

پیش پردازش های انجام شده در قطعه کد زیر صورت می گیرد:

```
68     for i in range(len(dataset)): #deploy phase
69         # for i in range(2): #test phase
70             #preprocess
71             .....normalized_content = preprocessor.fnormilizer(dataset.loc[i].content)
72             .....tokens = preprocessor.tokenizer(normalized_content)
73             .....tokens = preprocessor.filter_stopwords(tokens)
74             .....for j in range(len(tokens)):
75             .....tokens[j] = preprocessor.lemmetizer_and_stemmer(tokens[j])
76             #indexing
77             positional_index.add_and_merge_content(tokens,i)
78             file_to_write = open("positional_index.pickle", "wb")
79             pickle.dump(positional_index, file_to_write)
```

برای اینکار از کتابخانه hazm استفاده شده است. در مرحله اول متن را نرمال سازی می کنیم. نرمال سازی برای این می باشد که در سراسر متن به طور مثال نیم فاصله رعایت شود و به یک شکل باشد و این مرحله کلمه ای مانند "پیش نویس" را به "پیش نویس" تبدیل می کند. اگر این کار انجام نشود ممکن است یک جا "پیش نویس" به صورت جدا نوشته شده باشد و جزو کلمات مرتبط به حساب نیاید زیرا که "پیش نویس" در اصل با نیم فاصله می باشد.

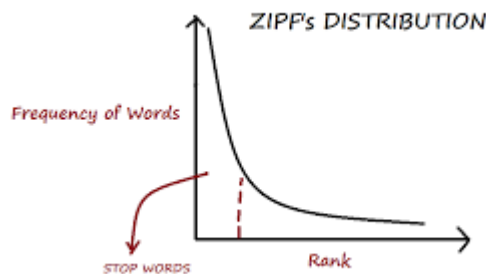
در مرحله بعدی توکن سازی انجام می دهیم تا متن به صورت کلمه کلمه در آید. این مرحله برای ساخت شاخص مکانی کاربرد دارد. در این مرحله جمله "من علی هستم" به توکن های "من" و "علی" و "هستم" تبدیل می شود.

در مرحله بعد کلمات پرتکرار زبان فارسی از متن حذف می شود. این اقدام از این جهت صورت می گیرد که چنین کلماتی تاثیری در مرتبط بودن محتوی متن ندارند و پردازش آن ها صرفا بار اضافی ای بر سیستم می باشد. در این مرحله کلماتی مانند "در"، "و"، "،"، "و ... حذف می شوند و سرعت ساخت شاخص به شدت افزایش پیدا می کند.

در مرحله آخر ریشه یابی صورت می گیرد. این امر از این جهت می باشد که هنگامی که کاربر کلمه ای مانند "درخت" را جستجو کرد متن هایی با محتوی "درخت ها" نیز به آن ها نمایش داده شود. زیرا که متن حاوی "درخت ها" قطعا به درخت مرتبط است و در صورت عدم ریشه یابی کاربر باید حتما کلمه درخت ها را جستجو کند تا سند به آن نمایش داده شود. همچنین در این مرحله فعل ها به صورت مصدر ماضی خود در می آیند. مانند "می روی" که تبدیل به "رفت" می شود.

-۲

به کمک قانون زیف می توان کلمات پرتکرار با بار اطلاعاتی کم را شناسایی کرد و آن ها را حذف کرد و به این صورت تحلیل های ما روی متن بسیار efficient تر خواهد بود.



بدون حذف کلمات پرتکرار:

طبق تکرار کلمات مرتب شده اند و ۲۰ تای اول آن ها به نمایش در آمده است. قانون زیف می گوید در یک متن طبیعی فراوانی i امین کلمه پرتکرار رابطه عکس با ضریب $1/i$ با پرتکرار ترین کلمه دارد.

```
PS C:\university\Information Retrieval\Project\information-retrieval> python main.py
[(('و', 130231), ('.', 99332), ('د', 95605), ('.', 82101), ('ب', 74502), ('ز', 56178), ('ا', 43182), ('ه', 39418), ('ب', 38300), (':', 35482), ('ن', 34354), ('ر', 33287), ('و', 32153), ('#', 21592), ('', 20294), ('', 20192), ('ر', 16992), ('ا', 15536), ('ب', 14845), ('آ', 13623))]
```

بدون حذف پرتکرار ترین کلمه "و" می باشد با مقدار 130231 و بعد از آن کلمه "در" با 95605 تکرار. طبق قانون زیف باید $k/2 = 95605$ باشد. و همچنین 74502 که مربوط به کلمه پرتکرار بعدی یعنی "به" است باید با $k/3$ برابر باشد. طبق رابطه اول $k=191210$ و طبق رابطه دوم $k=223506$. تقریباً قانون زیف رعایت شده است که علت این امر وجود کلمات پرتکرار می باشد.

با حذف کلمات پرتکرار:

```
ty\Information Retrieval\Project\information-retrieval> python main.py
[(('و', 12956), ('ا', 11213), ('https', 10873), ('farsnews', 9616), ('«', 8892), ('»', 8857), ('ش', 8537), ('م', 8005), ('ه', 7562), ('ز', 5133), ('ن', 5169), ('م', 5252), ('ک', 5268), ('ل', 5364), ('و', 5409), ('ج', 5685), ('ر', 5735), ('ل', 7147))]
```

با حذف کلمات پرتکرار نظم حاکم در متن طبیعی از بین می رود و قانون زیف رعایت نمی شود. در این حالت k برای $i=2$ برابر است با 22426 و برای $i=3$ برابر است با 32000 است. این امر نشان می دهد که ما توانسته ایم درصد زیادی از کلمات که بار اطلاعاتی کمی دارند را حذف کنیم.

-۳

Heaps' law: $M = kT^b$ ■

حالت با ریشه یابی: تعداد ۵۰۰

```
File Edit Selection View Go Run Terminal Help
main.py - information-retrieval - Visual Studio Code

EXPLORER
  INFORMATION-RETRIEVAL
    - $Report.docx
    - WRLO984.tmp
    IR-Project()]-Fall1400.pdf
    IR1_7k_news.xlsx
    main.py
    positional_index.pickle
    README.md
    Report.docx

main.py
165 count = 500
166 for i in range(count): #deploy phase
167     # for i in range(2): #test phase
168     #preprocess
169     normalized_content = preprocessor.fnormilizer(dataset.loc[i].content)
170     tokens = preprocessor.tokenizer(normalized_content)
171     tokens = preprocessor.filter_stopwords(tokens)
172     for j in range(len(tokens)):
173         tokens[j] = preprocessor.lemmetizer_and_stemmer(tokens[j])
174     #indexing
175     positional_index.add_and_merge_content(tokens,i)
176
177 sum =0
178 for key in positional_index.indexes.keys():
179     sum+= positional_index.indexes[key]['count']
180
181 print("sum for tokens for ",count," document is: ",sum," and length of vocabulary(M) is : ", len(list(positional_index.indexes.keys())
182
183 if __name__ == '__main__':
184     # main()
185     # check_zipfs Law()
186     check_heaps_Law()
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE
PS C:\University\Information Retrieval\Project\information-retrieval> python main.py
sum for tokens for 500 document is: 29378 and length of vocabulary(M) is : 7959
PS C:\University\Information Retrieval\Project\information-retrieval>

Ln 184, Col 21 Spaces: 4 UTF-8 CRLF Python Prettier
```

طبق محاسبات : $T=80000$ و $M=8000$

طبق معادله قانون هیپس : $k=30$ و $b=0.5$

حالت با ریشه یابی تعداد ۱۰۰۰:

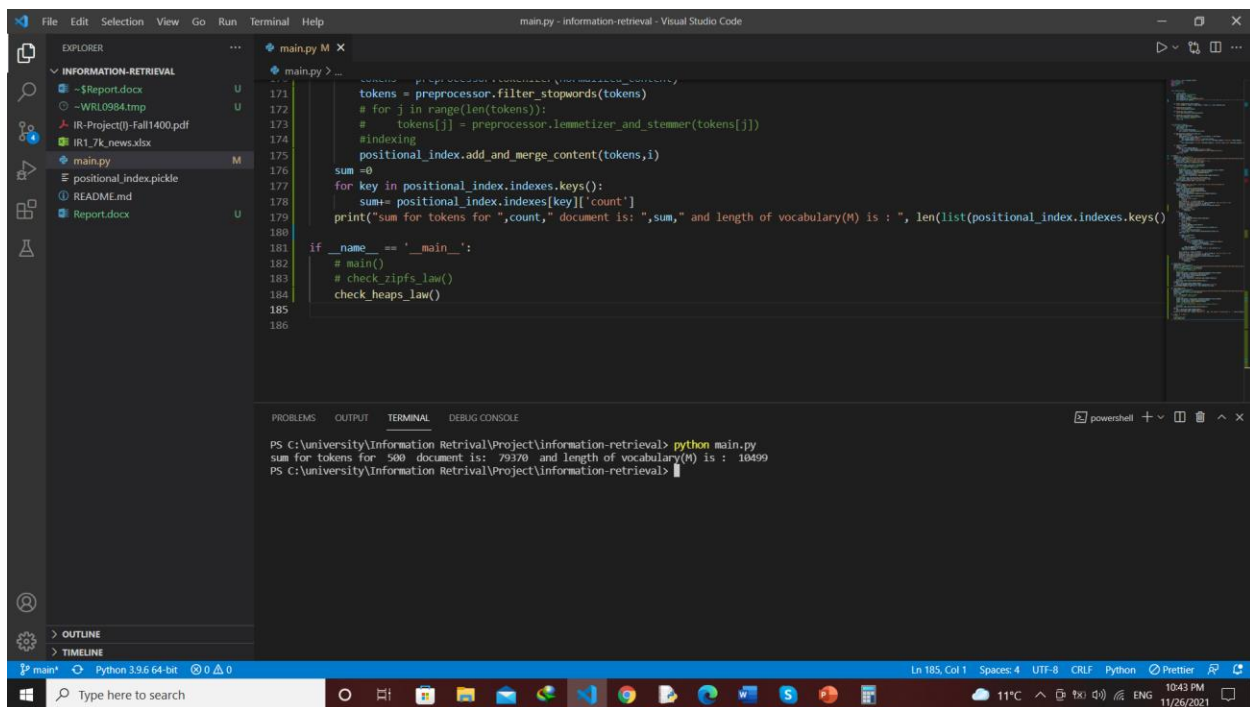
```
157 for key in positional_index.indexes.keys():
158     frequency.append((key, positional_index.indexes[key]['count']))
159 print(sorted(frequency, key=lambda x: x[1], reverse=True)[:20])
160
161 def check_heaps_law():
162     preprocessor = Preprocessor()
163     dataset = pd.read_excel("C:\\university\\Information Retrieval\\Project\\information-retrieval\\IR1_7k_news.xlsx")
164     positional_index = Positional_index(dataset)
165     count = 1000
166     for i in range(count): #deploy phase
167         # for i in range(2): #test phase
168             #preprocess
169             normalized_content = preprocessor.fnormalizer(dataset.loc[i].content)
170             tokens = preprocessor.tokenizer(normalized_content)
171             tokens = preprocessor.filter_stopwords(tokens)
172             for j in range(len(tokens)):
173                 tokens[j] = preprocessor.lemmetizer_and_stemmer(tokens[j])
174             #indexing
175             positional_index.add_and_merge_content(tokens, i)
176             sum += 0
177     for key in positional_index.indexes.keys():
```

PS C:\university\Information Retrieval\Project\information-retrieval> python main.py
sum for tokens for 1000 document is: 156073 and length of vocabulary(M) is : 11253
PS C:\university\Information Retrieval\Project\information-retrieval>

حال تعداد از ۵۰۰ به ۱۰۰۰ تبدیل شده است. تعداد توکن برابر ۱۵۶۰۷۳ می باشد. طبق قانون هیپز با اطلاعات موجود در قسمت قبل $k=30$ و $b=0.5$.
 $kT^b = 30 * 156073^{0.5} = 11850$. تعداد کلمات تخمینی
تعداد کلمات واقعی 11253.

به طور تقریبی میتوان گفت قانون هیپز رعایت شده است و به کمک آن توانستیم تعداد کلمات را تخمین بزنیم.

حالت بدون ریشه یابی:



```
171 tokens = preprocessor.filter_stopwords(tokens)
172 # for j in range(len(tokens)):
173 #     tokens[j] = preprocessor.lemmetizer_and_stemmer(tokens[j])
174 #indexing
175 positional_index.add_and_merge_content(tokens,i)
176
177 sum
178 for key in positional_index.indexes.keys():
179     sum+= positional_index.indexes[key]['count']
180
181 print("sum for tokens for ",count," document is: ",sum," and length of vocabulary(N) is : ", len(list(positional_index.indexes.keys()))
182
183 if __name__ == '__main__':
184     # main()
185     # check_zipfs_law()
186     check_heaps_law()
```

```
PS C:\University\Information Retrieval\Project\Information-retrieval> python main.py
sum for tokens for 500 document is: 29370 and length of vocabulary(N) is : 10499
PS C:\University\Information Retrieval\Project\Information-retrieval>
```

مشاهده می شود تعداد کلمات منحصر به فرد افزایش پیدا کرده است اما همچنان تعداد توکن ها ثابت مانده است. این امر نشان می دهد که بدون ریشه یابی تنها تعداد کلمات منحصر به فرد بیشتر می شود و توکن ها ثابت می ماند که این امر تنها بار حافظه ای ما را زیاد می کند.

۴-

در روند ریشه یابی فعل ها به صورت مصدر ماضی خود در می آیند و به طور مثال فعل "رفتم" به فعل "رفت" تبدیل می شود. در روند ریشه یابی یکی از اقدامات این می باشد که کلماتی که با "م" ختم می شوند، حرف "م" از انتهای آن ها حذف می شود. همچنین در روند ریشه یابی باید فعل "رفتی" به "رفت" تبدیل شود. در نتیجه ناخواسته طبق روند ریشه یابی کلماتی که فعل نیستند و انتهای آن ها حرف "ی" می باشد تغییر می کنند.همچنین فعل های امری مثل فعل "برویم" که به "رفت" تبدیل می شود.

نمونه ها:

سلامتی — سلامت

مسلمی - مسلم

تلاطم - تلاط

بہنام - ہنام

برای جلوگیری از این اتفاق باید قبل از ریشه یابی کلمات ، برچسب گذاری صورت گیرد. به این صورت اگر کلمه ای فعل باشد ریشه یابی صورت می گیرد اما اگر اسم باشد اتفاقی نمی افتد.

۵- از آنجایی که سند های مرتبط تعدادشان زیاد است تنها مرتبط ترین سند نمایش داده شده است. در صورت نیاز می توان لیست تمام سند های مرتب اولویت بندی شده را چاپ کرد.

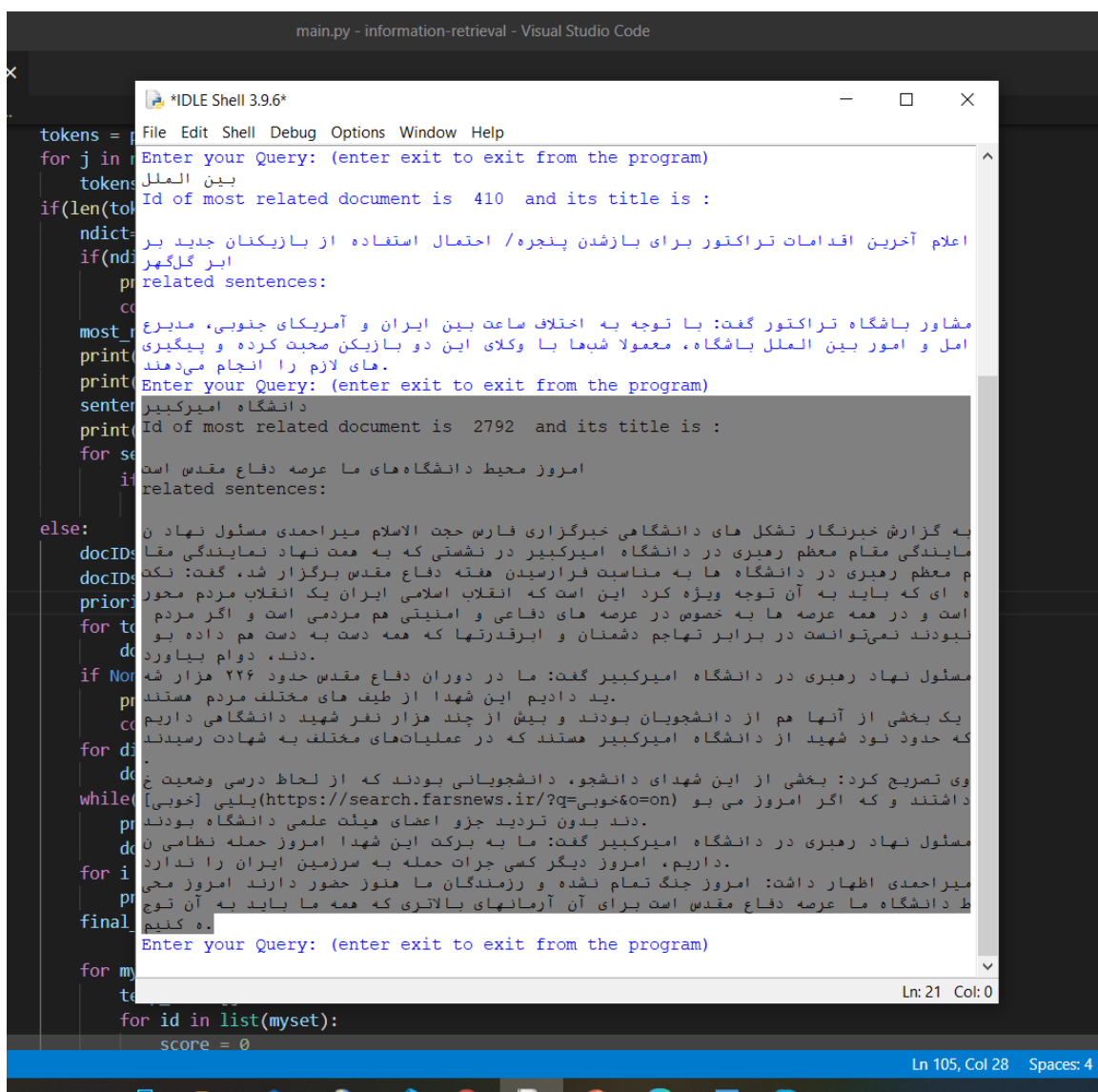
خروجی ها:

(الف) خروجی برای یک پیرسمان متداول تک کلمه ای: بین الملل

```
*IDLE Shell 3.9.6*
File Edit Shell Debug Options Window Help
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> exec(open("C:\\university\\Information Retrieval\\Project\\information-retrieval\\main.py").read())
Enter your Query: (enter exit to exit from the program)
اردک
Id of most related document is 5499 and its title is :
توصیه‌های غذایی در ماه رمضان و در شرایط پاندمی کرونا
related sentences:
گروه گوشت و تخم مرغ: این گروه منبع تامین پروتئین حیوانی و با کیفیت بالا هستند و شامل گوشت قرمز، مرغ، ماهی، میگو، بلدرچین، غاز، اردک، تخم مرغ و تخم سایر پرندگان هستند.
Enter your Query: (enter exit to exit from the program)
بین الملل
Id of most related document is 410 and its title is :
اعلام آخرین اقدامات تراکتور برای بازشدن پنجره / احتمال استفاده از بازیکنان جدید بر ابر گلگیر
related sentences:
مشاور باشگاه تراکتور گفت: با توجه به اختلاف ساعت بین ایران و آمریکای جنوبی، مدیرعامل و امور بین الملل باشگاه، معمولاً شبها با وکلای این دو بازیکن صحبت کرده و پیگیری های لازم را انجام می دهند.
Enter your Query: (enter exit to exit from the program)
```

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که در متن امر بین الملل آمده است که با بین الملل مرتبط است.

ب) متداول دو کلمه ای: دانشگاه امیرکبیر



```
tokens = [...]
for j in range(len(tokens)):
    tokens[j] = tokens[j].lower()
    tokens[j] = tokens[j].replace(' ', '')
    tokens[j] = tokens[j].replace('.', '')
    tokens[j] = tokens[j].replace('-', '')
    tokens[j] = tokens[j].replace('_', '')
    tokens[j] = tokens[j].replace('!', '')
    tokens[j] = tokens[j].replace('@', '')
    tokens[j] = tokens[j].replace('#', '')
    tokens[j] = tokens[j].replace('$', '')
    tokens[j] = tokens[j].replace('%', '')
    tokens[j] = tokens[j].replace('&', '')
    tokens[j] = tokens[j].replace("'", '')
    tokens[j] = tokens[j].replace('"', '')
    tokens[j] = tokens[j].replace('(', '')
    tokens[j] = tokens[j].replace(')', '')
    tokens[j] = tokens[j].replace('[', '')
    tokens[j] = tokens[j].replace(']', '')
    tokens[j] = tokens[j].replace('{', '')
    tokens[j] = tokens[j].replace('}', '')
    tokens[j] = tokens[j].replace(':', '')
    tokens[j] = tokens[j].replace(';')
    tokens[j] = tokens[j].replace(',')
```

File Edit Shell Debug Options Window Help

Enter your Query: (enter exit to exit from the program)

بین الملل

Id of most related document is 410 and its title is :

اعلام آخرین اقدامات تراکتور برای بازشدن پنجره/ احتمال استفاده از بازیکنان جدید بر

ابر گلگیر

related sentences:

مشاور باشگاه تراکتور گفت: با توجه به اختلاف ساعت بین ایران و آمریکای جنوبی، مدیر

امل و امور بین الملل باشگاه، معمولاً شبها با وکلای این دو بازیکن صحبت کرده و پیگیری

های لازم را انجام می‌دهند.

Enter your Query: (enter exit to exit from the program)

دانشگاه امیرکبیر

Id of most related document is 2792 and its title is :

امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است

related sentences:

به گزارش خبرنگار تشکل های دانشگاهی خبرگزاری فارس حجت الاسلام میراحمدی مسئول نهاد ن

مایندگی مقام معظم رهبری در دانشگاه امیرکبیر در نشستی که به همت نهاد نمایندگی مقام

م معظم رهبری در دانشگاه ها به مناسبت فرارسیدن هفته دفاع مقدس برگزار شد، گفت: نکت

ه ای که باید به آن توجه ویژه کرد این است که انقلاب اسلامی ایران یک انقلاب مردم محور

است و در همه عرصه ها به خصوص در عرصه های دفاعی و امنیتی هم مردمی است و اگر مردم

نی بودند نمی‌توانست در برابر تهاجم دشمنان و ابرقدرتها که همه دست به دست هم داده بو

دند، دوام بیاورند.

مسئول نهاد رهبری در دانشگاه امیرکبیر گفت: ما در دوران دفاع مقدس حدود ۲۲۶ هزار شه

ید دادیم این شهدا از طیف های 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

وی تصریح کرد: بخشی از این شهدای دانشجوی، دانشجویانی بودند که از لحاظ درسی وضعیت خ

داشتند و که اگر امروز می بو (https://search.farsnews.ir/?q=خوبی&o=on) یلی [خوبی]

دند بدون تردید جزو اعضای هیئت علمی دانشگاه بودند.

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

داریم، امروز دیگر کسی جرات حمله به سرزمین ایران را ندارد

میراحمدی اظهار داشت: امروز جنگ تمام نشده و رزمندگان ما هنوز حضور دارند امروز مح

ط دانشگاه ما عرصه دفاع مقدس است برای آن آرمانهای بالاتری که همه ما باید به آن توج

به کنیم.

Enter your Query: (enter exit to exit from the program)

Ln: 21 Col: 0

Ln 105, Col 28 Spaces: 4

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که در متن مسئول نهاد رهبری در دانشگاه امیرکبیر صحبتی کرده است و این مورد به دانشگاه امیرکبیر مرتبط است. و ترتیب رعایت شده است.

پ) متداول چندکلمه ای : جمهوری اسلامی ایران

```
main.py - information-retrieval - Visual Studio Code

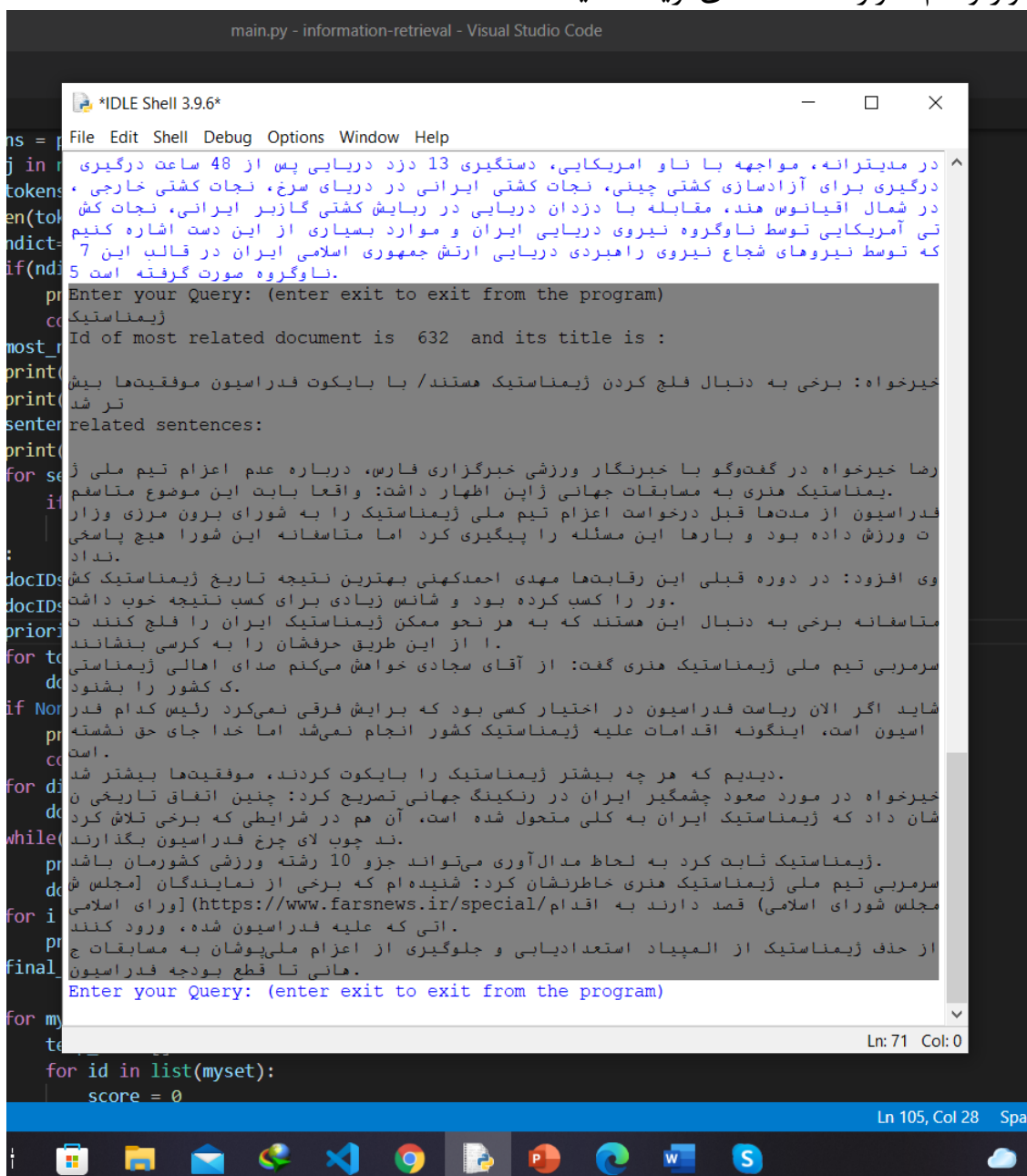
*IDLE Shell 3.9.6*
File Edit Shell Debug Options Window Help

Enter your Query: (enter exit to exit from the program)
جمهوری اسلامی ایران
Id of most related document is 3051 and its title is :
تمام اقیانوس-ها زیر عرشه شناورهای ایرانی/ امتداد «مکران» تا سواحل اروپا
related sentences:
به گزارش گروه دفاعی خبرگزاری فارس به نقل از روزنامه ایران، به تازگی ناوگروه 75 نی
ارتش جمهوری (https://www.farsnews.ir/special/روای دریایی [ارتش جمهوری اسلامی ایران]
.اسلامی ایران) از مأموریت بازگشت؛ مأموریتی که از سواحل جنوبی ایران آغاز شد
بیش از 12 سال است که از اعزام اولین ناوگروه ایرانی به آبهای آزاد می-گذرد و در طی
این سالها 75 ناوگروه عملیاتی نیروی دریایی ارتش جمهوری اسلامی برای مقابله با دزدا
ن دریایی و مأموریتهای دیگر به آبهای بین المللی رفته و پرچم سه رنگ جمهوری اسلامی
ایران را به امتداد درآوردند.
از سوی دیگر جمهوری اسلامی ایران از طریق دریای عمان به 20 درصد اقیانوسهای دنیا اشر
اف دارد و میتواند از این طریق ارتباط گسترده-ای با دیگر کشورهای جهان برقرار کند ک
ه این مستلزم این است که توسعه توان دریایی و بنادر مورد توجه قرار گیرد
منطقه خلیج عدن و دریای سرخ یکی از مهمترین خطوط مواصلاتی کشتیهای نفتکش و باری ایرا
نی بود که با شیوع پدیده دزدی دریایی با مشکلاتی مواجه شده بود لذا نیروی دریایی ارت
ش جمهوری اسلامی ایران برای مقابله با این موضوع وارد میدان و فاز عملیاتی شد
بی شک این حضور مقتدرانه نیروی دریایی ارتش تعجب و حتی تحسین کشورهای قدرتمند دریای
ی را نیز در پی داشت زیرا آنان می-دانند که این حضور شناورها و سامانه های ایرانی د
تحریمهای سن (https://search.farsnews.ir/?q=رغم&o=on) ر آبهای دوردست [علی رغم]
.گین در حوزه نظامی، نشان از توانمندی و قدرت دریایی ارتش جمهوری اسلامی ایران دارد
در طول این 12 سال و اعزام 75 ناوگروه، شناورهای ارتش جمهوری اسلامی ایران علاوه بر د
ریانوردی در همه اقیانوس-های کره زمین از جمله اقیانوس-های هند، آرام و اطلس، در مه
ترین گذرگاه-های آبی جهان نیز از جمله تنگه باب المندب، مالاکا، کانال سوئز، دماغه
... امید نیک و
حضور بی نقص شناورهای ایرانی در این آبها که برای نخستین بار صورت میگردت مورد توجه
رسانه های جهان قرار گرفت و پیام روشنی به دوستان و دشمنان جمهوری اسلامی ایران بود
و دست بلند و توانمند ارتش را در آبهای دوردست نمایان کرد بگونه ای که ناوهای ایرا
چندین کشور قرار گرفتند و مورد بحث ک لی در این مسیر مورد عکس برداری واحدهای نظامی
Ln: 46 Col: 285

Ln 105, Col 28 Spa
```

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که متن در ارتباط با جمهوری اسلامی ایران و ارتش آن می باشد و تعداد زیادی از این پرسمان با همان ترتیب در متن تکرار شده است و سیستم به خوبی توانسته این متن را در اولویت قرار دهد.

ت) دشوار و کم تکرار تک کلمه ای: ژیمناستیک



```
main.py - information-retrieval - Visual Studio Code

*IDLE Shell 3.9.6*
File Edit Shell Debug Options Window Help

Enter your Query: (enter exit to exit from the program)
ژیمناستیک
Id of most related document is 632 and its title is :
خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیتها بیش
تر شد
related sentences:
رضا خیرخواه در گفتوگو با خبرنگار ورزشی خبرگزاری فارس، درباره عدم اعزام تیم ملی ژ
یمناستیک هنری به مسابقات جهانی ژاپن اظهار داشت: واقعا بابت این موضوع متاسفم
فدراسیون از مدتها قبل درخواست اعزام تیم ملی ژیمناستیک را به شورای برون مرزی وزار
ت ورزش داده بود و بارها این مسئله را پیگیری کرد اما متاسفانه این شورا هیچ پاسخی
نداد.
وی افزود: در دوره قبلی این رقابتها مهدی احمدکهنی بهترین نتیجه تاریخ ژیمناستیک کش
ور را کسب کرده بود و شانس زیادی برای کسب نتیجه خوب داشت.
متاسفانه برخی به دنبال این هستند که به هر نحو ممکن ژیمناستیک ایران را فلج کنند ت
ا از این طریق حرفشان را به کرسی بنشانند.
سرمربی تیم ملی ژیمناستیک هنری گفت: از آقای سجادی خواهش میکنم صدای اهالی ژیمناستی
ک کشور را بشنود.
شاید اگر الان ریاست فدراسیون در اختیار کسی بود که برایش فرقی نمیکرد رئیس کدام فدر
اسیون است، اینگونه اقدامات علیه ژیمناستیک کشور انجام نمیشد اما خدا جای حق نشسته
است.
دیدیم که هر چه بیشتر ژیمناستیک را بایکوت کردند، موفقیتها بیشتر شد.
خیرخواه در مورد صعود چشمگیر ایران در رنکینگ جهانی تصریح کرد: چنین اتفاق تاریخی ن
شان داد که ژیمناستیک ایران به کلی متحول شده است، آن هم در شرایطی که برخی تلاش کرد
ند چوب لای چرخ فدراسیون بگذارند.
ژیمناستیک ثابت کرد به لحاظ مدال آوری میتواند جزو 10 رشته ورزشی کشورمان باشد.
سرمربی تیم ملی ژیمناستیک هنری خاطرنشان کرد: شنیده ام که برخی از نمایندگان [مجلس ش
مجلس شورای اسلامی] قصد دارند به اقدام (https://www.farsnews.ir/special/ [ورای اسلامی
اتی که علیه فدراسیون شده، ورود کنند.
از حذف ژیمناستیک از المپیاد استعدادیابی و جلوگیری از اعزام ملیپوشان به مسابقات ج
مانی تا قطع بودجه فدراسیون
Enter your Query: (enter exit to exit from the program)

Ln: 71 Col: 0

Ln 105, Col 28 Spac
```

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. متن کاملاً در ارتباط با ورزش ژیمناستیک می باشد.

ث) دشوار کم تکرار چند کلمه ای: واکسن آسترزنکا

```
main.py - information-retrieval - Visual Studio Code

*IDLE Shell 3.9.6*
File Edit Shell Debug Options Window Help
شان داد که ژیمناستیک ایران به کلی متحول شده است، آن هم در شرایطی که برخی تلاش کردند چوب لای چرخ فدراسیون بگذارند. ژیمناستیک ثابت کرد به لحاظ مدال آوری می تواند جزو 10 رشته ورزشی کشورمان باشد. سرمربی تیم ملی ژیمناستیک هنری خاطرنشان کرد: شنیده ام که برخی از نمایندگان [مجلس شورای اسلامی] قصد دارند به اقدام (https://www.farsnews.ir/special/ورای اسلامی) اتی که علیه فدراسیون شده، ورود کنند از حذف ژیمناستیک از المپاد استعدادیابی و جلوگیری از اعزام ملی پوشان به مسابقات جهانی تا قطع بودجه فدراسیون.
Enter your Query: (enter exit to exit from the program)
واکسن آسترزنکا
Id of most related document is 5825 and its title is :
نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
related sentences:
اما سوالی که معمولاً در چنین شرایطی برای افراد پیش می آید این است که بهترین واکسن ساخته شده برای پیشگیری از بیماری کووید-۱۹ چه واکسنی است؟
بهترین واکسن برای پیشگیری از بیماری کووید-۱۹ در دسترس ترین واکسن است میزان اثربخشی واکسن ها در پیشگیری از ابتلا به بیماری کووید-۱۹ در مطالعات انجام شده ممکن است تا حدی متفاوت باشد.
این تفاوت ها نباید در تصمیم گیری جامعه برای زمان تزریق واکسن تأثیرگذار باشد با تزریق همه واکسن ها ممکن است عوارض جدی به صورت نادر بروز کند، اما عواقب ابتلا به بیماری کووید-۱۹ از عوارض احتمالی و نادر واکسن ها بسیار بیشتر است و واکسیناسیون یک امر ضروری برای حفظ سلامت هر فرد و جامعه است.
اما، درباره زمان اثربخشی واکسن و زمان ماندگاری آن باید گفت ایمنی زایی واکسن دو تا سه هفته بعد از تزریق نوبت اول آغاز می شود و دو هفته بعد از تزریق نوبت دوم کامل می شود.
بنابراین ریسک ابتلا به بیماری کووید-۱۹ در روزهای ابتدایی بعد از تزریق واکسن وجود دارد و نباید نادیده گرفته شود.
ماندگاری ایمنی در برابر ویروس کرونا بعد از تزریق واکسن های کووید-۱۹ مشخص نیست، اما انتظار می رود دو هفته بعد از کامل شدن واکسیناسیون، فرد واکسینه شده حداقل شش ماه در برابر ویروس کرونا ایمنی داشته باشد.
**میزان اثر بخشی واکسن های کووید-۱۹ چقدر است؟*
همه واکسن های کووید-۱۹ به طور قابل توجهی ریسک تشدید بیماری و مرگ و میر ناشی از بیماری کووید-۱۹ را در صورت ابتلا کاهش می دهند.
این واکسن ها ریسک ابتلا (به صورت ناقل یا علائم خفیف و یا بدون علامت) به بیماری کووید-۱۹ را نیز کاهش می دهند، با این وجود اطلاعات در مورد میزان اثربخشی واکسن های کووید-۱۹ در برابر سوش های نوپدید ویروس کرونا محدود است.
**آیا واکسن های کووید-۱۹ موجود در بازار دارویی ایران محدودیت سنی دارد؟*
بله می تواند تزریق شود همه واکسن ها در کشور ما در سنین بالای ۱۸ سال

Ln: 90 Col: 44
```

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. متن در رابطه با واکسیانسون در زمان کرونا می باشد که واکسن آسترازنکا نیز یکی از این واکسن ها می باشد و کاملاً مرتبط است و در متن با همان ترتیب تکرار شده است.