سید پوریا احمدی 9723002

گزارش فاز اول

سوال 1)

استخراج توکن :

در این مرحله یک جمله را در نظر میگیریم، حال آن جمله را به کلمه های با معنی آن که سیستم بفهمد باید جمله را به کلمه های با معنی split می کنیم.

مثال :

['مسلم', 'مجدمی', 'در', 'گفت\u200cوگو', 'با', 'خبرنگار', 'ورزشی', 'خبرگزاری', 'فارس', 'در', 'مورد', 'شکست', '۳', 'بر', 'صفر', 'نفمسجدسلیمان', 'مقابل', 'سپاهان', 'اظهار', 'داشت']

نرمال سازی : در کل برای استاندارد سازی متن به کار می رود (به طور مثال حروف کوچک و بزرگ) تا از تصادفی بودن آن کم کند و باعث می شود که اطلاعات اضافه ای که با آن ها سروکار داریم تا حد زیادی کاهش پیدا کند و به efficiency کمک کند.

مثال : چ

'مسلم', 'مجدمی', 'در', 'گفت\u200cوگو', 'با', 'خبرنگار', 'ورزشی', 'خبرگزاری', 'فارس', 'در', 'مورد', 'شکست', '۳', 'بر', 'صفر', 'نفمسجدسلیمان', 'مقابل', 'سپاهان', 'اظهار', 'داشت']

حذف کلمات پر تکرار:

در ابتدا حجم زیادی از دیکشنری ما کاهش پیدا می کند همچنین با حذف این کلمات اطلاعات بی ارزش تر را حذف میکنیم تا بتوان روی اطلاعات اصلی تمرکز داشت.

مثال :

['مسلم', 'مجدمی', 'گفت\u200cوگو', 'خبرنگار', 'ورزشی', 'خبرگزاری', 'فارس', 'شکست', '۳', 'صفر', 'نفمسجدسلیمان', 'مقابل', 'سپاهان', 'اظهار']

ریشه یابی :

Lemmatization معمولاً به انجام درست کارها با استفاده از واژگان و تجزیه و تحلیل صرفی کلمات اشاره دارد که معمولاً با هدف حذف پایان های عطفی و برگرداندن شکل پایه یا فرهنگ لغت یک کلمه که به لم معروف است.

['مسلم', 'مجدمی', 'گفت\u200cوگو', 'خبرنگار', 'ورزش', 'خبرگزاری', 'فارس', 'شکست', '۳', 'صفر',
'نفمسجدسلیمان', 'مقابل', 'سپاهان', 'اظهار']

سوال2)

با توجه به اینکه این قانون با توجه به کلمات پرتکرار بیان شده و برای stopword ها برقرار است (مثلا کلمه to و the) در زبان انگلیسی پر تکرار ترین ها هستند و این قانون در صورت حذف نکردن آنها برقرار است اما در صورت حذف کلمات پر تکرار، کلماتی که بیشترین تکرار را دارند اختلاف frequency بسیار کمتری با حالت قبلی دارند. قبل از حذف کلمات پر تکرار این قانون بهتر برقرار بود و بهتر روی نمودارش fit می شود اما بعد از حذف آن دچار نقض هایی می شود.

سوال3)

اعداد k =42 و b=0.48 برای دو حالت انتخاب شد و روی مقادیر بدست آمد هر چه جلوتر می رویم پیش بینی بد تر می شود شد.

در حالت بدون ریشه یابی برای کل داده ها:

تعداد ترم ها = 1865128

سایز کلمات = 70043

با استفاده از اعداد بالا مقدار پیش بینی شده : 42972

در حالت با ریشه یابی برای کل داده ها:

تعداد ترم ها = 1865128

سایز کلمات = 48428

با استفاده از اعداد بالا مقدار پیش بینی شده : 42972

دیده می شود در حالتی که ریشه یابی انجام می شود این قانون بهتر fit می شود.

سوال4)

به سه موردی که برخورد شد و به مشکل خوردم:

خوزستان تبدیل به خوز شد، بازی تبدیل شد به باز ، تیم تبدیل شد به ت.

این موارد باعث از دست رفتن اطلاعات زیادی می شوند به خصوص مواردی که کم تکرار هستند.

در کل خیلی از کلمات ی آخر حذف می شود که ممکن است به طور کلی اطلاعات زیادی را از دست بدهیم. مثل بازی

به طور کلماتی که با ان تمام می شوند را حذف میکند مثل خوزستان

سوال 5)

بین‌الملل

بله با پرسمان کاربر به طور کامل منطبق هست و طبیعی می باشد زیر کلمه ای متداول است(هر ده جواب برگردانده شده بین‌الملل را دارند و به نوعی به کلمه بین‌الملل مرتبطند (به علت تکرار زیاد جملات آورده نشد)

```
enter your query: بین‌الملل
0
و ماجرای قرارداد با شرکت اسرائیلی VAR واکنش تند فدراسیون فوتبال به حواشی
==============
1
توضیحات مسؤول مسابقات لیگ یک درباره شایعه سکته ناظر بازی
==============
2
داور بین‌المللی هندی مسلمان شد +عکس
==============
3
جوابیه باشگاه آلومینیوم نسبت به محرومیت 2 بازیکن ملی‌پوش
==============
4
اعلام زمان انتخابات کمیته ملی المپیک/ صالحی‌امیری: اختلافی با فدراسیون فوتبال نداریم/ دنبال انتقام از فدراسیونی نیستیم
==============
5
دستگیری چند معترض در یونان/مخالفت با المپیک پکن همچنان ادامه دارد
==============
6
انتقاد عجیب هندبالی‌ها از پروفسور/فرناندز مدرس است یا مربی؟
==============
7
آمادگی قوه قضائیه برای حمایت حقوقی و معنوی از ورزشکاران/ اژه‌ای: به دعاوی حقوقی ورزشی به صورت تخصصی رسیدگی خواهد شد
==============
8
اسکی صاحب کرسی جهانی شد
==============
9
چرا کمیته بین‌المللی المپیک مخالف برگزاری جام جهانی فوتبال هر 2 سال یکبار است؟
==============
```

دانشگاه امیر کبیر

-1

نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید

جمله های دارای امیرکبیر:

28. فرهاد رحمتی

هیات علمی دانشگاه امیرکبیر،معاون سابق مالی اداری دانشگاه صنعتی امیر کبیر و

سابقه مدیریت در وزات عتف

به طور کلی به پرسمان کاربر مربوط نیست

-2

نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور

جمله های دارای امیرکبیر:
امضاکنندگان نامه:

بسیج دانشگاه تهران
بسیج دانشگاه صنعتی شریف
بسیج دانشگاه امیرکبیر
بسیج دانشگاه علم و صنعت
بسیج دانشگاه شهید بهشتی
بسیج دانشگاه خواجه نصیر الدین طوسی
بسیج دانشگاه عالمه طباطبایی (ره)

تا حدی مرتبط است

-3

بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
جمله دارای امیرکبیر:

1. جمعیت افغانستانی‌های مقیم ایران

2. بسیج دانشجویی دانشگاه صنعتی شریف

3. بسیج دانشجویی دانشگاه شهید بهشتی(ره)

4. بسیج دانشجویی دانشگاه امیرکبیر

5. بسیج دانشجویی دانشگاه مالک اشتر

6. بسیج دانشجویی مدرسه عالی شهید مطهری

7. موسسه رسانه‌ای ناصرین

8. موسسه بین‌المللی معبر

تا حدی مرتبط است

-4

امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است
جمله ها :

مسئول نهاد رهبری در دانشگاه امیرکبیر گفت: ما در دوران دفاع مقدس حدود ۲۲۶ هزار شهید دادیم این شهدا از طیف‌های مختلف مردم هستند. یک بخشی از آنها هم از دانشجویان بودند و بیش از چند هزار نفر شهید دانشگاهی داریم که حدود نود شهید از دانشگاه امیرکبیر هستند که در عملیات‌های مختلف به شهادت رسیدند.

مسئول نهاد رهبری در دانشگاه امیرکبیر گفت: ما به برکت این شهدا امروز حمله نظامی

نداریم، امروز دیگر کسی جرات حمله به سرزمین ایران را ندارد. البته این به معنای

این نیست که جنگ تمام شده، امروز عرصه جنگ تفاوت پیدا کرده و ما امروز شاهد جنگ

اقتصادی، جنگ فرهنگی و... هستیم.

مرتبط به دانشگاه امیرکبیر

-5

باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
جمله ها :

وی افزود: بیش از ۹۰ شهید دوران دفاع مقدس از دانشگاه امیرکبیر بودند، یعی در زمانی که دانشگاه حدود ۱۲۰۰ دانشجو داشت دانشگاهیان ایین دانشگاه این گونه در خدمت جبهه و جنگ بودند.

کاملا مرتبط و خبر کلا درباره دانشگاه امیر کبیر است

از این مرحله موارد دیگر دانشگاه امیر کبیر ندارند و فقط درباره دانشگاه جستجو انجام شده که فقط تیتر خبر آورده شده:

-6

دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی

-7

تجلیل دانشگاه آزاد از قهرمانان المپیک و کشتی‌گیران مدال‌آور+ تصاویر

-8

یزدانی: فکر می‌کردم امتیاز فینال جهانی را به من می‌دهند/سبک حریف روسی تغییر کرده بود

-9

میزبانان لیگ کشتی معرفی شدند/ زمان شروع مسابقات مشخص شد

-10

انتقاد عجیب هندبالی‌ها از پروفسور/فرناندز مدرس است یا مربی؟

همه مواردی که فقط دانشگاه داشتند نا مربوط بودند

```
enter your query: دانشگاه امیرکبیر
0
نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید
==============
1
نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور
==============
2
بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
==============
3
امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است
==============
4
باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
==============
5
دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی
==============
6
تجلیل دانشگاه آزاد از  قهرمانان المپیک و کشتی‌گیران مدال‌آور+ تصاویر
==============
7
یزدانی: فکر می‌کردم امتیاز فینال جهانی را به من می‌دهند/سبک حریف روسی تغییر کرده بود
==============
8
میزبانان لیگ کشتی معرفی شدند/ زمان شروع مسابقات مشخص شد
==============
9
انتقاد عجیب هندبالی‌ها از پروفسور/فرناندز مدرس است یا مربی؟
==============
```

دانشگاه صنعتی امیرکبیر

-1

باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد

جمله:

ه گزارش خبرنگار تشکل‌های دانشگاهی خبرگزاری فارس، سیداحمد معتمدی رئیس [دانشگاه
در (https: //search. farsnews. ir/?q=دانشگاه صنعتی امیرکبیر&o=on) [صنعتی امیرکبیر
نشستی که به همت نهاد نمایندگی مقام معظم رهبری در دانشگاه‌ها به مناسبت فرارسیدن
هفته دفاع مقدس برگزار شد درباره نقش آموزش عالی در حفظ و نگهداری خاطرات دوران
دفاع مقدس و یاد و خاطره شهدای جنگ تحمیلی گفت: این موضوع خیلی مهم هست که بدانیم
در تاریخ انقلاب و (https: //www. farsnews. ir/special/دانشگاهیان) [دانشگاهیان]
دوران دفاع مقدس چه نقشی داشتند و جدای از آن وظیفه اصلی که از آنها انتظار

مرتبط

-2

دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام

جمله:
** :اصلاحات مربوط به دانشگاه صنعتی امیرکبیر**

تا حدی مرتبط

در همه موارد 3 تا 9 دانشگاه صنعتی برگردانده شده و در مورد آخر دانشگاه امیر کبیر تیتر ها در زیر قابل مشاهده است:

```
enter your query: دانشگاه صنعتی امیرکبیر
0
باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
==============
1
دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
==============
2
دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
==============
3
سیدمحسن دهنوی عضو هیئت امنای صندوق نوآوری و شکوفایی شد
==============
4
وزیر علوم: علم و عقل دو بال دانایی است/ علم باید برای جامعه ثروت‌آفرین باشد
==============
5
حجت‌الاسلام رستمی فقدان فعال دانشجویی دانشگاه شریف را تسلیت گفت
==============
6
سیدرضا مرتضوی و مهدی دوستی استانداران اصفهان و هرمزگان شدند
==============
7
بیانیه دانشجویان شمال‌غرب کشور در محکومیت اظهارات مقامات جمهوری آذربایجان
==============
8
پیام وزیر علوم در پی شهادت دانشجوی جهادی دانشگاه خواجه نصیر
==============
9
امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است
==============
```

که مرتبط نیستند به جز مورد آخر

ژیمناستیک

8 مورد برگردانده شد که همه مرتبطند

اما به علت موارد زیاد تکرار در متن فقط تیتر آورده شده که در شکل زیر قابل مشاهده است.

```
enter your query: ژیمناستیک
0
خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیت‌ها بیشتر شد
==============
1
هشدار هیات ژیمناستیک تهران در خصوص سالن‌های مختلط و اقدامات غیراخلاقی
==============
2
سوتزنی| تذکر و برخورد با سالن‌های ورزشی مختلط توسط وزارت ورزش
==============
3
دبیر مجمع فدراسیون ژیمناستیک مشخص شد
==============
4
ثبت نام ۱۳ نامزد برای پست ریاست فدراسیون ژیمناستیک + اسامی
==============
5
جزییات تعطیلی ورزش ایران تا پایان تیرماه+ تصویر
==============
6
!دبیر: اگر من در مباحث فنی ۱۰ باشم، درستکار ۱۰۰ است/ بنا کاملا بر اساس چرخه انتخابی عمل کرد
==============
7
جزییات تعطیلی‌های ورزش ایران تا ۹ مهر ۱۴۰۰/ کدام فعالیت‌های ورزشی در تهران ممنوع است؟
==============
```

واکسن آسترازنکا

1-و 2-

محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد

دو بار این خبر در دادهها بوده است

جمله ها:

مهرداد جمال ارونقی در گفت‌وگو با ایسنا تصریح کرد: محموله جدید واکسن کرونا ساعت
از مبداء ایتالیا وارد فرودگاه امام خمینی شد که حاوی یک میلیون و ۴۰۰ هزار ۱۰
دوز واکسن آسترازنکا از مجموعه کوواکس است
مرتبط است

-3

مهم‌ترین سلاح مبارزه با کرونا

جمله:

علاوه بر مورد زیر دو بار دیگر نیز تکرار شده

افرازی ادامه داد: انواع واکسن‌های تجاری کووید -۱۹ غیرقابل تعویض است و فرد**
بعد از دریافت نوبت اول واکسن حتما باید از همان نوع واکسن در نوبت دوم استفاده
نماید. فاصله دوز اول و دوم واکسن در برندهای مختلف متفاوت است که در مورد
واکسن‌های اسپوتنیک (فاصله بین دو نوبت ۲۸ روز است که در شرایط خاص امکان
فاصله‌گذاری بین ۲۱ تا ۹۰ روز هم امکان‌پذیر است.)، کووکسین و سینوفارم ۲۸ روز و
* .برای واکسن آسترازنکا ۱۲ هفته است
مرتبط است

-4

در این مورد کلمه مانند بین واکسن و آسترازنکا هست که با حذف کلمات پر تکرار در سند ها این مورد هم به ما برگردانده می شود

واکسیناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسیناسیون

جمله:

واکسنی مانند آسترازنکا هم از مکانیزم وکتورهای ادنوویروس استفاده
.می‌کند. هیچکدام از این واکسن‌ها باعث ابتلا به کرونا نمی‌شوند

تا حدی مرتبط

بقیه موارد نیز به همین صورت است به جز مورد آخر که فقط واکسن را دارد و آسترازنکا ندارد که مرتبط نیست

لیست تیتر ها:

```
enter your query: واکسن آسترازنکا
0
محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
==============
1
محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
==============
2
مهم‌ترین سلاح مبارزه با کرونا
==============
3
واکسیناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسیناسیون
==============
4
نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
==============
5
واکسیناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسیناسیون
==============
6
نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
==============
7
واکسن‌های کرونا با چه داروهایی تداخل دارند؟
==============
8
امکان ایجاد لخته خون در واکسن  آسترازنکا چقدر است؟
==============
9
تارتار: 3 امتیاز باارزش در آبادان بدست آوردیم/نفت این فصل از سال گذشته بهتر است
==============
```