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

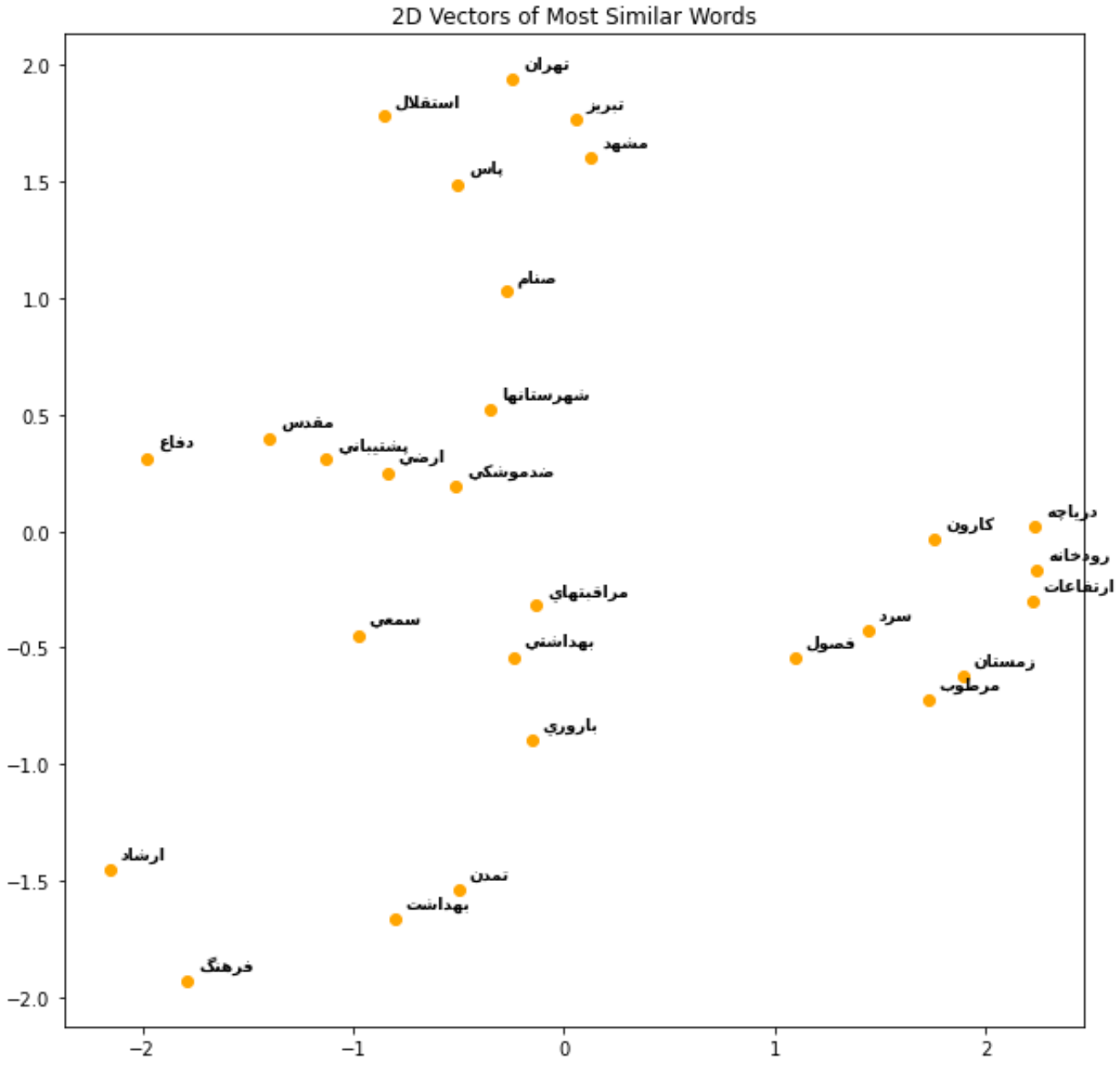
گام سوم: یافتن اسناد مشابه

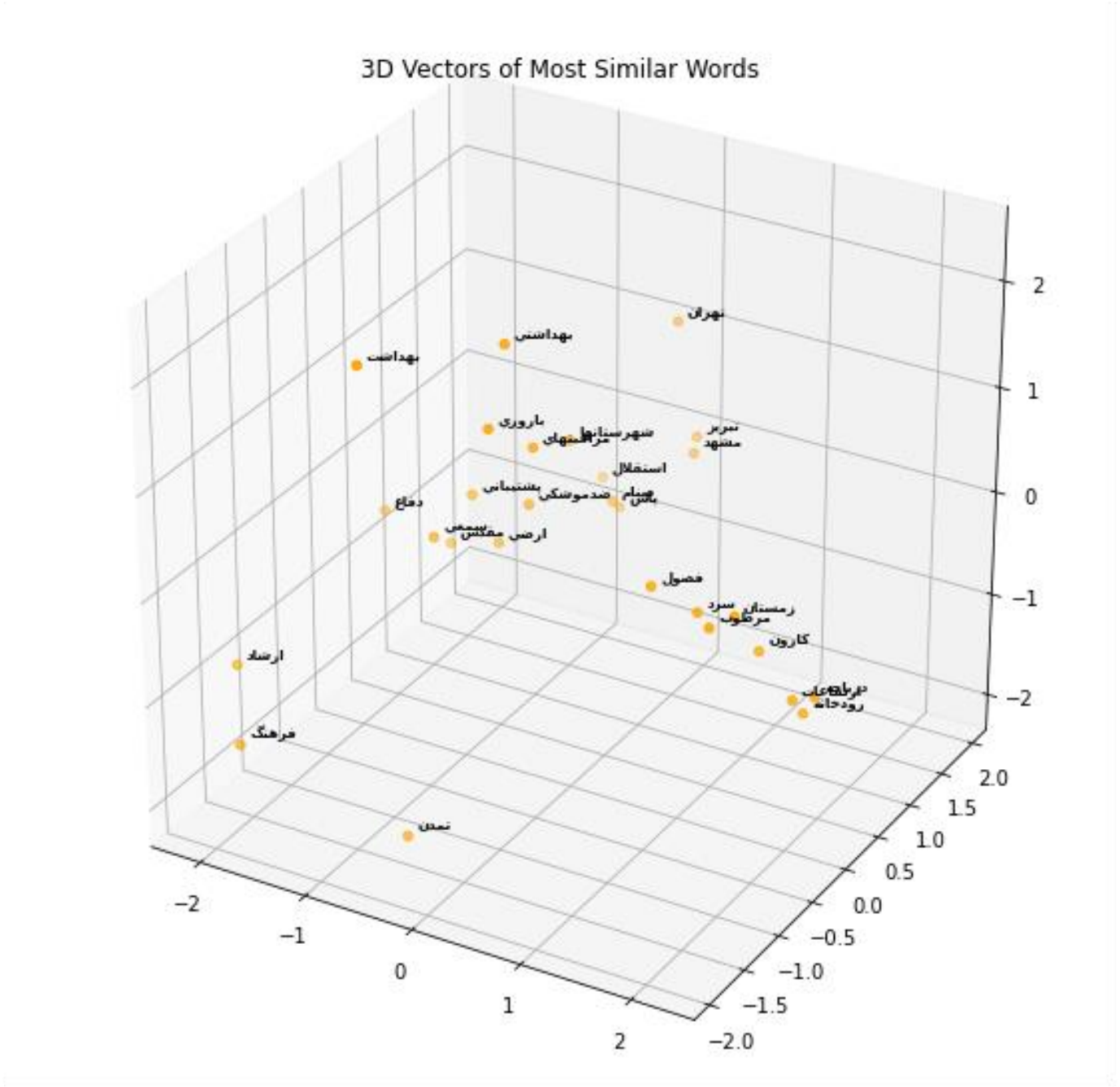
	مدل Word2Vec و TF-IDF		مدل Doc2Vec	
سند	شبیه‌ترین سند	امتیاز کسینوسی	شبیه‌ترین سند	امتیاز کسینوسی
Doc1	Doc13	98.76%	Doc13	79.60%
Doc3	Doc19	99.61%	Doc20	83.68%
Doc5	Doc26	98.96%	Doc26	81.29%
Doc25	Doc679	100%	Doc679	95.02%
Doc36	Doc7	98.73%	Doc550	84.04%

گام چهارم: بررسی کلمات مشابه

	کلمه اول	امتیاز کلمه اول	کلمه دوم	امتیاز کلمه دوم	کلمه سوم	امتیاز کلمه سوم
تهران	تبریز	۶۱/۲۵٪	شهرستانها	۶۰/۱۲٪	مشهد	۵۹/۹۵٪
بهداشت	باروری	۷۸/۱۶٪	بهداشتی	۷۴/۱۳٪	مراقبتهای	۷۳/۶۲٪
دفاع	مقدس	۶۹/۹۳٪	ضدموشکی	۶۶/۲۴٪	پشتیبانی	۶۴/۴۶٪
رودخانه	دریاچه	۸۳/۸۸٪	کارون	۸۲/۰۹٪	ارتفاعات	۸۰/۸۱٪
سرد	زمستان	۷۸/۶۰٪	مرطوب	۷۷/۷۳٪	فصول	۷۷/۳۳٪
فرهنگ	ارشاد	۷۷/۱۹٪	تمدن	۶۹/۸۶٪	سمعی	۶۷/۳۷٪
استقلال	پاس	۶۷/۹۸٪	ارضی	۶۷/۸۳٪	صنام	۶۶/۶۰٪

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


2D Vectors of Most Similar Words
تهران
استقلال
تبریز
مشهد
پاس
صنام
شهرستانها
مقدس
دفاع
پشتیبانی
ارضی
ضدموشکی
دریاچه
کارون
رودخانه
ارتفاعات
مراقبتهاي
سرد
سمعي
بهداشتي
فصول
زمستان
مرطوب
باروري
ارشاد
تمدن
بهداشت
فرهنگ
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2

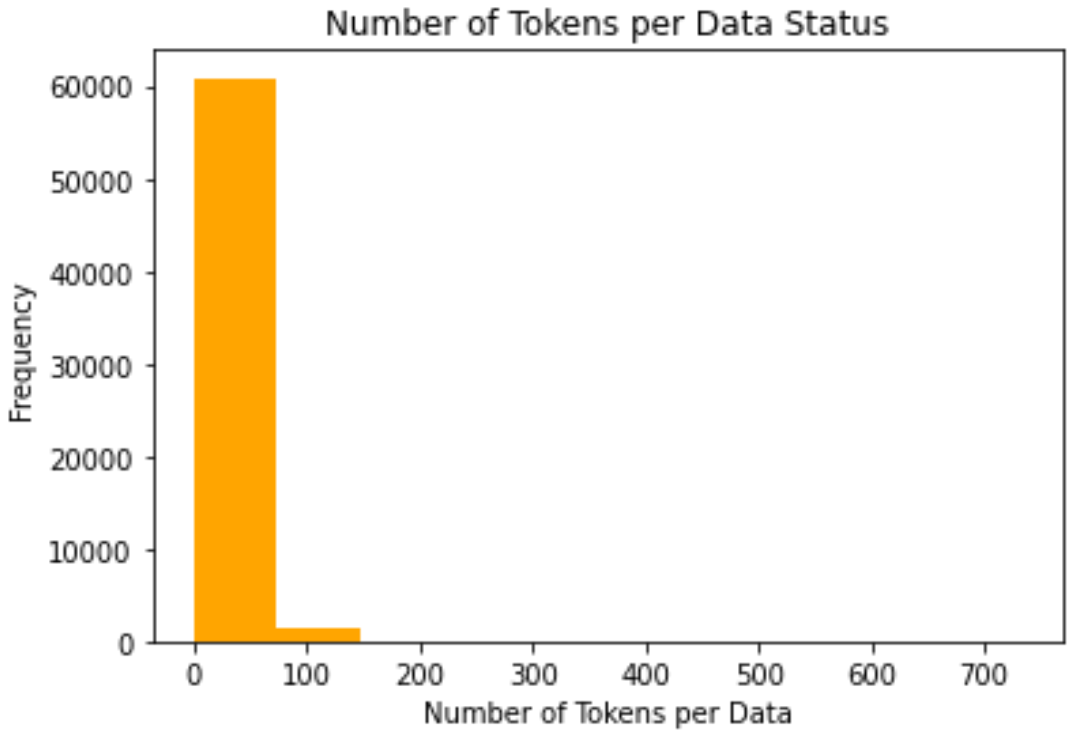

3D Vectors of Most Similar Words
بهداشت
بهداشتی
تهران
ناروری
استقلال
دفاع
فصول
کاران
ارشاد
فرهنگ
تمدن
2
1
0
−1
−2
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

- 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 ۸۰٪ 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.

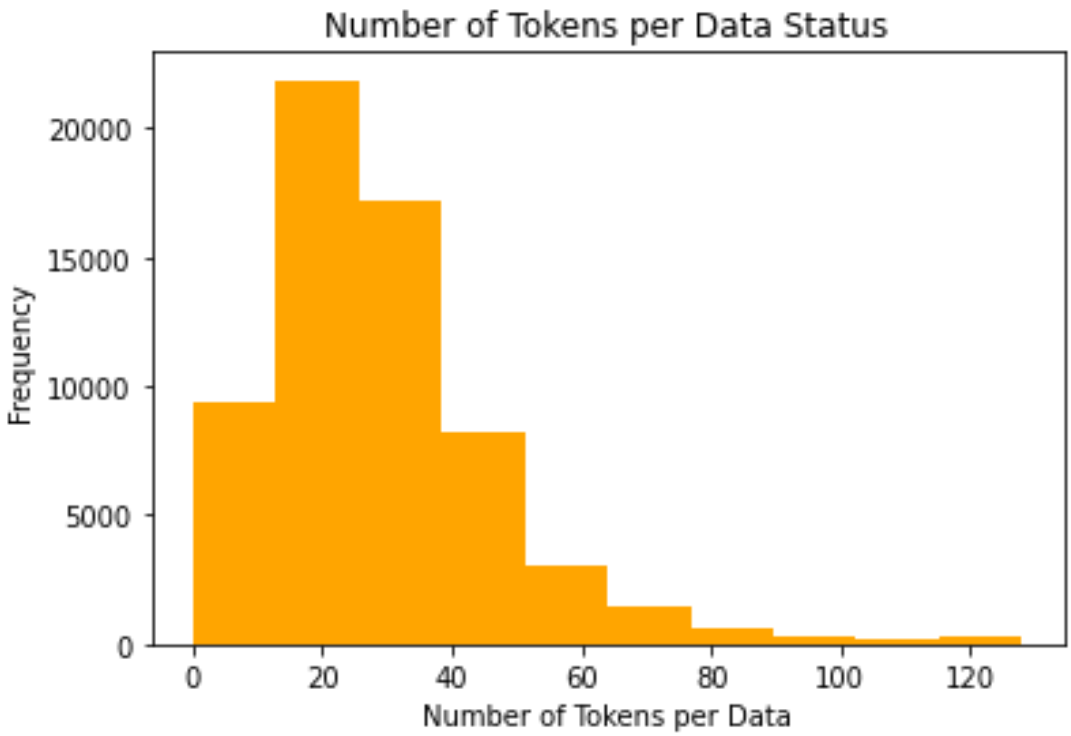
بخش دوم: تشخیص اجزای سخن (POS)

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 ۷۳۴ 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 ۳۰ 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 ۱۲۸ 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 ۶۴ 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:

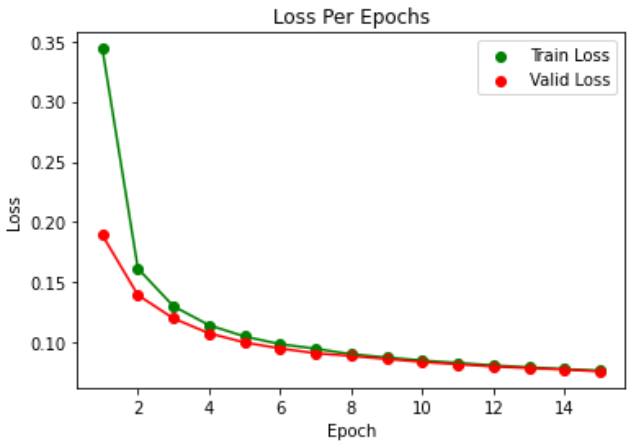
مجموعه‌داده	صحت (Accuracy)
آموزش (Train)	۹۴٫۹۷٪
اعتبارسنجی (Valid)	۹۴٫۸۱٪
آزمون (Test)	۹۴٫۸۰٪

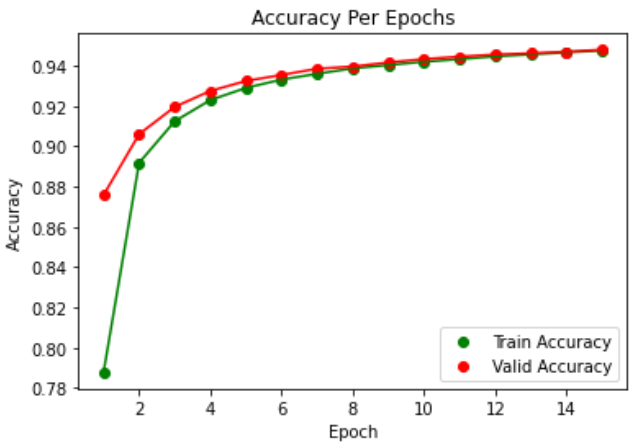
اگر هم بخواهیم برای هر تگ POS دقتی ارائه دهیم، می‌توانیم از معیارهایی نظیر Precision و Recall استفاده بکنیم که نتایج آن در جدول زیر به طور کامل ارائه شده است:

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

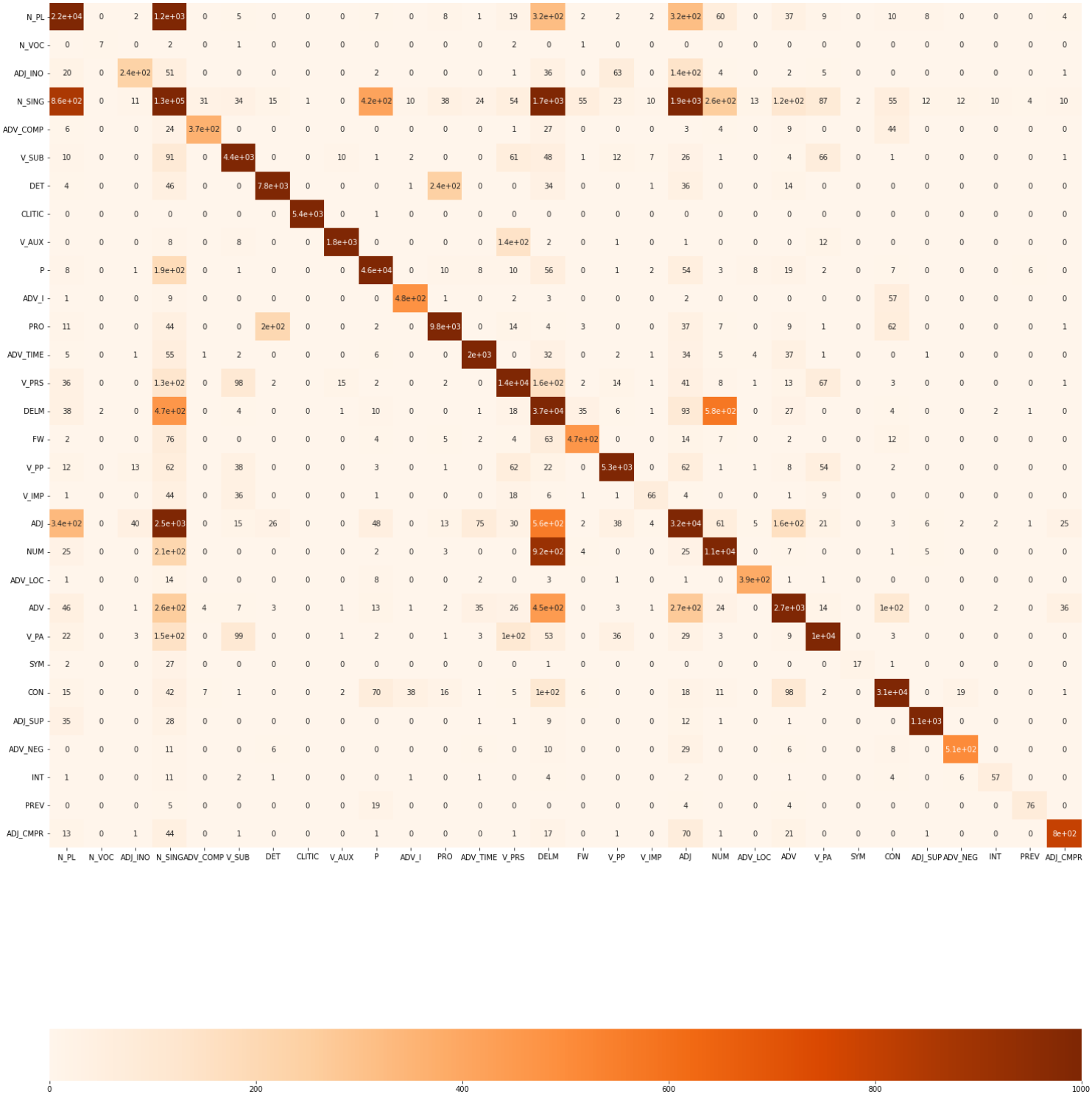
93	96	V_PP
35	68	V_IMP
89	90	ADJ
89	91	NUM
92	92	ADV_LOC
67	81	ADV
95	96	V_PA
35	89	SYM
98	98	CON
92	96	ADJ_SUP
86	92	ADV_NEG
62	78	INT
70	86	PREV
82	90	ADJ_CMPR

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	۱۶۷۳
ADJ	N_SING	۱۹۱۱
N_SING	ADJ	۲۴۸۰

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 ۹۵٪ است که کاملا منطبق بر عملکرد مدل است اما تعداد تگ‌های آن به طور عمومی زیاد است به گونه‌ای که حدود ۱۳۰ 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	۲۷	۵۶٫۲۵٪
درست	ADJ	ADJ_INO	۱۳۹	۲۴٫۸۶٪
درست	N_SING	V_IMP	۴۴	۲۳٫۴۰٪
پیش‌بینی‌شده	N_VOC	DELM	۲	۲۲٫۲۲٪

در این جدول به عنوان مثال ۲۷ 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 ۵۶٫۲۵٪ از تگ‌های درست SYM اشتباهی N_SING تشخیص داده شوند! مقدار Recall پایین ۳۵٪ ای SYM هم این موضوع را تایید می‌کند.