دانشکده مهندسی کامپیوتر

مبانی پردازش زبان طبیعی

گزارش پروژه

فاز۲ - تحلیل، استخراج ویژگی و رده‌بندی

مهدی امیری

فهرست مطالب

۱ لینک به ریپوی پروژه در GitHub

https://github.com/AmiriShavaki/IUST-NLP-Project

۲ آماده‌سازی محیط مجازی برای اجرای پروژه

شما می‌توانید به کمک conda و از طریق فایل requirements.txt که در root پروژه در اختیار شما قرار داده شده، به راحتی و فقط با اجرای دستور زیر تمامی پکیج‌های موردنیاز برای اجرای پروژه را نصب کنید: conda create –name <envname> –file requirements.txt

۳ آموزش و تحلیل بردار کلمات (word2vec)

در این‌جا معیار بکارگرفته‌شده برای شباهت‌سنجی بین بردارها، شباهت کسینوسی بوده است که از رابطه $\frac{A.B}{|A||B|}$ بدست می‌آید.

۱.۳ شبیه‌ترین کلمات مشترک بین جفت لیبل‌ها

	۱ستاره و ۲ستاره	۱ستاره و ۳ستاره	۱ستاره و ۴ستاره	۱ستاره و ۵ستاره	۲ستاره و ۳ستاره	۲ستاره و ۴ستاره	۲ستاره و ۵ستاره	۳ستاره و ۴ستاره	۳ستاره و ۵ستاره	۴ستاره و ۵ستاره
کلمه	obstructive	scroller	nighttime	offerings	refunded	twich	luxe	grew	ither	companions
شباهت	0.87	0.94	0.79	0.76	0.93	0.97	0.93	0.93	0.90	0.76

۲.۳ متفاوت‌ترین کلمات مشترک بین جفت لیبل‌ها

	۱ستاره و ۲ستاره	۱ستاره و ۳ستاره	۱ستاره و ۴ستاره	۱ستاره و ۵ستاره	۲ستاره و ۳ستاره	۲ستاره و ۴ستاره	۲ستاره و ۵ستاره	۳ستاره و ۴ستاره	۳ستاره و ۵ستاره	۴ستاره و ۵ستاره
کلمه	baffling	ha	manages	science	tnx	reccomended	tnx	bye	firmware	bye
شباهت	-0.56	-0.41	-0.37	-0.17	-0.34	-0.23	-0.60	-0.23	-0.25	-0.47

نکته جالب تفاوت کمتر بردارها در جفت (۱ستاره، ۵ستاره) نسبت به بقیه جفت‌ها است. در این جفت حداکثر تفاوت برداری 0.17 بوده در حالی که همانطور که در جدول می‌بینید در بقیه جفت‌ها میزان تفاوت متفاوت‌ترین بردارها بیشتر است.

کلمه "tnx" در جفت (۵ستاره و ۲ستاره) بیشترین تفاوت برداری را داشته است. احتمالا بخاطر اینکه در داده‌های ۵ستاره این کلمه همراه با رضایت‌مندی و خوشحالی کاربر بیان شده در حالی که در کلاس ۲ستاره این کلمه توأم با نارضایتی و شاید با لحن کنایه‌آمیز بیان شده است و تفاوت برداری احتمالا ناشی از همین تفاوت معنی در بکارگیری کلمه بوده است.

در جفت (۳ستاره و ۴ستاره) کلمه ی "reccomended" که دارای غلط املایی است دارای بیشترین تفاوت برداری است. احتمالا تعداد رخ دادن این کلمه با این غلط املایی خاص در بین داده‌ها کم بوده و وقتی بطور تصادفی در هر دو کلاس ۳ستاره و ۴ستاره از این غلط املایی یافت شده، مدل تعداد زیادی از این کلمه با این شکل نگارشی را در این دو کلاس ندیده که بتواند تشابهات معنایی آن‌ها را درک کند و احتمالا در همین تعداد محدودی که رخ داده هم context معنایی متن در دو کلاس متفاوت بوده.

۴ tokenization

۱.۴ بخش۱ داده‌ها

نتیجه آموزش روی کلاس‌های ۲،۳،۴،۵ ستاره و تست روی کلاس ۱ستاره:

	vocabsize=50	vocabsize=800	vocabsize=5000	vocabsize=15000
UNK percentage	2.02	4.90	7.29	7.67

۲.۴ بخش۲ داده‌ها

نتیجه آموزش روی کلاس‌های ۱،۳،۴،۵ ستاره و تست روی کلاس ۲ستاره:

	vocabsize=50	vocabsize=800	vocabsize=5000	vocabsize=15000
UNK percentage	1.73	4.33	6.44	6.78

۳.۴ بخش۳ داده‌ها

نتیجه آموزش روی کلاس‌های ۱،۲،۴،۵ ستاره و تست روی کلاس ۳ستاره:

	vocabsize=50	vocabsize=800	vocabsize=5000	vocabsize=15000
UNK percentage	1.70	4.28	6.32	6.65

۴.۴ بخش۴ داده‌ها

نتیجه آموزش روی کلاس‌های ۱،۲،۳،۵ ستاره و تست روی کلاس ۴ستاره:

	vocabsize=50	vocabsize=800	vocabsize=5000	vocabsize=15000
UNK percentage	1.77	4.38	6.56	6.91

۵.۴ بخش۵ داده‌ها

نتیجه آموزش روی کلاس‌های ۱،۲،۳،۴ ستاره و تست روی کلاس ۵ستاره:

	vocabsize=50	vocabsize=800	vocabsize=5000	vocabsize=15000
UNK percentage	2.07	4.91	7.41	7.87

۶.۴ انتخاب بهترین تنظیم Tokenizer

طبق خروجی‌هایی که در جدول‌های بالا می‌بینید حالت‌های با vocabsize کوچک‌تر تعداد توکن <UNK> کمتری در هنگام تست ایجاد می‌کرده‌اند. احتمالا دلیل بهتر عمل کردنشان این است

که هر چقدر vocabsize کوچک‌تر باشد مدل زیرکلمه‌های کوچکتر و عمومی‌تری را انتخاب می‌کند و در نتیجه در مواجهه با کلمات بیشتر و کلی‌تری از آن‌ها می‌تواند استفاده کند.

بهترین عملکرد طبق جداول بالا را مدل آموزش دیده روی کلاس‌های ۱،۲،۴،۵ با vocabsize=50 داشته است پس در فایل run-phase2.bat دستور منتقل کردن این مدل به پوشه models به صورت hard-code اضافه شده است.