محمدرضا غفرانی
۴۰۰۱۳۱۰۷۶
۲۴ اردیبهشت ۱۴۰۱

پردازش زبان طبیعی
تمرین دوم

بخش اول

گام سوم

در جدول ۱ شبیه‌ترین مستند به اسناد گفته شده مشاهده می‌شود. به طور کلی درصد شباهت نسبت داده شده توسط ترکیب دو روش TF-IDF و Word2vec بیشتر از مدل Doc2vec است.

Table 1: شناسه‌های شبیه‌ترین مستندات به اسناد گفته شده

	TF-IDF & Word2vec		Doc2Vec	
	Similar Doc	Similarity	Similar Doc	Similarity
Doc1	Doc165	0.98	Doc33	0.69
Doc3	Doc19	0.99	Doc19	0.85
Doc5	Doc26	0.98	Doc0	0.62
Doc25	Doc679	1	Doc679	0.99
Doc36	Doc7	0.98	Doc0	0.61

همچنین روش TF-IDF&Word2vec مستندات شبیه‌تری را خروجی برمی‌گرداند. برای مثال برای مستند Doc5 روش TF-IDF&Word2evc مستند با شناسه Doc26 را بازگردانده است در حالی که روش Doc2vec مستند Doc0 را بازگردانده است. با بررسی این مستندات مشخص است که مستند Doc26 از نظر محتوایی نسبت به سند Doc0 به مستند Doc5 بسیار شبیه‌تر است.

گام چهارم

در جدول ۲ شبیه‌ترین کلمه به هر یک از کلمات داده شده مشاهده می‌شود. کلمه‌های مشابه و درصد شباهت هر کلمه به نظر معقول می‌رسد. در کلماتی نظیر «استقلال» که کلمه چندین معنا دارد، کلمات از حوزه‌های مختلف آورده شده است.

جدول ۲: شبیه‌ترین واژگان به کلمات داده شده

تهران	
کلمه	درصد شباهت
کرج	۶۰.۰
اصفهان	۶۰.۰
تبریز	۵۹.۰

بهداشت	
کلمه	درصد شباهت
باروری	۷۹.۰
مراقبت‌های	۷۵.۰
بهداشتی	۷۳.۰

دفاع	
کلمه	درصد شباهت
مقدس	۶۸.۰
پشتیبانی	۶۶.۰
ضدموشکی	۶۴.۰

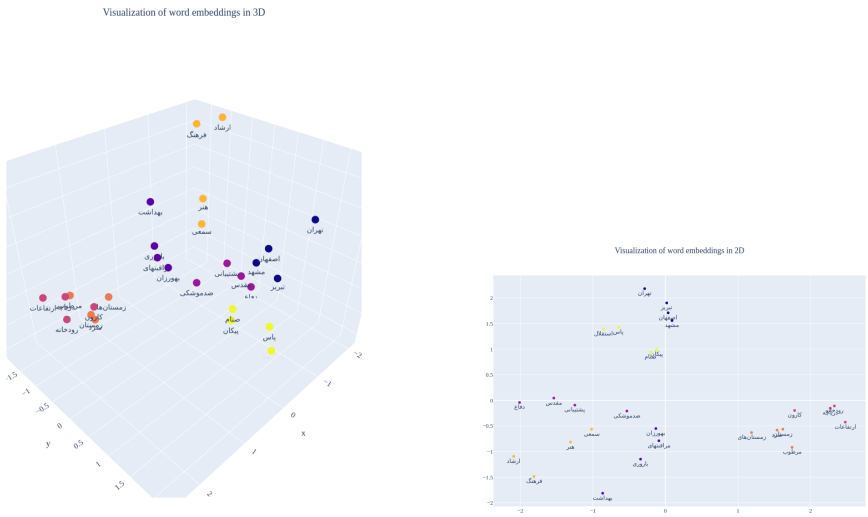
رودخانه	
کلمه	درصد شباهت
دریاچه	۸۶.۰
کارون	۸۵.۰
ارتفاعات	۸۴.۰

سرد	
کلمه	درصد شباهت
زمستان‌های	۷۸.۰
زمستان	۷۸.۰
مرطوب	۷۷.۰

فرهنگ	
کلمه	درصد شباهت
ارشاد	۷۹.۰
سمعی	۷۱.۰
تجسمی	۶۸.۰

استقلال	
کلمه	درصد شباهت
پاس	۶۸.۰
ارضی	۶۸.۰
پیکان	۶۷.۰

در شکل‌های ۱ نمودار بردار تعبیه کلمات با کاهش ابعاد بردار به ۲ و ۳ بعد رسم شده است. همان‌طور که انتظار داشتیم کلمات مشابه به یکدیگر در کنار یکدیگر قرار دارند. همچنین کلماتی از دسته‌های مختلف که به هم نزدیک هستند نیز در نزدیکی یکدیگرند قرار گرفته‌اند. با افزایش بعد داده‌ها می‌توان با دقت بیشتر شهود بهتری را می‌توان از کلمات به دست آورد.

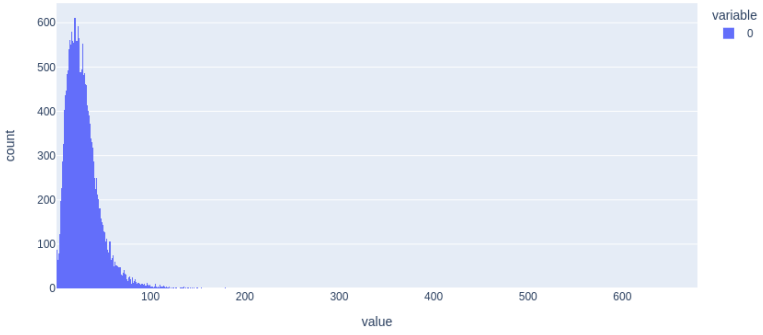


شکل ۱: نمودار تعبیه کلمات مدل با کاهش ابعاد به ۲ و ۳ بعد

بخش دوم

گام اول

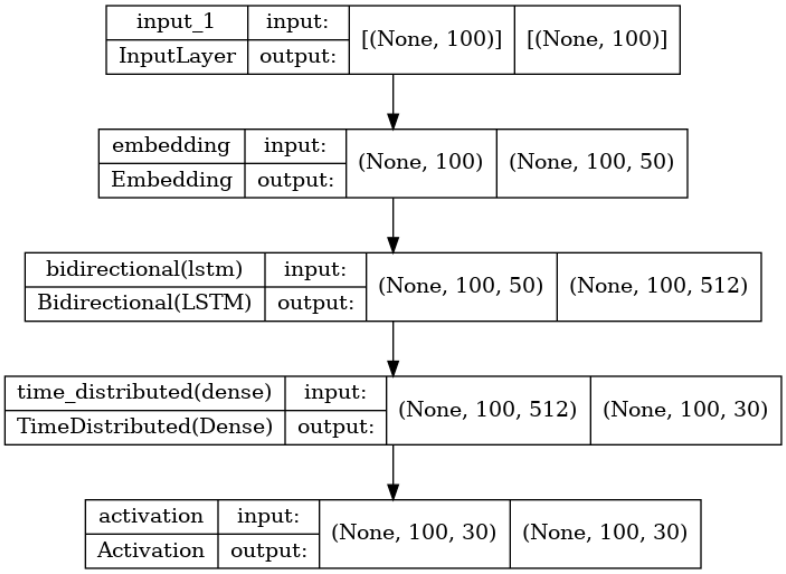
در این سوال داده‌ها را به این صورت مدل کرده‌ایم که یک جمله به تمام تگ‌های POS آن به مدل داده می‌شود و انتظار داریم که مدل بر اساس تعبیه‌های کلمه خروجی POS آن را تولید کند. طبیعتا طول بعضی از جمله‌ها کمتر است به انتهای این جملات padding اضافه می‌کنیم تا برای مدل قابل آموزش باشد. در این حالت با توجه به توضیح داده‌ها که در شکل ۲ آورده شده است، می‌توان جملات با طول بیشتر از ۱۰۰ را حذف کرد. چرا که علی‌رغم طول زیاد، فراوانی کمی دارند.



شکل ۲: نمودار توزیع اندازه طول جملات

نمودار کلی مدل در شکل ۳ آورده شده است. مدل GloVe داده شده به صورت یک لایه

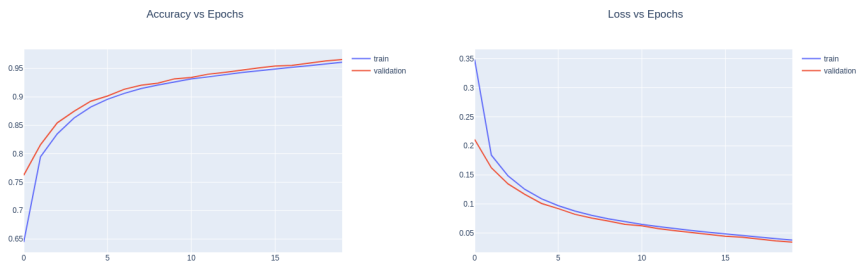
Embedding آورده شده است. در ادامه خروجی لایه تعبیه کلمه به لایه LSTM داده شده و در ادامه با استفاده از لایه Dense خروجی نهایی مدل تولید می‌شود. این مدل در داده‌های آموزش و ارزیابی به صحت 0.96 و در داده‌های تست به صحت 0.93 می‌رسد.



شکل ۳: ساختار کلی مدل LSTM

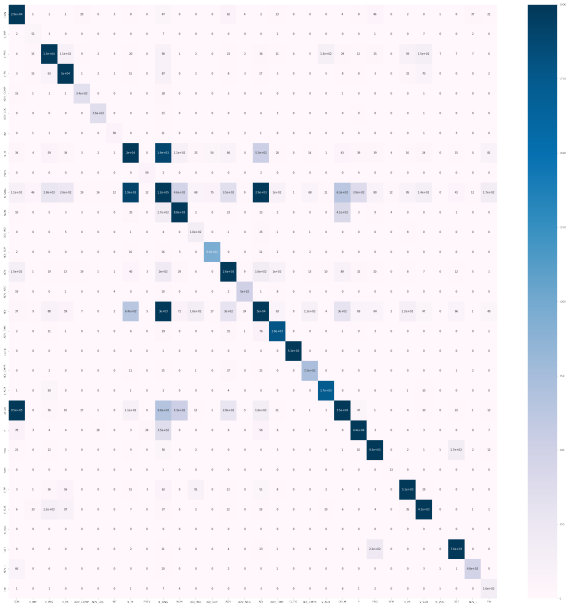
گام دوم

در شکل ۴ نمودار خطای مدل بر روی داده‌های آموزشی و ارزیابی در هنگام طی هر گام یادگیری آورده شده است. آموزش مدل در طی ۱۵ گام یادگیری و با نرخ یادگیری 10^{-3} انجام شده است. در گام‌های نهایی آموزش، خطای مدل در داده‌های ارزیابی چندان تغییراتی نداشته در نتیجه می‌توان گفت که آموزش مدل به حالت اشباع رسیده بوده است و ممکن بود که اگر آموزش را بیشتر ادامه می‌دادیم مدل بر روی داده‌ها بیش‌برازش می‌شد. روند رشد نمودار صحت نیز تقریبا به همان صورت است با این تفاوت که در گام‌هایی نهایی مدل همچنان به صورت بسیار ضعیف در حال بهبود بوده است.



شکل ۴: نمودار تغییرات خطا و صحت در طی گام‌های مختلف یادگیری

در شکل ۵ ماتریس درهم‌ریختگی مدل مشاهده می‌شود. همان‌طور که مشاهده می‌شود درصد خوبی از تگ‌های DELM برچسب CON خورده است. دلیل شباهت بسیار زیاد این دو برچسب است. هر دو این برچسب‌ها در انتهای جمله آمده و مشخص می‌کنند که جمله قبلی به پایان رسیده است. علاوه بر این مشکل، مدل برچسب N_SING را به اشتباه برچسب N_PL و ADJ تشخیص می‌دهد. برچسب زدن داده‌های N_SING و N_PL به جای یکدیگر به دلیل ساختار زبان فارسی و وجود کلماتی نظیر جمع مکسر است که در ظاهر ساده هستند اما در واقع جمع هستند. همچنین درصد زیادی از داده‌های ADJ برچسب N_SING زده می‌شود. دلیل این مشکل نیز مشابه مشکل قبلی است. در فارسی عبارت‌های مضاف و مضاف‌الیه و موصوف و صفت بسیار شبیه به هم هستند در نتیجه طبیعی است که عبارت‌های اسمی و صفت به جای هم تشخیص داده شوند.



شکل ۵: ماتریس درهم‌ریختگی شبکه عصبی

در نهایت در جدول ۳ درصد صحت تشخیص هر تگ آورده شده است. همان‌طور که مشاهده می‌شود به جز تگ CON برای باقی تگ‌ها معیار صحت ۱۰۰ درصد محاسبه شده است. این اتفاق گر چه در نگاه اول خوشحال‌کننده است اما لزوما به معنی عملکرد بهینه مدل نیست. چرا که طبق جدول درهم‌آمیختگی که پیش‌تر رسم شد، مدل مشکلات جدی دارد. با بررسی جدول درهم‌آمیختگی مشخص می‌شود که علت پایین بودن

Table 3: جدول صحت تشخیص هر تگ POS

Tag	Accuracy
CON	0.28
V_IMP	1
V_PRS	1
V_PA	1
ADV_COMP	1
ADV_LOC	1
INT	1
N_PL	1
PREV	1
N_SING	0.99
NUM	1
ADJ_INO	1
ADJ_SUP	1
ADV	1
ADV_NEG	1
ADJ	0.99
ADV_TIME	1
CLITIC	1
ADJ_CMPR:	1
V_AUX	1
DELM	0.28
P	1
PRO	1
SYM	1
V_PP	1
V_SUB	1
N_VOC	1
DET	1
ADV_I	1
FW	1