محاسبه tf-idf:

برای محاسبه بردارهای tf-idf اسناد می‌بایست ابتدا term frequency و doc frequency را به ازای کلمات و اسناد متفاوت بدست می‌آوردیم که برای این کار همانند حالتی که inverted index را می‌سازیم زمانی که یک کلمه در یک سند دیده می‌شود می‌توان مقادیر tf و df متناسب با آن کلمه و سند را بروزرسانی کرد تا درنهایت با تمام شدن دیدن اسناد مقادیر تمامی tf و dfها را داشته باشیم که بعد از این مرحله با توجه به مقادیر بدست آمده و رابطه tf-idf نمایش برداری اسناد را بسازیم و در اختیار داشته باشیم که برای این کار هم فقط درایه‌هایی را نگه می‌داریم که مقداری ناصفر دارند به شکل یک دیکشنری نگه می‌داریم

بعد از بررسی خبر و با توجه به مضمون آن که در رابطه با لژیونرهای والیبالیست هست و کلمات پرارزش و کم‌ارزش بدست آمده از این خبر که در ذیل آمده است می‌توان گفت کلمات پرارزش تشخیص داده شده از تیم‌های والیبال یا اسم والیبالیست‌ها هستند به میزان خوبی مرتبط و دارای اهمیت هستند و کلمات کم‌ارزش هم تقریبا جزو کلمات عامیانه‌تر هستند

('سال', 0.4177)	('بدجین', 4.7412)
('کار', 0.5464)	('لوبه', 4.4402)
('حضور', 0.5794)	('بلهاتوف', 4.4402)
('گذشته', 0.613)	('میرزاجانپور', 4.2641)
('شرکت', 0.7687)	('اولیشتن', 4.1392)

اما در مقابل با بررسی خبر می‌توان گفت کلماتی که در این خبر پرارزش شناخته شدند به دلیل مرتبط بودن با مضمون خبر نیست و دلیل آن اشتباه در نوشتار و در نتیجه متعارف نبودن آن‌ها که در مقدار df بالایی را می‌گیرند و در مجموع هم در آن خبر مقدار tf-idf بالایی می‌گیرند

('توان', 0.4991)	('ودرحال', 6.1684)
('روز', 0.5413)	('دودختر', 6.1684)
('ایران', 0.5663)	('ومدام', 4.7412)
('حضور', 0.5794)	('وتشنج', 4.74122)
('مردم', 0.6518)	('واریزگردد', 4.7412)

پاسخ‌گویی به پرسمان:

حال برای پاسخگویی به پرسمان کاربر ابتدا باید نمایش برداری پرسمان را محاسبه کنیم و سپس مشابهت پرسمان با اسناد را محاسبه کنیم که برای این کار ابتدا هم نمایش بردار پرسمان را به حالت یکه در می‌آوریم و حال با توجه به ایده index elimination شباهت‌ها را با بردار یکه شده اسنادی که در inverted index حداقل یکی از کلمات داخل پرسمان حضور پیدا کردند با استفاده از کسینوسی محاسبه می‌کنیم و درنهایت با استفاده از heap تعدادی از مرتبطترین نتایج را بازمی‌گردانیم

به عنوان مثال ده نتیجه برتر برای پرسمان «سازمان لرزه‌نگاری تهران» در ذیل آمده است

کلاله لرزید	زمین لرزه خانه زنیان را لرزاند
زلزله بیش از ۹۲۰ بار "ایران" را لرزاند	زلزله چهار و یک دهم ریشتری در هرمزگان
رویدر لرزید	زمین لرزه ۳.۵ ریشتری مسجدسلیمان را لرزاند
زمین لرزه در درز	زمین لرزه ای به بزرگی ۳.۱ ریشتر سردشت را لرزاند
زلزله صحنه را لرزاند	زمین لرزه سردشت را لرزاند

محاسبه champion lists:

حال برای ساختن champion listها با استفاده از tf-idfهای ساخته شده در مرحله قبل به ازای هر کلمه ابتدا مقادیر tf-idfهای متناظر با آن کلمه و اسنادی که در آن‌ها حضور داشته است را از بزرگ به کوچک مرتب می‌کنیم و تعدادی از ابتدا این لیست مرتب شده را به عنوان champion list آن کلمه درنظر می‌گیریم که این تعداد را برابر با فرجه دوم تعداد کل اسناد درنظر گرفته‌ام

در زمانی هم که بخواهیم به پرسمانی از کاربر پاسخ دهیم به جای اینکه در میان اسنادی که حداقل در یکی از inverted indexهای کلمات پرسمان حضور دارد شباهت را محاسبه کنیم و درنهایت مرتبطترین اسناد را برگردانیم در میان اسنادی که در حداقل یکی از champion listهای توکن‌های پرسمان حضور داشتند شباهت‌ها را محاسبه می‌کنیم و مرتبطترین اسناد را برمی‌گردانیم با این کار زمان جستجو برای همان پرسمان بالا از حدود 20ms به 4.5ms می‌رسد و اسناد بازگردانده شده هم کاملا شبیه و به همان ترتیب حالت قبل خواهند بود اما در حالت کلی می‌توان گفت با

این کار ممکن است بعضی از اسنادی که در حالت قبل به عنوان مرتبطترین اسناد شناخته شده بود را بازنگرداند و champion list کلمات ذکر شده هم در ذیل آمده است:

- ابریشم

نمایشگاه هنری کودکان ایرانی و چینی به توسعه تبادلات فرهنگی 2 کشور می‌انجامد
سمفونی صلح کودکان چینی و ایرانی نواخته شد
مذاکره برای دیپلماسی حمل‌ونقل منطقه‌ای در سومین اجلاس راه ابریشم تفلیس

- فرهنگیان (این کلمه توسط stemmer به «فرهنگ» تبدیل می‌شود)

پاتوق فرهنگی کرج اینجاست/ حکایت مشتریانی که مخاطب شدند
پژوهش‌های دانشگاه فرهنگیان در حوزه آموزش است
ضرورت خردورزی جمعی سینماگران و علمای دینی

- آرامش

اعجاز ابراز محبت در کاهش و حل اختلافات زوجین
تحولات لبنان پس از استعفای حریری/ آغاز رایزنی میشل عون برای انتخاب نخست وزیر جدید
دو مسیر زیارتی برای صلح؛ از کربلا تا سانتیاگو