر نشدند در نشست خبری بعد از بازی حاضر شوند. ربیعی البته به خبرنگاران گفت من تمایل داشتم به نشست بیایم اما فرصت نشد. انتهای پیام/

اعلام زمان نشست خبری گل‌محمدی پیش از بازی با آلومینیوم

به گزارش خبرگزاری فارس، نشست خبری سرمربی تیم فوتبال پرسپولیس پیش از دیدار با آلومینیوم اراک فردا (پنجشنبه) از ساعت ۱۳ آغاز می‌شود. این نشست به میزبانی باشگاه پرسپولیس و در مجموعه ورزشی نفت تهرانسر برگزار می‌شود. رسول خطیبی، سرمربی تیم میهمان به این نشست خبری نمی‌رسد. انتهای پیام/

منصوریان در نشست خبری بازی با استقلال حاضر نمی شود

به گزارش خبرنگار ورزشی خبرگزاری فارس، علیرضا منصوریان در نشست خبری پیش از بازی با استقلال شرکت نخواهد کرد. از آنجا که تمرین تیم نفت آبادان با ساعت نشست خبری همزمان شده وی قصد دارد به صورت ویدیویی در خصوص بازی با نفت توضیحاتی را ارائه کند. دیدار تیم های استقلال و نفت آبادان از ساعت 20 فردا برگزار می شود. انتهای پیام/

هر سه سند دریافتی مرتبط با کوئری هستند. در هر سه سند کلمه نشست خبری در تیتر هم آمده که نشان می‌دهد که این کلمه در آن سند، جزو کلمات مورد بحث است. سند دوم و سوم cosine similarity برابری داشتند در حالی که در اولی کلمه نشست خبری یک بار بیشتر تکرار شده است و علت این است که در دومی در یک متن کوتاه دو بار این کلمات آمده اند و این نشان می‌دهد که لزوما تکرار بیشتر منجر به امتیاز بالاتر نمی‌شود و هر دو جنبه تکرار و idf سنجیده می‌شود و علاوه بر این نرمال‌سازی می‌شود.

پ) یک پرسمان دشوار و کم تکرار تک کلمه‌اي

کوئری : کریسمس

ستاره اسپانیایی؛ هدیه کریسمس گواردیولا به ژاوی+عکس

به گزارش خبرگزاری فارس، فران تورس ستاره اسپانیایی باشگاه منچستر سیتی با قراردادی تا سال 2027 به باشگاه بارسلونا پیوست. انتقال این بازیکن 21 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/

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

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 2022 از تاریخ 28 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/

کشتار در ورزشگاه فوتبال در آستانه سال جدید

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 6 نفر دیگر زخمی شدند. رسانه‌های محلی ذکر می‌کنند که این جنایت ناشی از نزاع بین 2 جناح جنایتکار در برزیل بوده و بر همین اساس 3 نفر بازداشت شده‌اند. مقامات ایالت سئارا تنها توانسته‌اند 2 نفر از قربانیان را شناسایی کنند که یکی از آنها 21 ساله و یکی دیگر 26 ساله هستند. هر 2 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/

همانطور که می‌بینیم، چون کلمه کریسمس خیلی پرکاربرد نیست، تعداد تکرار آن در داکیومنت‌های با امتیاز بالا بسیار کم است. داکیومنت اول تنها یک بار در آن کلمه کریسمس آمده است اما در امتیاز بالاتری قرار دارد چرا که یک کلمه با idf بالا در یک متن کوتاه ظاهر شده است پس یعنی مرتبط تر است.

ت) یک پرسمان دشوار و کم تکرار چند کلمه‌اي

کوئری : مجمع عمومی حزب

مجمع عمومی سالیانه حزب ندای ایرانیان دوباره به تعویق افتاد

به گزارش خبرنگار گروه سیاسی خبرگزاری فارس، مجمع عمومی سالیانه حزب ندای ایرانیان که قرار بود 13اسفندماه برگزار شود،به زمانی دیگر موکول شد. بنا بر اعلام مسئولین این حزب اصلاح‌طلب، برخی مشکلات اجرایی و در راستای هرچه بهتر، برگزاری این مراسم به زمان دیگری موکول شده که متعاقبا اعلام می‌شود. پیش از این اعلام شده بود که این کنگره سالیانه که ماهیتی انتخاباتی دارد، از ساعت 9 صبح 13 اسفندماه در «هتل فردوسی» تهران آغاز می‌شود و در نوبت صبح با حضور خبرنگاران و سخنرانی مدعوین و در نوبت عصر با رای‌گیری برای انتخاب اعضای شورای مرکزی این حزب به کار خود پایان خواهد داد. همچنین این دومین بار است که این مجمع عمومی به تعویق می افتد و قبل‌تر مقرر بود که 29 بهمن برگزار شود. انتهای پیام/

کنگره حزب جمهوریت بدون حضور دو سوم اعضای شورای مرکزی برگزار شد

به گزارش خبرنگار گروه سیاسی خبرگزاری فارس، سومین مجمع عمومی سالیانه حزب جمهوریت ایران اسلامی عصر امروز (جمعه) با حضور اعضای شورای مرکزی این حزب و همچنین حضور مجازی سایر اعضای حزب برگزار شد. بنا بر پیگیری‌های خبرنگار فارس، کنگره امروز این حزب در حالی برگزار می شود که دو سوم اعضای شورای مرکزی شامل رحمت الله بیگدلی، قربانعلی قائمی، مجید نصیرپور و چند نفر دیگر از اعضای شورای مرکزی در کنگره حاضر نیستند و به‌زودی با انتشار بیانیه ای استعفای خود را از عضویت در این حزب اعلام خواهند کرد. جمهوریت حزبی با محوریت رسول منتجب نیا است که در پی اختلافات درون حزبی حزب اعتماد ملی و مخصوصا اختلاف با مهدی کروبی دبیرکل سابق این حزب، منفک و تبدیل به حزبی جدید شد. انتهای پیام/

گزارش مصوبات آخرین جلسه کمیسیون ماده 10 احزاب

به گزارش خبرنگار گروه سیاسی خبرگزاری فارس، محسن اسلامی مدیرکل سیاسی وزارت کشور ظهر امروز (دوشنبه) در نشستی با اصحاب رسانه با اشاره به جلسه کمیسیون ماده ۱۰ احزاب اظهار داشت: اولین مصوبه امروز تقاضای تاسیس حزب همدلی و توسعه ملی بود که در جلسه کمیسیون ماده ۱۰ احزاب طرح و بررسی شد و کمیسیون با فعالیت این حزب در گستره استانی منوط به تغییر عنوان موافقت کرد. وی افزود: مصوبه دوم تقاضای تاسیس حزب جمعیت معلمان و فرهیختگان ایران اسلامی بود که در کمیسیون طرح و با فعالیت آن در گستره استانی و منوط به تغییر عنوان موافقت کرد. اسلامی ادامه داد: گزارش برگزاری کنگره مجمع نیروهای خط امام سومین مصوبه بود که مورد تایید قرار گرفت. همچنین مصوبه بعدی گزارش برگزاری مجمع عمومی انجمن اسلامی دانشگاهیان اصفهان بود که در جلسه مطرح و مورد تایید فراز گرفت. وی خاطرنشان کرد: کمیسیون براساس مصوبات ۱۸و۱۹ موظف شد نسبت به اعمال فرایند رصد و پایش و نظارت در عمل به مصوبات کمیسون ماده ۱۰ احزاب در حیطه احزاب ملی و استانی اقدام لازم را در دستور کار قرار دهد. دبیر کمیسیون ماده ۱۰ احزاب درباره تعبیه محلی برای برگزاری قانونی تجمعات، گفت: در این رابطه و در راستای تعامل بین وزارت کشور و مجلس مدتی است جلسات کارشناسی تشکیل شده تا جوانب مختلف موضوع مورد بررسی قرار گیرد.همچنین مجلس با هماهنگی وزارت کشور در حال تدوین قانون جامعی در این رابطه است. اسلامی در پاسخ به سوالی درباره اظهارات اخیر نماینده مجلس در کمیسیون احزاب درباره غیرقانونی بودن انتخاب آذر منصوری به عنوان دبیرکل حزب اتحاد ملت، گفت: هنوز گزارش مجمع عمومی حزب اتحاد ملت در کمیسیون ماده ۱۰ مطرح نشده است. انتهای پیام/

همانطور که می‌بینیم، نتایج به دست آمده تا حد زیادی مرتبط با این عبارت کم تکرار هستند.