گزارش پروژه پنجم
(مبانی هوش محاسباتی)

مهدی طاهری ۴۰۰۱۲۶۲۱۳۷

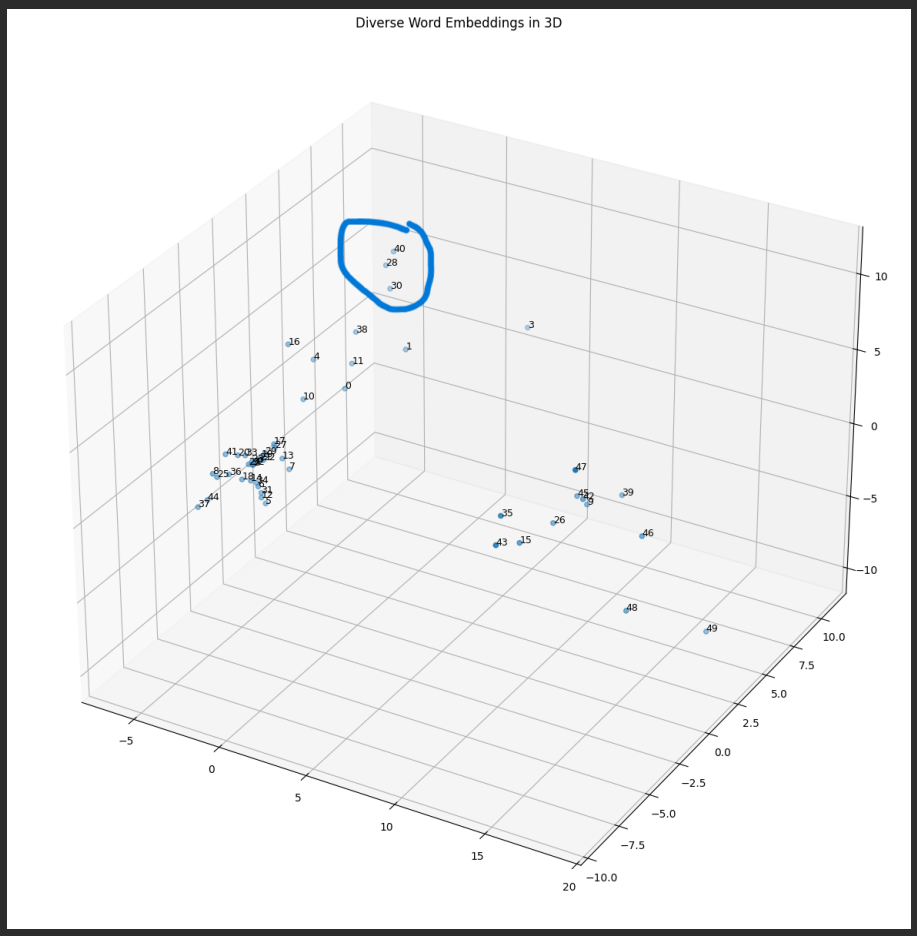
ایمان موقرمقدم ۴۰۰۱۲۶۲۱۱۵

آذر ۱۴۰۳

۱) فاز اول

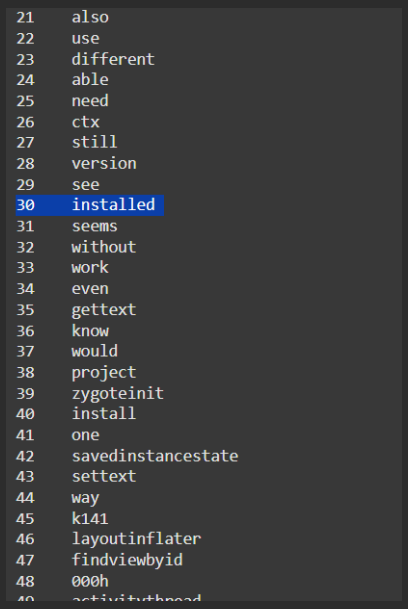
در این فاز stopWord ها و همچنین نقطه گذاری ها از دیتاست حذف شد و همچنین ستون title با ستون body ادغام شد .
همچنین lemmatize برای جایگذاری کلمات با ریشه آنها استفاده شد.

۲) فاز دوم

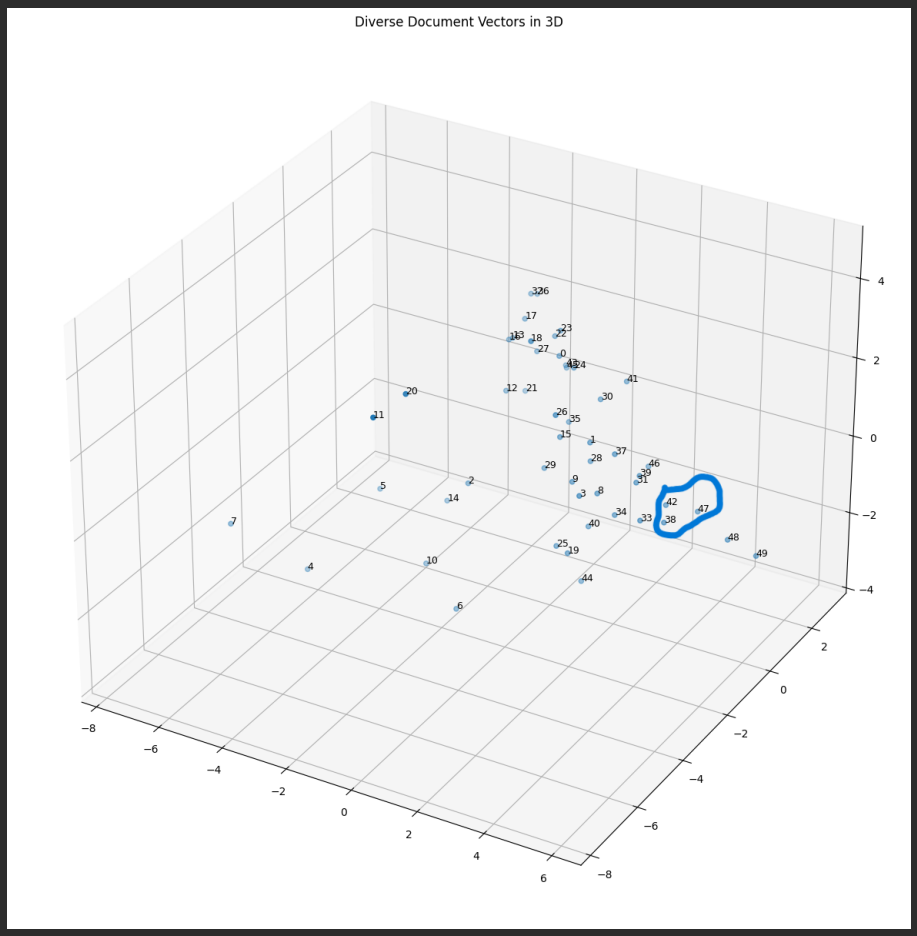


شکل ۱: هر شماره نشان دهنده یک کلمه است

همانطور که مشاهده میشود شماره های ۴۰، ۲۸، ۳۰ درباره مفهوم مشترکی هستند.



شکل ۲: کلمات مرتبط با هر لیبل



شکل ۳: هر شماره نشان دهنده یک title است

همانطور که مشاهده میشود شماره های ۴۲و ۳۸و ۴۷ درباره آرایه صحبت میکنند.

```
How to substract some values in different line using awk?
What is the output of this C program?
What is the Compatible version of Spring-security with spring-4.3.0-release?
How to sort matrix diagonally
how to understand initialization of 3d arrays?
count the ocurrence of a number and put the counters in a list (scheme)
Moving values in a column to a specific row.
Why variable y is 0, is not 2?
SVG Path Data Regex C#
Clean data in scala
Lists in Python - Each array in different row
Generate a specific dictionary type from list in python
ordering array of numbers returns invalid result
Vector with 1 2 3 4 5 2 3 4 5 6 3 4 5 6 7 4 5 6 7 8 5 6 7 8 9 with the commands rep() and seq()?
How to remove repetition in output results in this code.
how to split cvs file
Is spark 1.6.3 will support Kafka 1.0.0
How do you write (9.5*(4.5)-2.5*3)/(45.5-3.5) in Java?
Whats happening here?
Adding '' to every character of a tuple sublist
Vectorisation of for loop with multiple conditions
How can I sequence a character vector where "0.1" is distinguished from "0.10"?
how to compare arrays above and below python
Changing one inner list's element changes all inner lists python
No matching distribution found for django
Split the array values into 3 columns in php?
Can't install Django 2.0 by pip
Python- generate a symmetric list of list with half of nxn list of list data
Uniqe ID in list of indexed 2D/3D Image masks (or 2D/3D matrix) in R
how to define array in python
Finding connected components in a pixel-array
```

شکل ۴: title مرتبط با هر لیبل

۳) فاز سوم

با دادن کوئری ماشین لرنینگ title های مرتبط با آن پیدا میشوند:

```
Similarity: 0.84, Title: Hyperparameter Tuning of Tensorflow Model
Similarity: 0.80, Title: Questions about hyperparameter tuning in Keras/Tensorflow
Similarity: 0.78, Title: Machine learning query
Similarity: 0.78, Title: Which deep learning library support the compression of the deep learning models to be used on the phones?
Similarity: 0.78, Title: What are some machine learning algorithms
```

شکل ۵

لیبل پردیکشن knn بر اساس کلمات مشترکی که حداقل در ۲ تا همسایه مشاهده میشود ، ساخته شده است. به عنوان مثال اگر تگ یک داده تست بصورت زیر باشد :

<python><list><array>

و اگر ۴ همسایه نزدیک آن بصورت زیر باشند:

neighbor۱:<python><c#><java>
neighbor۲:<ML><pandas><python>
neighbor۳:<python><c#><c++>
neighbor۴:<apple><microsoft><samsung>

لیبل پردیکشن بصورت زیر ساخته میشود :

<python><c#>

در پایان دقت مدل به ۸۸ رسید.

	prediction	original
0	<android><google-apps-script><google-chrome><g...	<youtube><schema><google-search><structured-da...
1	<haskell><pointers><function><c><c++><computer...	<types><f#>
2	<python><swift><struct><javascript><android><k...	<angularjs>
3	<perl><ffmpeg><python-3.x><file-io><python><cs...	<matlab>
4	<swift><xcode-ui-testing><ios><swiftui><dart><...	<uinavigationbar><ios13><uinavigationbarappear...

شکل ۶: Unsuccessful Predictions

با توجه به اینکه هر داکیومنت صرفا با یک میانگین گیری از بردارهای کلمات نشان داده میشود ، احتمال زیادی وجود دارد که تگ ها به اشتباه پیش بینی شوند و به عنوان مثال اگر بردارهای کلمات را یک بعدی در نظر بگیریم ، میانگین ۱۰۰ و ۱۰۵ و ۹۵ مساوی با ۱۰۰ میشود و همینطور میانگین -۲۰۰ و ۱۰۰ و ۴۰۰ نیز ۱۰۰ میشود که چنین مواردی اجتناب ناپذیر است.

همچنین باید این را در نظر گرفت که مدل w۲vec به تنهایی برای پردازش زبان طبیعی کافی نیست اما ترکیب آن با سایر روش ها مثل شبکه های عصبی میتواند نتیجه مطلوبی بدهد.