تمرین اول

پردازش زبان‌های طبیعی

هانیه صحرانورد - ۹۸۴۴۳۱۳۷ - December 7, 2020

تعریف مسئله:

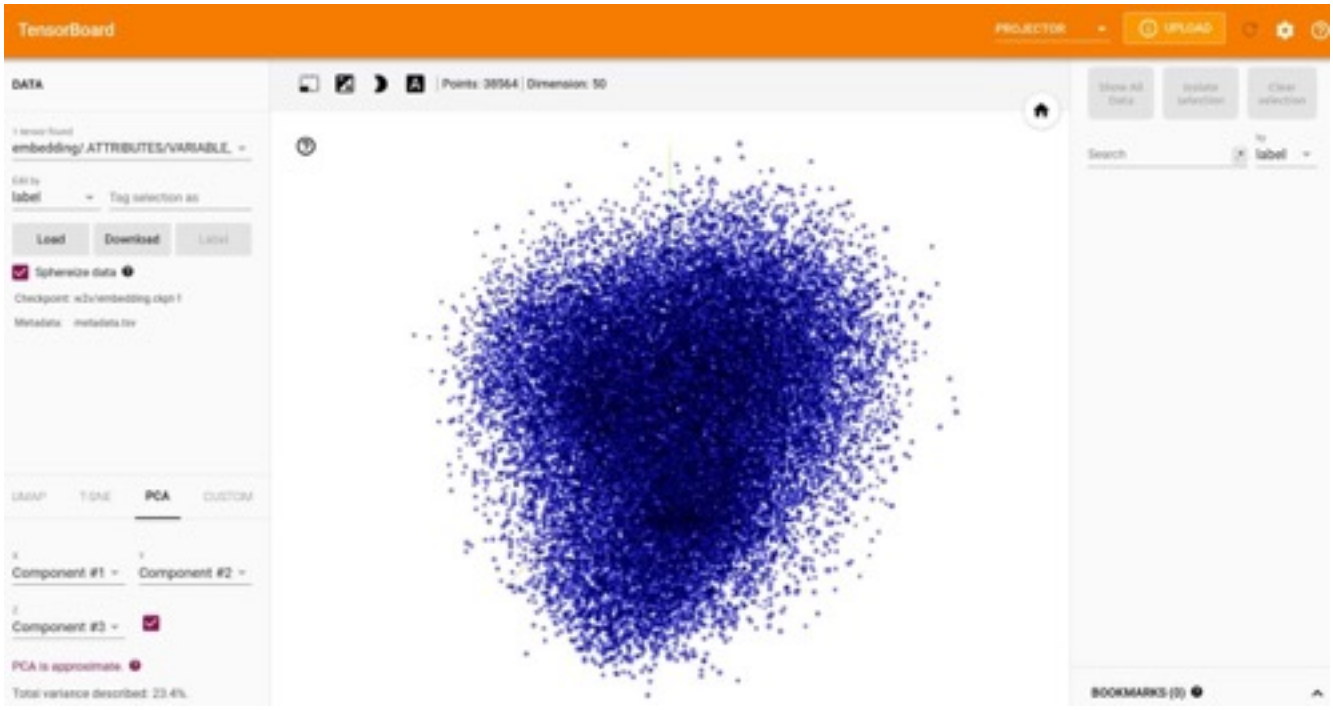
در این تمرین می‌خواهیم با استفاده از ابزارهای موجود برای پردازش زبان طبیعی، مجموعه دادگانی به زبان فارسی را جداسازی کرده و نهایتا مدل‌های آموزش دیده را مورد ارزیابی قرار دهیم. مدل‌های مورد استفاده به شرح زیر می‌باشند:

word2vec CBOW(D = 50, min_count = 20, Window Size = 5)
word2vec CBOW (D = 50, min_count = 20, Window Size = 10)
word2vec CBOW (D = 100, min_count = 20, Window Size = 5)
word2vec CBOW (D = 100, min_count = 20, Window Size = 10)
FastText (D = 50, min_count = 20, sg =1)
FastText (D = 50, min_count = 20, sg= 0)
FastText (D = 100, min_count = 20, sg =1)
FastText (D = 100, min_count = 20, sg =0)

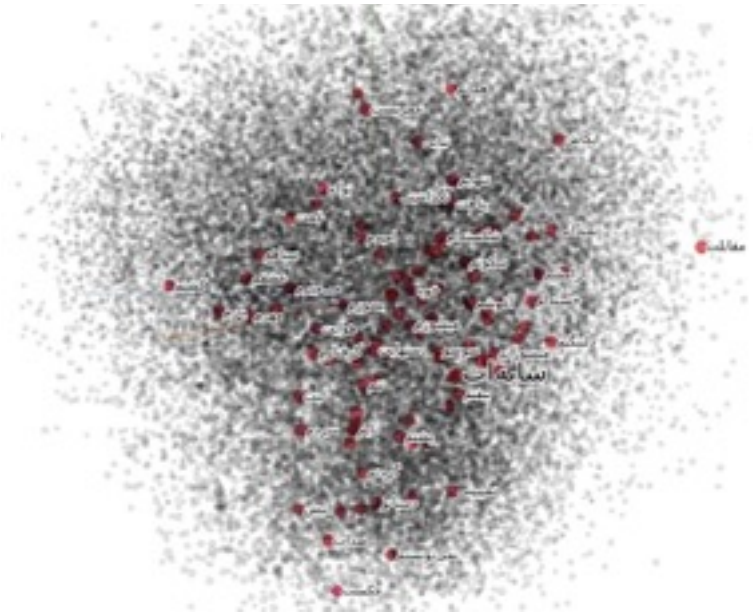
تمامی مدل ها در ۱۲ تکرار(iteration) آموزش داده شده اند.

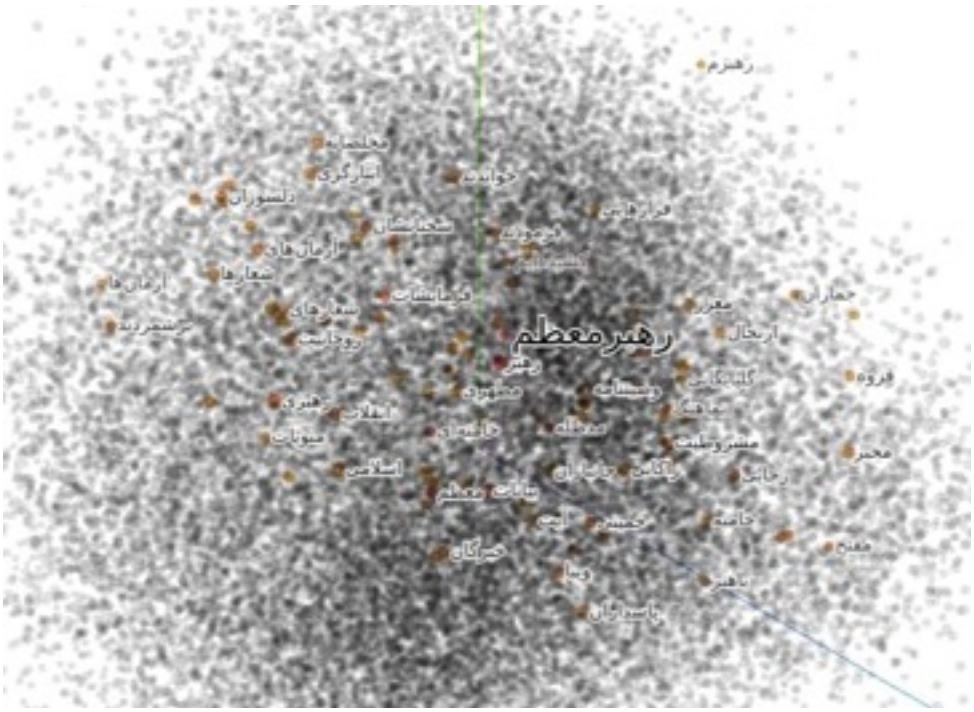
تصویرسازی یکی از مدل‌های word to vector:

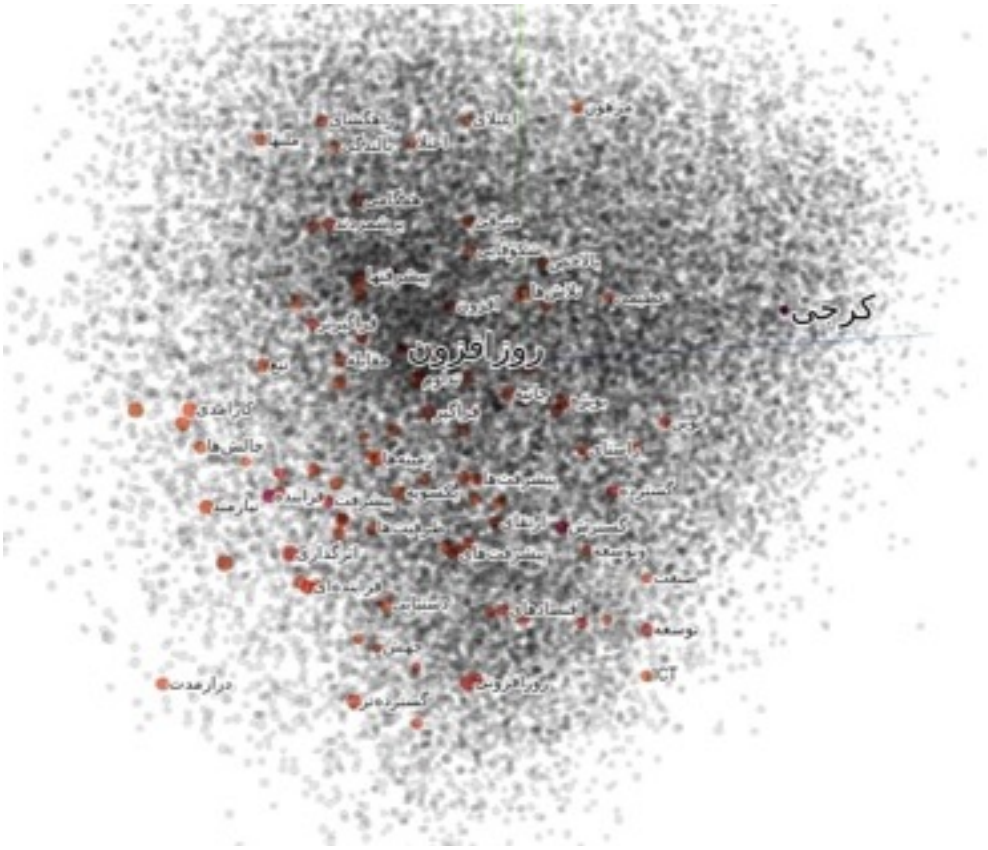
برای تصویرسازی از مدل‌ها، لازم است که دامنه هر بردار کلمه به ۲ یا ۳ بعد کاهش یابد تا بتوان تصویر آن را مشاهده نمود، به این ترتیب از PCA یا همان projector استفاده می‌کنیم. همانطور که در کد تمرین قابل مشاهده است، با استفاده از ابزار tensor board تنها لازم از است که لاگ مدل آموزش دیده شده را به عنوان ورودی tensor board داده و تصویر سازی مورد نظر خود را مشاهده کنیم:

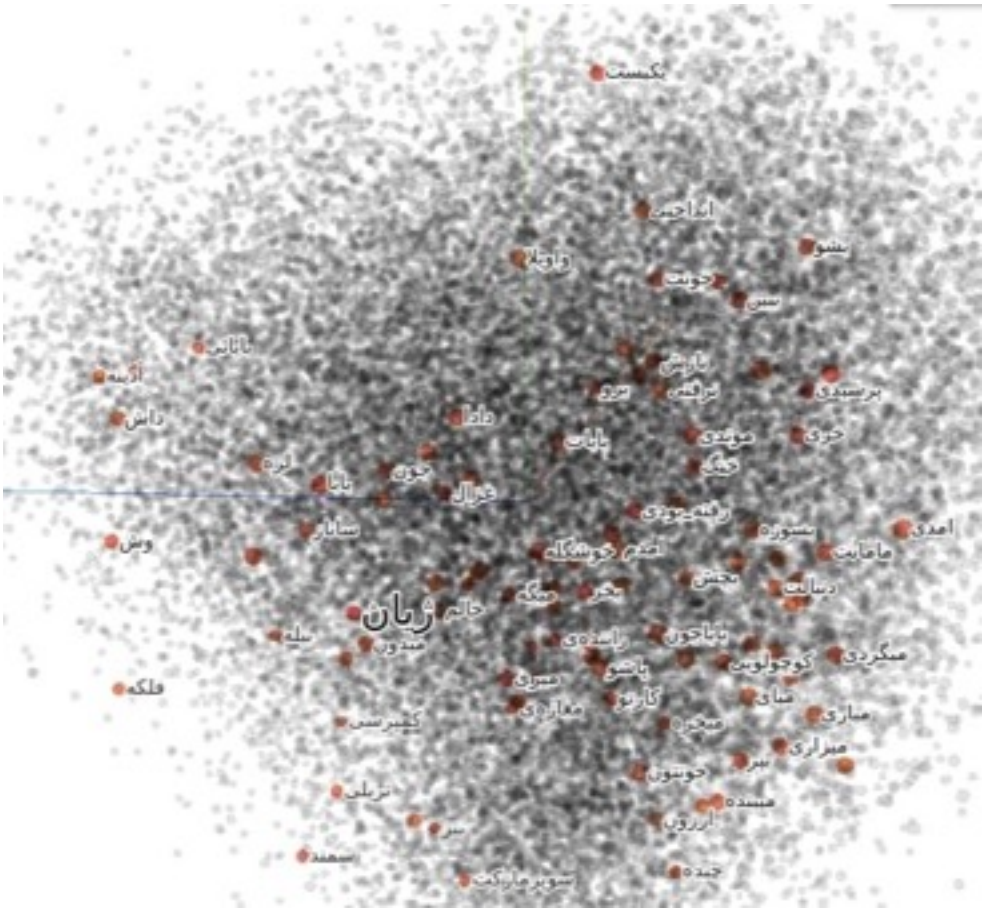


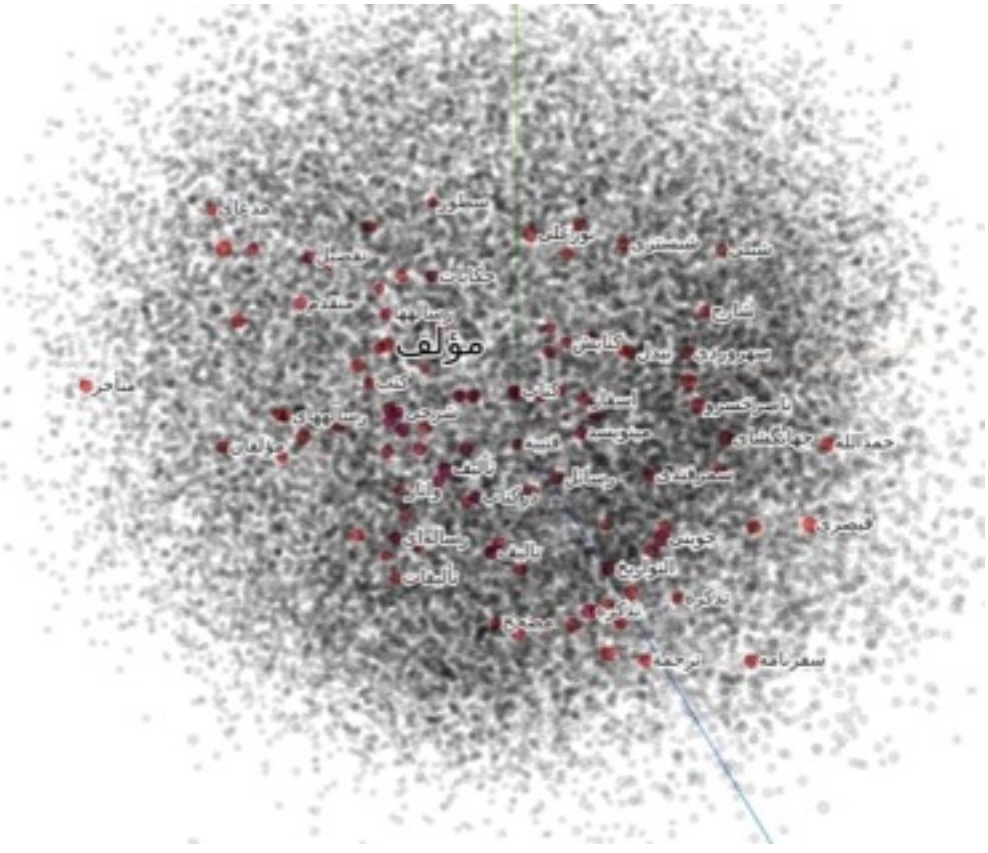
برای مثال تصاویر زیر مربوط به بردارهای کلمات متفاوت است که توسط این ابزار و این روش، به دست آمده اند:




رهبر معظم


روزافزون
کرجی


زبان


مؤلف

دقت مدل‌ها بر اساس ۴ معیار:

ارزیابی ۸ مدل آموزش دیده شده توسط پیکره آزمون با معیار بیشترین مشابهت بین ۴ رده صورت گرفته است. کلماتی با معانی مشابه در فضای نگاشت شده، به یکدیگر نزدیک هستند (همانند کلماتی که در قسمت تصویرسازی به عنوان نمونه گذاشته شده است). باید با داشتن ۲ کلمه‌ی ورودی و یک کلمه‌ی خروجی، کلمه‌ی خروجی مناسبی را پیدا کرد که رابطه‌ی بین هر ۴ کلمه از جهت فاصله‌ی بین بردارها درست باشد. مدل مجموعه‌ای از کلمات را به ترتیب فاصله آنها باز می‌گرداند. برای ارزیابی مشابهت بین ۵ کلمه، ۵ کلمه‌ی اولی که مدل به عنوان خروجی احتمالی می‌دهد را بررسی می‌کنیم، در صورتی که کلمه مورد نظر حاضر بود، به عنوان نتیجه درست در نظر می‌گیریم.

نتایج و تحلیل:

Models	Top-1	Top-5	Top-10	Top-20
fasttext-100-0	0.0810	0.2972	0.3243	0.36486
fasttext-100-1	0.08108	0.3918	0.4189	0.4459
fasttext-50-0	0.0810	0.3243	0.3648	0.4054
fasttext-50-1	0.1081	0.2432	0.2837	0.3513
w2v-100-10	0.0405	0.2432	0.3378	0.3648
w2v-100-5	0.0675	0.2702	0.3513	0.4054
w2v-50-10	0.0540	0.1756	0.2567	0.2837
w2v-50-5	0.0540	0.2162	0.2702	0.3513

همانطور که قابل مشاهده است، با افزایش بازه‌ی ارزیابی، مدل‌ها از دقت بالاتری برخوردار چرا که احتمال انتخاب کلمه‌ی خروجی درست، افزایش می‌یابد و این اتفاق در تمامی مدل‌ها می‌افتد. بنابراین مدل ایده‌آل مدلی خواهد بود که در ارزیابی TOP-1 بتواند پاسخ درستی به کلمات ورودی بدهد.

بعد: با دقت بر روی نتایج حاصل شده از مدل‌های آموزش دیده شده و مقایسه آنها در هر نوع w2v و fasttext در می‌یابیم که به صورت میانگین، مدل‌هایی که دارای ابعاد ۱۰۰ بوده اند از دقت بالاتری برخوردار اند که این نتیجه مورد انتظار می باشد.

نوع: روش جداسازی fasttext برتری قابل توجهی به w2v دارد که در جدول بالا این برتری به صورت نسبی قابل مشاهده است. علت این رخداد تعداد iteration پایین برای آموزش این مدل‌هاست که بیشتر شدن آن در حدود زمانی و سخت افزاری این تمرین گنجیده نمی‌شد.

مقدار SG: استفاده از روش skipgram بر روش w2v برتری نسبی دارد.

اندازه پنجره: با مقایسه مدل‌های w2v و اندازه پنجره در آنها در می‌یابیم که سایز پنجره ۵ برای این پیکره و مدل‌های آموزش دیده، مناسبتر عمل کرده است. این درحالی است که در بسیاری از نتایج بر روی پیکره‌های دیگر، اندازه پنجره بزرگتر برتری نسبی یا مرزی دارد.

انتظار می‌رود که با افزایش تعداد iteration ها و آزمایش اعداد بیشتری برای پارامترهای ذکر شده‌ی بالا، دقت مدلهای fast text به مراتب بالاتر از word to vector باشد.

با تشکر