گزارش تمرین سری سوم درس بازیابی پیشرفته‌ی اطلاعات
سامانه‌ی بازیابی برای شبکه‌های اجتماعی

مصطفی اوجاقی۹۷۱۰۵۷۸۲، دانیال عرفانیان۹۷۱۱۰۱۵۵، کیان باختری۹۷۱۱۰۰۲۵

استاد درس: دکتر احسان‌الدین عسگری

۹ تیر ۱۴۰۱

چکیده: در این تمرین مجموعه‌ی داده‌هایی از شبکه‌ی اجتماعی توییتر مورد تحلیل و بررسی قرار گرفته و سیستم بازیابی اطلاعات برای آن پیاده‌سازی شده است. کدهای این تمرین، در کنار این گزارش، در یک ژوپیتر نوت‌بوک با نام HW3 موجود است.

۱ جمع‌آوری داده

دادگان استفاده شده در این تمرین از شبکه‌ی اجتماعی توییتر به دست آمده‌اند. شیوه‌ی جمع‌آوری داده به این صورت بوده است که به علت تعلق نگرفتن دولوپر اکانت توییتر به اعضای گروه، به سراغ وب سایت vicinitas رفتیم و حدود بیست هزار توییت از این سایت دانلود کردیم.

شیوه‌ی کار وب سایت به این صورت است که در هر کوئری یک هشتگ یا کلمه‌ی کلیدی را از کاربر دریافت می‌کند و حدود دو هزار توییت مرتبط یا آن کوئری که در ده روز اخیر توییت شده‌اند را در یک فایل اکسل تحویل می‌دهد. ما در حدود بیست کوئری به سایت دادیم و حدود چهل هزار توییت دریافت کردیم که پس از پیش‌پردازش در حدود بیست هزار توییت یکتا در مجموعه‌ی دادگان ما باقی ماند. کلمات کلیدی‌ای که ما به سایت دادیم اکثراً کلیدواژه‌ی حوزه‌ی تکنولوژی بودند مانند اسم شرکت‌های بزرگ و یا رویدادهای کامپیوتری معروف. همچنین چند کوئری مرتبط با مسائل سیاسی و اجتماعی هم به مجموعه اضافه شدند. فایل‌های خام اکسل که مستقیم از سایت دریافت شده‌اند در پوشه‌ای به نام Excels در پوشه‌ی اصلی پروژه موجود هستند.

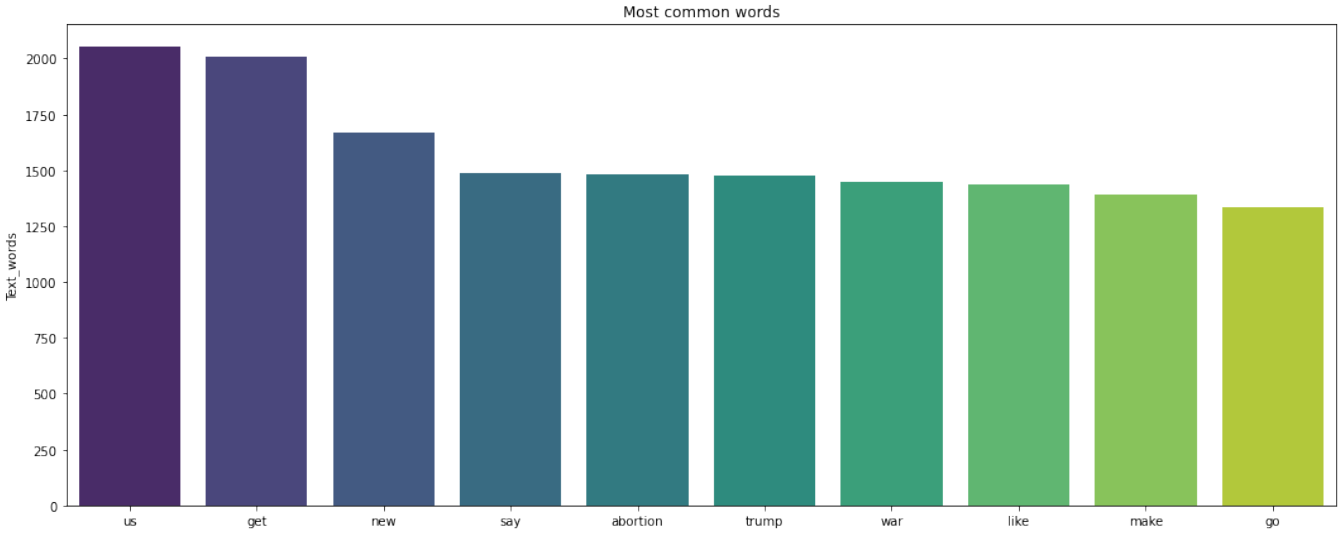
۲ پیش‌پردازش

توییت‌های جمع‌آوری شده حاوی زبان‌های مختلف، اموجی، کلمات قصار، هشتگ، منشن و بسیاری از موارد دیگر بودند که با استفاده از پیش‌پردازش یا حذف شدند و یا به فرمت قابل قبولی تبدیل شدند. در خط لوله‌ی پیش‌پردازش (preprocessing pipeline) به تربیت از موارد زیر استفاده شده است:

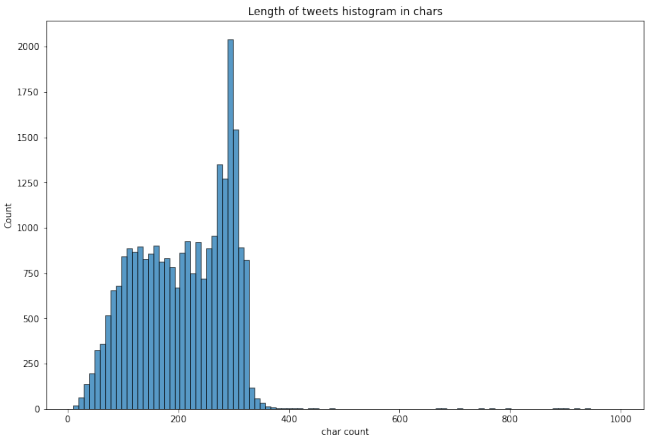
- **حذف توییت‌های با زبان غیر انگلیسی**
- **حذف توییت‌های تکراری**
- **حذف خطوط اضافه و فاصله‌های سفید طولانی**
- **حذف منشن‌ها و هشتگ‌ها**
- **حذف تگ‌های html**
- **حذف علائم نگارشی**
- **حذف لینک‌ها و هایپرلینک‌ها**
- **حذف حروف لهجه‌دار**
- **تبدیل حروف بزرگ به کوچک**
- **کاهش حروف مکرر (پیاپی) به یک حرف**
- **گسترش عبارات فشرده**
- **حذف کلمات ایست**
- **تصحیح املای کلمات**

پس از اجرای این مراحل، با استفاده از کتابخانه‌ی nltk این توییت‌ها lemmatize و tokenize شدند تا آماده‌ی نهفته‌سازی (embedding) شوند.

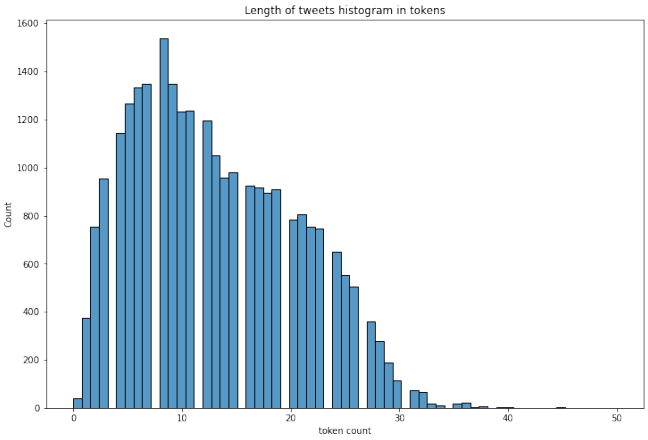
در شکل‌های ۱ و ۲ می‌توان کلمات پربسامد در این توییت‌ها را به ترتیب در نمودار ستونی و در ابر کلمات مشاهده کرد. همچین در شکل‌های ۳ و ۴ توزیع توییت‌ها بر اساس طولشان قابل مشاهده است.



شکل ۱: نمودار ستونی کلمات پر بسامد در توییت‌ها



شکل ۳: توزیع توییت‌ها بر اساس تعداد کاراکترها



شکل ۴: توزیع توییت‌ها بر اساس تعداد توکن‌ها پس از پیش‌پردازش

شکل ۲: ابر کلمات پر بسامد در توییت‌ها

۳ سامانه‌ی بازیابی

۳-۱ tf-idf

در این بخش از امبدینگ tf-idf برای بازنمایی کوئری‌ها و توییت‌ها استفاده شده است. با ورود هر کوئری، پس از گذشتن از خط لوله‌ی پیش پردازش، بازنمایی کوئری محاسبه شده و از طریق فاصله‌ی زاویه‌ای با توییت ها مقایسه می‌شود. در نهایت ده توییت که بیشترین ارتباط را به کوئری دارند باز گردانده می‌شوند.

۳-۲ boolean search

در این بخش از امبدینگ boolean برای بازنمایی کوئری‌ها و توییت‌ها استفاده شده است که هر متن تبدیل به یک بردار دودویی می‌شود. با ورود هر کوئری، پس از گذشتن از خط لوله‌ی پیش پردازش، بازنمایی کوئری با استفاده از بازنمایی دودویی محاسبه شده و از طریق مقایسه‌ی تعداد بیت‌های یکسان با توییت ها مقایسه می‌شود. در نهایت ده توییت که بیشترین ارتباط را به کوئری دارند باز گردانده می‌شوند

۳-۳ FastText

در این بخش از امبدینگ FastText برای بازنمایی کوئری‌ها و توییت‌ها استفاده شده است. هر کوئری، پس از گذشتن از خط لوله‌ی پیش پردازش، با استفاده از مدل از پیش ترین شده‌ی FastText بازنمایی‌اش محاسبه شده و از طریق فاصله‌ی کسینوسی با توییت‌ها مقایسه می‌شود. در نهایت ده توییت که بیشترین ارتباط را به کوئری دارند باز گردانده می‌شوند

۳-۴ Transformer

بازنمایی مبتنی بر مدل‌های ترنسفورمری را با استفاده از مدل از پیش ترین شده‌ی SentenceTransformer به دست آوردیم. این مدل توییت‌های پیش‌پردازش شده را ورودی می‌گیرد و برای هر کدام یک بازنمایی مبتنی بر ترنسفورمر ارائه می‌دهد. هر کوئری نیز به این مدل داده می‌شود تا بازنمایی‌اش به دست آید و از طریق ضرب داخلی فاصله‌اش تا توییت‌ها محاسبه شده و توییت‌های مرتبط‌تر به دست آیند.

۴ ارزیابی

ارزیابی این چهار سامانه‌ی مختلف بازیابی به این صورت انجام شد که تعداد ده کوئری تنظیم شد تا به هر چهار سامانه داده شود. سپس نتایج از طریق درست شدن gold standard توسط انسان و با استفاده از معیار MRR ارزیابی شدند. این ارزیابی به این صورت انجام شد که برای هر ده کوئری و هر چهار مدل، نتایج بازگردانده شده بررسی شد و مشخص شد که هر توییت مرتبط به کوئری هست یا خیر. البته این ارزیابی در دو مرحله انجام شد: ابتدا برای مدل‌های غیر ترنسفورمری در لوکال و برای مدل ترنسفورمری در گوگل کولب. این جدایی دو علت دارد، یکی حجم بالای مدل ترنسفومری و زمان طولانی لازم برای ساخته شدن بازنمایی‌ها توسط سی‌پی‌یو و دیگری نیز عدم دسترسی به وب سایت

	models	scores
0	FastText	0.900000
1	BooleanSearch	0.850000
2	TFIDFSearch	0.766667

شکل ۵: نتایج ارزیابی MRR روی مدل‌های FastText، tf-idf، BooleanSearch

	models	scores
0	TransformerSearch	0.95

شکل ۶: نتایج ارزیابی MRR روی مدل SentenceTransformer

hugging face که از مدل‌های از پیش ترین شده‌ی کتابخانه‌ی trans-formers میزبانی می‌کند به خاطر مسائل تحریمی و ارور ۴۰۳ روی لوکال. هر دوی این مشکلات با اجرای برنامه روی گوگل کولب حل شدند. نتایج ارزیابی در شکل‌های ۵ و ۶ قابل مشاهده هستند. هنگام ارزیابی این نکته رعایت شد که برچسب‌ها مستقل از مدل ایجاد شوند تا از ایجاد بایاس به نفع مدل‌های قوی‌تر جلوگیری شود. همان طور که در جداول پیدا است، نتایج بسیار مطابق با انتظار بوده است.

کوئری‌هایی که برای ارزیابی استفاده شده اند به شرح زیر می‌باشند و برچسب‌هایی که برای هر توییت در نظر گرفته شده در نوت‌بوک‌ها موجود هست.

۱. how to became full stack developer
۲. Microsoft Internet Explorer age
۳. Tesla price
۴. Macbook M2 Price
۵. Abortion rights
۶. Ukraine war
۷. Raisi
۸. Harry potter
۹. Covid vaccine
۱۰. Meta VR

۵ فایل‌ها و پوشه‌ها

در کنار این گزارش، یه پوشه با نام Excels قرار دارد که حاوی خروجی‌های خام از وب سایت هنگام جمع‌آوردی داده است. یک فایل all.csv موجود هست که نسخه‌ی ذخیره شده‌ای از مجموعه‌ی داده‌ها

است که پیش‌پردازش شده و آماده‌ی استفاده‌ی مدل‌ها می‌باشند. یک فایل contraction map نیز موجود هست که برای گسترش واژگان فشرده هنگام پیش‌پردازش کاربرد دارد. یک ژوپیتر نوت‌بوک با نام HW3 و نوت‌بوک دیگری با نام HW3-Transformer موجود هست که در اولی ارزیابی مربوط به سه مدل اول و در دومی ارزیابی مربوط به مدل ترنسفورمری انجام شده است.