محمد مهدی نظری – ۹۹۳۱۰۶۱

* در این پروژه فقط از ۲۰۱ داکیومنت استفاده شده

در موارد بررسی شده ابتدا positional index و champion list آورده شده است.

```
Positional Index of :'15', [61, 6]:'12', [43]:'9', [8]:'7', [184, 151, 122, 94, 78, 46, 17]:'3', [40]:'2', [40, 33]:'1', [10, 4]:'0', 147]: 122', 148', [10, 6]:'12', [40, 8]:'9', [25, 4]:'35', [28, 4]:'15', [54, 28, 10, 4]:'35', [28, 5]: 129', [95, 41]:'23', [17, 6]:'16', [6], [75]:'76', [11]:'74', [40, 8]:'9', [31], 263, 128, 44, 13]:'66', [102]:'64', [44]:'35', [54, 9, 71], 368, 322, 309, 256, 64]:'55', [37]:'48', [26], [75]:'76', [11]:'74', [40, 8]:'81', [16]:'82', [213, 190, 183, 102, 187, 140], 313, 129, 117, 99, 56, 55, 12]: 181', [10], 18, 11]:'88', [19]:'79', 194', [389]:'91', [4]:'90', [572, 557, 546, 537, 434, 369, 343, 271, 140, 12]:'89', [12]:'88', [61, 5]:'87', [768, 270]:'86', [442, 427, 294, 213, 38, 14]:'131', [8]:'113', [5]:'112', [11]:'109', [8]:'107', [13]:'165', [9]:'146', [13]:'142', [72, 10]:'150', [57, 37, 19]:'96', [60]:'95', [10], [8]:'172', [46, 21]:'117', [55, 9]:'179', [72]:'167', [39]:'165', [9]:'146', [13]:'142', [72, 10]:'150', [5]:'135', [84, 14, 6]:'134', [61], [9]:'148', [10], [9]:'148', [10]:'168', [9]:'144', [13]:'142', [72, 10]:'150', [5]:'135', [84, 14, 6]:'134', [61], [9]:'148', [10], [9]:'148', [10]:'168', [9]:'148', [10]:'168', [9]:'148', [10]:'168', [9]:'148', [10]:'168', [9]:'172', [10]:'168', [9]:'138', [84, 14]:'172', [9]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]:'168', [9]:'172', [10]
```

۶ کلمه ای که قرار است مورد بررسی قرار دهیم برای هرکدام فیلد های positional index علاوه بر مقادیر فاز ۱ داری یک لیست برای امتیاز ff-idf برای هر داکیومنتی که در آن قرار دارد است بعلاوه امتیاز گفته شده داکیومنت های موجود در champion list هر کدام هم مشخص است . در این فاز ۳ داکیومنت با بیشترین امتیاز شباهت به ترتیب عکس نمایش داده میشوند.

۱ و ۳ –

در این قسمت علاوه بر نتیجه کویری، لیست مقایسه شباهت برای کویری برای هر داکیومنت(۰ تا ۲۰۱) آورده شده است (تعداد صفر ها زیاد است چونکه با استفاده از index elimination و champion list تعداد محاسبات به حداقل رسیده است)

معيار شباهت : كسينوسي

الف)

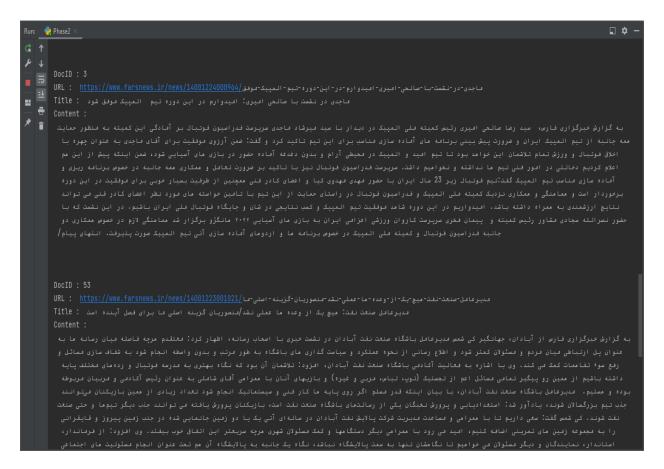
عبارت: "فوتبال"

DocID: 85

نا -فوتبال-و-رسمكتبال-MMA-با-بردر آسدترین-ورزشكار ان-جهان-آشنا-شوبد-از/URL: https://www.farsnews.ir/news/14001223000292

تا فوتبال و بسکتبال MMA با پردرآمدترین ورزشکاران جهان آشنا شوید؛ از : Title

Content :



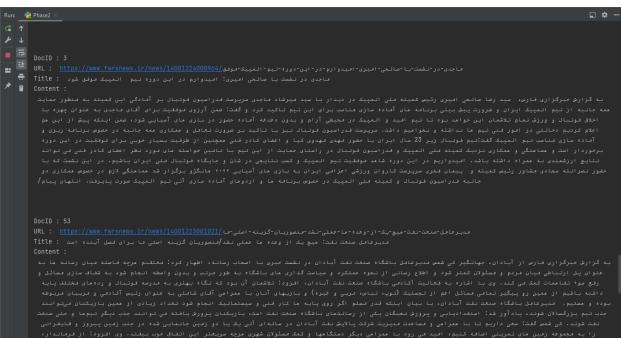
کلمه فوتبال چندبار دیده میشود پس هر ۳ داکیومنت مرتبط است.

در عبارت تک کلمه ای امتیاز های غیر صفر در لیست امتیاز ها همان داکیومنت های champion list هستند که در این حالت ۳ تای بالاتر آن یعنی ۳ و ۵۳ و ۸۵ آورده شده (۸۱ و ۸۹ آورده نشده)

ر ا

عبارت: "ليگ برتر فوتبال"

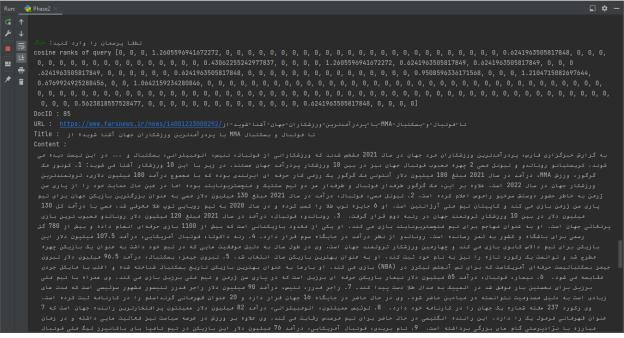


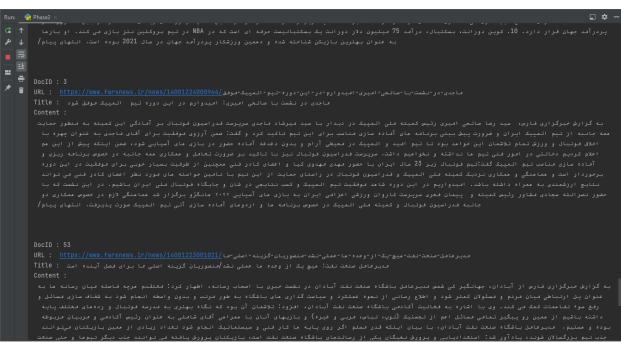


در هر سه عبارت فوتبال دیده میشود اما لیگ یا برتر در هر سه نیست و محاسبه امتیاز برای داکیومنت های بیشتری انجام شده(اجتماع champion list های سه عبارت) و خروجی این حالت با حالت الف یکی میشود به دلیل الگوریتم tf-idf که نشان دهنده امتیاز است که دقت کمی دارد .

ر (

عبارت : "مايكل"

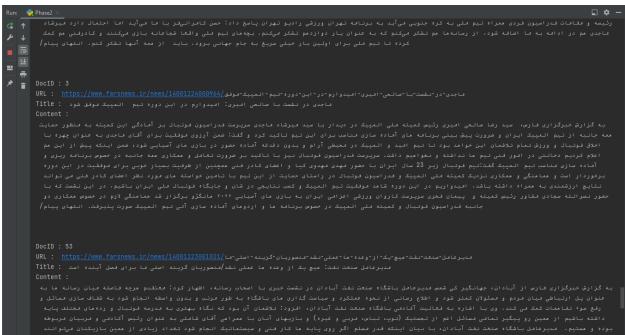




در این حالت چونکه در champion list عبارت مایکل فقط یک داکیومنت داریم . برای تولید ۳ خروجی نیاز داریم که امتیاز حاصل با داکیومنت های دیگر را بررسی کرده و در نظر میگیریم اما میدانیم این حالت بسیار دقت را پایین میاورد و فقط داکیومنت اول مرتبط است و دو داکیومنت دیگر کاملا بی ربط هستند.

عبارت: " مايكل جردن آمريكا"





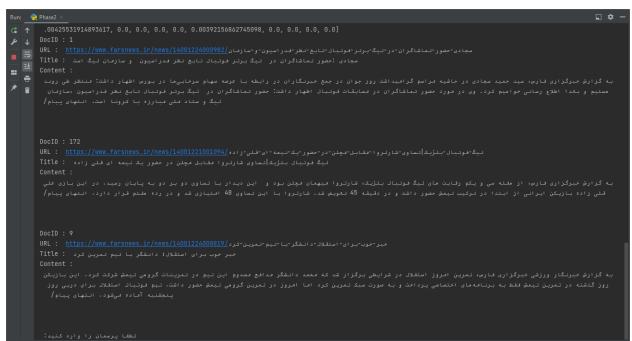
در این حالت هم مانند قسمت قبل بدلیل ضعف الگوریتم صرفا خروجی بی ربط داریم. تنها تفاوت این است که اضافه شدن عبارت آمریکا که در داکیومنت های بیشتری نسبت به مایکل و جردن آمده است باعث میشود لیست محاسبه امتیازات منطقی تر شود.

معيار شباهت : ژاكارد

در حالت مشابهت ژاکارد بدلیل زمان به نسبت بالای محاسبه اجتماع و اشتراک لیست ها زمان بیشتری نسبت مشابهت کسینوسی طی میکنیم. در این حالت هم ابتدا لیست امتیاز ژاکار برای همه ی داکیومنت ها آورده شده. همانطور که در نتایج مشاهده میکنیم بدلیل محاسبه امتیاز همه داکیومنت ها جواب درست تری نسبت به مشابهت کسینوسی داریم اما زمان خیلی بیشتری صرف محاسبات میکنیم.

الف)

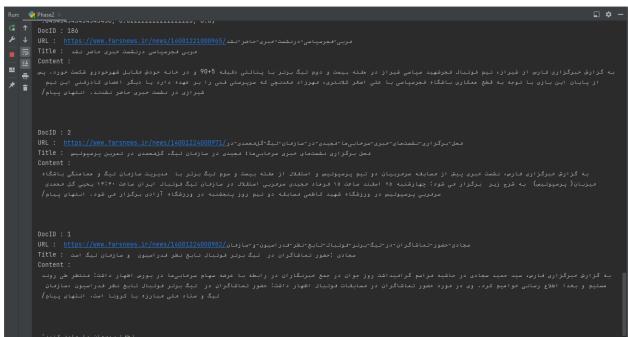
عبارت: "فوتبال"



همانطور که مشاهده میکنیم هر سه خروجی مرتبط هستند.

عبارت: "ليگ برتر فوتبال"





همانطور که مشاهده میکنیم هر سه خروجی مرتبط هستند.

پ)

عبارت: "مايكل"





تفاوت این حالت شباهت ژاکارد و کسینوسی این است که برای ژاکارد فقط یک امتیاز محاسبه شده و بقیه خروجی را به ترتیب از داکیومنت و چاپ میکنیم اما در مشابهت کسینوسی داکیومنت های دیگر را از نظر امتیازی بررسی کردیم و فقط بر اساس ترتیب شماره داکیومنت نمایش نمیدهیم.

ت)

عبارت: "مايكل جردن آمريكا"







در این حالت به دلیل اضافه شدن عبارت آمریکا خروجی منطقی تری داریم و امتیاز های بیشتری محاسبه شده ،

– ۲

ب)

عبارت: "ليگ برتر فوتبال"







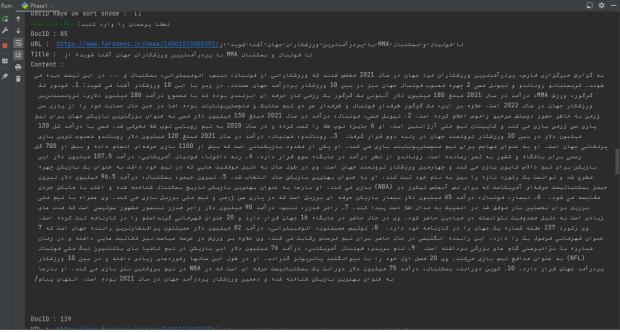




در این حالت چونکه وجود کلمات بصورت AND در داکیومنت را بررسی میکنیم خروجی منطقی تری میدهد چونکه وجود همزمان کلمات در یک داکیونت را بررسی میکند اما در دو روش دیگر صرفا امتیاز هر کلمه بررسی شده و حضور همزمان همه کلمات اهمیت خاصی ندارد . حالت مکانی برای عبارت یا phrase ها مناسب تر است.

ت)

عبارت: "مايكل جردن آمريكا"











خروجی این حالت به دلیل نادر بودن کویری فرق چندانی با دو حالت قبلی نمیکند و هر سه نتایج ضعیفی ارایه میدهند.