به نام خدا

گزارش بهبود برنامه استخراج کلید واژه

استاد: دکتر سعیده ممتازی

دانشجو: محمد قربانی

توضیح: نسخه اولیه پروژه مشکل تولید عبارات یکریخت را داشت که از نظر معنایی و لفظی بسیار به یکدیگر نزدیک بودند. به همین دلیل استاد پیشنهاد نمودند که هر عبارت‌کلیدی که انتخاب شد قبل از اینکه به خروجی رود با عبارت‌های کلیدی استخراج شده‌ی قبلی مقایسه شود اگر دارای کلمات یکسان بودند، بررسی شود که در صورتی که عبارت فعلی حداقل ۸۰ درصد امتیاز آن عبارت را دارد در خروجی نمایش داده شود در غیر این صورت از آن صرف نظر شود.

تغییرات کد: تغییرات داده شده در کد به صورت زیر می‌باشد:

```
161
162     def get_keyphrases(self, file_name, phrase_weights, number=10):
163         out_file = open('./Output/PhraseOut_' + file_name, 'w', encoding="utf8")
164         node_weight = OrderedDict(sorted(phrase_weights.items(), key=lambda t: t[1], reverse=True))
165         confirmed = 0
166         outputed_phrase = []
167         for i, (key, value) in enumerate(node_weight.items()):
168             if confirmed > number:
169                 break
170             """------>Below lines has been added<------"""
171             phrase = list()
172             phrase.append((key, value))
173             check_result = self.check_equal_phrases(outputed_phrase, phrase)
174             if check_result == False:
175                 outputed_phrase.append((key, value))
176                 out_file.write(key + ' - ' + str(value) + '\n')
177                 confirmed = confirmed + 1
178         out_file.close()
179
```

کد اضافه شده قبل از قرار دادن عبارت در فایل، آن را بررسی می‌کند که با عبارت‌کلیدی که قبل از آن چاپ شده‌است چه قدر مشترک است.

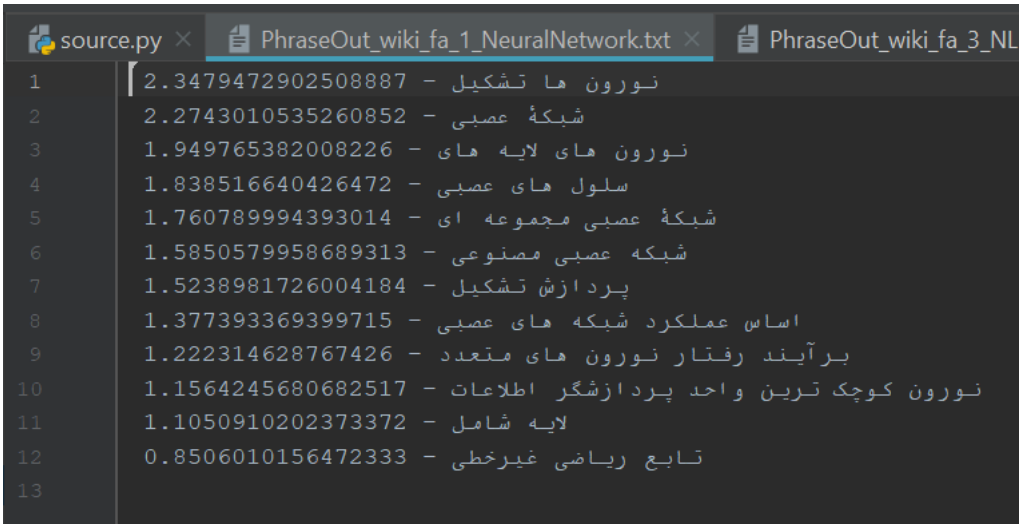
```
145     def check_equal_phrases(self, confirmed_keyphrases, candid_keyphrase):
146         keyphrase_splited = candid_keyphrase[0][0].split()
147         if len(confirmed_keyphrases) == 0:
148             return False
149         else:
150             for (key, value) in confirmed_keyphrases:
151                 current_keyphrase = key.split()
152                 common_words = set(keyphrase_splited).intersection(set(current_keyphrase))
153                 if len(common_words) > 1:
154                     # If candid phrase has score greater than %80 of confirmed phrase we use that
155                     if candid_keyphrase[0][1] > (80 / 100) * value:
156                         return False
157                     else:
158                         return True
159             return False
```

در قسمت بالا تابع اضافه شده را مشاهده می‌کنیم. در صورتی که عبارت فعلی با عبارت‌های کلیدی قبلی ۲ یا بیشتر کلمه مشترک داشته باشد بررسی می‌کنیم که اگر بیش از ۸۰ درصد امتیاز عبارت قبلی را داشته باشد به خروجی اضافه می‌شود در غیر این صورت از این عبارت صرفنظر می‌کنیم.

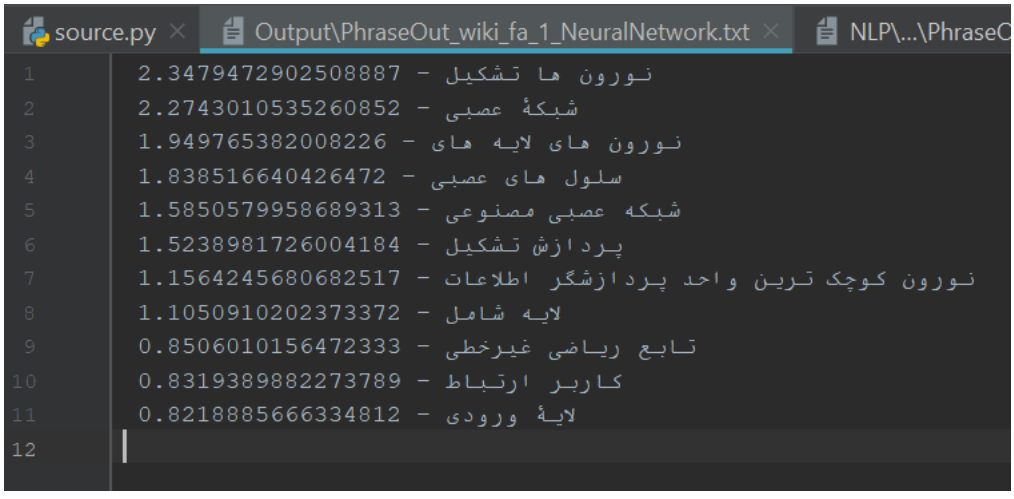
تغییرات مشاهده شده بعد از اضافه کردن کد: (اثرات تغییرات ایجاد شده در فایل سوم بیش از دو فایل اول مشاهده شد)

فایل اول:

قبل



بعد



فایل دوم:

قبل

source.py | PhraseOut_wiki_fa_2_DataBase.txt

```
4.826393493273273 - پایگاه داده های
4.003776527306387 - پایگاه داده ها
3.9984717340187492 - پایگاه داده ای
3.7133735377302584 - طراح پایگاه
3.6152713266811283 - تصمیم گیری پایگاه
3.2224043895897783 - اطلاعات سازمان
3.09074823297544 - سیستم مدیریت پایگاه
2.8513342994077724 - تولید مدل
2.494506247790386 - سیستم مدیریت پایگاه دادههم
2.4478161990909846 - طراحی مدل
2.3754469792594533 - توصیف انواع پایگاه داده ها
2.323695803020238 - منعکس کننده ساختار اطلاعات
```

بعد

source.py | PhraseOut_wiki_fa_2_DataBase.txt

```
4.826393493273273 - پایگاه داده های
4.003776527306387 - پایگاه داده ها
3.9984717340187492 - پایگاه داده ای
3.7133735377302584 - طراح پایگاه
3.6152713266811283 - تصمیم گیری پایگاه
3.2224043895897783 - اطلاعات سازمان
3.09074823297544 - سیستم مدیریت پایگاه
2.8513342994077724 - تولید مدل
2.494506247790386 - سیستم مدیریت پایگاه دادههم
2.4478161990909846 - طراحی مدل
2.323695803020238 - منعکس کننده ساختار اطلاعات
```

فایل سوم:

قبل

source.py | PhraseOut_wiki_fa_3_NLP.txt

```
4.433850985642123 - زبان طبیعی
3.8855069323215354 - پردازش زبان طبیعی
3.512580956988304 - درک زبان طبیعی
3.4558133979446604 - کاربردهای پردازش زبان طبیعی
3.2863219601747957 - زبان طبیعی انسانی
3.2007168970497775 - پردازش زبان گفتاریو زبان نوشتاری
2.8851040067651113 - پردازش زبان های طبیعی
2.754715624732765 - کاربردهای گفتاری پردازش زبان
2.3452284465816566 - کاربردهای نوشتاری
2.3426395858514426 - پردازش زبان های طبیعی عبارت
2.3211348267115053 - علوم رایانه
2.309534117152534 - کاربردهای متنوع پردازش زبان های طبیعی

```

بعد

source.py | PhraseOut_wiki_fa_3_NLP.txt

```
4.433850985642123 - زبان طبیعی
3.8855069323215354 - پردازش زبان طبیعی
3.2007168970497775 - پردازش زبان گفتاریو زبان نوشتاری
2.3452284465816566 - کاربردهای نوشتاری
2.3211348267115053 - علوم رایانه
2.1520420075110613 - دانش زبان شناسان
1.6249528426858961 - کاربردهای گفتاری
1.5659936672218266 - پردازش اطلاعات زبانی
1.41984372147448 - سیستم های آموزش
1.320720140035594 - سیستم های پرسش
1.3071537765452772 - سیستم های کنترلی

```