پروژه بازیابی اطلاعات

فاز سوم

تهیه کننده:

سارا تاجرنیا

استاد مربوطه:

دکتر نیک آبادی

بهار ۱۴۰۰

قسمت های اضافه شده کد Boolean Query به شرح زیر است:

```
index_name = 'index_name'
es = Elasticsearch("http://elastic:D2Fhnh0Swa-8EPT81yf6@localhost:9200")
```

```
query= {
        "bool": {
          "should": [
              {
                  #TODO: add a match query structure ==> use for normal words
                "match": {
                    "content": ""
                }
              },

              {
                  #TODO: add a match pharse query strucutre ==> use for words in <"">
                "match_phrase": {
                    "content": ""
                }
              },

          ],
          "must_not": [
              {
                  #TODO: add a match or match pharse query structure ==> use for words after <!>
                "match": {
                    "content": ""
                }
              }
          ],
        },
    }
```

- قسمت اول برای کلمات منطبق
- قسمت دوم برای عبارات منطبق
- قسمت سوم برای کلمات نامنطبق

گزارش

۱. ساخت شاخص در این فاز (با استفاده از Bulk API) را با ساخت شاخص مکانی در فاز یک از نظر زمانی (پیاده سازی و زمان اجرا) مقایسه نمایید.

۲. پرسمان‌های زیر را در نظر بگیرید

الف) یک پرسمان دشوار (مانند "تحریم هسته‌ای" آمریکا ! ایران)

ب) یک پرسمان از کلمات نادر (مانند اورشلیم ! صهیونیست)

برای هر کدام از موارد فوق پرسمان مورد استفاده در فاز یک را تکرار کنید و عملکرد دو موتور بازیابی را از نظر سرعت بازیابی اسناد و کیفیت رتبه‌بندی اسناد مرتبط مقایسه نمایید.

۳. با ذکر علت بیان کنید شما به عنوان کاربر استفاده از کدام مدل را ترجیح می‌دهید.

توجه: برای بررسی دقت رتبه‌بندی، بررسی سه سند اول کافیست.

1 - مدت زمان پیاده سازی و زمان اجرا در این فاز به دفعات کوتاه تر و بهینه تر از فاز ۱ است به طوری که در این فاز زمان برابر 4.4 ثانیه است در حالی که در فاز یک بالای 2.5 دقیقه است.

2 -

الف) پاسخ پرسمان به "تحریم هسته‌ای" آمریکا ! ایران

```
query= {
        "bool": {
          "should": [
              {
                  #TODO: add a match query structure ==> use for normal words
                "match": {
                    "content": "آمریکا"
                }
              },

              {
                  #TODO: add a match pharse query strucutre ==> use for words in <"">
                "match_phrase": {
                    "content": "تحریم هسته ای"
                }
              },

          ],
          "must_not": [
              {
                  #TODO: add a match or match pharse query structure ==> use for words after <!>
                "match": {
                    "content": "ایران"
                }
              }
          ],
        },
    }
```

قطعه کد:
پاسخ:

```
✓ 0.2s

171 results in 0.052 s:
https://www.farsnews.ir/news/14001214000789/نماینده-کلیمیان-در-مجلس-دهم--کارتل‌ها-بر-تصمیمات-هیأت-حاکمه-آمریکا
https://www.farsnews.ir/news/14001217000665/خبائث‌های-آمریکا-در-برجام-روسیه-و-چین-را-هم-به-این-کشور-بدبین-کرد
https://www.farsnews.ir/news/14001213000089/آمریکا-با-بحران-آفرینی-به-حیات-خود-ادامه-می-دهد
https://www.farsnews.ir/news/14001211000321/آمریکا-دولت‌های-متحد-خود-را-در-زمان-اضطرار-تنها-می‌گذارد
https://www.farsnews.ir/news/14001211000898/سود-مافیای-اسلحه‌سازی-آمریکا-در-ناامن-بودن-جهان-است
https://www.farsnews.ir/news/14001214000825/آمریکا-رژیمی-مافیایی-است-و-مردم-در-تصمیمات-حاکمان-آن-جایگاهی-ندارند
https://www.farsnews.ir/news/14001212000725/آمریکا-برای-فروش-تسلیحات-خود-به-ایجاد-ناامنی-و-بحران-آفرینی-نیاز-دارد
https://www.farsnews.ir/news/14001204000444/آمریکایی-در-جام-باشاردوغو-جردن-باروز-هم-می‌آید-۲۸
https://www.farsnews.ir/news/14001213000328/کارتل‌های-اقتصادی-رؤسای-جمهور-آمریکا-را-تعیین-می‌کنند
https://www.farsnews.ir/news/14001211000220/نماینده-کلیمیان-در-مجلس-منافع-رژیم-مافیایی-آمریکا-در-ایجاد-ناامنی-است
```

ب) اورشلیم ! صهیونیست
قطعه کد:

```
query= {
        "bool": {
          "should": [
              {
                  #TODO: add a match query structure ==> use for normal words
                "match": {
                    "content": "اورشلیم"
                }
              },

              {
                  #TODO: add a match pharse query strucutre ==> use for words in <"">
                "match_phrase": {
                    "content": ""
                }
              },

          ],
          "must_not": [
              {
                  #TODO: add a match or match pharse query structure ==> use for words after <!>
                "match": {
                    "content": "صهیونیست"
                }
              }
          ],
        },
    }
```

پاسخ:

```
0 results in 0.003 s:
```

-3
در فاز یک زمان تحلیل و ذخیره اطلاعات ده ها برابر از زمان مورد نیاز برای این موارد در فاز سه بود همچنین مدت زمان پاسخ گویی همان گونه که در مثال ها پیدا ست 0.2 s و 0.003 s است در حالی که در فاز یک این زمان چندین ثانیه است.

سند اول:
https://www.farsnews.ir/news/14001214000789/نماینده-کلیمیان-در-مجلس-دهم--کارتل‌ها-بر-تصمیمات-هیأت-حاکمه-آمریکا

**نماینده کلیمیان در مجلس دهم: کارتل‌ها بر تصمیمات هیأت حاکمه آمریکا تأثیرگذار هستند

نماینده ایرانیان کلیمی در مجلس دهم گفت: کارتل‌ها در آمریکا همواره بر تصمیمات هیأت حاکمه آمریکا تأثیرگذار هستند که از جمله آن می‌توان به تأثیرات و نقش هالیوود در تصمیمات و سیاست‌های اقتصادی هیأت حاکمه حاکم اشاره کرد.

به دفعات زیادی کلمه آمریکا در آن تکرار شده به طوری که با وجود اینکه عبارت "تحریم هسته‌ای" در آن نیامده اما در رتبه اول قرار دارد(کلمه ایران هم نیامده!).

سند دوم:
https://www.farsnews.ir/news/14001217000665/خباثت‌های-آمریکا-در-برجام-روسیه-و-چین-را-هم-به-این-کشور-بدبین-کرد

**خباثت‌های آمریکا در برجام روسیه و چین را هم به این کشور بدبین کرد

عضو کمیسیون اجتماعی مجلس گفت: بدعهدی و خباثت آمریکا در ماجرای برجام روسیه و چین را هم به این کشور بدبین کرد و امروز این دو کشور هم از آمریکا تضمین می‌خواهند.

در این سند هم مانند قبلی صرفا تعداد زیادی کلمه آمریکا داریم که سبب ارتباط بین query و document میشود.

سند سوم:
**آمریکا با بحران‌آفرینی به حیات خود ادامه می دهد

عضو شورای مرکزی جبهه پایداری گفت: آمریکا با بحران‌آفرینی به حیات خود ادامه می دهد و با ایجاد شرارت احساس تداوم زندگی دارد که البته این هم یکی از نشانه های افول هرچه نزدیکتر آمریکاست

این سند هم مانند قبلی هاست که تنها با وجود تعدد کلمه آمریکا مرتبط است.

من به عنوان کاربر ترجیح میدهم از این راه استفاده کنم چرا که سرعت آن بسیار بالا تر است در حالی که مرتبط بودن query به document آنقدر هم متفاوت نیستند.

قسمت های اضافه شده کد Spelling_Correction به شرح زیر است:

```
sc_mapping = {
    "settings": {
    "index": {
      #TODO: define your analyzers here
      "analysis": {
        "analyzer": {
          "my_custom_analyzer": {
            "type": "custom",
            "tokenizer": "standard",
            "char_filter":
            {
              "zero_width_spaces":
              { "type": "mapping",
              "mappings": [ "\\u200C=>\\u0020"] } },

            "filter":
            {
              "type": "shingle",
              "min_shingle_size": 2,
              "max_shingle_size": 3
            }
          }
        }
      }
    },
    "mappings": {
        "properties": {
            #TODO: define your fields here
          "content_title": {
            "type": "text",
            "fields": {
              "trigram": {
                "type":     "text",
                "analyzer": "my_custom_analyzer"
              }
            }
          }
        }
    }
}
```

```
def get_suggestions(text , index_name):
    resp = es.search(index=index_name,suggest={
        "text": text,
        "simple_phrase": {
            "phrase": {
                "smoothing" : {
                    "laplace" : {
                    "alpha" : "0.7"
                    }
                },
                "field": "content_title.trigram" ,
                "size": "4",
                "confidence": "1",
                "real_word_error_likelihood": "0.95",
                "max_errors": "3",
                "direct_generator": [ {
                    "field": "content_title",
                    "prefix_length": "2"
                }
                ]
            }
        }
    },size=0)
    return dict(resp)
```

قسمتی از پاسخ ها در قسمت اول:

```
[{'text': 'لیک برتر فوطبال', 'offset': 0, 'length': 15, 'options': [{'text': 'لیک برتر فوتبال', 'score':
3.8619805e-06}, {'text': 'لیک برتر فوتسال', 'score': 1.1990304e-06}, {'text': 'لیک برتر فوتبالی', 'score':
7.257021e-07}, {'text': 'لیک برتر فوتبالش', 'score': 2.5862772e-07}]}]
========================
[{'text': 'تورنومنت شش جانبه', 'offset': 0, 'length': 17, 'options': []}]
========================
[{'text': 'طبعیض نژادی', 'offset': 0, 'length': 11, 'options': [{'text': 'طبیعی نژادی', 'score': 5.851293e-
05}, {'text': 'طبیعت نژادی', 'score': 1.9974208e-05}, {'text': 'طبعاً نژادی', 'score': 1.21406565e-05},
{'text': 'طبعی نژادی', 'score': 4.8353113e-06}]}]
========================
[{'text': 'اردوی طیم امیذ', 'offset': 0, 'length': 14, 'options': [{'text': 'اردوی طیم امید', 'score':
5.5828457e-07}, {'text': 'اردوی طیم امیر', 'score': 4.45046e-07}, {'text': 'اردوی طیم امین', 'score':
2.8232267e-07}, {'text': 'اردوی طیم امیه', 'score': 4.3622247e-08}]}]
========================
[{'text': 'جام ملب های آشیا', 'offset': 0, 'length': 16, 'options': [{'text': 'جام ملب های آشنا', 'score':
4.4530008e-08}, {'text': 'جام ملب های آشکار', 'score': 3.3973766e-08}, {'text': 'جام ملب های آشوب', 'score':
2.9898516e-08}, {'text': 'جام ملب های آشیل', 'score': 1.758879e-08}]}]
========================
[{'text': 'کنره سیاسی آمریکا', 'offset': 0, 'length': 17, 'options': [{'text': 'کنگره سیاسی آمریکا', 'score':
1.9478046e-05}, {'text': 'کناره سیاسی آمریکا', 'score': 1.036026e-05}, {'text': 'کنده سیاسی آمریکا', 'score':
6.8913428e-06}, {'text': 'کنته سیاسی آمریکا', 'score': 3.1233722e-06}]}]
========================
[{'text': 'انغلاب اشلامی ایران', 'offset': 0, 'length': 19, 'options': [{'text': 'انقلاب اشکالی ایران', 'score':
2.7467715e-05}, {'text': 'انقلاب اشخاصی ایران', 'score': 2.1585096e-05}, {'text': 'انقلاب اشغالی ایران',
'score': 2.0130883e-05}, {'text': 'انقلاب اشرافی ایران', 'score': 1.9509167e-05}]}]
========================
[{'text': 'فدراصیون فوتبال ایران', 'offset': 0, 'length': 21, 'options': [{'text': 'فدراسیون فوتبال ایران',
'score': 0.0002310098}, {'text': 'فدارسیون فوتبال ایران', 'score': 2.6549564e-05}, {'text': 'فدراسیونی فوتبال
ایران', 'score': 2.5080411e-05}, {'text': 'فدراسیون\u200c فوتبال ایران', 'score': 1.130555e-05}]}]
========================
[{'text': 'لایهه مجلص خبرگان', 'offset': 0, 'length': 17, 'options': [{'text': 'لایحه مجلس خبرگان', 'score':
3.4283275e-05}, {'text': 'لایه مجلس خبرگان', 'score': 7.206687e-06}, {'text': 'لایحه مجدد خبرگان', 'score':
5.220123e-06}, {'text': 'لایحه مجاز خبرگان', 'score': 4.6150353e-06}]}]
```

کد با کمک reverse کردن کلمه:

```
sc_mapping_v1 = {
    "settings": {
    "index": {
      #TODO: define your analyzers here
      "analysis": {
        "analyzer": {
          "my_custom_analyzer": {
            "type": "custom",
            "tokenizer": "standard",
            "char_filter":
            {
              "zero_width_spaces":
              { "type": "mapping",
              "mappings": [ "\\u200C=>\\u0020"] } },

            "filter":
            {
              "type": "shingle",
              "min_shingle_size": 2,
              "max_shingle_size": 3
            }
          },
        "analyzer_reverse" :{
            "type": "custom",
            "tokenizer": "standard",
            "char_filter":{
            "zero_width_spaces":
              { "type": "mapping",
                "mappings": [ "\\u200C=>\\u0020"] } ,
                "filter" : ["reverse"]

            }
          }
        }
      }
    },
      "mappings": {
          "properties": {
              #TODO: define your fields here
            "content_title": {
              "type": "text",
              "fields": {
                "trigram2": {
                  "type":     "text",
                  "analyzer": "analyzer_reverse"
                }
              }
            }
          }
      }
}
```

```
def get_suggestions_v1(text , index_name):
    resp = es.search(index=index_name,suggest={
        "text": text,
        "simple_phrase": {
            "phrase": {
                "smoothing" : {
                    "laplace" : {
                    "alpha" : "0.7"
                    }
                },
                "field": "content_title" ,
                "size": "4",
                "confidence": "1",
                "real_word_error_likelihood": "0.95",
                "max_errors": "3",
                "direct_generator": [ {
                    "field": "content_title",
                    "prefix_length": "2",
                    "suggest_mode" : "always"
                }, {
                    "field" : "content_title.trigram2",
                    "pre_filter" : "analyzer_reverse",
                    "post_filter" : "analyzer_reverse"
                }
                ]
            }
        }
    },size=0)
    return dict(resp)
```

قسمتی از پاسخ ها در این قسمت با reverse کردن کلمه:


```
[{'text': 'لیک برتر فوطبال', 'offset': 0, 'length': 15, 'options': [{'text': 'لیک برتر فوتبال',
'score': 3.8619805e-06}, {'text': 'لیک برتر فوتسال', 'score': 1.1990304e-06}, {'text': 'لیک برتر
فوتبالی', 'score': 7.257021e-07}, {'text': 'لیک برتر فوتبالش', 'score': 2.5862772e-07}]}]
========================
[{'text': 'تورنومنت شش جانبه', 'offset': 0, 'length': 17, 'options': [{'text': 'تورنمنت شش جانبه',
'score': 2.314735e-06}, {'text': 'تورنمنت شش جانبی', 'score': 9.497933e-07}, {'text': 'تورنمنت شش
جاذبه', 'score': 7.0017546e-07}, {'text': 'تورنمنتی شش جانبه', 'score': 6.8675297e-07}]}]
========================
[{'text': 'طبعیض نژادی', 'offset': 0, 'length': 11, 'options': [{'text': 'طبیعی نژاد', 'score':
0.00023461529}, {'text': 'طبیعت نژاد', 'score': 8.0089216e-05}, {'text': 'طبیعی نژادی', 'score':
5.851293e-05}, {'text': 'طبیعی نژادیان', 'score': 5.3357377e-05}]}]
========================
[{'text': 'اردوی طیم امیذ', 'offset': 0, 'length': 14, 'options': [{'text': 'اردوی طیم امید',
'score': 5.5828457e-07}, {'text': 'اردوی طیم امیر', 'score': 4.45046e-07}, {'text': 'اردوی طیم
امین', 'score': 2.8232267e-07}, {'text': 'اردوی طیم امیه', 'score': 4.3622247e-08}]}]
========================
[{'text': 'جام ملب های آشیا', 'offset': 0, 'length': 16, 'options': [{'text': 'جام ملب های آسیا',
'score': 1.3091295e-07}, {'text': 'جام ملب های آشنا', 'score': 4.4530008e-08}, {'text': 'جام ملب های
آشکار', 'score': 3.3973766e-08}, {'text': 'جام ملب های آشوب', 'score': 2.9898516e-08}]}]
========================
[{'text': 'کنره سیاسی آمریکا', 'offset': 0, 'length': 17, 'options': [{'text': 'کنگره سیاسی آمریکا',
'score': 1.9478046e-05}, {'text': 'کاره سیاسی آمریکا', 'score': 1.1162472e-05}, {'text': 'کناره
سیاسی آمریکا', 'score': 1.036026e-05}, {'text': 'کوره سیاسی آمریکا', 'score': 8.849013e-06}]}]
========================
[{'text': 'انغلاب اشلامی ایران', 'offset': 0, 'length': 19, 'options': [{'text': 'انقلاب اسلامی ایران',
'score': 0.00033255838}, {'text': 'انتخاب اسلامی ایران', 'score': 0.0001565941}, {'text': 'انقلابی
اسلامی ایران', 'score': 0.00014381837}, {'text': 'انتصاب اسلامی ایران', 'score': 5.6628993e-05}]}]
========================
[{'text': 'فدراصیون فوتبال ایران', 'offset': 0, 'length': 21, 'options': [{'text': 'فدراسیون فوتبال
ایران', 'score': 0.0002310098}, {'text': 'فدارسیون فوتبال ایران', 'score': 2.6549564e-05}, {'text':
'فدراسیونی فوتبال ایران', 'score': 2.5080411e-05}, {'text': 'فدراسیون\u200cفوتبال ایران', 'score':
1.130555e-05}]}]
========================
[{'text': 'لایهه مجلس خبرگان', 'offset': 0, 'length': 17, 'options': [{'text': 'لایحه مجلس خبرنگار',
'score': 0.0001269758}, {'text': 'لایحه مالی خبرنگار', 'score': 4.3848308e-05}, {'text': 'لایحه مجلس
خبرگان', 'score': 3.4283275e-05}, {'text': 'لایه مجلس خبرنگار', 'score': 2.6691581e-05}]}]
```

کد های با کمک synonum ها:

```
        "analyzer_synonym" :{
          "type": "custom",
          "tokenizer": "whitespace",
          "char_filter":
          {
            "zero_width_spaces":
            {"type": "mapping",
              "mappings": [ "\\u200C=>\\u0020"] } },
            "filter": [ "synonym" ]
          }
        },
        "filter": {
          "synonym": {
            "type": "synonym",
            "synonyms_path": "synonyms.txt"
          }
        }
      }
    }
  },
    "mappings": {
        "properties": {
            #TODO: define your fields here
          "content_title": {
            "type": "text",
            "fields": {
              "trigram3": {
                "type":     "text",
                "analyzer": "analyzer_synonym"
              }
            }
          }
        }
    }
}
```

```
def get_suggestions_v2(text , index_name):
    resp = es.search(index=index_name,suggest={
        "text": text,
        "simple_phrase": {
            "phrase": {
                "smoothing" : {
                    "laplace" : {
                    "alpha" : "0.7"
                    }
                },
                "field": "content_title" ,
                "size": "4",
                "confidence": "1",
                "real_word_error_likelihood": "0.95",
                "max_errors": "3",
                "direct_generator": [ {
                    "field": "content_title",
                    "prefix_length": "2",
                    "suggest_mode" : "always"
                }, {
                    "field" : "content_title.trigram3",
                    "post_filter" : "analyzer_synonym"
                }
                ]
            }
        }
    },size=0)
    return dict(resp)
```

خروجی برخی قسمت ها که نسبت به قبلی تغییر کرده:

```
[{'text': 'طبعیض نژادی', 'offset': 0, 'length': 11, 'options': [{'text': 'اصل نژاد', 'score':
0.00038619604}, {'text': 'طبیعی نژاد', 'score': 0.00023461529}, {'text': 'حالت نژاد', 'score':
0.00018953791}, {'text': 'اصل نژادی', 'score': 9.631709e-05}]}]
==================================
[{'text': 'جام ملب های آشیا', 'offset': 0, 'length': 16, 'options': [{'text': 'جام ملب های دوست',
'score': 9.197914e-08}, {'text': 'جام ملب های شناخت', 'score': 6.662835e-08}, {'text': 'جام ملب های
خویش', 'score': 4.5126097e-08}, {'text': 'جام ملب های خودی', 'score': 3.5414516e-08}]}]
==================================
[{'text': 'کنره سیاسی آمریکا', 'offset': 0, 'length': 17, 'options': [{'text': 'حاشیه سیاسی آمریکا',
'score': 4.0848423e-05}, {'text': 'جانب سیاسی آمریکا', 'score': 1.8157003e-05}, {'text': 'دامن سیاسی
آمریکا', 'score': 1.8092589e-05}]}]
==================================
```

۲. تحقیق کنید که افزایش یا کاهش فیلد max_errors چه تاثیری بر روی کیفیت پیشنهاددهندهها و مدت زمان پاسخ درخواست می‌گذارد و نتیجه را در گزارش خود شرح دهید. برای این بخش مقدار آن را سه در نظر بگیرید.

با افزایش max_errors میزان خطا در بازیابی اطلاعات زیاد میشود اما در عوض برای query هایی که کمیاب تر هستند میتواند منابع که حداقل میزان ارتباط را هم دارند برگرداند به جای آن که هیچ اطلاعاتی ندهد.
حداکثر درصد عباراتی که برای ایجاد تصحیح به عنوان غلط املایی در نظر گرفته می شوند. این روش یک مقدار شناور در محدوده [0..1) را به عنوان کسری از عبارت های پرس و جو واقعی یا یک عدد >=1 را به عنوان تعداد مطلق عبارت های پرس و جو می پذیرد. پیش فرض روی 1.0 تنظیم شده است، به این معنی که فقط اصلاحات با حداکثر یک عبارت غلط املایی برگردانده می شوند. توجه داشته باشید که تنظیم بیش از حد آن می تواند بر عملکرد تأثیر منفی بگذارد. مقادیر کم مانند 1 یا 2 توصیه می شود. در غیر این صورت ممکن است زمان صرف شده در تماس های پیشنهادی از زمان صرف شده در اجرای پرس و جو بیشتر شود.

۳. بررسی کنید که دو فیلد confidence و real_word_error_likelihood چه تاثیری بر امتیاز عبارت ورودی و عبارت‌های پیشنهادی دارند. مقادیری که برای این دو عبارت در نظر می‌گیرید را گزارش کنید و دلیل آن را شرح دهید.

Confidence :
سطح اطمینان عاملی را برای امتیاز عبارات ورودی تعریف می کند که به عنوان آستانه ای برای سایر نامزدهای پیشنهادی استفاده می شود. فقط داوطلبانی که نمرات بالاتر از حد نصاب کسب کرده باشند در نتیجه شرکت خواهند کرد. به عنوان مثال، سطح اطمینان 1.0 تنها پیشنهادهایی را برمی گرداند که امتیاز بالاتری از عبارت ورودی دارند. اگر روی 0.0 تنظیم شود، N کاندیدای برتر برمی گردند. پیش فرض 1.0 است.

اما ما این مقدار را برابر با ۱ یعنی پیش فرض گذاشتیم تا بهترین کاندیدا را برای ما برگرداند.

real_word_error_likelihood :
احتمال غلط املایی یک اصطلاح حتی اگر اصطلاح در فرهنگ لغت وجود داشته باشد. پیش فرض 0.95 است، به این معنی که 5٪ از کلمات واقعی غلط املایی دارند.

ما مقدار 95% را گذاشتیم چرا که طبق آمار هم چیزی حدود 5% کلمات غلط املایی دارند.(البته مثال هایی که برای این قسمت زده شده مقدار زیادی از آنها غلط است)

گزارش

عملکرد سیستم را به ازای جملات قرار داده شده در ژوپیتر نوت بوک بررسی کنید. برای هر عبارت ورودی، حداکثر ۵ عبارت پیشنهادی را نمایش دهید.

```
"طبعیض نژادی",
"اردوی طیم امیذ",
"جام ملب های آشیا",
"کنره سیاسی آمریکا",
"انغلاب اشلامی ایران",
```

خروجی برای عبارات زیر:

```
[{'text': 'طبعیض نژادی', 'offset': 0, 'length': 11, 'options': [{'text': 'طبیعی نژادی',
'score': 5.851293e-05}, {'text': 'طبیعت نژادی', 'score': 1.9974208e-05}, {'text': 'طبعاً
نژادی', 'score': 1.21406565e-05}, {'text': 'طبعی نژادی', 'score': 4.8353113e-06}]}]
========================
[{'text': 'اردوی طیم امیذ', 'offset': 0, 'length': 14, 'options': [{'text': 'اردوی طیم امید',
'score': 5.5828457e-07}, {'text': 'اردوی طیم امیر', 'score': 4.45046e-07}, {'text': 'اردوی طیم
امین', 'score': 2.8232267e-07}, {'text': 'اردوی طیم امیه', 'score': 4.3622247e-08}]}]
========================
[{'text': 'جام ملب های آشیا', 'offset': 0, 'length': 16, 'options': [{'text': 'جام ملب های
آشنا', 'score': 4.4530008e-08}, {'text': 'جام ملب های آشکار', 'score': 3.3973766e-08},
{'text': 'جام ملب های آشوب', 'score': 2.9898516e-08}, {'text': 'جام ملب های آشیل', 'score':
1.758879e-08}]}]
========================
[{'text': 'کنره سیاسی آمریکا', 'offset': 0, 'length': 17, 'options': [{'text': 'کنگره سیاسی
آمریکا', 'score': 1.9478046e-05}, {'text': 'کناره سیاسی آمریکا', 'score': 1.036026e-05},
{'text': 'کنده سیاسی آمریکا', 'score': 6.8913428e-06}, {'text': 'کنته سیاسی آمریکا', 'score':
3.1233722e-06}]}]
========================
[{'text': 'انغلاب اشلامی ایران', 'offset': 0, 'length': 19, 'options': [{'text': 'انقلاب اشکالی
ایران', 'score': 2.7467715e-05}, {'text': 'انقلاب اشخاصی ایران', 'score': 2.1585096e-05},
{'text': 'انقلاب اشغالی ایران', 'score': 2.0130883e-05}, {'text': 'انقلاب اشرافی ایران',
'score': 1.9509167e-05}]}]
========================
```

ارسالی، در کنار مولدی که در قسمت قبل تعریف شد، مولد دیگری تعریف کنید که ورودی کاربر را به توکن‌های وارونه تبدیل کند و پس از تولید کلمات پیشنهادی، آن‌ها را قبل از رفتن به فاز امتیازدهی، وارونه کند. (راهنمایی: از فیلد های pre_filter و post_filter استفاده کنید). پس از اعمال تغییرات ذکر شده، بررسی کنید که آیا بهبودی در پیشنهادهای تولید شده ایجاد شده است یا خیر؟

بله پاسخ ها به نسبت دقیق تر شده اند.

قسمت های اضافه شده کد SM به شرح زیر است:

```
source_code = "double tf = Math.sqrt(doc.freq); double idf =
Math.log((field.docCount+1.0)/(term.docFreq+1.0)) + 1.0; double norm =
1/Math.sqrt(doc.length); return query.boost * tf * idf * norm;"
```

```
# closing the index
es.indices.close(index=sm_index_name)

# applying the settings
es.indices.put_settings(index=sm_index_name,
                        settings={
                            "similarity": {
                                "default": {
                                  "type": "scripted",
                                  "script": {
                                    # TODO : uncomment the code bellow and pass the suitable parameter
                                    "source": source_code
                                  }
                                }
                            }
                        }
                    )

# reopening the index
es.indices.open(index=sm_index_name)
```

گزارش

۱. به پرسمان‌ها در حالت‌های زیر پاسخ دهید:

توجه: query خود را به شکل match کوئری برای فیلد content اخبار بزنید.

الف) یک پرسمان دشوار و کم تکرار تک کلمه‌ای

ب) یک پرسمان دشوار و کم تکرار چند کلمه‌ای

۲. در هر حالت پرسمانی که در فاز۲ استفاده کردید را تکرار کنید. نتایج بازگردانده شده را از نظر میزان ارتباط مقایسه و تحلیل نمایید.

- 1

الف) "استراتژی"

```
استراتژی
https://www.farsnews.ir/news/14000826000200/ضعف‌های-استراتژی-توسعه-صنعتی-از-منظر-مرکز-پژوهش‌های-مجلس
https://www.farsnews.ir/news/14001119000709/مدیر-اجرایی-بارسلونا-استعفا-کرد
https://www.farsnews.ir/news/14001111000712/فان-مارویک-می-خواهیم-3-امتیاز-را-کسب-کنیم-در-دبی-بدشانس-بودیم
https://www.farsnews.ir/news/14001124000896/عالمی-روند-استقلال-از-هر-نظر-برای-قهرمانی-ایده آل-است-نیمه-اول-ذوب
https://www.farsnews.ir/news/14001025000364/چرا-آمریکایی‌ها-باید-منتظر-انتقام-از-قاتلان-شهید-سلیمانی-باشند
https://www.farsnews.ir/news/14000825000108/قرائت-گزارش-وضعیت-شاخص‌های-اقتصادی-در-دولت-روحانی-در-مجلس-افزایش-نرخ
https://www.farsnews.ir/news/14001124000170/اظهارات-سرمربی-سانتاکلارا-در-خصوص-مصدومیت-ستاره-کشورمان
https://www.farsnews.ir/news/14000725000333/نامه-60-نماینده-به-رئیس‌جمهور-دولت-برای-استفاده-حداکثری-از-پیمان
https://www.farsnews.ir/news/14001220000825/مهدی-پور-بهترین-کار-جمع-آوری-امتیاز-است-برای-حفظ-صدرنشینی-به-مصاف
https://www.farsnews.ir/news/14001211000719/هاشمیان--یک-روز-مربی-تیمی-در-بوندس-لیگا-می‌شوم
```

ب) "استراتژی دشوار پیروزی"

```
استراتژی دشوار پیروزی
https://www.farsnews.ir/news/14001202000108/طارمی-مهمترین-چیز-پیروزی-پورتو-است-نه-گلزنی-من
https://www.farsnews.ir/news/14000826000200/ضعف‌های-استراتژی-توسعه-صنعتی-از-منظر-مرکز-پژوهش‌های-مجلس
https://www.farsnews.ir/news/14001201001098/منچستر-یونایتد-رکورددار-بیشترین-پیروزی-در-تاریخ-لیگ-برتر-انگلیس
https://www.farsnews.ir/news/14001028001105/ارزیابی-مربی-تیم-ملی-هندبال-از-کسب-نخستین-پیروزی--صادقی-استرالیا
https://www.farsnews.ir/news/14001124000170/اظهارات-سرمربی-سانتاکلارا-در-خصوص-مصدومیت-ستاره-کشورمان
https://www.farsnews.ir/news/14001123000247/اظهارات-جالب-گواردیولا-در-خصوص-یورگن-کلوپ
https://www.farsnews.ir/news/14000927000692/مربی-استقلال-ملاثانی-امیدوارم-بازی-با-پدیده-شروع-موفقیت-های-ما-باشد
https://www.farsnews.ir/news/14000826000899/آیت‌الله-مجتهد-شبستری-از-مبارزان-راه-دشوار-نهضت-پیروزی-انقلاب-بود
https://www.farsnews.ir/news/14001119000709/مدیر-اجرایی-بارسلونا-استعفا-کرد
https://www.farsnews.ir/news/14001216000182/واکنش-سرمربی-منچستریونایتد-به-شکست-در-دربی
```

2 - به نسبت میزان ازتباط میتوان گفت تفاوت آنچنانی نداشتند اما این کد کمی دقیق تر نسبت به فاز قبل نتایج را برگرداند به طوری که تنها تعداد کمی متفاوت بودند.

برخی قسمت های اضافه شده کد KNN (edited) به شرح زیر است:

```
from elasticsearch import Elasticsearch
from elasticsearch import helpers
import json
from tqdm import tqdm
from gensim.models import Word2Vec
import numpy as np
import random
import pandas as pd
from hazm import *
from hazm import stopwords_list
import string
import sklearn
# import whatever you need for your implementation
```

```
def filter_doc(contents):
    normalizer = Normalizer()
    stemmer = Stemmer()
    contents = normalizer.normalize(contents)
    contents = word_tokenize(contents)
    j = len(contents) - 1
    while j > -1:
        contents[j] = stemmer.stem(contents[j])
        j -= 1
    return contents
```

```
cols = [1]
df = pd.read_excel('IR01_3_test_4k.xlsx', usecols=cols)
#print(df['content'][0])
dataset = []
stopwords = stopwords_list()
normalizer = Normalizer()
for i in range(len(df['content'])):
    contents = df['content'][i]
    contents = filter_doc(contents)
    dataset.append(contents)
```

```
from copy import deepcopy
data_tmp = []
for i in tqdm(range(len(data))):
    tmp_doc = data
    doc = dict()
    # filter_doc: method for preprocessing a doc.
    doc['content'] = " ".join(filter_doc(tmp_doc['content'][i]))
    doc['vec'] = list(doc_vectors[i])
    doc['category'] = tmp_doc['category'][i]
    data_tmp.append(doc)
```

گزارش

۱. در این قسمت، بازیابی نتایج پرسمان را بر روی اسنادی که توسط KNN برچسب زده‌ایم انجام خواهیم داد. برای هر پرسمان علاوه بر متن پرسمان، برچسب مورد نظر خود را نیز مشخص می‌کنیم تا تنها اسنادی که حاوی برچسب مد نظر ما هستند در نتایج مشاهده کنیم. در این قسمت سه پرسمان چند کلمه‌ای در حوزه ورزشی، اقتصادی و سلامت مشخص کنید و نتایج بازیابی را بررسی و تحیلی کنید. برای مثال یک پرسمان چند کلمه ای در حوزه ی اخبار ورزشی خواهیم داشت: "نتایج مسابقات لیگ برتر فوتبال ایران" برچسب: "ورزشی". جهت بررسی عملکرد سیستم ۵ سند اول بازیابی شده را باز کرده و مشخص کنید که آیا به پرسمان ارتباطی دارد؟ همچنین در حوزه‌ی مد نظر قرار دارد؟ در صورتی که هر دو شرط مذکور رعایت شود سیستم بازیابی عملکرد قابل قبولی دارد.

استقلال در لیگ برتر فوتبال:

```
https://www.farsnews.ir/news/14001219000609/آی-تی-سی-مدافع-برزیلی-استقلال-صادر-شد
https://www.farsnews.ir/news/14001023000660/استقلال-بهترین-خط-دفاع-لیگ-در-پایان-نیم-فصل-رکورد-جدید-شاگردان-مجیدی
https://www.farsnews.ir/news/14001212000761/حضور-مدیرعامل-گل‌گهر-در-جمع-شاگردان-قلعه‌نویی-قبل-از-مصاف-با-استقلال
https://www.farsnews.ir/news/14001125000682/تمرین-فردای-استقلال-تعطیل-شد
https://www.farsnews.ir/news/14001128000942/غیبت-یک-بازیکن-استقلال-در-سفر-به-تبریز
https://www.farsnews.ir/news/14001214000755/از-استقلال-20-بازی-بدون-شکست-و-در-مسیر-قهرمانی-عکس-AFC-تمجید
https://www.farsnews.ir/news/14001215000216/مجیدی-به-رکورد-قابل-توجه-قلعه‌نویی-رسید
https://www.farsnews.ir/news/14001112000417/زمان-برگزاری-نشست-خبری-مجیدی-مشخص-شد
https://www.farsnews.ir/news/14001224000971/محل-برگزاری-نشست‌های-خبری-سرخابی‌ها-مجیدی-در-سازمان-لیگ-گل‌محمدی-در
https://www.farsnews.ir/news/14001008000943/دیدار-استقلال-و-فولاد-تماشاگر-داشت-تصاویر
```

بالا رفتن تورم سالانه کشور:

```
https://www.farsnews.ir/news/14000727000517/تجمع-جمعی-از-مردان-مقابل-مجلس-برای-کاهش-میزان-مهریه
https://www.farsnews.ir/news/14000916000722/میرکاظمی-اعدادی-که-درباره-افزایش-پلکانی-حقوق‌ها-مطرح-می-شود-صحت-ندارد
https://www.farsnews.ir/news/14001120000134/400-گزارش-مرکز-پژوهش‌های-مجلس-درباره-تورم-سه‌ماهه-تابستان
https://www.farsnews.ir/news/14000820000214/جزئیات-نشست-مجمع-نمایندگان-سه-استان-با-رئیس-جمهور-رئیسی-به-جای
https://www.farsnews.ir/news/14001015000141/یکی-از-دلایل-ایجاد-حقوق‌های-نجومی-تنوع-در-پرداخت-حقوق-است
https://www.farsnews.ir/news/14001211000261/سقف-معافیت-مالیاتی-کارمندان-دولت-مشخص-شد
https://www.farsnews.ir/news/14000826000067/یوسفی-رشد-تولید-ناخالص-ملی-در-دولت‌های-یازدهم-و-دوازدهم-فقط-2-درصد
https://www.farsnews.ir/news/14000818000406/درآمدهای-دامی-گلستان-از-استان-خارج-شده-و-دست-دلالان-می‌افتد
https://www.farsnews.ir/news/14001124000157/تشکیل-کارگروه-برای-بررسی-نقش-فدراسیون-اتومبیلرانی-در-کاهش-تصادفات
https://www.farsnews.ir/news/14001205000232/عضو-حزب-مردم‌سالاری-تصور-رفع-مشکلات-کشور-با-لغو-تحریم‌ها-خوشبینانه
```

بهداشت دهان و دندان:

```
https://www.farsnews.ir/news/14001222000916/خالفت-کمیسیون-بهداشت-و-درمان-مجلس-با-ادغام-آموزش-پزشکی-در-آموزش-عالی
https://www.farsnews.ir/news/14001119000454/پرتاب-آب-دهان-در-دربی-میلان--مهاجم-اینتر-محروم-می‌شود-عکس-و-فیلم
https://www.farsnews.ir/news/14001116000877/پیشنهاد-9-هزار-میلیاردی-تومانی-کمیسیون-بهداشت-مجلس-برای-حوزه-دارو-در
https://www.farsnews.ir/news/14001021000100/پزشکیان-به-صورت-مشروط-از-توضیحات-وزیر-بهداشت-قانع-شد
https://www.farsnews.ir/news/14000812000294/لایحه-مقاوله‌نامه-ایمنی-و-بهداشت-شغلی-تصویب-شد
https://www.farsnews.ir/news/14000909000770/بررسی-آخرین-وضعیت-بودجه-400-وزارت-بهداشت-ارز-ترجیحی-دارو-و-تجهیزات
https://www.farsnews.ir/news/14000918000472/ارتباط-تصویری-سرلشکر-موسوی-با-پرستاران-ارتش
https://www.farsnews.ir/news/14001008000327/2-سوال-یک-نماینده-از-وزیر-بهداشت-اعلام-وصول-شد
https://www.farsnews.ir/news/14000811000094/وزیر-بهداشت-به-سوالات-نمایندگان-درباره-واکسن-توجهی-نکرده-عین‌اللهی
https://www.farsnews.ir/news/14000927000610/مشکل-مسکن-یک-پرسپولیسی-رفع-شد-گلایه‌های-عجیب-پیشکسوت-سرخ‌پوشان
```

در اسناد به دست آمده میزان ranking به درستی بیان شده به طوری که برای مثال در قسمت که عبارت "بهداشت دهان و دندان" وجود دارد به وضوح میزان شباهت query با document در سند اول نسبت به بعدی ها بیشتر است همچنین تقریبا همه ی اسناد بازگردانده شده در حوزه ای که مشخص شده بود قرار دارند.
پس میتوان گفت این نوع بازیابی نتایج به خوبی عمل میکند.