به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش زبان طبیعی

استاد ممتازی

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: آشنایی با روش‌های بازنمایی کلمات

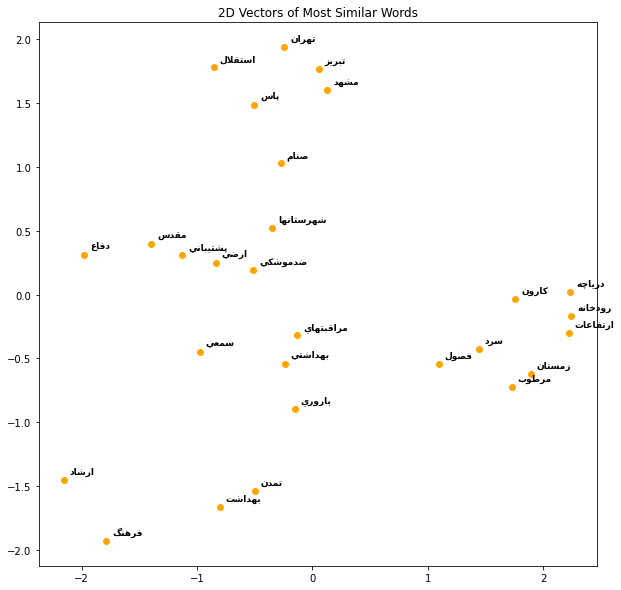
گام سوم: یافتن اسناد مشابه

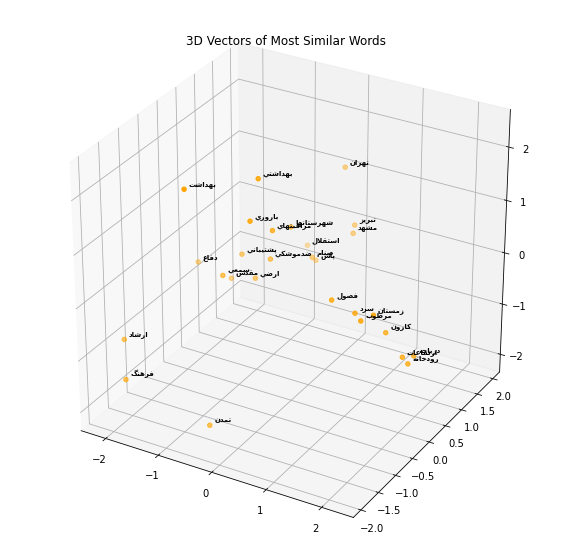
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | مدل Word2Vec و TF-IDF | | مدل Doc2Vec | |
| سند | شبیه‌ترین سند | امتیاز کسینوسی | شبیه‌ترین سند | امتیاز کسینوسی |
| Doc1 | Doc13 | 98.76% | Doc13 | 79.60% |
| Doc3 | Doc19 | 99.61% | Doc20 | 83.68% |
| Doc5 | Doc26 | 98.96% | Doc26 | 81.29% |
| Doc25 | Doc679 | 100% | Doc679 | 95.02% |
| Doc36 | Doc7 | 98.73% | Doc550 | 84.04% |

گام چهارم: برسی کلمات مشابه

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | کلمه اول | امتیاز کلمه اول | کلمه دوم | امتیاز کلمه دوم | کلمه سوم | امتیاز کلمه سوم |
| تهران | تبریز | 61.25٪ | شهرستانها | 60.12٪ | مشهد | 59.95٪ |
| بهداشت | باروری | 78.16٪ | بهداشتی | 74.13٪ | مراقبتهای | 73.62٪ |
| دفاع | مقدس | 69.93٪ | ضدموشکی | 66.24٪ | پشتیبانی | 64.46٪ |
| رودخانه | دریاچه | 83.88٪ | کارون | 82.09٪ | ارتفاعات | 80.81٪ |
| سرد | زمستان | 78.60٪ | مرطوب | 77.73٪ | فصول | 77.33٪ |
| فرهنگ | ارشاد | 77.19٪ | تمدن | 69.86٪ | سمعی | 67.37٪ |
| استقلال | پاس | 67.98٪ | ارضی | 67.83٪ | صنام | 66.60٪ |

در دو صفحه بعد نمودار دو بعدی و سه بعدی درخواست‌شده ارائه شده است:





* قبل از بیان هر نکته باید گفت که موارد مطرح‌شده به طور کلی برقرار است و ممکن است برای برخی از داده‌ها فاصله در نمودار به طرز غیرمنتظره‌ای زیاد شود؛ این امر به دلیل آن است که فضا کاهش بعد شدیدی را تجربه کرده است و طبیعتا اگر تمام ابعاد قابل حفظ و نمایش بود، موارد استثنا دیده نمی‌شد. باتوجه به کاهش ابعاد حتی برخی از نقاط در نمودار دوبعدی با نمودار سه بعدی فاصله کاملا متفاوتی دارند؛ مانند «بهداشت» و «تمدن» که در نمودار دوبعدی در کنار هم و در نمودار سه بعدی با فاصله زیاد قرار گرفته‌اند.
* به طور کلی می‌توان دید که هر کلمه با سه کلمه نزدیک به خود در فضای هندسی نزدیک افتاده است. مثلا «رودخانه»، «دریاچه»، «کارون» و «ارتفاعات» به خوبی در کنار هم قرار گرفته‌اند. این کنار هم بودن تا حد زیادی تابع امتیازهای موجود هم هست. مثلا برای «رودخانه» و کلمات مشابه آن امتیازها همگی بالای 80٪ است و نزدیکی بیشتر در مورد آن‌ها دیده می‌شود.
* به طور کلی می‌توان دید که از سه کلمه نزدیک به هر کلمه آن کلمه که در رتبه بهتری قرار داشته است، در اینجا هم به آن نزدیک‌تر است. مثلا برای «بهداشت» به ترتیب کلمات «باروری»، «بهداشتی» و «مراقبتها» به آن نزدیک است. به عنوان مثالی دیگر برای کلمه «فرهنگ»، کلمه «ارشاد» با اختلاف مشابه‌ترین کلمه است و در نمودار هم از دو کلمه مشابه دیگر نزدیک‌تر به «فرهنگ» قرار دارد.
* در نمایش‌های مختلف برخی از کلمات به دسته‌ای دیگر از کلمات نزدیک شده است. مثلا در نمودار دو بعدی «استقلال» به «تهران» نزدیک است. این مورد می‌تواند به دلیل شهر تیم فوتبال استقلال باشد.
* در دو نمودار یک شکاف و فاصله میان دسته کلمات «رودخانه» و «سرد» و کلمات مشابهشان با بقیه کلمات دیده می‌شود. این نشان می‌دهد که احتمالا در فضای اصلی هم این دسته از کلمات فاصله زیادی با بقیه دارند که به نظر می‌رسد از نظر مفهومی کاملا طبیعی باشد.به طور مشابه مفاهیم انتزاعی «فرهنگ» و «ارشاد» و «تمدن» به دور از بقیه کلمات افتاده‌اند.

بخش دوم: تشخیص اجزای سخن (POS)

پاسخ