١.

عملیات های گام پیش پردازش به کمک کتابخانه ی Hazm:

```
def preprocess(raw_document):
normalizer = Normalizer()
stemmer = Stemmer()
lemmatizer = Lemmatizer()
stopwordslist = set(stopwords_list())
stopwordslist.update(['.', ':', '?', '!', '/', '*', '[', ']', '{', '}',
 ,'\"','(',')',''])
tokens = []
words = []
words_temp = word_tokenize(normalizer.normalize(raw_document.content))
for word in words temp:
    if word not in stopwordslist:
        temp = lemmatizer.lemmatize(word)
        temp = stemmer.stem(temp)
        words.append(temp)
for word in set(words):
    positions = [i + 1 for i, x in enumerate(words) if x == word]
    tokens.append([word, positions, len(positions)])
return tokens
```

نرمال سازی: به منظور یکسان سازی انجام میشود. به طور مثال علامت جمع یک کلمه اگر با فاصله از آن آمده باشد با حالتی که به آن چسبیده باشد یکسان باشد. یا حالت جدا و چسبیده در کلمات مرکب یکی حساب شود.

ساختن توکن: پس از نرمال سازی، متن آماده ی استخراج توکن می باشد. در این مرحله کلمات متن استخراج شده در لیست words_temp اضافه می شود.

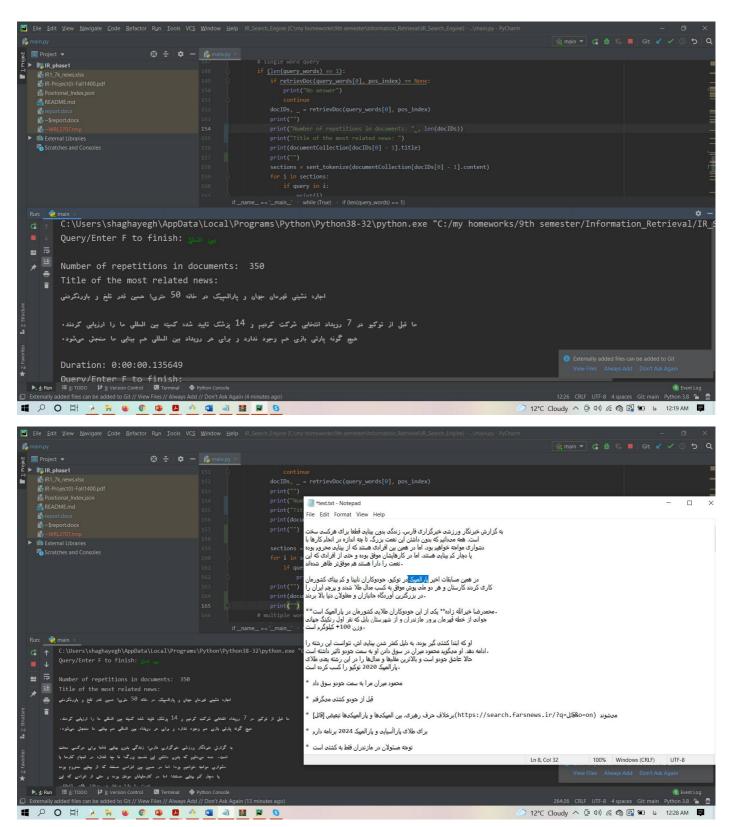
حذف کلمات یرتکرار: کلماتی که تکرار آن ها زیاد است از ارزش کمتری برخوردار هستند و علاوه بر آن وقت زیادی صرف جستجو روی آن ها می شود. بنابراین در مرحله ی پیش پردازش خوب است توکن های پر تکرار مثل علامت ها، حروف اضافه و افعالی مانند است و ... حذف شوند.

ریشه پابی: برای آنکه زمانی که کاربر کلمه ی درخت ها را سرچ میکند سند حاوی کلمه درخت هم برگردانده شود یا زمان های مختلف یک فعل در پیدا کردن سند دخیل باشد توکن ها باید ریشه یابی شوند.

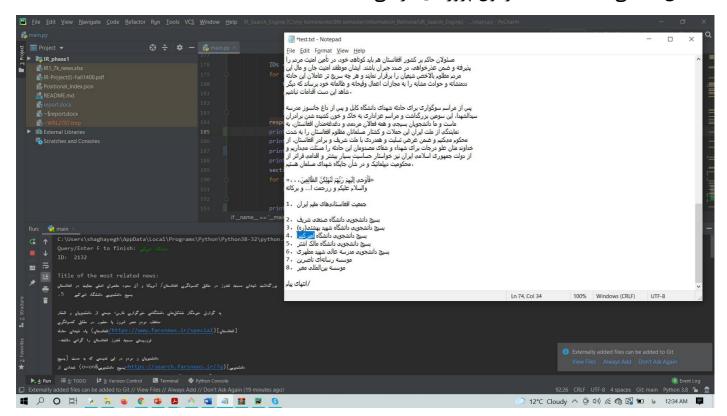
۴. در مواری شاهد تبدیلاتی مثل مهارت به مهار، شکافت به شکاف، درختان به درخ هستیم که تبدیلات مطلوبی نیستند. به طور مثال شکافت در عبارت شکافت هسته ای به اشتباه با شکاف تبدیل می شود.

۵.

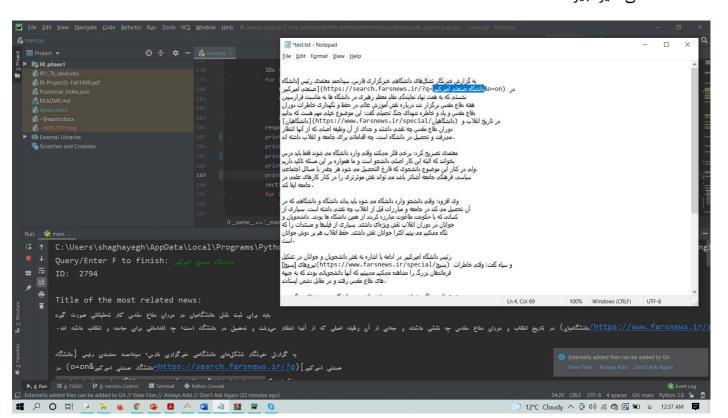
الف.



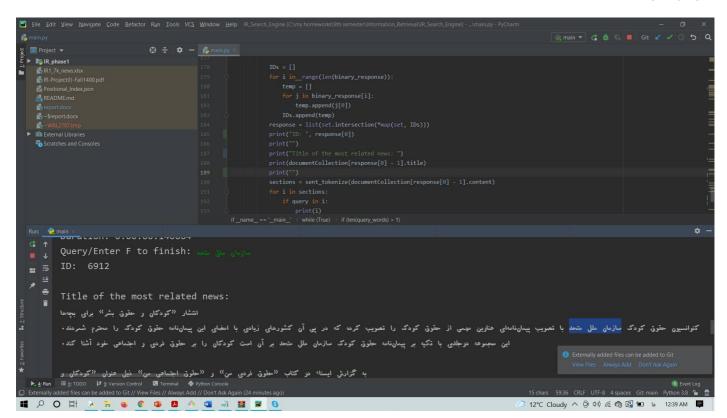
با توجه به محتوی سند، سند مرتبطی را به عنوان خروجی نمایش داده است.



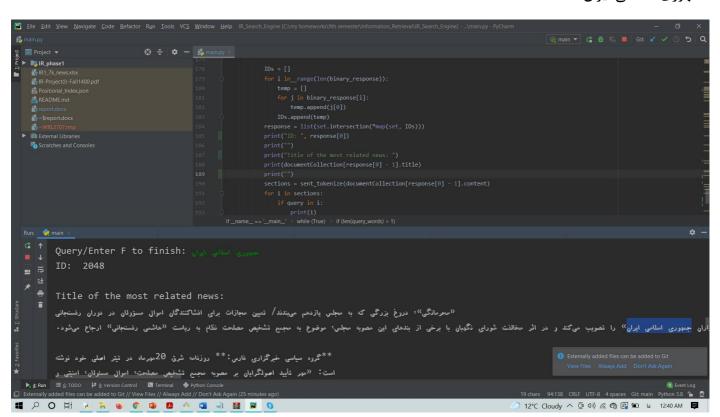
دانشگاه صنعتی امیر کبیر:

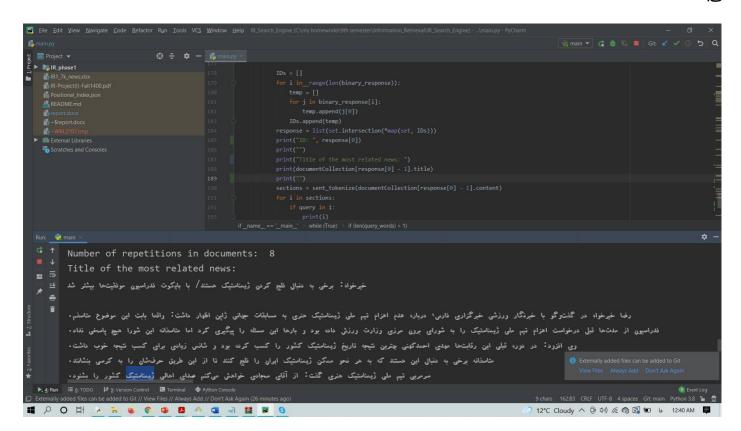


سازمان ملل متحد:

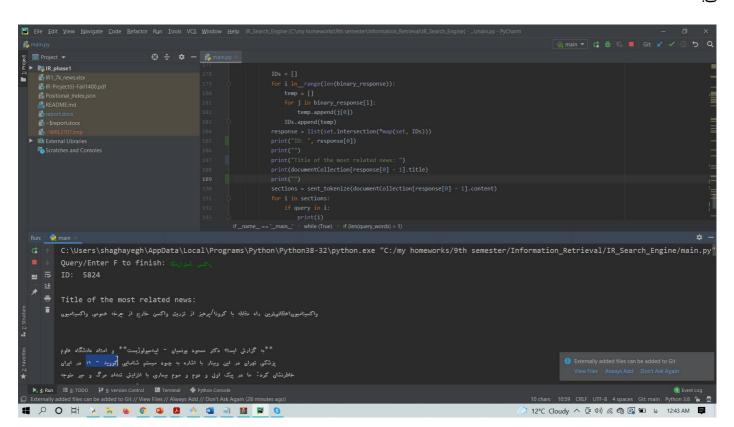


جمهوري اسلامي ايران:





ث.



سیستم تا حد خوبی حالت های مختلف را بازیابی می کند.