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

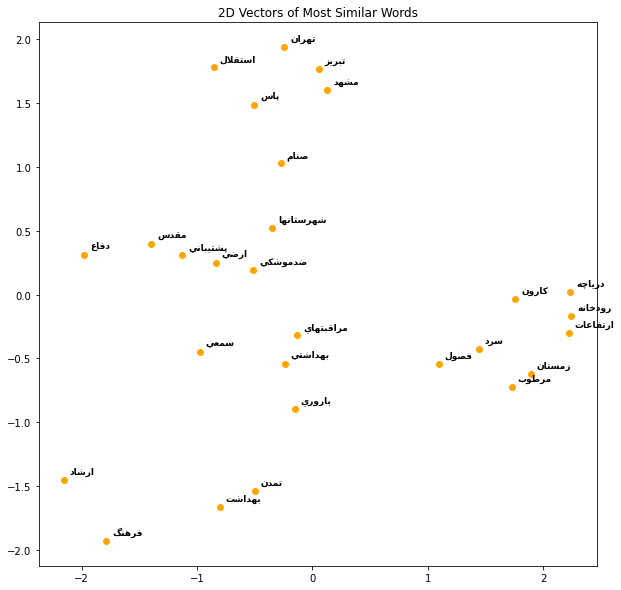
گام سوم: یافتن اسناد مشابه

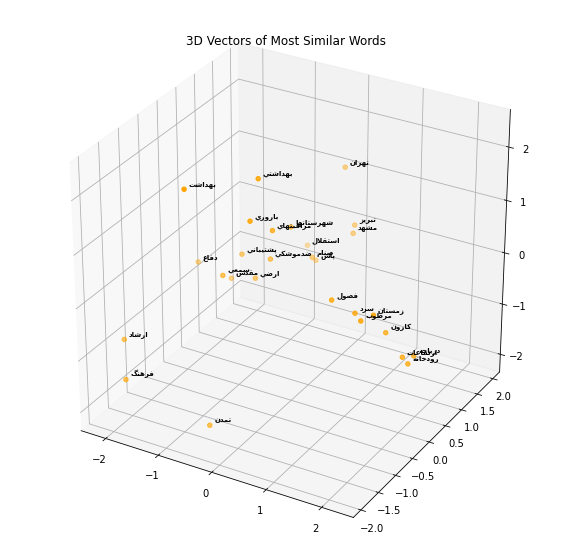
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | مدل Word2Vec و TF-IDF | | مدل Doc2Vec | |
| سند | شبیه‌ترین سند | امتیاز کسینوسی | شبیه‌ترین سند | امتیاز کسینوسی |
| Doc1 | Doc13 | 98.76% | Doc13 | 79.60% |
| Doc3 | Doc19 | 99.61% | Doc20 | 83.68% |
| Doc5 | Doc26 | 98.96% | Doc26 | 81.29% |
| 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:





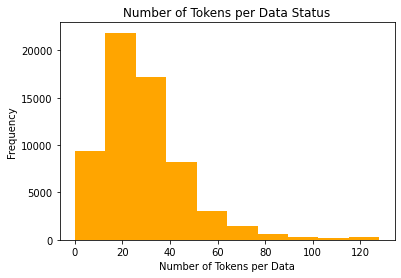
* قبل از بیان هر نکته باید گفت که موارد مطرح‌شده به طور کلی برقرار است و ممکن است برای برخی از داده‌ها فاصله در نمودار به طرز غیرمنتظره‌ای زیاد شود؛ این امر به دلیل آن است که فضا کاهش بعد شدیدی را تجربه کرده است و طبیعتا اگر تمام ابعاد قابل حفظ و نمایش بود، موارد استثنا دیده نمی‌شد. باتوجه به کاهش ابعاد 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)

گام اول: ایجاد شبکه‌ی عصبی حافظه کوتاه-مدت بلند دوطرفه

ورودی شبکه BiLSTM‌ باید دنباله‌هایی با اندازه‌های یکسان باشد؛ لذا باید طول کوتاه‌ترین دنبال‌ها را با حاشیه‌گذاری (Padding) به طول بلندترین آن برسانیم. بررسی‌های من نشان می‌دهد در سه مجموعه‌داده، داده‌هایی با 734 توکن وجود دارند. اما به طور کلی جملات با تعداد توکن بالا بسیار کم هستند. به بیان دقیق‌تر در هر سه مجموعه میانگین تعداد توکن‌ها به ازای هر جمله کمتر از 30 توکن است! در نمودار زیر که فراوانی داده‌ها با اندازه‌های مختلف توکن را نشان می‌دهد به خوبی مشخص است که بیشتر داده‌ها تعداد نسبتا کمی توکن دارند.

طبیعی است که در این شرایط حاشیه‌گذاری باعث پیچیده‌شدن شدید مدل و مشکلات حافظه و زمان اجرا می‌شود؛ بنابراین من تصمیم گرفتم که داده‌های بلند را به تعداد کوچک‌تری داده بشکنم. در نمودار زیر مجددا نمودار فراوانی داده‌ها با اندازه‌های مختلف ترسیم شده است با این تفاوت که این بار تنها در محدوده 128 توکن است:



با نگاه به این نمودار من ۶۴ توکن را به عنوان حد بالای تعداد توکن یک داده در نظر گرفتم. بررسی‌های من نشان می‌دهد تنها ۴.۳٪ از داده‌ها بیشتر از ۶۴ توکن دارند بنابراین دقت کلی کاهش چندانی نخواهد داشت ولی آموزش بسیار آسان‌تر خواهد شد.

نهایتا توجه کنید که برای محاسبه Accuracy خروجی مدل به ازای توکن‌ها حاشیه‌گذاری حذف شده است و تنها دقت به ازای توکن‌ها اصلی محاسبه شده است.

شبکه‌ای که من برای آموزش استفاده کردم متشکل از یک BiLSTM‌ با 64 واحد و یک لایه Dense برای تولید تگ خروجی است. از تابع فعال‌سازی Softmax برای دریافت نتایج بهتر در لایه Dense خروجی کمک گرفته شده است.

برای آموزش هم از پانزده گام استفاده شده است و برای جلوگیری از بیش‌برازش احتمالی از یک EarlyStopping Callback بهره گرفتم.

گام دوم: ارزیابی شبکه‌ی عصبی حافظه کوتاه-مدت بلند دوطرفه

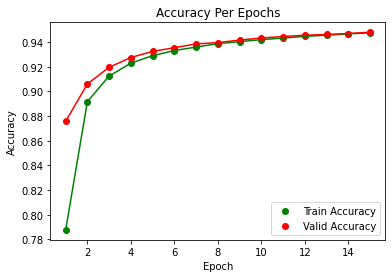
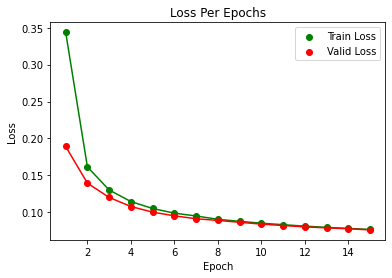
طبیعتا نمی‌توان معیار Accuracy را به ازای هر برچسب جداگانه حساب کرد. اگر بخواهیم به ازای کل تگ‌ها حساب کنیم، نتایج جدول زیر بدست می‌آید:

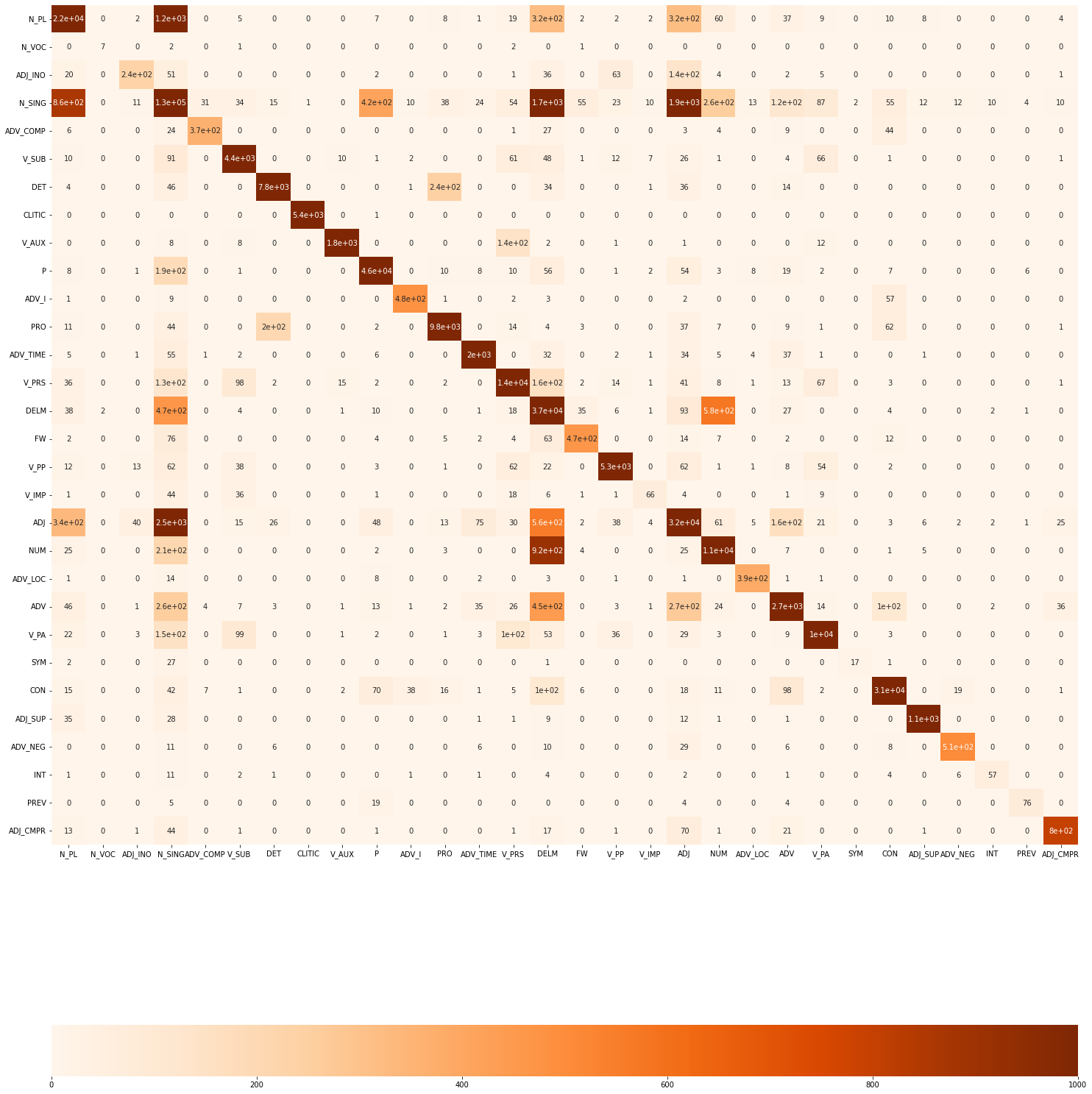
|  |  |
| --- | --- |
| مجموعه‌داده | صحت (Accuracy) |
| آموزش (Train) | ۹۴.۹۷٪ |
| اعتبارسنجی (Valid) | ۹۴.۸۱٪ |
| آزمون (Test) | ۹۴.۸۰٪ |

اگر هم بخواهیم برای هر تگ POS دقتی ارائه دهیم، می‌توانیم از معیارهایی نظیر Precision و Recall‌ استفاده بکنیم که نتایج آن در جدول زیر به طور کامل ارائه شده است:

|  |  |  |
| --- | --- | --- |
| تگ | Precision | Recall |
| N\_PL | 93 | 91 |
| N\_VOC | 77 | 53 |
| ADJ\_INO | 76 | 42 |
| N\_SING | 95 | 95 |
| ADV\_COMP | 89 | 75 |
| V\_SUB | 92 | 92 |
| DET | 96 | 95 |
| CLITIC | 99 | 99 |
| V\_AUX | 98 | 91 |
| P | 98 | 99 |
| ADV\_I | 90 | 86 |
| PRO | 96 | 96 |
| ADV\_TIME | 92 | 91 |
| V\_PRS | 96 | 95 |
| DELM | 88 | 95 |
| FW | 80 | 70 |
| V\_PP | 96 | 93 |
| V\_IMP | 68 | 35 |
| ADJ | 90 | 89 |
| NUM | 91 | 89 |
| ADV\_LOC | 92 | 92 |
| ADV | 81 | 67 |
| V\_PA | 96 | 95 |
| SYM | 89 | 35 |
| CON | 98 | 98 |
| ADJ\_SUP | 96 | 92 |
| ADV\_NEG | 92 | 86 |
| INT | 78 | 62 |
| PREV | 86 | 70 |
| ADJ\_CMPR | 90 | 82 |

در دو نمودار زیر تغییرات خطا و صحت برای مجموعه آموزش و اعتبارسنجی آورده شده است. نکته‌ی عجیب آن است که عملکرد مدل در ابتدا بر روی مجموعه اعتبارسنجی از مجموعه آموزشی بهتر بوده است! نکته دیگری هم که می‌توان در آن دید این است که روند آموزش مدل تقریبا متوقف شده است و ادامه دادن آموزش دیگر موثر نبوده است.



نمودار درهم‌ریختگی برای داده‌های آزمون در ادامه آورده شده است:

چهار جفت تگ هستند که بالای هزار مرتبه باهم دیگر اشتباه گرفته شده‌اند که اسامی آن‌ها در جدول زیر آورده شده است:

|  |  |  |
| --- | --- | --- |
| تگ پیش‌بینی‌شده | تگ درست | تعداد اشتباه |
| N\_SING | N\_PL | ۱۱۶۰ |
| DELM | N\_SING | 1673 |
| ADJ | N\_SING | 1911 |
| N\_SING | ADJ | 2480 |

طبیعتا مدل برای داشتن Accuracy بهتر باید بتواند تفاوت میان این جفت‌ها را بهتر تشخیص دهد. البته عاملی که شاید از آن غفلت شده باشد آن است که برخی از این تگ‌ها بسیار پرتکرار هستند و یک درصد خطای کوچک روی آن هم تواند تعداد خطای بالایی را ایجاد کند. به عنوان واضح‌ترین مثال، N\_SING در هر چهار سطر جدول فوق آمده است ولی دارای Precision و Recall ای معادل 95٪ است که کاملا منطبق بر عملکرد مدل است اما تعداد تگ‌های آن به طور عمومی زیاد است به گونه‌ای که حدود 130 هزار پیشبینی درست برای آن ثبت شده است.

باتوجه به آنچه گفته شد می‌توانیم ماتریس درهم‌ریختگی را سطری یا ستونی نرمال کنیم و بر مبنای درصد خطا نظر بدهیم. در جدول زیر چهار جفت از بزرگترین درصدهای خطا آوررده شده است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| نوع نرمال‌سازی | تگ پیش‌بینی‌شده | تگ درست | تعداد اشتباه | درصد اشتباه |
| درست | N\_SING | SYM | 27 | 56.25٪ |
| درست | ADJ | ADJ\_INO | 139 | 24.86٪ |
| درست | N\_SING | V\_IMP | 44 | 23.40٪ |
| پیش‌بینی‌شده | N\_VOC | DELM | ۲ | 22.22٪ |

در این جدول به عنوان مثال 27 تا از تگ‌های SYM به اشتباه N\_SING تشخیص داده شده‌اند. این مقدار اگرچه زیاد نیست ولی برای تگ بسیار کم تکرار SYM زیاد است و این مسئله باعث شده است تا 56.25٪ از تگ‌های درست SYM اشتباهی N\_SING تشخیص داده شوند! مقدار Recall پایین 35٪ ای SYM هم این موضوع را تایید می‌کند.