گزارش کار پروژه درس بازیابی اطلاعات
سارا سید عزیزی – ریحانه باقری

فاز اول:
برای شروع ابتدا با استفاده از دستور pip install میایم hazm رو نصب می کنیم. بعد باید فایل json رو اپلود میکنیم:

```
[ ]  from google.colab import files
     uploaded = files.upload()
```

Choose Files IR_data_ne…12k (1).json

- **IR_data_news_12k (1).json**(application/json) - 131550996 bytes, last modified: 1/24/2023 - 100% done

```
Saving IR_data_news_12k (1).json to IR_data_news_12k (1).json
```

میایم لودش میکنم و به دیتافریم تبدیلش میکنیم که بتونیم راحت تر باهاش کار کنیم.

```
[ ]  f = open('IR_data_news_12k (1).json')
```

```
[ ]  data = json.load(f)
```

get json object as a dictionary

```
[ ]  f.close()
```

The close() method closes an open file. You should always close your files, in some cases, due to buffering, changes made to a file may not show until you close the file.

```
df = pd.DataFrame(data).T
df.to_csv('IR_data_news_12k (1).json', sep='\t', index = False)
```

بعدم یه head زدم ببینم درست نمایش میده یا نه:

```
df.head(10)
```

	title	content	tags	date	url	category
0	اعلام زمان قرعه کشی جام باشگاه های فوتسال آسیا	به گزارش خبرگزاری فارس، کنفدراسیون فوتبال\n... اس	اعلام زمان, قرعه‌کشی, قرعه‌کشی جام,] ... قرعه‌کشی	3/15/2022 5:59:27 PM	https://www.farsnews.ir/news/14001224001005/...ا	sports
1	سجادی :حضور تماشاگران در لیگ برتر فوتبال ...تایع	به گزارش خبرگزاری فارس، سید حمید سجادی\n ...در حا	[سجادی, لیگ, فدراسیون, وزیر ورزش]	3/15/2022 5:30:07 PM	https://www.farsnews.ir/news/14001224000982/...سج	sports
2	محل برگزاری نشست‌های خبری سرخابی‌ها؛ ... مجیدی در	به گزارش خبرگزاری فارس، نشست خبری پیش\n ...از مسا	دربی 94, محل برگزاری, خبری سرخابی‌ها,] ... ,مجیدی	3/15/2022 5:20:01 PM	https://www.farsnews.ir/news/14001224000971/...مج	sports
3	ماجدی در نشست با صالحی امیری: امیدوارم ... در این	به گزارش خبرگزاری فارس، سید رضا\n ...صالحی امیری	کمیته امداد امام خمینی (ره), کمیته ملی ...المپیک	3/15/2022 5:18:00 PM	https://www.farsnews.ir/news/14001224000964/...ما	sports
4	لیگبرتر بسکتبال| نخستین پیروزی شهرداری ...گرگان	به گزارش خبرنگار ورزشی خبرگزاری فارس،\n ...در نخس	بسکتبال, لیگ برتر بسکتبال, شهرداری گرگان,] ...تیم	3/15/2022 5:16:41 PM	https://www.farsnews.ir/news/14001224000947/...لی	sports
5	مسابقات تنیس روی میز فیدر قطر| هر4 بانوی ...ملی پ	به گزارش خبرگزاری فارس، چهار بانوی تنیس\n ... روی	مهشید اشتری, تنیس روی میز, ملی پوشان] ...تنیس روی	3/15/2022 5:15:34 PM	https://www.farsnews.ir/news/14001224000916/...مس	sports
6	اعلام برنامه نشست خبری گل محمدی/ مجیدی ...هم باید	به گزارش خبرگزاری فارس و به نقل از سایت\n ...باش	باشگاه استقلال, یحیی گل محمدی, فرهاد ...مجیدی, ن	3/15/2022 3:57:47 PM	https://www.farsnews.ir/news/14001224000865/...ا	sports
7	احضار مدیران پرسپولیس به کمیته انضباطی ...پیش از	به گزارش خبرنگار ورزشی خبرگزاری فارس،\n ... کمیته	باشگاه پرسپولیس, احضار مدیران پرسپولیس] ...به کمی	3/15/2022 3:43:18 PM	https://www.farsnews.ir/news/14001224000842/...اح	sports
8	مدیر پیدستاورد، رئیس شد/ روزگار تلخ‌تر از ...تلخ	به گزارش خبرنگار ورزشی خبرگزاری فارس،\n ...مجمع ا	والیبال, پیمان رضایی, تیم والیبال شهرداری] ...ارو	3/15/2022 3:35:21 PM	https://www.farsnews.ir/news/14001224000811/...مد	sports
9	خبر خوب برای استقلال؛ دانشگر با تیم تمرین کرد	به گزارش خبرنگار ورزشی خبرگزاری فارس،\n ... تمرین	دانشگر با تیم تمرین کرد, محمد دانشگر,] [استقلال	3/15/2022 3:21:04 PM	https://www.farsnews.ir/news/14001224000819/...خب	sports

حالا باید متد های بخش پیش پردازش رو تعریف کنیم:

```
[ ]  def normalize_text(text):
       normalizer = Normalizer()
       return normalizer.normalize(text)


     def remove_stopwords(text, stopword_list):
       stopwords_dict = Counter(stopword_list)
       result_text = ' '.join([word for word in text.split() if word not in stopwords_dict])
       return result_text

     def tokenize_text(text):
       return word_tokenize(text)


     def stem_text(text):
       stemmer = Stemmer()
       return stemmer.stem(text)
```

برای هر کدوم تو کامنتای کد مثال زدیم که واضح باشه، در کل بخوایم تعریف مختصری از هر کدوم داشته باشیم:

متد نرمالایزر:
مثلا توی متون فارسی گاهی بعضی از کلمات با ی فارسی و بعضی کلمات با ئ نوشته میشوند، میایم با نرمالایزر اینارو یکی میکنیم.

متد برای حذف stopped words :
اینجا ما اومدیم یک لیست از این واژه های پرتکرار و یه سری علامتا البته به جز اونایی که تو کوئری زدن استفاده میشه رو از یک گیت فچ کردیم و بعد با متدش اومدیم از اسنادمون حذفشون کردیم. جزئیات بیشتر کد هم پایین میذاریم.

توکنایزر:
برای توکن توکن کردن تمام محتوای اسناد است، کلمه کلمه میشه و میریزمشون تو یه لیست جدید.

متد برای stem کردن:
اینجا میایم ریشه یابی میکنیم، مثلا رفتم با رفتش یه ریشه دارند، میایم کلمه رو به ریشه میرسونیم.

خب اول برای نرمالایز کردن طبق تعریفی که بالا کردم:

```
[ ]  content_text = list(df['content'])
     normal_text = [None] * len(content_text)

     for i, text in enumerate(content_text):
       normal_text[i] = normalize_text(text)
```

اینجا میایم از بخش content داده هارو نرمالایز میکنیم و تو یه لیست جدید میریزیم. ما فقط با همین بخش content کار کردیم.
بعد برای پاک کردن اون stopped words ها میایم یه متد مینویسیم که فچ کنه، بعدا بهش ورودی میدیم:

```
def fetch_stopwords(user, repo_name, path_in_repo):


    url = f'https://api.github.com/repos/{user}/{repo_name}/contents/{path_in_repo}'
    req = requests.get(url)
    if req.status_code == requests.codes.ok:
      req = req.json()  # the response is a JSON
      content = base64.b64decode(req['content']).decode("utf-8").split()


    return content
```

```
[ ]  stop_words = fetch_stopwords(user ='ziaa', repo_name = 'Persian-stopwords-collection',
                                   path_in_repo = 'Stopwords/Mojiry/PersianStopWords.txt')
     stop_words.remove('!')
     stop_words.remove('"')
```

خب بعدم که ورودیاشو دادیم ازش استفاده میکنیم:

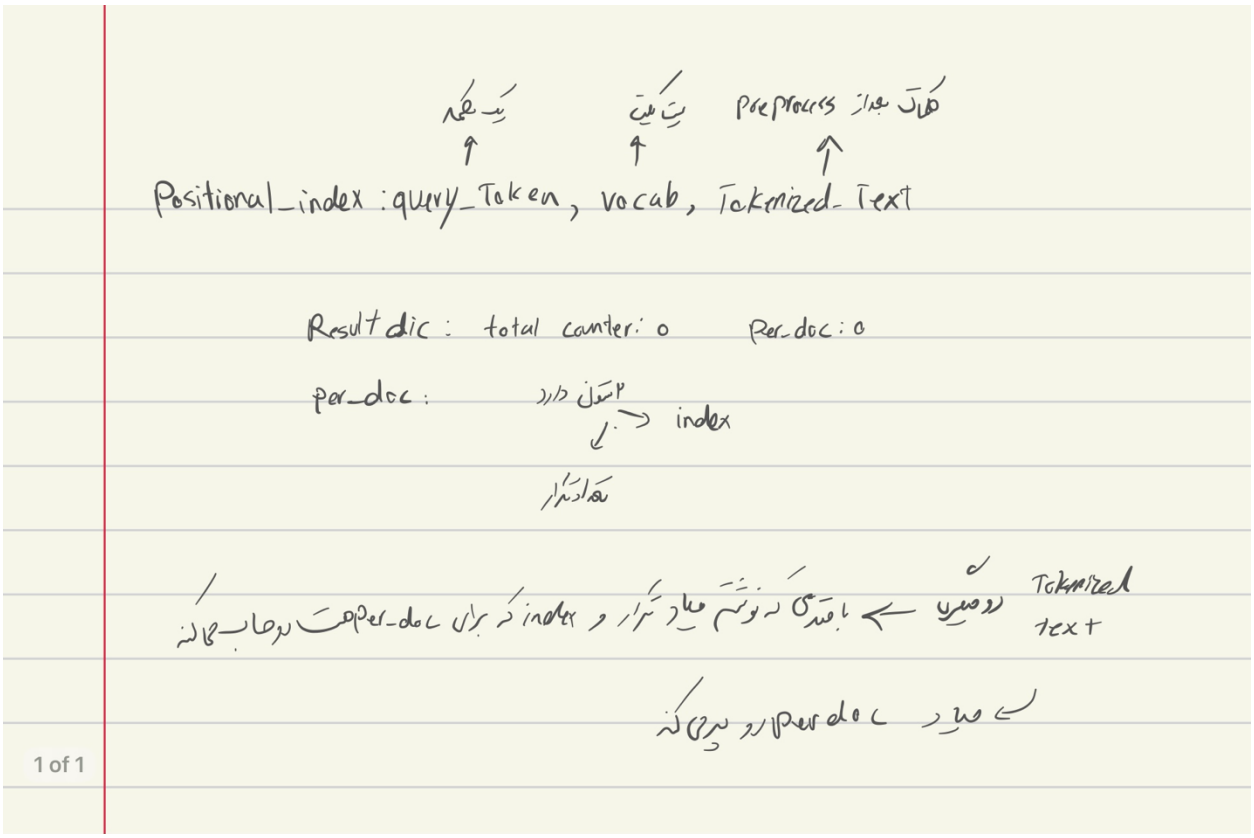
```
stemmed_tokenized_text = copy.deepcopy(tokenized_text)

for i, token_list in enumerate(stemmed_tokenized_text):
  for j, token in enumerate(token_list):
    stemmed_tokenized_text[i][j] = stem_text(token)
```

بعد یه متد زدیم برای اینکه بعدا ازش استفاده کنیم، عملا میاد یکتا میکنه، یعنی یه لیست داریم که کلمات به شکل یکتا داخلش هستن:

```
def build_vocabulary(tokenized_text):
  vocab = []
  for doc_tokens in tokenized_text:
    doc_vocab = set(doc_tokens)
    if len(vocab) != 0:
      vocab.update(doc_vocab)
    else:
      vocab = doc_vocab
  return list(vocab)
```

حالا یه فانکشن داریم که تعداد هارو توش هندل میکنیم:




```python
def positional_index(query_token, vocab, tokenized_text):
  if query_token not in vocab:
    print(f'"{query_token}" is out of vocabulary')
    return None

  else:
    result_dict = {'total_occ':0, 'per_doc_occ':None}
    per_doc_occ = pd.DataFrame(columns = ['occ_nums', 'indices'])

    for i, token_list in enumerate(tokenized_text):
      occ_nums, indices = find_occurences(query_token, token_list)
      per_doc_occ.loc[i] = [occ_nums, indices]

    result_dict['per_doc_occ'] = per_doc_occ
    result_dict['total_occ'] = sum(per_doc_occ['occ_nums'])

  return result_dict
```

واسه اینکه فهمش اسون تر باشه بالا نوشتم دقیقا به چه شکل هست و چیا بر میگردونه.
همون مراحل پیش پردازش رو بعد برای کوئری هم میزنیم. داستان اینه که این واسه ریحانه گاهی این فچ کار نمیکرد واسه همین یه لیست stopped word هم تعریف کردیم مشکل نباشه. تو کوئری توی یه سری از بخشا از اون یکی استفاده کردیم.

حالا قراره ریز شیم رو بخش کوئری:

برای اینکه بتونیم بعدا حالات مختلف مثل and و not و ... رو هندل کنیم میایم چند تا فانکشن اول مینویسیم. خب فانکشن اول میاد چک میکنه ببینه کدوم داک ها حاوی اون کلمه ای که پاس میدیم هستن، فانشکن دوم هم دقیقا عکسش عمل میکنه، اونایی که نیستن:

```
[26] def find_containing_docs(p_index):
       return np.array(p_index['per_doc_occ'][p_index['per_doc_occ']['occ_nums'] > 0].index)
```

Find not containing documents

```
def find_not_containing_docs(p_index):
  return np.array(p_index['per_doc_occ'][p_index['per_doc_occ']['occ_nums'] == 0].index)
```

این تابع لیستی از اندیسهای مربوط به اسناد حاوی چند کلمه مختلف را دریافت و اسنادی رو که حاوی همه آن کلمات باشند برمیگردونه. این تابع برای زمانی که کاربر یک عبارت حاوی چند کلمه را داخل "" جستجو میکنه کاربردیه.

Find intersect

```
def find_intersect(iter_of_iters):
  common_elements = set(iter_of_iters[0])
  for item in iter_of_iters:
    common_elements = common_elements.intersection(item)
  return np.array(list(common_elements))
```

intersection return similarity between sets.

```
find_intersect([[2,5,1200],[5,1200,1402],[18,1200]])
```

```
[ ] def find_difference(iter1, iter2):
      return [x for x in iter1 if x not in iter2]
```

Find co-occurences

```
[ ] def find_co_occurences(p_index_list, common_docs):
      accepted_docs = []

      # for every document that has ALL the search terms
      for doc_id in common_docs:

        # where each of them occurred in the doc
        doc_indices = [p_index['per_doc_occ'].iloc[doc_id]['indices'] for p_index in p_index_list]

        # the indices of the first query word occurence in doc
        first_term_idx = doc_indices[0]

        # Now check if the other words happened right after it
        term_counter = 1
        for idx in first_term_idx:
          for i, idx_list in enumerate(doc_indices[1:]):
            if (idx + i + 1) in idx_list:
              term_counter += 1

        if term_counter == len(p_index_list):
          accepted_docs.append(doc_id)

      return accepted_docs
```

این یکی تابع، difference در واقع میاد and not میکنه، اولیو داشته باشه دومیو نه. بعد هم تابع find co میاد اون اسنادیو پیدا میکنه که ترتیب هم دقیق رعایت میکنند:

حالا ۳ تا عملگر داریم، یکی and یکی or و یکی not. برای هر کدوم هم علامتش رو مشخص کردیم. اول and رو پیاده سازی میکنیم:

AND

```
def find_best_results_AND(query, vocab = vocab, tokenized_text = stemmed_tokenized_text, stop_words = stop_words):
  # q_tokens = query.split()

  q_tokens = preprocess_query(query, stop_words)[1:-1]
  q_pindex = [None]*len(q_tokens)
  containing_docs = [None]*len(q_tokens)
  common_docs = []

  for i, qt in enumerate(q_tokens):
    # populate positional index for every query term
    q_pindex[i] = positional_index(qt, vocab, tokenized_text)

    # indicies of documents in which the query terms occurred
    containing_docs[i] = find_containing_docs(q_pindex[i])

  common_docs = find_intersect(containing_docs)
  top_docs = find_co_occurences(q_pindex, common_docs)
  return top_docs
```

صرفا از تمامی تابع هایی که بالا توضیح دادیم میایم استفاده میکنیم. برای هر ترم ایندکسشو در میاریم، با استفاده از تابعی که بالاتر نوشتیم داک هایی ک حاوی ترم ها هستنو در میاریم، با اینترسکت میایم اونایی که حاوی همه کلمات هستند رو در میاریم و بعد هم با اخرین تابعی که نوشتیم دقیقا اونایی ک طبق ترتیب کوئری هستند رو استخراج میکنیم.

تابع بعدی:

OR

```
def find_best_results_OR(query, vocab = vocab, tokenized_text = stemmed_tokenized_text, stop_words = stop_words):
  # q_tokens = query.split()
  q_tokens = preprocess_query(query, stop_words)
  q_pindex = [None]*len(q_tokens)
  containing_docs = []
  occs = []

  for i, qt in enumerate(q_tokens):
    q_pindex[i] = positional_index(qt, vocab, tokenized_text)
    cont_docs = np.array(find_containing_docs(q_pindex[i]))
    containing_docs.extend(cont_docs)
    occs.extend(q_pindex[i]['per_doc_occ']['occ_nums'].iloc[cont_docs])

  occs = np.array(occs)
  containing_docs = np.array(containing_docs)
  idx = (-occs).argsort()
  return list(containing_docs[idx])
```

اینجا دیگه مثل بالا لازم نیست همرو داشته باشه اما سورت میکنیم که اگر داشت اول نشونش بده.
برای بخش آخر هم صرفا از not_containing_docs استفاده میکنیم تا اختلاف رو بدست بیاریم و حذف کنیم.

جواب سوالات فاز ۱:

سوال ۱: این سوال رو در صفحات قبل به طور کامل توضیح دادیم و دلایلش رو شرح دادیم. ما توی پیش پردازش اومدیم ۴ تا کار انجام دادیم. اول اینکه اومدیم یه نرمال سازی انجام دادیم، طبق مثالی ک زدیم مثلا ی فارسی با ئ رو یکی میکنیم تو عملیات نرمال سازی. برای اینکار از hazm استفاده کردیم. بعد اومدیم یه سری کلمات و نشانه های خاص رو حذف کردیم. دو تا تکنیک به ذهنمون رسید، یکی همینکه لیستشون کنیم دستی و بعد از تو داک ها حذف کنیم و نتیجه رو بریزیم تو یه لیست جدید، یکی ام اینکه بیایم فچ کنیم به یه ریسورز که خودش اومده دسته بندی و مشخص کرده این stopped word ها رو. بعد اومدیم توکن توکنش کردیم و در نهایت ام ریشه یابی کردیم. مثلا کشاورزان و کشاورزی جفتش میشه کشاورز.

سوال ۲:

نداریم

سوال ۳:

نداریم

سوال ۴:

خیلی جاها مشکل این رو داشتیم که ت اخر فعل ها حذف میشد. یکسری از م ها هم که نیاز بود باشه حذف میشدند اشتباها. در کل ما که با hazm کار کردیم عمدتا با فعل ها مشکل داشتیم مثلا است میشد اس.

سوال ۵:

```
user_query = 'تحریم آمریکا علیه ایران'
show_results(user_query)
```

```
--------------------------------------------------

https://www.farsnews.ir/news/14000801000437/نامه-جمعی-از-اساتید-و-متخصصان-آقای-رئیس‌جمهور-در-گام-دوم-انقلاب-به
نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید

--------------------------------------------------

https://www.farsnews.ir/news/14001124000522/مصاحبه-فارس-با-کارشناس-فوتبال-آسیا|-از-میراث-بزرگ-کی‌روش-و-قدرت-ایران
مصاحبه فارس با کارشناس فوتبال آسیا| از میراث بزرگ کی‌روش و قدرت ایران با اسکوچیچ تا انقلاب برانکو در عمان

--------------------------------------------------

https://www.farsnews.ir/news/14001030000671/رییس-جمهور-در-مسکو-از-تاکید-پوتین-بر-توسعه-روابط-با-ایران-تا-تشویق
رییس جمهور در مسکو/ از تاکید پوتین بر توسعه روابط با ایران تا تشویق رییسی در دومای روسیه

--------------------------------------------------

https://www.farsnews.ir/news/14001018000637/مانور-موشکی-ایران-در-عین‌الاسد-گزارشی-از-اثر-بازدارندگی-موشک‌های
مانور موشکی ایران در عین‌الاسد/ گزارشی از اثر بازدارندگی موشک‌های ایرانی

--------------------------------------------------

https://www.farsnews.ir/news/14000820000224/شهید-طهرانی-مقدم-و-قدرت-بازدارندگی-ایران-پس-از-انقلاب-مواجهه-موشکی-با
شهید طهرانی مقدم و قدرت بازدارندگی ایران پس از انقلاب/ مواجهه موشکی با قدرت‌های جهانی

--------------------------------------------------

https://www.farsnews.ir/news/14001215000516/میزگرد-تلاطم-در-اوکراین|-ائتلاف-اقتصادی-حول-محور-دلار-فرو-می‌پاشد
میزگرد «تلاطم در اوکراین»| ائتلاف اقتصادی حول محور دلار فرو می‌پاشد

--------------------------------------------------

https://www.farsnews.ir/news/14001208000502/دستاورد-سفرهای-خارجی-رئیسی-و-غفلت-ساله-روحانی-از-همسایگان
دستاورد سفرهای خارجی رئیسی و غفلت 8 ساله روحانی از همسایگان
```

```
user_query = 'آمریکا !ایران'
show_results(user_query)
```

```
https://www.farsnews.ir/news/14001223000292/تا-فوتبال-و-بسکتبال-MMA-با-پردرآمدترین-ورزشکاران-جهان-آشنا-شوید-از
تا فوتبال و بسکتبال MMA با پردرآمدترین ورزشکاران جهان آشنا شوید؛ از

------------------------------------------------

https://www.farsnews.ir/news/14001221000109/بمب-خبری-بازگشت-مسی-به-بارسلونا-خنثی-شد-پای-جام-جهانی-درمیان-است
بمب خبری بازگشت مسی به بارسلونا خنثی شد/پای جام جهانی درمیان است

------------------------------------------------

https://www.farsnews.ir/news/14001220000373/منچستریونایتد-به-دنبال-بهترین-بازیکن-دنیا--زوج-رونالدو--لواندوفسکی
منچستریونایتد به دنبال بهترین بازیکن دنیا / زوج «رونالدو - لواندوفسکی» در اولدترافورد +عکس

------------------------------------------------

https://www.farsnews.ir/news/14001218000297/حضور-بازیکن-اوکراینی-در-لیگ-ایران-پس-از-جنگ-با-روسیه
حضور بازیکن اوکراینی در لیگ ایران پس از جنگ با روسیه

------------------------------------------------

https://www.farsnews.ir/news/14001217000994/راهیابی-بانوی-سنگنورد-کشورمان-به-مسابقات-جهانی-آمریکا
راهیابی بانوی سنگنورد کشورمان به مسابقات جهانی آمریکا

------------------------------------------------

https://www.farsnews.ir/news/14001214000887/مدرس-جهانی-تنیس-روی-میز-صحبت‌های-عیلقارداشی-بیشتر-شوخی-است-تا-جدی-در
مدرس جهانی تنیس روی میز: صحبت‌های عیلقارداشی بیشتر شوخی است تا جدی/در همه زمینه‌ها سقوط کردیم

------------------------------------------------

https://www.farsnews.ir/news/14001214000964/لیگبرتر-بسکتبال|مهرام-در-غیاب-کاپیتان-و-بازیکن-آمریکایی-پیروز-شد
لیگبرتر بسکتبال|مهرام در غیاب کاپیتان و بازیکن آمریکایی پیروز شد/دومین باخت پیاپی مس
```

```
user_query = '"کنگره ضدتروریست"'
show_results(user_query)
```

```
https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین
توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

------------------------------------------------
```

```
user_query = ' اورشلیم !صهیونیست'
show_results(user_query)
```

```
"اورشل" is out of vocabulary
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-37-3aaa6d7ac266> in <module>
      1 user_query = ' اورشلیم !صهیونیست'
----> 2 show_results(user_query)

 2 frames
<ipython-input-26-2e2098648110> in find_containing_docs(p_index)
      1 def find_containing_docs(p_index):
----> 2   return np.array(p_index['per_doc_occ'][p_index['per_doc_occ']['occ_nums'] > 0].index)

TypeError: 'NoneType' object is not subscriptable

SEARCH STACK OVERFLOW
```

این اورشلیم کلا نیستش اصلا برای همین نمیاره.

فاز ۲:
در این فاز اسناد را به شکل برداری درآورده سپس با استفاده از تابع کسینوس و تعریف آن مشابهت با کوئری را بررسی میکنیم
و امتیازات را محاسبه میکنیم و در گام بعد نیز از championlist استفاده کرده ایم :
الف) یک پرسمان از کلمات ساده و متداول تک کلمه ای
به عنوان مثال کلمه فوتسال
ابتدا در دیکشنری:زمان پاسخ 3 ثانیه

```
rank: 1
url
https://www.farsnews.ir/news/14001224001005/اعلام-زمان-قرعه-کشی-جام-باشگاه-های-فوتسال-آسیا
title
اعلام زمان قرعه کشی جام باشگاه های فوتسال آسیا
content

که هر سه دوره به فینال مسابقات راه پیدا کرده و یک عنوان قهرمانی و دو مقام دومی بدست آورده است. انتهای پیام (AFC) به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا



rank: 2
url
https://www.farsnews.ir/news/14001224000800/تورنمنت-شش-جانبه-تایلند|-پیروزی-پر-گل-فوتسالیست‌های-جوان-ایران-برابر
title
تورنمنت شش جانبه تایلند| پیروزی پر گل  فوتسالیست‌های  جوان ایران برابر مغولستان
content

یر ۱۹ سال کشورمان که در این تورنمنت حضور دارد در اولین بازی خود امروز  به مصاف مغولستان رفت و توانست این تیم را با نتیجه ۸ بر صفر شکست دهد. انتهای پیام



rank: 3
url
https://www.farsnews.ir/news/14001223001099/قرارداد-شمسایی-و-اعضای-کادر-فنی-تیم-ملی-فوتسال-نهایی-شد
title
قرارداد شمسایی و اعضای کادر فنی تیم ملی فوتسال نهایی شد
content

اعضا قرار گرفت و بر همین اساس شب گذشته این قرارداد از سوی ذبحسابی و اعضای هیات رییسه فدراسیون نهایی و  جهت ابلاغ به سرمربی تیم ملی ارسال شد. انتهای پیام
```

✓ 3s completed at 5:25 AM

سپس در لیست :champion

```
rank: 1
url
https://www.farsnews.ir/news/14001207000286/تشریح-جزئیات-دریافت-مجوز-حرفه‌ای-برای-تیم‌های-فوتسال-و-تشکیل-کمیته
title
تشریح جزئیات دریافت مجوز حرفه‌ای برای تیم‌های فوتسال و تشکیل کمیته ویژه در هفته‌های پایانی  لیگ
content

رزشی خبرگزاری فارس، آرش جابری درباره اینکه آیا بحث لزوم حرفه‌ای سازی باشگاه‌های فوتسال جدی است، اظهار داشت: هرچند مکتوب تاکنون چیزی به ما ابلاغ نشده اما قطعا



rank: 2
url
https://www.farsnews.ir/news/14001119001101/تصمیمات-سازمان-لیگ-فوتسال-راهگشا-یا-بلای-جان
title
تصمیمات سازمان لیگ فوتسال راهگشا یا بلای جان؟
content

تصمیماتی که در فوتسال گرفته می شود بی توجهی زیادی می شود و سازمان لیگ فوتسال با تصمیمات یک نفره موجب انتقادات خیلی از اهالی این رشته شده است.  انتهای پیام
```

✓ 0s completed at 5:28 AM

تحلیل:در حالتی که از لیست چمپیون استفاده شده ، زمان کمتری صرف شده و همچنین با توجه به متن ، ارتباط بیشتری با کوئری دارند
فوتسال در اسنادی که طول بیشتری داشته اند ، دارای termfrequency بالا تری است و در لیست چمپیون در اوایل قرار میگیرد
منتها سرعت پاسخ آن بیشتر بوده است.
ب)
یک پرسمان از عبارات ساده و متداول چند کلمه ای
مثل فوتسال آسیا

```
rank: 1
url
https://www.farsnews.ir/news/14001029000320/حسن-زاده-فدراسیون-فوتبال-شمسایی-و-تیم-ملی-را-حمایت-کند-فعلا-قصد
title
حسن زاده: فدراسیون فوتبال شمسایی و تیم ملی را حمایت کند/ فعلا قصد خداحافظی ندارم
content

ند خداحافظی ندارم. من تا جایی که بتوانم به باشگاهی که در آن حضور دارم و تیم ملی  کمک می کنم و همیشه مثل یک سرباز برای تیم ملی بازی کردم و شرایط بدنیم خوب و مناسب است.  انتهای پیام



rank: 2
url
https://www.farsnews.ir/news/14001224001005/اعلام-زمان-قرعه-کشی-جام-باشگاه-های-فوتسال-آسیا
title
اعلام زمان قرعه کشی جام باشگاه های فوتسال آسیا
content
```

0s completed at 5:32 AM

در این حالت ، اسنادی که دارای طول کمتری هستند در این رنک قرار میگیرند
و امتیاز بیشتری را دارا هستند .
در رنک های آخر اسناد طولانی تری قرار گرفتند و شامل تمام کلمات کوئری هستند.
سرعت در جمپیون لیست بهتره

پ) یک پرسمان دشوار و کم تکرار تک کلمه ای
کلمه واکسن در دیکشنری

```
print(df['content'].iloc[totalCosine[i][0]] , end='\n')
```

```
rank: 1
url
https://www.farsnews.ir/news/14001223000837/معاون-وزارت-بهداشت-خبر-داد-فقط-دربی-تهران-با-تماشاگر-نیست
title
معاون وزارت بهداشت خبر داد:فقط دربی تهران با تماشاگر نیست
content

نی آمده، می‌توانند کنار هم بنشینند و مشکلی ندارد چون آن‌ها با هم زندگی می‌کنند. منوط به اینکه واکسن کامل (2دوز) زده باشند و بیمار هم نباشند. انتهای پیام



rank: 2
url
https://www.farsnews.ir/news/14001221000302/-گفت-و-گو-با-ساکت-ترین-مربی-ایران|-حسینی-با-جدایی-گل‌محمدی-برزخی-شدم
title
گفت و گو با ساکت ترین مربی ایران| حسینی: با جدایی گل‌محمدی برزخی شدم/ قبول نکردم در پرسپولیس تست بدهم
content

ی بهتر، همراه با سلامتی و اتفاقات خوب باشد. مردم خوب ما لیاقت بهترین ها و زندگی بهتر را دارند. فیلم این گفت و گو را در زیر ببینید:      انتهای پیام



rank: 3
url
https://www.farsnews.ir/news/14001220000209/مهدی-حضور-تماشاگران-در-دربی-مصوبه-اولیه-شده-است-5-30-زمان-مناسبی
title
مهدی: حضور تماشاگران در دربی مصوبه اولیه شده است/ 15:30 زمان مناسبی برای برگزاری بازی نیست
content

ست که ساعت 15:30 ساعت مناسبی نیست. درخواست ما این است که برگزاری بازی به سمت عصر برود و امیدواریم این مساعدت صورت بگیرد. انتهای پیام
```

0s completed at 5:50 AM

کلمه واکسن در لیست champion

```
rank: 1
url
https://www.farsnews.ir/news/14001016000298/عضو-هیئت-علمی-دانشگاه-علوم-پزشکی-اصرار-برای-وارد-کردن-واکسن-خاص
title
عضو هیئت علمی دانشگاه علوم پزشکی: اصرار برای وارد کردن واکسن خاص، منطقی نبود
content

، نیز داشتند، گفت: اینکه عده‌ای خواهان واردکردن این واکسن‌ها بودند رانمی‌توان رد کرد اما، هیچ وقت نمی‌توانیم بگوییم که واکسن فایزر یا مدرنا به شکل قطعی از فوت افراد جلوگیری می‌کند. انتهای پیام



rank: 2
url
https://www.farsnews.ir/news/14000914000784/رئیسی-گلایه‌های-مردم-را-بحق-می‌دانم-ماهانه-۱۰-هزار-میلیارد-تومان-از
title
رئیسی: گلایه‌های مردم را بحق می‌دانم/ ماهانه ۱۰ هزار میلیارد تومان از فاکتورهای دولت قبل را پرداخت می‌کنیم
content

خواهیم کرد، تصریح کرد: ساعتی نیست که از دغدغه های مردم غافل شوم، مبالغه هم نمی‌کنم. از لحظه ای که از بعد از نماز صبح کار را آغاز می‌کنم دغدغه های مردم را پیگیری می‌کنم. انتهای پیام



rank: 3
url
https://www.farsnews.ir/news/14000823000234/از-دولت-حرکت-از-واکسن‌سازان-همّت-اولویت-رئیسی-برای-حمایت-از
title
از دولت حرکت از واکسن‌سازان همّت/ اولویت رئیسی برای حمایت از تولیدکنندگان داخلی واکسن
content

گ ها، گذشته خودشان را جبران کنند تا ایران نیز بتواند به غیر از تأمین نیاز داخل به فکر صادر کردن واکسن به کشورهای دیگر باشد تا به جای اروپ، اروپایی برای کشور داشته باشد. انتهای پیام
```

تحلیل: تعداد اسناد برابر آورده میشود اما زمانی که برای champion بوده بیشتر شده است نزدیک 2 برابر

ت)یک پرسمان دشوار و کم تکرار چند کلمه ای
مثلا واکسن آسترازنکا

```
url
https://www.farsnews.ir/news/14000815000496/یادداشت|-ماسک‌هایی-که-بوی-رفاقت-نمی‌دهد
title
یادداشت| ماسک‌هایی که بوی رفاقت نمی‌دهد
content

ار ماسک در برابر جمعیت 85 میلیون نفری جمهوری اسلامی ایران رقمی ناچیز و غیر قابل توجه است. ضمن آنکه باید بر این نکته نیز توجه شود که برای همین جمعیت 85 میلیون نفری با وجود کارشکنی



rank: 2
url
https://www.farsnews.ir/news/14000811000114/قرائت-گزارش-بررسی-عملکرد-حسن-روحانی-در-ستاد-کرونا-مردم-از-عملکرد-دولت
title
قرائت گزارش بررسی عملکرد حسن روحانی در ستاد کرونا/ مردم از عملکرد دولت سایق علیه کرونا ناراضی هستند
content

یس جمهور و اظهارات جالب و ضد ونقیض ایشان به عنوان رئیس ستاد مقابله با کرونا از اسفند ۱۳۹۸ به شرح زیر بوده است: ۶ اسفند ۱۳۹۸: "تنبه‌ای که نرسید" همانگونه که از بیماری دشوار آنفلوانزای



rank: 3
url
https://www.farsnews.ir/news/14001118000684/-پسری-با-کفش‌های-کتانی-خاطره‌انگیز|-رحیمی-غریبانه-وارد-ایران-شدیم
title
پسری با کفش‌های کتانی خاطره‌انگیز| رحیمی: غریبانه وارد ایران شدیم/ برخی دنبال حرف‌های خالمزنکی هستند
content

م. روحیه بازیکنان بسیار تغییر کرده است. فارس: صحبت پایانی... باید از پاکدل و تمام کارمندان فدراسیون تشکر کنم که در این مدت تلاش کردند تا بهترین شرایط برای ملی‌پوشان فراهم شود. انتهای پیام
```

تحلیل:
در قسمت اول (استفاده از دیکشنری) بعضی از اسناد کلمه آسترازنکا را ندارند و زمان طولانی تری برای آوردن صرف شده
در قسمت لیست champion هم سرعت بالا تری برای پاسخ داشته و هم اینکه میزان شباهت بیشتری دارند.
مقایسه با فاز اول:
در فاز اول چون بر اساس آنهایی که تمام کلمات را دارا هستند در اولویت قرار داده ایم، سرعت فاز 1 از سرعت فاز 2 مربوط به قسمت استفاده از دیکشنری بالا تره اما نسبت به لیست champion کمتر بوده.

پس برای سرعت ابتدا champion سپس فاز 1