بسم اللّه الرحمن الرحیم

تمرین سری سوم

یافتن عبارات مشابه در متون دینی با استفاده از جاگذاری کلمات

پردازش زبان‌های طبیعی

بهار ۱۴۰۱

فهرست

یافتن عبارات مشابه در متون دینی با استفاده از جاگذاری کلمات

تکلیف سری سوم

شناسنامه

تمرین سری سوم	
۱۴۰۱–۱۴۰۰	دانشگاه صنعتی شریف – دانشکده کامپیوتر
مشخصات درس	
عنوان درس	پردازش زبان‌های طبیعی
استاد مربوطه	آقای دکتر احسان‌الدین عسگری
استاد حل تمرین	تیم اساتید حل تمرین
مشخصات گروه	
نام و نام خانوادگی	امیر پورمند پویا خانی مهدی آخی
مشخصات نویسنده و ارسال کننده گزارش	مهدی آخی – ۹۹۲۰۱۴۹۸
نشانی الکترونیکی	mahdiakhi@ce.sharif.edu
مشخصات سند	
عنوان	گزارش تمرین سری سوم
تاریخ تحویل	۱۴۰۱/۳/۱۹

تعریف مسئله

برای این که بتوان از علوم داده در پردازش متن و به طور خصوص پردازش متون زبان طبیعی استفاده کرد نیاز است تا آنها را به فرمی قابل استفاده و قابل درک برای کامپیوترها تبدیل کرد نیاز است تا حروف و علائم نگارشی به عدد تبدیل شوند(Word to Vec). برای این کار روش‌های مختلفی وجود دارد که به صورت کلی به دودسته تقسیم می‌شوند: دسته اول روش‌هایی هستند که خاصیت‌های آماری کلمات را به عدد مدل می‌کنند مانند TF-IDF و دسته دوم روش‌هایی هستند که خواص معنایی کلمات را در خود دارند مانند FastText یا BERT. هر دوی این دسته‌ها کاربردهای گوناگونی دارند مانند حدس کلمه، تکمیل جمله، ترمیم جملات، پیدا کردن جملات مشابه(متنی و معنایی). در این تکلیف سعی داریم تا با بهره‌گیری از هر دو دسته عملیات پیدا کردن عبارات مشابه در متون دینی(قرآن کریم، نهج‌البلاغه و صحیفه سجادیه) را پیاده‌سازی کنیم. آن چه در ادامه این گزارش آمده شرح عملیات و تکنیک‌های استفاده شده در این مسیر است.

یافتن عبارات مشابه در متون دینی با استفاده از جاگذاری کلمات

تکلیف سری سوم

داده‌ها و پیش‌پردازش

تمامی داده‌ها از گیت‌هاب درس به آدرس زیر برداشته شده‌اند.

```
git clone https://github.com/language-ml/course-nlp-ir-1-text-exploring
```

لود کردن آیات قرآن و عبارات نهج البلاغه چالش خاصی نداشت ولی لود کردن عبارات صحیفه سجادیه نیاز به پیش پردازش‌های دیگری هم داشت زیرا آیات با شماره آیه شروع می‌شدند و بعضی از خطوط نیز اصلا شامل عبارتی نبودند. سپس داده‌ها با استفاده از کتابخانه camel_tools که یک کتابخانه استاندارد برای پردازش متن در زبان عربی بصورت عام هست، پیش پردازش شدند.از آنجایی که متون ثابتی در مجموعه داده وجود داشت عملیات پیش‌پردازش شامل مراحل عمومی بود. از جمله:

- حذف اعراب و تنوین‌ها
- حذف برخی علائم قرائت
- حذف کلمات با اندازه کوچکتر از ۳
- یک شکل کردن حروفی مانند «ک» و «ی» عربی

. در اینجا یک نمونه از داده‌های پیش پردازش شده برای هر کدام از ۳ رفرنس اصلی نشان داده شده است.

```
*Quranic(pure): وَإِذَا سَأَلَكَ عِبَادِي عَنِّي فَإِنِّي قَرِيبٌ ۖ أُجِيبُ دَعْوَةَ الدَّاعِ إِذَا دَعَانِ ۖ فَلْيَسْتَجِيبُوا لِي وَلْيُؤْمِنُوا بِي لَعَلَّهُمْ يَرْشُدُونَ
*Quranic(Processed): واذا سالك عبادي عني فاني قريب ۖ اجيب دعوه الداع اذا دعان ۖ فليستجيبوا لي وليؤمنوا بي لعلهم يرشدون

*Nahj(pure): وَ قَالَ ع لِلظَّالِمِ البَادِي غداً بِكَفِّهِ عَضَّةٌ
*Nahj(Processed): و قال ع للظالم البادى غدا بكفه عضه

*Sahifa(pure):  حَمْداً يَرْتَفِعُ مِنَّا إِلَى أَعْلَى عِلِّيِّينَ فِي كِتَابٍ مَرْقُومٍ يَشْهَدُهُ الْمُقَرَّبُونَ.
*Sahifa(Processed):  و كاشف في الدعاء اليک حامته
```

بعد از عملیات پیش‌پردازش نوبت به توکن کردن جملات(داکیومنت‌ها) می‌رسد. در تصویر زیر خروجی نمونه این مرحله را مشاهده می‌کنید. می‌توانید ببینید که در بردار زیر کلمات کمتر از سه حرف حذف شده‌اند.

```
Tokenized:
 *Quranic Example: ['قاني', 'الدع', 'ربه', 'كلمات', 'فلب', 'عليه', 'انه', 'التواب', 'الرحيم']
*Nahj Example: ['قبح', 'الله', 'مصلكه', 'فعل', 'فعل', 'السلده', 'فرار', 'العبيد', 'فما', 'انطق', 'مادحه', 'حتى', 'اسكته', 'صدق', 'واصفه', 'حتى', 'بكته', 'الفم', 'لاخذنا', 'ميسوره', 'انتظرنا', 'بماله', 'وفوره']
*Sahifa Example: ['قبح', 'الله', 'مصلكه', 'فعل', 'فعل', 'السلده', 'فرار', 'العبيد', 'فما', 'انطق', 'مادحه', 'حتى', 'اسكته', 'صدق', 'واصفه', 'حتى', 'بكته', 'الفم', 'لاخذنا', 'ميسوره', 'انتظرنا', 'بماله', 'وفوره']
```

روش اول – یافتن عبارات مشابه با استفاده از TF-IDF

در این بخش در ابتدا از داده‌های پیش‌پردازش شده در مرحله قبل استفاده می‌کنیم به طوری‌که آنها را توسط تابع نرمال‌سازی که داشتیم، نرمال می‌کنیم.

```
#Normalize all verses from quran, Nahj, and Sahife
#_____________________________________________________________________________________________
verse_dict_nrmlz = {k:normalize_arabic(v) for k,v in tqdm.tqdm(verse_complete_dict.items())}
nahj_dict_nrmlz = {k:normalize_arabic(v) for k,v in tqdm.tqdm(nahj_complete_dict.items())}
sahife_dict_nrmlz = {normalize_arabic(v) for v in tqdm.tqdm(sahife_complete_dict)}
#_____________________________________________________________________________________________
```

```
100%|██████████| 6236/6236 [00:00<00:00, 11174.55it/s]
100%|██████████| 800/800 [00:00<00:00, 3515.74it/s]
100%|██████████| 924/924 [00:00<00:00, 7923.86it/s]
```

سپس دیکشنری‌های نرمال‌شده و نرمال‌نشده را به لیست تبدیل کرده و سپس لیست‌ها را به دیتافریم‌های کتابخانه Pandas تبدیل می‌کنیم تا بتوانیم ساده‌تر با آنها کار کنیم.

```
# Converting standard and normalize dictionaries(index:verse) from quran, Nahj, and Sahife to Pandas DataFrame
# with column names: num and verse
#_____________________________________________________________________________________________
df_verse_complete_dict = pd.DataFrame(list(verse_complete_dict.items()), columns = ['num','verse'])
df_verse_complete_dict_nrmlz = pd.DataFrame(list(verse_dict_nrmlz.items()), columns = ['num','verse'])
df_nahj_complete_dict = pd.DataFrame(list(nahj_complete_dict.items()), columns = ['num','verse'])
df_nahj_complete_dict_nrmlz = pd.DataFrame(list(nahj_dict_nrmlz.items()), columns = ['num','verse'])
df_sahife_complete_dict = pd.DataFrame(list(sahife_complete_dict), columns = ['verse'])
df_sahife_complete_dict_nrmlz = pd.DataFrame(list(sahife_dict_nrmlz), columns= ['verse'])
#_____________________________________________________________________________________________
```

سپس توسط کتابخانه scikit-learn به تعداد ۶ تا آبجکت از وکتورایزرهای TF-IDF میسازیم و توسط نام‌گذاری آن‌هارا برای استفاده‌های بعدی متمایز می‌کنیم.

```
# Create TfidfVectorizer object (first three ones are Word level and others are Character Level)
#_____________________________________________________________________________________________
vectorizer_quran = TfidfVectorizer(lowercase = False)
vectorizer_nahj = TfidfVectorizer(lowercase = False)
vectorizer_sahife = TfidfVectorizer(lowercase = False)
vectorizer_quran_char = TfidfVectorizer(lowercase = False,analyzer='char')
vectorizer_nahj_char = TfidfVectorizer(lowercase = False,analyzer='char')
vectorizer_sahife_char = TfidfVectorizer(lowercase = False,analyzer='char')
#_____________________________________________________________________________________________
```

سپس این آبجکت‌هارا روی ستون verse که ستونی هست که آیه‌ها و عبارات کتب دینی را در خود دارند، به عبارتی fit کرده و سپس داده‌هارا به آن embedding تبدیل می‌کنیم. خروجی این کار ۶ ماتریس می‌شود که با توجه به ویژگی‌های مشخص شده، ابعاد متفاوتی دارند. این ماتریس‌ها به تعداد آیه/عبارت کتب دینی سطر و به تعداد واحدهای متفاوت و متمایز در این متون ستون دارند. برای مثال ۳ ماتریس اول مربوط به بدست آوردن تعبیه برای کتب قرآن، نهج‌البلاغه و صحیفه سجادیه است به طوری که روی کلمات این کتب این فضای تعبیه ساخته شود. لذا برای قرآن به تعداد کلمات منحصربه‌فرد آن یعنی ۱۴۶۵۸ تا، برای نهج‌البلاغه نیز به همین منوال ۲۱۰۹۴ تا و برای صحیفه سجادیه نیز ۷۷۱۱ ستون در ماتریس متناظرشان در نظر گرفته می‌شود. ۳ ماتریس بعدی متعلق به همین ۳ کتاب اما در حالتی هستند که روی کاراکترهای متون تعبیه ساخته می‌شود. لذا برای قرآن ۴۰، نهج‌البلاغه ۳۶ و صحیفه سجادیه ۴۷ ستون در نظر گرفته می‌شود(بخاطر کاراکترهای خاص در متون تعداد ستون‌ها متفاوت است).

```
# Generate matrices of word vectors
#_____________________________________________________________________________________________
tfidf_matrix_quran = vectorizer_quran.fit_transform(df_verse_complete_dict_nrmlz['verse'])
tfidf_matrix_nahj = vectorizer_nahj.fit_transform(df_nahj_complete_dict_nrmlz['verse'])
tfidf_matrix_sahife = vectorizer_sahife.fit_transform(df_sahife_complete_dict_nrmlz['verse'])
tfidf_matrix_quran_char = vectorizer_quran_char.fit_transform(df_verse_complete_dict_nrmlz['verse'])
tfidf_matrix_nahj_char = vectorizer_nahj_char.fit_transform(df_nahj_complete_dict_nrmlz['verse'])
tfidf_matrix_sahife_char = vectorizer_sahife_char.fit_transform(df_sahife_complete_dict_nrmlz['verse'])
#_____________________________________________________________________________________________

# Print shapes of tfidf_matrices
#_____________________________________________________________________________________________
print("Shape of tfidf_matrix(Word-Level) of quran is: ",tfidf_matrix_quran.shape)
print("Shape of tfidf_matrix(Word-Level) of Nahj is: ",tfidf_matrix_nahj.shape)
print("Shape of tfidf_matrix(Word-Level) of Sahife is: ",tfidf_matrix_sahife.shape)
print("Shape of tfidf_matrix of quran is: ",tfidf_matrix_quran_char.shape)
print("Shape of tfidf_matrix of Nahj is: ",tfidf_matrix_nahj_char.shape)
print("Shape of tfidf_matrix of Sahife is: ",tfidf_matrix_sahife_char.shape)
#_____________________________________________________________________________________________
```

```
Shape of tfidf_matrix(Word-Level) of quran is:  (6236, 14658)
Shape of tfidf_matrix(Word-Level) of Nahj is:  (800, 21094)
Shape of tfidf_matrix(Word-Level) of Sahife is:  (924, 7711)
Shape of tfidf_matrix of quran is:  (6236, 40)
Shape of tfidf_matrix of Nahj is:  (800, 36)
Shape of tfidf_matrix of Sahife is:  (924, 47)
```

سپس یک لیست از زوج (ایندکس:آیه) از هر سه کتاب استخراج می‌کنیم زیرا در ادامه نیاز به جستجو آیه بر اساس ایندکس یا برعکس می‌شود. در نهایت در بخش بعد ورودی مدل را مشخص کرده تا در ادامه از هر کتاب، ۱۰ آیه/جمله مشابه از لحاظ معیار فاصله فضای تعبیه ساخته شده توسط وکتورایزر TF-IDF نشان داده‌می‌شود. برای محاسبه MRR به جهت مقایسه این روش شباهت‌یابی و دو روش بعدی، ترتیب ۱۰ آیه/جمله مشابه را بهم ریخته و به دو نفر اعضای تیم که این خروجی‌هارا ندیده‌اند، داده می‌شود تا به صورت دستی مشابه‌ترین آیه/جمله مد نظرشان را اعلام کنند.

ورودی که برای روش اول یعنی TF-IDF آزمایش شد، یکی از آیه‌های قرآن یعنی آیه ۴ سوره ناس: «مِنْ شَرِّ الْوَسْوَاسِ الْخَنَّاسِ» بود. توجه شود که چون یکی از آیه‌های قرآن را به عنوان ورودی مشخص کردیم، طبیعی است که برترین و مشابه‌ترین آیه قرآنی با این آیه، خودش باشد لذا برای قرآن، ما ۹ آیه مشابه بعدی را برای محاسبه دستی استفاده کردیم. ۹ آیه برتر در شباهت به ترتیب عبارتند از:

۱- مِنْ شَرِّ مَا خَلَقَ

۲- وَمِنْ شَرِّ النَّفَّاثَاتِ فِی الْعُقَدِ

۳- وَمِنْ شَرِّ غَاسِقٍ إِذَا وَقَبَ

۴- وَمِنْ شَرِّ حَاسِدٍ إِذَا حَسَدَ

۵- إِنَّ شَرَّ الدَّوَابِّ عِنْدَ اللَّهِ الَّذِینَ کَفَرُوا فَهُمْ لَا یُؤْمِنُونَ

۶- إِنَّ شَرَّ الدَّوَابِّ عِنْدَ اللَّهِ الصُّمُّ الْبُکْمُ الَّذِینَ لَا یَعْقِلُونَ

۷- إِنَّ الَّذِینَ کَفَرُوا مِنْ أَهْلِ الْکِتَابِ وَالْمُشْرِکِینَ فِی نَارِ جَهَنَّمَ خَالِدِینَ فِیهَا ۚ أُولَٰئِکَ هُمْ شَرُّ الْبَرِیَّةِ

۸- الَّذِینَ یُحْشَرُونَ عَلَىٰ وُجُوهِهِمْ إِلَىٰ جَهَنَّمَ أُولَٰئِکَ شَرٌّ مَکَانًا وَأَضَلُّ سَبِیلًا

۹- فَوَقَاهُمُ اللَّهُ شَرَّ ذَٰلِکَ الْیَوْمِ وَلَقَّاهُمْ نَضْرَةً وَسُرُورًا

نفر اول آیه شماره ۲ و نفر دوم آیه شماره ۴ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Quran}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{2}+\frac{1}{4}\right] = 0.375$$

به همین ترتیب آیه/جمله‌های برتر در نهج‌البلاغه عبارت بودند از:

۱- وَ إِنّ عَمَلَکَ لَیسَ لَکَ بِطُعمَةٍ وَ لَکِنّهُ فِی عُنُقِکَ أَمَانَةٌ وَ أَنتَ مُستَرعًی لِمَن فَوقَکَ لَیسَ لَکَ أَن تَفتَاتَ فِی رَعِیّةٍ وَ لَا تُخَاطِرَ إِلّا بِوَثِیقَةٍ وَ فِی یَدَیکَ مَالٌ مِن مَالِ اللّهِ عَزّ وَ جَلّ وَ أَنتَ مِن خُزّانِهِ حَتّی تُسَلّمَهُ إِلَیّ وَ لَعَلّی أَلّا أَکُونَ شَرّ وُلَاتِکَ لَکَ وَ السّلَامُ

۲- فَاخفِض لَهُم جَنَاحَکَ وَ أَلِن لَهُم جَانِبَکَ وَ ابسُط لَهُم وَجهَکَ وَ آسِ بَینَهُم فِی اللّحظَةِ وَ النّظرَةِ حَتّی لَا یَطمَعَ العُظَمَاءُ فِی حَیفِکَ لَهُم وَ لَا یَیأَسَ الضّعَفَاءُ مِن عَدلِکَ عَلَیهِم فَإِنّ اللّهَ تَعَالَی یُسَائِلُکُم مَعشَرَ عِبَادِهِ عَنِ الصّغِیرَةِ مِن أَعمَالِکُم وَ الکَبِیرَةِ وَ الظّاهِرَةِ وَ المَستُورَةِ فَإِن یُعَذّب فَأَنتُم أَظلَمُ وَ إِن یَعفُ فَهُوَ أَکرَمُ وَ اعلَمُوا عِبَادَ اللّهِ أَنّ المُتّقِینَ ذَهَبُوا بِعَاجِلِ الدّنیَا وَ آجِلِ

یافتن عبارات مشابه در متون دینی با استفاده از جاگذاری کلمات

تکلیف سری سوم

الآخِرَةِ فَشَارَكُوا أهلَ الدّنيَا فِی دُنيَاهُم وَ لَم يُشَارِكُوا أهلَ الدّنيَا فِی آخِرَتِهِم سَكَنُوا الدّنيَا بِأفضَلِ مَا سُكِنَت وَ أكَلُوهَا بِأفضَلِ مَا أُكِلَت فَحَظُوا مِنَ الدّنيَا بِمَا حظِيَ بِهِ المُترَفُونَ وَ أخَذُوا مِنهَا مَا أخَذَهُ الجَبَابِرَةُ المُتَكَبّرُونَ ثُمّ انقَلَبُوا عَنهَا بِالزّادِ المُبَلّغِ وَ المَتجَرِ الرّابِحِ أصَابُوا لَذّةَ زُهدِ الدّنيَا فِی دُنيَاهُم وَ تَيَقّنُوا أنّهُم جِيرَانُ اللّهِ غَداً فِی آخِرَتِهِم لَا تُرَدّ لَهُم دَعوَةٌ وَ لَا يَنقُصُ لَهُم نَصِيبٌ مِن لَذّةٍ فَاحذَرُوا عِبَادَ اللّهِ المَوتَ وَ قُربَهُ وَ أعِدّوا لَهُ عُدّتَهُ فَإنّهُ يأتِي بِأمرٍ عَظِيمٍ وَ خَطبٍ جَلِيلٍ بِخَيرٍ لَا يَكُونُ مَعَهُ شَرّ أبَداً أو شَرّ لَا يَكُونُ مَعَهُ خَيرٌ أبَداً فَمَن أقرَبُ إلَى الجَنّةِ مِن عَامِلِهَا وَ مَن أقرَبُ إلَى النّارِ مِن عَامِلِهَا وَ أنتُم طُرَدَاءُ المَوتِ إن أقَمتُم لَهُ أخَذَكُم وَ إن فَرَرتُم مِنهُ أدرَكَكُم وَ هُوَ ألزَمُ لَكُم مِن ظِلّكُم المَوتُ مَعقُودٌ بِنَوَاصِيكُم وَ الدّنيَا تُطوَى مِن خَلفِكُم فَاحذَرُوا نَاراً قَعرُهَا بَعِيدٌ وَ حَرّهَا شَدِيدٌ وَ عَذَابُهَا جَدِيدٌ دَارٌ لَيسَ فِيهَا رَحمَةٌ وَ لَا تُسمَعُ فِيهَا دَعوَةٌ وَ لَا تُفَرّجُ فِيهَا كُربَةٌ وَ إنِ استَطَعتُم أن يَشتَدّ خَوفُكُم مِنَ اللّهِ وَ أن يَحسُنَ ظَنّكُم بِهِ فَاجمَعُوا بَينَهُمَا فَإنّ العَبدَ إنّمَا يَكُونُ حُسنُ ظَنّهِ بِرَبّهِ عَلَى قَدرِ خَوفِهِ مِن رَبّهِ وَ إنّ أحسَنَ النّاسِ ظَنّاً بِاللّهِ أشَدّهُم خَوفاً لِلّهِ وَ اعلَم يَا مُحَمّدَ بنَ أبِی بَكرٍ أنّی قَد وَلّيتُكَ أعظَمَ أجنَادِی فِی نفسِی أهلَ مِصرَ فَأنتَ مَحقُوقٌ أن تُخَالِفَ عَلَى نَفسِكَ وَ أن تُنَافِحَ عَن دِينِكَ وَ لَو لَم يَكُن لَكَ إلّا سَاعَةٌ مِنَ الدّهرِ وَ لَا تُسخِطِ اللّهَ بِرِضَا أحَدٍ مِن خَلقِهِ فَإنّ فِی اللّهِ خَلَفاً مِن غَيرِهِ وَ لَيسَ مِنَ اللّهِ خَلَفٌ فِی غَيرِهِ صَلّ الصّلَاةَ لِوَقتِهَا المُؤَقّتِ لَهَا وَ لَا تُعَجّل وَقتَهَا لِفَرَاغٍ وَ لَا تُؤَخّرهَا عَن وَقتِهَا لِاشتِغَالٍ وَ اعلَم أنّ كُلّ شَیءٍ مِن عَمَلِكَ تَبَعٌ لِصَلَاتِكَ وَ مِنهُفَإنّهُ لَا سَوَاءَ إمَامُ الهُدَى وَ إمَامُ الرّدَى وَ وَلِیّ النّبِیّ وَ عَدُوّ النّبِیّ وَ لَقَد قَالَ لِی رَسُولُ اللّهِ-ص إنِیّ لَا أخَافُ عَلَى أمّتِی مُؤمِناً وَ لَا مُشرِكاً أمّا المُؤمِنُ فَيَمنَعُهُ اللّهُ بِإيمَانِهِ وَ أمّا المُشرِكُ فَيَقمَعُهُ اللّهُ بِشِركِهِ وَ لَكِنِیّ أخَافُ عَلَيكُم كُلّ مُنَافِقِ الجَنَانِ عَالِمِ اللّسَانِ يَقُولُ مَا تَعرِفُونَ وَ يَفعَلُ مَا تُنكِرُونَ

۳- وَ قَالَ ع شَرّ الإخوَانِ مَن تُكُلّفَ لَهُ

۴- وَ قَالَ ع يأتِي عَلَى النّاسِ زَمَانٌ لَا يَبقَى فِيهِم مِنَ القُرآنِ إلّا رَسمُهُ وَ مِنَ الإسلَامِ إلّا اسمُهُ وَ مَسَاجِدُهُم يَومَئِذٍ عَامِرَةٌ مِنَ البِنَاءِ خَرَابٌ مِنَ الهُدَى سُكّانُهَا وَ عُمّارُهَا شَرّ أهلِ الأرضِ مِنهُم تَخرُجُ الفِتنَةُ وَ إلَيهِم تأوِی الخَطِيئَةُ يَرُدّونَ مَن شَذّ عَنهَا فِيهَا وَ يَسُوقُونَ مَن تَأخّرَ عَنهَا إلَيهَا يَقُولُ اللّهُ سُبحَانَهُ فَبِی حَلَفتُ لَأبعَثَنّ عَلَى أولَئِكَ فِتنَةً تَترُكُ الحَلِيمَ فِيهَا حَيرَانَ وَ قَد فَعَلَ وَ نَحنُ نَستَقِيلُ اللّهَ عَثرَةَ الغَفلَةِ

۵- وَ قَالَ ع فَاعِلُ الخَيرِ خَيرٌ مِنهُ وَ فَاعِلُ الشّرّ شَرّ مِنهُ

۶- وَ قَالَ ع المَرأةُ شَرّ كُلّهَا وَ شَرّ مَا فِيهَا أنّهُ لَا بُدّ مِنهَا

۷- وَ قَالَ ع مَا خَيرٌ بِخَيرٍ بَعدَهُ النّارُ وَ مَا شَرّ بِشَرّ بَعدَهُ الجَنّةُ وَ كُلّ نَعِيمٍ دُونَ الجَنّةِ فَهُوَ مَحقُورٌ وَ كُلّ بَلَاءٍ دُونَ النّارِ عَافِيَةٌ

۸- إنّ اللّهَ بَعَثَ مُحَمّداً-ص نَذِيراً لِلعَالَمِينَ وَ أمِيناً عَلَى التّنزِيلِ وَ أنتُم مَعشَرَ العَرَبِ عَلَى شَرّ دِينٍ وَ فِی شَرّ دَارٍ مُنِيخُونَ بَينَ حِجَارَةٍ خُشنٍ وَ حَيّاتٍ صُمّ تَشرَبُونَ الكَدِرَ وَ تَأكُلُونَ الجَشِبَ وَ تَسفِكُونَ دِمَاءَكُم وَ تَقطَعُونَ أرحَامَكُم الأصنَامُ فِيكُم مَنصُوبَةٌ وَ الآثَامُ بِكُم مَعصُوبَةٌ فَنَظَرتُ فَإذَا لَيسَ لِی مُعِينٌ إلّا أهلُ بَيتِی فَضَنَنتُ بِهِم عَنِ المَوتِ وَ أغضَيتُ عَلَى القَذَى وَ شَرِبتُ عَلَى

الشّجَا وَ صَبَرتُ عَلَى أخذِ الكَظمِ وَ عَلَى أمَرّ مِن طَعمِ العَلقَمِ ومنها وَ لَم يُبَايِع حَتّى شَرَطَ أن يُؤتِيَهُ عَلَى البَيعَةِ ثَمَناً فَلَا ظَفِرَت يَدُ البَائِعِ وَ خَزِيَت أمَانَةُ المُبتَاعِ فَخُذُوا لِلحَربِ أهبَتَهَا وَ أعِدّوا لَهَا عُدّتَهَا فَقَد شَبّ لَظَاهَا وَ عَلَا سَنَاهَا وَ استَشعِرُوا الصّبرَ فَإِنّهُ أدعَى إِلَى النّصرِ

٩- فَإِنّکَ قَد جَعَلتَ دِينَکَ تَبَعاً لِدُنيَا امرِئٍ ظَاهِرٍ غَيّهُ مَهتُوكٍ سِترُهُ يَشِينُ الكَرِيمَ بِمَجلِسِهِ وَ يُسَفّهُ الحَلِيمَ بِخِلطَتِهِ فَاتّبَعتَ أثَرَهُ وَ طَلَبتَ فَضلَهُ اتّبَاعَ الكَلبِ لِلضّرغَامِ يَلُوذُ بِمَخَالِبِهِ وَ يَنتَظِرُ مَا يُلقَى إِلَيهِ مِن فَضلِ فَرِيسَتِهِ فَأذهَبتَ دُنيَاکَ وَ آخِرَتَکَ وَ لَو بِالحَقّ أخَذتَ أدرَکتَ مَا طَلَبتَ فَإِن يمُكَنّيَ اللّهُ مِنکَ وَ مِن ابنِ أبِي سُفيَانَ أجزِكُمَا بِمَا قَدّمتُمَا وَ إِن تُعجِزَا وَ تَبقَيَا فَمَا أمَامَكُمَا شَرّ لَكُمَا وَ السّلَامُ

١٠- صَابَكُم حَاصِبٌ وَ لَا بقَيَ مِنكُم آثِرٌ أ بَعدَ إِيمَانِي بِاللّهِ وَ جِهَادِي مَعَ رَسُولِ اللّهِ-ص أشهَدُ عَلَى نفَسِي بِالكُفرِ لَقَد ضَلَلتُ إِذاً وَ ما أنَا مِنَ المُهتَدِينَفَأوبُوا شَرّ مَآبٍ وَ ارجِعُوا عَلَى أثَرِ الأعقَابِ أمَا إِنّكُم سَتَلقَونَ بعَدِي ذُلّا شَامِلًا وَ سَيفاً قَاطِعاً وَ أثَرَةً يَتّخِذُهَا الظّالِمُونَ فِيكُم سُنّةً

نفر اول آیه شماره ۸ و نفر دوم آیه شماره ۶ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Nahj}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{8}+\frac{1}{6}\right] = 0.146$$

به همین ترتیب آیه/جمله‌های برتر در صحیفه سجادیه عبارت بودند از:

١- فَسُبْحَانَکَ! مَا أبْيَنَ كَرَمَکَ فِي مُعَامَلَةِ مَنْ أطَاعَکَ أوْ عَصَاکَ: تَشْكُرُ لِلْمُطِيعِ مَا أنْتَ تَوَلّيْتَهُ لَهُ، وَ تُمْلِي لِلْعَاصِي فِيمَا تَمْلِكُ مُعَاجَلَتَهُ فِيهِ.

٢- وَ أجْزِلْ لَنَا فِيهِ مِنَ الْحَسَنَاتِ، وَ أخْلِنَا فِيهِ مِنَ السَّيِّئَاتِ، وَ امْلَأ لَنَا مَا بَيْنَ طَرَفَيْهِ حَمْداً وَ شُكْراً وَ أجْراً وَ ذُخْراً وَ فَضْلًا وَ إِحْسَاناً.

٣- وَ امْنُنْ عَلَيَّ بِالصِّحَّةِ وَ الْأمْنِ وَ السَّلَامَةِ فِي دِينِي وَ بَدَنِي، وَ الْبَصِيرَةِ فِي قَلْبِي، وَ النَّفَاذِ فِي أمُورِي، وَ الْخَشْيَةِ لَکَ، وَ الْخَوْفِ مِنْکَ، وَ الْقُوَّةِ عَلَى مَا أمَرْتَنِي بِهِ مِنْ طَاعَتِکَ، وَ الاجْتِنَابِ لِمَا نَهَيْتَنِيإِذَا انْصَرَفَ مِنْ صَلَاتِهِ قَامَ قَائِماً ثُمَّ اسْتَقْبَلَ الْقِبْلَةَ، وَ فِي يَوْمِ الْجُمُعَةِ، فَقَالَ:) عَنْهُ مِنْ مَعْصِيَتِکَ.

۴- وَ أَنْ نَتَقَرَّبَ إِلَيْکَ فِيهِ مِنَ الْأَعْمَالِ الزَّاکِيَةِ بِمَا تُطَهِّرُنَا بِهِ مِنَ الذُّنُوبِ، وَ تَعْصِمُنَا فِيهِ مِمَّا نَسْتَأْنِفُ مِنَ الْعُيُوبِ، حَتَّى لَا يُورِدَ عَلَيْکَ أَحَدٌ مِنْ مَلَائِکَتِکَ إِلَّا دُونَ مَا نُورِدُ مِنْ أَبْوَابِ الطَّاعَةِ لَکَ، وَ أَنْوَاعِ الْقُرْبَةِ إِلَيْکَ.

۵- وَ أَسْتَحْمِلُکَ مِنْ ذُنُوبِي مَا قَدْ بَهَظَنِي حَمْلُهُ، وَ أَسْتَعِينُ بِکَ عَلَى مَا قَدْ فَدَحَنِي ثِقْلُهُ.

۶- وَ لَا تَذَرْنِي فِي طُغْيَانِي عَامِهاً، وَ لَا فِي غَمْرَتِي سَاهِياً حَتَّى حِينٍ، وَ لَا تَجْعَلْنِي عِظَةً لِمَنِ اتَّعَظَ، وَ لَا نَکَالًا لِمَنِ اعْتَبَرَ، وَ لَا فِتْنَةً لِمَنْ نَظَرَ، وَ لَا تَمْکُرْ بِي فِيمَنْ تَمْکُرُ بِهِ، وَ لَا تَسْتَبْدِلْ بِي غَيْرِي، وَ لَا تُغَيِّرْ لِي اسْماً، وَ لَا تُبَدِّلْ لِي جِسْماً، وَ لَا تَتَّخِذْنِي هُزُواً لِخَلْقِکَ، وَ لَا سُخْرِيّاً لَکَ، وَ لَا تَبَعاً إِلَّا لِمَرْضَاتِکَ، وَ لَا مُمْتَهَناً إِلَّا بِالانْتِقَامِ لَکَ.

۷- اللَّهُمَّ اجْعَلْ مَا يُلْقِي الشَّيْطَانُ فِي رُوعِي مِنَ التَّمَنِّي وَ التَّظَنِّي وَ الْحَسَدِ ذِکْراً لِعَظَمَتِکَ، وَ تَفَکُّراً فِي قُدْرَتِکَ، وَ تَدْبِيراً عَلَى عَدُوِّکَ، وَ مَا أَجْرَى عَلَى لِسَانِي مِنْ لَفْظَةِ فُحْشٍ أَوْ هُجْرٍ أَوْ شَتْمِ عِرْضٍ أَوْ شَهَادَةِ بَاطِلٍ أَوِ اغْتِيَابِ مُؤْمِنٍ غَائِبٍ أَوْ سَبِّ حَاضِرٍ وَ مَا أَشْبَهَ ذَلِکَ نُطْقاً بِالْحَمْدِ لَکَ، وَ إِغْرَاقاً فِي الثَّنَاءِ عَلَيْکَ، وَ ذَهَاباً فِي تَمْجِيدِکَ، وَ شُکْراً لِنِعْمَتِکَ، وَ اعْتِرَافاً بِإِحْسَانِکَ، وَ إِحْصَاءً لِمِنَنِکَ.

۸- وَ نَعُوذُ بِکَ أَنْ نَنْطَوِيَ عَلَى غِشِّ أَحَدٍ، وَ أَنْ نُعْجِبَ بِأَعْمَالِنَا، وَ نَمُدَّ فِي آمَالِنَا

۹- اللَّهُمَّ صَلِّ عَلَى مُحَمَّدٍ وَ آلِهِ، وَ اکْفِنِي مَا يَشْغَلُنِي الِاهْتِمَامُ بِهِ، وَ اسْتَعْمِلْنِي بِمَا تَسْأَلُنِي غَداً عَنْهُ، وَ اسْتَفْرِغْ أَيَّامِي فِيمَا خَلَقْتَنِي لَهُ، وَ أَغْنِنِي وَ أَوْسِعْ عَلَيَّ فِي رِزْقِکَ، وَ لَا تَفْتِنِّي بِالنَّظَرِ، وَ أَعِزَّنِي وَ لَا تَبْتَلِيَنِّي بِالْکِبْرِ، وَ عَبِّدْنِي لَکَ وَ لَا تُفْسِدْ عِبَادَتِي بِالْعُجْبِ، وَ أَجْرِ لِلنَّاسِ عَلَى يَدَيَ الْخَيْرَ وَ لَا تَمْحَقْهُ بِالْمَنِّ، وَ هَبْ لِي مَعَالِيَ الْأَخْلَاقِ، وَ اعْصِمْنِي مِنَ الْفَخْرِ.

۱۰- حَتَّى اسْتَتَبَّ لَهُ مَا حَاوَلَ فِي أَعْدَائِکَ

نفر اول آیه شماره ۷ و نفر دوم آیه شماره ۲ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Sahifa}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{7}+\frac{1}{2}\right] = 0.321$$

بنابراین اگر روش TF-IDF Word-Level را بخواهیم با یک MRR نشان دهیم، با میانگین گیری از این سه عدد بدست آمده داریم:

$$MRR_{TF-IDF} = \frac{0.375 + 0.146 + 0.321}{3} = 28\%$$

به همین ترتیب برای حالتی که روی فضای کاراکترها تعبیه‌سازی کنیم، نفر اول برای هر ۳ کتب، آیه‌های ۳ و ۵ و ۹ و نفر دوم آیه‌های ۴ و ۳ و ۳ را انتخاب کردند لذا داریم:

$$\mathrm{MRR}_{\mathrm{Quran}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{3}+\frac{1}{4}\right] = 0.291$$

$$\mathrm{MRR}_{\mathrm{Nahj}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{5}+\frac{1}{3}\right] = 0.266$$

$$\mathrm{MRR}_{\mathrm{Sahifa}_{\mathrm{TF-IDF}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{9}+\frac{1}{3}\right] = 0.222$$

پس داریم:

$$MRR_{TF-IDF} = \frac{0.291 + 0.266 + 0.222}{3} = 26\%$$

روش دوم – استفاده از FASTTEXT

در روش دوم برای محاسبه شباهت بین جملات از مدل FastText استفاده کردیم. این مدل توسط شرکت فیسبوک توسعه داده شده و هم اکنون از پراستفاده‌ترین ترنسفورمرهای دنیای پردازش زبان طبیعی است. در این روش نیز ازهمان موارد پیش‌پردازش ذکر شده در بخش اول استفاده شده است. برای بهره‌گیری از قابلیت‌های این ابزار از کتابخانه Gensim استفاده کردیم که یک کتابخانه برای Topic Modeling می‌باشد. همچنین برای تولید وکتورها و تعبیه آنها از مدل پیش‌آموزش داده شده عربی فیسبوک استفاده کردیم.

Load pretrained arabic vector (downlaoded from fasttext data storage)

```
[ ]  1  model = KeyedVectors.load_word2vec_format(datapath(f'{vectorModelDir}cc.ar.300.vec'), binary=False)
```

بعد از لود کردن مدل نوبت به تولید وکتور برای کلمات موجود مجموعه داده‌ها می‌رسد(قرآن کریم، نهج البلاغه و صحیفه سجادیه). برای این کار توکن‌های هر آیه/جمله را به مدل داده تا برای آنها وکتور تولید کند. این کار با استفاده از تابع زیر صورت می‌گیرد:

```
"""
get vector for each token from pretrained arabic model
if each token doesnt exist in model we assign 0 to its vector
@param tokens an array of sentence word
@return an array of tuples with format (token, vector)
"""
def W2V4corpus(token,models):
  w2vToken = token
  k=0
  for i in range(len(token)):
    for j in range(len(token[i])):
      word = token[i][j]
      models = model
      if word in models:
        w2vToken[i][j] = (word,models[word])
      else:
        # list of words that dosent exist in the vector
        print(word)
        k=k+1
        w2vToken[i][j] = (word,0)
  print(f'number of words that dosent exist in the pretrained fasttext model: {k}')
  return w2vToken
```

بعضی از لغات در مجموعه داده ما بودند که در مدل فیسبوک برای آن تعبیه‌سازی‌ای وجود نداشت. البته بسیاری از آنها به دلیل نوشتار نادرست با این مشکل روبه رو شده بودند(مانند چسبیدن حرف «و» به کلمه بعدی) که باعث شد ما آنها را

نادیده بگیریم. متاسفانه مدل FastText مانند مدل BERT از توکنایزر پیشرفته و مختص به خود بهره نمی‌برد و استفاده از ابزارهای متفاوت توکنایزیشن نیز برای ما دردی را دوا نکرد به همین دلیل تصمیم به چشم پوشی از این کلمات گرفتیم.

بعد از تشکیل وکتور برای کلمات حاضر در آیات/جملات نوبت به تشکیل وکتور برای آیات/جملات می‌رسد. برای این کار می‌توانستیم از دو روش استفاده کنیم. اول از طریق APIهای خود FastText و دوم ساخت آن به صورت دستی. برای استفاده از FastText نیاز داشتیم تا یک مدل Doc2Vec برای زبان عربی در دسترس داشته باشیم که بعد از جست و جوهای فراوان پیدا نشد(تقریبا فقط برای زبان انگلیسی موجود بود). پس به سراغ راه دوم، یعنی ساخت وکتور به صورت دستی، رفتیم. برای این کار وکتور جمله را میانگین وکتور کلمات سازنده آن در نظر گرفتیم. در قطعه کد زیر توابع مورد نیاز برای این مورد آورده شده است:

```python
# calculate avg of word vectors in sentence
def calculateSentenceVector(wordVector):
  return np.mean( np.array(wordVector), axis=0 )

# convert doc to vector
def doc2vec(corpus):
  temp = []
  for i in range(len(corpus)):
    temp.append(calculateSentenceVector([l[1] for l in corpus[i]]))
  return temp
```

پس از این مرحله ما وکتور را برای تمام آیات و جملات موجود در مجموعه داده در دسترس داریم. پس نوبت به کوئری زدن و یافتن موارد مشابه است. برای کوئری ورودی نیز تمام مراحل پیش‌پردازش و وکتور کردن که در این بخش ذکر کردیم صادق است. بعد از تشکیل وکتور برای جمله ورودی، با استفاده از محاسبه کسینوس زاویه بین وکتور جمله ورودی(کوئری) و تک تک جملات موجود در مجموعه داده، موارد محاسبه شده به صورت نزولی مرتب شده و ۱۰تای اول به عنوان ۱۰ جمله با بالاترین میزان تشابه انتخاب می‌شوند.

```python
"""
calculate similarity between input query and all words in vecList using cosine similarity
@param query the input query vector
@param vecList a list of vectors(words)
@return a list of similarity rank
"""

def calculateSimilarity(qurey, vecList):
  temp=[]
  k=0
  for i in vecList[1]:
    print(k)
    k = k+1
    temp.append(cosine_similarity([qurey], [i]))
  return temp
```

تابع زیر وظیفه یافتن مشابه‌ترین‌ها را بر عهده دارد. این امکان وجود دارد که در ورودی تابع مقرر کنیم تا در کدام متون دنبال جملات مشابه بگردیم.

```
"""
return similarity matrix
@param quran if true search in quranic verse to find similar verse
@param nahj if true search in nahj ol balaghe verse to find similar verse
@param sahifa if true search in sahifa verse to find similar verse
"""
def mostSimilar(queryVector, quran=True, nahj=False, sahifa=False):
  similarVector=[]
  if quran:
    similarVector.append(calculateSimilarity(queryVector, verse_complete_dict_nrmlze))
  if nahj:
    similarVector.append(calculateSimilarity(queryVector, nahj_complete_dict_nrmlze))
  if sahifa:
    similarVector.append(calculateSimilarity(queryVector, nahj_complete_dict_nrmlze))
  return similarVector
```

```
# a list for similarity between input query and all sentence
similarity = mostSimilar(query, quran=True)
```

```
#sort similarity and select top 10
sortedSimilarity = sorted(similarity, key=lambda x: x[1])
sortedSimilarity[:10]
```

برای ارزیابی مدل طبق روش MRR ما یک آیه را به عنوان ورودی به مدل دادیم و سپس خروجی زیر را به عنوان ۱۰ آیه با بیشترین تشابه دریافت کردیم(بدیهی است که شبیه‌ترین عبارت به یک آیه، خود آیه است):

ورودی:

```
['واذ قلنا للملائكه اسجدوا لادم فسجدوا الا ابليس ابي واستكبر وكان من الكافرين']
```

خروجی:

سپس برای ارزیابی مدل(نابود شدیم پای این قسمت تا جملات راحت پیدا کنیم که راحت‌تر بشه مقایسه کرد!!!)، به مانند بخش TF-IDF این ۱۰ آیه را به صورت نامرتب به ۲ عضو دیگر دادیم تا مشابه‌ترین را مشخص کنند که هر دو نفر جایگاه ۱ را حدس زدند. پس داریم:

$$\mathrm{MRR}_{\mathrm{Quran}_{\mathrm{FastText}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}[1+1] = 1.0$$

برای داده‌های نهج البلاغه هم داریم:

ورودی: الا و ان الشیطان قد جمع حزبه و استجلب خیله و رجله و ان معی لبصیرتی ما لبست علی نفسی و لا لبس علی و ایم الله لافرطن لهم حوضا انا ماتحه لا یصدرون عنه و لا یعودون الیه

خروجی:

۱. الا و ان الشیطان قد جمع حزبه و استجلب خیله و رجله و ان معی لبصیرتی ما لبست علی نفسی و لا لبس علی و ایم الله لافرطن لهم حوضا انا ماتحه لا یصدرون عنه و لا یعودون الیه
۲. و ان علی من الله جنه حصینه فاذا جاء یومی انفرجت عنی و اسلمتنی فحینئذ لا یطیش السهم و لا یبرا الکلم

۳. الحمد لله الذی لا تواری عنه سماء سماء و لا ارض ارضا منها و قد قال قائل انک علی هذا الامر یا ابن ابی طالب لحریص فقلت بل انتم و الله لاحرص و ابعد و انا اخص و اقرب و انما طلبت حقا لی و انتم تحولون بینی و بینه و تضربون وجهی دونه فلما قرعته بالحجه فی الملا الحاضرین هب کانه بهت لا یدری ما یجیبنی به اللهم انی استعدیک علی قریش و من اعانهم فانهم قطعوا رحمی و صغروا عظیم منزلتی و اجمعوا علی منازعتی امرا هو لی ثم قالوا الا ان فی الحق ان تاخذه و فی الحق ان تترکه فخرجوا یجرون حرمه رسول الله-ص کما تجر الامه عند شرائها متوجهین بها الی البصره فحبسا نساءهما فی بیوتهما و ابرزا حبیس رسول الله-ص لهما و لغیرهما فی جیش ما منهم رجل الا و قد اعطانی الطاعه و سمح لی بالبیعه طائعا غیر مکره فقدموا علی عاملی بها و خزان بیت مال المسلمین و غیرهم من اهلها فقتلوا طائفه صبرا و طائفه غدرا فوالله لو لم یصیبوا من المسلمین الا رجلا واحدا معتمدین لقتله بلا جرم جره لحل لی قتل ذلک الجیش کله اذ حضروه فلم ینکروا و لم یدفعوا عنه بلسان و لا بید دع ما انهم قد قتلوا من المسلمین مثل العده التی دخلوا بها علیهم

۴. من عبد الله علی امیر المؤمنین الی القوم الذین غضبوا لله حین عصی فی ارضه و ذهب بحقه فضرب الجور سرادقه علی البر و الفاجر و المقیم و الظاعن فلا معروف یستراح الیه و لا منکر یتناهی عنه اما بعد فقد بعثت الیکم عبدا من عباد الله لا ینام ایام الخوف و لا ینکل عن الاعداء ساعات الروع اشد علی الفجار من حریق النار و هو مالک بن الحارث اخو مذحج فاسمعوا له و اطیعوا امره فیما طابق الحق فانه سیف من سیوف الله لا کلیل الظبه و لا نابی الضریبه فان امرکم ان تنفروا فانفروا و ان امرکم ان تقیموا فاقیموا فانه لا یقدم و لا یحجم و لا یؤخر و لا یقدم الا عن امری و قد اثرتکم به علی نفسی لنصیحته لکم و شده شکیمته علی عدوکم

۵. ایها الناس غیر المغفول عنهم و التارکون الماخوذ منهم ما لی اراکم عن الله ذاهبین و الی غیره راغبین کانکم نعم اراح بها سائم الی مرعی وبی و مشرب دوی و انما هی کالمعلوفه للمدی لا تعرف ما ذا یراد بها اذا احسن الیها تحسب یومها دهرها و شبعها امرها و الله لو شئت ان اخبر کل رجل منکم بمخرجه و مولجه و جمیع شانه لفعلت و لکن اخاف ان تکفروا فی برسول الله-ص الا و انی مفضیه الی الخاصه ممن یؤمن ذلک منه و الذی بعثه بالحق و اصطفاه علی الخلق ما انطق الا صادقا و قد عهد الی بذلک کله و بمهلک من یهلک و منجی من ینجو و مال هذا الامر و ما ابقی شیئا یمر علی راسی الا افرغه فی اذنی و افضی به الی ایها الناس انی و الله ما احثکم علی طاعه الا و اسبقکم الیها و لا انهاکم عن معصیه الا و اتناهی قبلکم عنها

۶. اتق الله الذی لا بد لک من لقائه و لا منتهی لک دونه و لا تقاتلن الا من قاتلک و سر البردین و غور بالناس و رفه فی السیر و لا تسر اول اللیل فان الله جعله سکنا و قدره مقاما لا ظعنا فارح فیه بدنک و روح ظهرک فاذا وقفت حین ینبطح السحر او حین ینفجر الفجر فسر علی برکه الله فاذا لقیت العدو فقف من اصحابک وسطا و لا تدن من

القوم دنو من یرید ان ینشب الحرب و لا تباعد عنهم تباعد من یهاب الباس حتی یاتیک امری و لا یحملنکم شنانهم علی قتالهم قبل دعائهم و الاعذار الیهم

۷. ما هی الا الکوفه اقبضها و ابسطها ان لم تکونی الا انت تهب اعاصیرک فقبحک الله و تمثل بقول الشاعر لعمر ابیک الخیر یا عمرو اننی || علی وضر من ذا الاناء قلیل ثم قال ع انبئت بسرا قد اطلع الیمن و انی و الله لاظن ان هؤلاء القوم سیدالون منکم باجتماعهم علی باطلهم و تفرقکم عن حقکم و بمعصیتکم امامکم فی الحق و طاعتهم امامهم فی الباطل و بادائهم الامانه الی صاحبهم و خیانتکم و بصلاحهم فی بلادهم و فسادکم فلو ائتمنت احدکم علی قعب لخشیت ان یذهب بعلاقته اللهم انی قد مللتهم و ملونی و سئمتهم و سئمونی فابدلنی بهم خیرا منهم و ابدلهم بی شرا منی اللهم مث قلوبهم کما یماث الملح فی الماء اما و الله لوددت ان لی بکم الف فارس من بنی فراس بن غنم هنالک لو دعوت اتاک منهم || فوارس مثل ارمیه الحمیم ثم نزل ع من المنبر

۸. ۱۱

۹. و قد امرت علیکما و علی من فی حیزکما مالک بن الحارث الاشتر فاسمعا له و اطیعا و اجعلاه درعا و مجنا فانه ممن لا یخاف وهنه و لا سقطته و لا بطؤه عما الاسراع الیه احزم و لا اسراعه الی ما البطء عنه امثل

۱۰. و اعلم ان البصره مهبط ابلیس و مغرس الفتن فحادث اهلها بالاحسان الیهم و احلل عقده الخوف عن قلوبهم و قد بلغنی تنمرک لبنی تمیم و غلظتک علیهم و ان بنی تمیم لم یغب لهم نجم الا طلع لهم اخر و انهم لم یسبقوا بوغم فی جاهلیه و لا اسلام و ان لهم بنا رحما ماسه و قرابه خاصه نحن ماجورون علی صلتها و مازورون علی قطیعتها فاربع ابا العباس رحمک الله فیما جری علی لسانک و یدک من خیر و شر فانا شریکان فی ذلک و کن عند صالح ظنی بک و لا یفیلن رایی فیک و السلام

در این جا نفر اول جایگاه ۳ و نفر دوم جایگاه ۱ را حدس زد پس باز هم داریم:

$$\mathrm{MRR}_{\mathrm{Nahj}_{\mathrm{FastText}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{3}+1\right] = 0.665$$

برای صحیفه سجادیه نیز همین عملیات صورت گرفت:

ورودی: حمدا تقر به عیوننا اذا برقت الابصار، و تبیض به وجوهنا اذا اسودت الابشار.

خروجی:

۱. حمدا تقر به عیوننا اذا برقت الابصار، و تبیض به وجوهنا اذا اسودت الابشار.

۲. حمدا نعمر به فیمن حمده من خلقه، و نسبق به من سبق الی رضاه و عفوه.

۳. یا من تحمد الی عباده بالاحسان و الفضل، و غمرهم بالمن و الطول، ما افشی فینا نعمتک، و اسبغ علینا منتک، و اخصنا ببرک!
۴. اللهم و قد اشرف علی خفایا الاعمال علمک، و انکشف کل مستور دون خبرک، و لا تنطوی عنک دقائق الامور، و لا تعزب عنک غیبات السرائر
۵. و هب لی التطهیر من دنس العصیان، و اذهب عنی درن الخطایا، و سربلنی بسربال عافیتک، و ردنی رداء معافاتک، و جللنی سوابغ نعمائک، و ظاهر لدی فضلک و طولک
۶. و انا ابرا الیک من ان استکبر، و اعوذ بک من ان اصر، و استغفرک لما قصرت فیه، و استعین بک علی ما عجزت عنه.
۷. و الذین عرفتهم مثاقیل المیاه، و کیل ما تحویه لواعج الامطار و عوالجها
۸. و اذا هممنا بهمین یرضیک احدهما عنا، و یسخطک الاخر علینا، فمل بنا الی ما یرضیک عنا، و اوهن قوتنا عما یسخطک علینا
۹. لا یخیب منک الاملون، و لا ییاس من عطائک المتعرضون، و لا یشقی بنقمتک المستغفرون.
۱۰. و اجزل لنا فیه من الحسنات، و اخلنا فیه من السیئات، و املا لنا ما بین طرفیه حمدا و شکرا و اجرا و ذخرا و فضلا و احسانا.

نفر اول جایگاه ۱ و نفر دوم نیز جایگاه ۲ را حدس زد که داریم:

$$\mathrm{MRR}_{\mathrm{Sahifa}_{\mathrm{FastText}}} = \frac{1}{2}\sum_{\mathrm{i}=1}^{2}\frac{1}{\mathrm{Rank}^{\mathrm{i}}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{2}+1\right] = 0.75$$

و در کل هم شاخص MRR را به صورت زیر برای مدل FastText داریم:

$$MRR_{TF-IDF} = \frac{0.75 + 0.665 + 1}{3} = 80\%$$

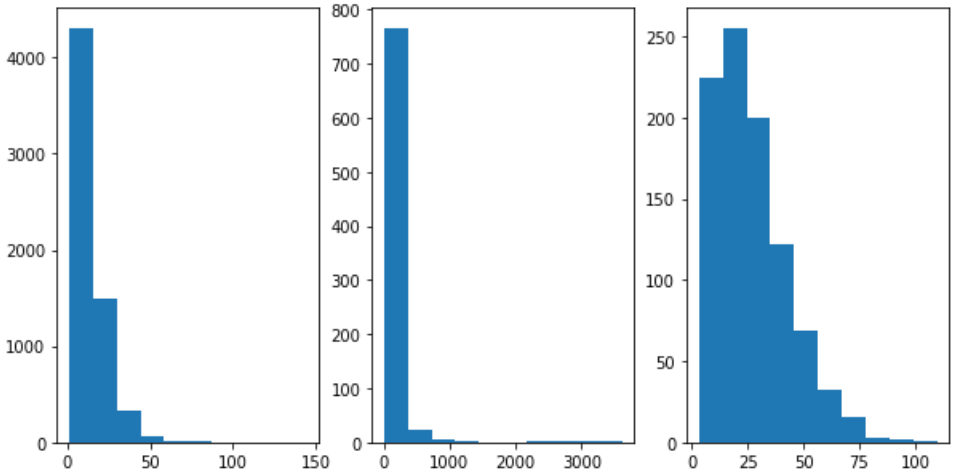
روش سوم – استفاده از شبکه‌های FINE-TUNED BERT

آموزش دادن مدل برت

برای لود کردن مدل برت از کتابخانه transformers و از یکی از مدل آموزش دیده روی زبان عربی (البته بیس ورژن آن) استفاده شده است.

```
from transformers import AutoTokenizer, AutoModel
import torch
device = 'cuda' if torch.cuda.is_available() else 'cpu'

# Mini:   asafaya/bert-mini-arabic
# Medium: asafaya/bert-medium-arabic
# Base:   asafaya/bert-base-arabic
# Large:  asafaya/bert-large-arabic
# https://www.youtube.com/watch?v=jVPd7lEvjtg

tokenizer = AutoTokenizer.from_pretrained("asafaya/bert-base-arabic")
model = AutoModel.from_pretrained("asafaya/bert-base-arabic")

model=model.to(device)

model.eval()

device
```

ابتدا یک پلات کلی برای هیستوگرام تعداد توکن‌ها در هر جمله از قران و صحیفه سجادیه و نهج البلاغه کشیده شد. به نظر میآید مناسب‌ترین عدد ۱۲۸ باشد.



بدین ترتیب با استفاده از توکنایزر برت، جملاتی که طول کمتر از ۱۲۸ دارند پد می‌شوند و آن‌ها که طول بیشتر از ۱۲۸ دارند بریده می‌شوند که در کد زیر نشان داده شده است:

```
def tokenize(sentence_list):
    return tokenizer.batch_encode_plus(sentence_list,
                    add_special_tokens=True,
                    max_length=128,
                    padding='max_length',
                    return_attention_mask=True,
                    truncation=True,
                    return_tensors='pt')
```

تابع بعدی که در اینجا مهم است، تابع get_mean_embedding است. این تابع جملات خام و مدل برت را به عنوان ورودی میگیرد و یک امبدینگ به ازای هر جمله خروجی میدهد. برای بدست آوردن خروجی مدل برت از last_hidden_state استفاده شده است سپس روی کلماتی که واقعا در جمله بوده اند میانگین گرفته شده است. این مسئله به شدت اهمیت دارد زیرا در غیراینصورت بردارهای میانگین تقریبا شبیه به هم خواهند شد و اطلاعات خاصی را نشان نخواهند داد.

```
import numpy as np
import torch
from torch.utils.data import DataLoader,TensorDataset

def get_mean_embedding(model,sentences,batch_size=128):
    encoding= tokenize(sentences)

    dataset=TensorDataset(encoding.input_ids, encoding.attention_mask)
    dataloader = DataLoader(dataset,batch_size=batch_size
                            ,drop_last=False,shuffle=False)

    predictions = []
    with torch.no_grad():
        for (input_ids, attention_mask) in dataloader:
            prediction = model(input_ids.to(device),
                               attention_mask.to(device))
            predictions.append(prediction.last_hidden_state)

    predictions = torch.cat(predictions,dim=0)

    attention_mask=encoding.attention_mask.unsqueeze(-1) \
                   .expand(predictions.shape).float()
    # for filtering out unnecessary ones
    mask_embeddings=predictions.cpu() * attention_mask

    summed = torch.sum(mask_embeddings,dim=1)
    count = torch.clamp(attention_mask.sum(dim=1),min=1e-9)
    mean_embedding = summed / count

    return mean_embedding
```

سپس بردار میانگین جملات مختلف با یکدیگر مقایسه شده و شبیه‌ترین موارد برگردانده می‌شود. در اینجا از کتابخانه sklearn استفاده شده است.

```python
from sklearn.metrics.pairwise import cosine_similarity

def get_most_similar(mean_embedding,compare_verse_no):
    mean_embedding = mean_embedding.numpy()
    similarity=cosine_similarity(
        [mean_embedding[compare_verse_no]],
        mean_embedding
    ).squeeze()

    most_similar=similarity.argsort()[-10:][::-1]
    return most_similar
```

تا اینجا می‌توان یک سری نتایج بدست آورد ولی برای این که مدل را به خوبی فاین تیون کنیم ابتدا از تسک masked language modelling استفاده کردیم. این تسک بدین صورت است که ۱۵ درصد کلمات در هر جمله‌mask می‌شوند سپس مدل باید روی این کلمات آموزش ببیند. در واقع بدین صورت دارد کلمات تخصصی و پرکاربرد هر متن را یاد می‌گیرد که میتواند دقت بالاتری بدهد.

```python
from transformers import BertTokenizer, BertForMaskedLM
import torch
from transformers import AdamW
device = 'cuda' if torch.cuda.is_available() else 'cpu'

# Mini:   asafaya/bert-mini-arabic
# Medium: asafaya/bert-medium-arabic
# Base:   asafaya/bert-base-arabic
# Large:  asafaya/bert-large-arabic

tokenizer = BertTokenizer.from_pretrained("asafaya/bert-base-arabic")
model = BertForMaskedLM.from_pretrained("asafaya/bert-base-arabic")

model=model.to(device)
model.train()

optim = AdamW(model.parameters(), lr=5e-5)
```

سپس مدل با یک حلقه معمولی‌ترین می‌شود. تفاوت مدل ما با مدل اصلی برت (که نتایج بهتری هم می‌گیرد) این است که در هر ۵ ایپاک داده‌های مسک شده مجددا ساخته می‌شوند تا مدل روی مقدار داده بیش‌برازش نشود و بتواند قسمت‌های مختلف متن را یاد بگیرد. در نهایت در بخش بعد ورودی مدل را مشخص کرده تا در ادامه از هر کتاب، ۱۵ آیه/جمله مشابه توسط مدل فاین‌تیون شده bert مشخص شود.

برای محاسبه MRR به جهت مقایسه این روش شباهت‌یابی و دو روش بعدی، ترتیب ۱۵ آیه/جمله مشابه را بهم ریخته و به دو نفر اعضای تیم که این خروجی‌هارا ندیده‌اند، داده می‌شود تا به صورت دستی مشابه‌ترین آیه/جمله مدنظرشان را

اعلام کنند. ورودی که برای روش سوم یعنی Fine-tuned BERT آزمایش شد، یکی از آیه‌های قرآن یعنی آیه ۲۴ سوره بقره: «فان لم تفعلوا ولن تفعلوا فاتقوا النار التی وقودها الناس والحجاره ۪ اعدت للکافرین» بود. توجه شود که چون یکی از آیه‌های قرآن را به عنوان ورودی مشخص کردیم، طبیعی است که برترین و مشابه‌ترین آیه قرآنی با این آیه، خودش باشد لذا برای قرآن، ما ۱۴ آیه مشابه بعدی را برای محاسبه دستی استفاده کردیم. ۱۴ آیه برتر در شباهت به ترتیب عبارتند از:

۱- یا ایها الذین امنوا قوا انفسکم واهلیکم نارا وقودها الناس والحجاره علیها ملائکه غلاظ شداد لا یعصون الله ما امرهم ویفعلون ما یؤمرون

۲- یا قوم ادخلوا الارض المقدسه التی کتب الله لکم ولا ترتدوا علی ادبارکم فتنقلبوا خاسرین

۳- لن تنالوا البر حتی تنفقوا مما تحبون ۜ وما تنفقوا من شیء فان الله به علیم

۴- وما کان صلاتهم عند البیت الا مکاء وتصدیه ۜ فذوقوا العذاب بما کنتم تکفرون

۵- ااشفقتم ان تقدموا بین یدی نجواکم صدقات ۜ فاذ لم تفعلوا وتاب الله علیکم فاقیموا الصلاه واتوا الزکاه واطیعوا الله ورسوله ۜ والله خبیر بما تعملون

۶- وجعلوا لله اندادا لیضلوا عن سبیله ۗ قل تمتعوا فان مصیرکم الی النار

۷- الذین کفروا لهم عذاب شدید ۪ والذین امنوا وعملوا الصالحات لهم مغفره واجر کبیر

۸- والذین کفروا بعضهم اولیاء بعض ۜ الا تفعلوه تکن فتنه فی الارض وفساد کبیر

...

۱۳- قل للذین کفروا ستغلبون وتحشرون الی جهنم ۜ وبئس المهاد

۱۴- واتقوا فتنه لا تصیبن الذین ظلموا منکم خاصه ۪ واعلموا ان الله شدید العقاب

نفر اول آیه شماره ۱ و نفر دوم آیه شماره ۲ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Quran}_{\mathrm{BERT}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[1+\frac{1}{2}\right] = 0.75$$

به همین ترتیب آیه/جمله‌های برتر در نهج‌البلاغه عبارت بودند از:

۱-لا تلقین طلحه فانک ان تلقه تجده کالثور عاقصا قرنه یرکب الصعب و یقول هو الذلول و لکن الق الزبیر فانه الین عریکه فقل له یقول لک ابن خالک عرفتنی بالحجاز و انکرتنی بالعراق فما عدا مما بدا

۲- ما کنت تصنع بسعه هذه الدار فی الدنیا و انت الیها فی الاخره کنت احوج و بلی ان شئت بلغت بها الاخره تقری فیها الضیف و تصل فیها الرحم و تطلع منها الحقوق مطالعها فاذا انت قد بلغت بها الاخره فقال له العلاء یا امیر المؤمنین اشکو الیک اخی عاصم بن زیاد قال و ما له قال لبس العباءه و تخلی عن الدنیا قال علی به فلما جاء قال یا عدی نفسه لقد استهام بک الخبیث ا ما رحمت اهلک و ولدک ا تری الله احل لک الطیبات و هو یکره ان تاخذها انت اهون علی الله من ذلک قال یا امیر المؤمنین هذا انت فی خشونه ملبسک و جشوبه ماکلک قال ویحک انی لست کانت ان الله تعالی فرض علی ائمه العدل ان یقدروا انفسهم بضعفه الناس کیلا یتبیغ بالفقیر فقر

۳- و ساله رجل ان یعرفه الایمان فقال ع اذا کان الغد فاتنی حتی اخبرک علی اسماع الناس فان نسیت مقالتی حفظها علیک غیرک فان الکلام کالشارده ینقفها هذا و یخطئها هذا

...

۱۴- ایها الناس من عرف من اخیه وثیقه دین و سداد طریق فلا یسمعن فیه اقاویل الرجال اما انه قد یرمی الرامی و تخطئ السهام و یحیل الکلام و باطل ذلک یبور و الله سمیع و شهید اما انه لیس بین الحق و الباطل الا اربع اصابع الباطل ان تقول سمعت و الحق ان تقول رایت

۱۵- اما بعد فان صلاح ابیک غرنی منک و ظننت انک تتبع هدیه و تسلک سبیله فاذا انت فیما رقی الی عنک لا تدع لهواک انقیادا و لا تبقی لاخرتک عتادا تعمر دنیاک بخراب اخرتک و تصل عشیرتک بقطیعه دینک و لئن کان ما بلغنی عنک حقا لجمل اهلک و شسع نعلک خیر منک و من کان بصفتک فلیس باهل ان یسد به ثغر او ینفذ به امر او یعلی له قدر او یشرک فی امانه او یؤمن علی جبایه فاقبل الی حین یصل الیک کتابی هذا ان شاء الله

نفر اول آیه شماره ۳ و نفر دوم آیه شماره ۱ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Nahj}_{\mathrm{BERT}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[1+\frac{1}{3}\right] = 0.665$$

به همین ترتیب آیه/جمله‌های برتر در صحیفه سجادیه عبارت بودند از:

۱- فختم بنا علی جمیع من ذرا، و جعلنا شهداء علی من جحد، و کثرنا بمنه علی من قل.

۲- هذا مقام من اعترف بسبوغ النعم، و قابلها بالتقصیر، و شهد علی نفسه بالتضییع.

۳- اللهم انک انزلته علی نبیک محمد- صلی الله علیه و اله- مجملا، و الهمته علم عجائبه مکملا، و ورثتنا علمه مفسرا، و فضلتنا علی من جهل علمه، و قویتنا علیه لترفعنا فوق من لم یطق حمله.

...

۱۴- الحمد لله الذی هدانا لحمده، و جعلنا من اهله لنکون لاحسانه من الشاکرین، و لیجزینا علی ذلک جزاء المحسنین.

۱۵- السلام علیک ما کان اطولک علی المجرمین، و اهیبک فی صدور المؤمنین!

نفر اول آیه شماره ۲ و نفر دوم آیه شماره ۲ را به عنوان مشابه‌ترین آیه برگزیدند. بنابراین داریم:

$$\mathrm{MRR}_{\mathrm{Sahifa}_{\mathrm{BERT}}} = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{\mathrm{Rank}^{i}_{\mathrm{choice}}} = \frac{1}{2}\left[\frac{1}{2}+\frac{1}{2}\right] = 0.5$$

بنابراین اگر روش Fine-tuned Bert را بخواهیم با یک MRR نشان دهیم، با میانگین گیری از این سه عدد بدست آمده داریم:

$$MRR_{BERT} = \frac{0.75 \ + \ 0.665 \ + \ 0.5}{3} = 63.8\%$$

البته مدل ما در سایت Huggingface نیز آپلود شده است و بصورت انلاین از طریق این لینک قابل استفاده است پیشنهاد می‌کنیم حتما به آن سری بزنید.

یافتن عبارات مشابه در متون دینی با استفاده از جاگذاری کلمات

تکلیف سری سوم