دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گزارش پروژه درس مباحثی در علوم‌کامپیوتر

سیستم پیشنهاد هشتگ

دانشجویان:

مهسا دیباجی

علیرضا طغیانی

پرنیان تقی پور

استاد:

دکتر اکبری

پاییز ۹۹

فهرست مطالب

مقدمه

توصیه هشتگ، به ویژه با افزایش علاقه به استفاده از رسانه های اجتماعی مانند توییتر در دهه گذشته، یک کار اساسی است. سیستم های پیشنهادی هشتگ هنگام نوشتن توییت به طور خودکار هشتگ را به کاربر پیشنهاد می دهند. (السینی و همکاران، ۲۰۲۰) در توییتر، کاربران توئیت‌هایی می‌نویسند که پیام کوتاه هستند و بیش از ۲۸۰ حرف ندارند. هشتگ کلمه‌ای است که با نماد # پیشوند می‌شود و می‌توان یک یا چند هشتگ را در یک توییت وارد کرد. برخی از افراد برای دسته بندی توییت‌های خود از هشتگ استفاده می‌کنند. برخی دیگر از هشتگ‌ها برای برچسب گذاری مطالب مرتبط با بلایا یا رویدادهای خاص مانند انتخابات استفاده می کنند. هشتگ‌ها باعث می شوند توییت‌ها و موضوعات پرطرفدار توسط دیگر کاربران مرتبط به راحتی قابل جستجو باشند. از آنجا که هشتگ‌ها توسط هیچ کاربر یا گروهی ثبت و کنترل نمی شوند ، پیدا کردن هشتگ‌های مناسب برای توییت‌های خود برای برخی از کاربران سخت خواهد بود. (کایوی و همکاران، ۲۰۱۱)

تعریف مسئله

در مسئله مورد بررسی، تلاش بر این است که با داشتن مجموعه‌ای از توییت‌های کاربران و هشتگ‌های استفاده شده در آن‌ها، سیستمی ارائه دهیم که برای توییت‌های جدید کاربران به طور خودکار با استفاده از روش‌های پردازش زبان طبیعی و یادگیری ماشین هشتگ‌های مناسب را پیشنهاد دهد.

مجموعه داده

در مرحله اول ۲۵۰۰۰ داده از توییت‌های کاربران در سال ۲۰۲۰ توسط api توییتر جمع آوری شد. سپس به منظور اینکه دقت مدل بهبود یابد و برای توییت‌های جدید نیز پیشنهاد مناسب‌تری بدهد ۳۰۰۰ توییت دیگر با موضوعات مشخص جمع آوری شد.

در مرحله پیش پردازش، توییت‌های تکراری حذف و سپس در توییت‌های باقی مانده ایست‌واژه‌ها[۱]، لینک‌ها، نام های کاربری و نمادهای ریتوییت حذف شدند. هشتگ‌های هر توییت استخراج شد و پس از آن هرکلمه با ریشه خود جایگزین شد و عددها و کاراکترهای نامربوط و کلمات با طول کوتاه(کمتر از ۳) از متن توییت ها حذف شدند.

---

۱ Stopwords

روش

برای این پروژه از دو روش برای پیشنهاد هشتگ استفاده شد. در روش اول الگوریتم Multinomial Naïve Bayes به کار رفت. ابتدا به دلیل تعداد زیاد ویژگی‌ها و برای کاهش سایز بردار ویژگی توسط CountVectorizer ۲۰۰۰ ویژگی پرتکرار انتخاب شدند. سپس این روش بر اساس نتایج حاصل از مدل Multinomial Naïve Bayes که برای اسناد متنی با واژگان بزرگ و داده های پراکنده استاندارد است، توصیه هایی را ارائه می دهد. در این مدل، رتبه بندی هشتگ بستگی به احتمال پسین هشتگ H_i با داشتن توییتی که از مجموعه ای از کلمات t_j که هر کدام با تکرار f_{tj} تشکیل شده است، دارد.

روش دوم الگوریتم HF-IHU بود که یک روش ایده گرفته از tf-idf است. در این الگوریتم ابتدا دو ساختمان داده از رابطه‌ی کلمات و هشتگ‌ها از هر دو سمت ایجاد میکنیم. به این ترتیب که کدام کلمات به هشتگ به خصوصی مربوط میشوند و برای رابطه‌ی هر کلمه با هر هشتگ تعداد تکرار رخداد آن‌ها با یکدیگر را در ساختمان داده‌های مربوطه نگه میداریم.

برای امتیاز دهی به هر هشتگ برای هر توییت از روابط زیر استفاده میکنیم.

$$\mathrm{hf}_{t,h} = \frac{\mathrm{THFM}[t][h]}{\sum_{h'} \mathrm{THFM}[t][h']}$$

$$\mathrm{ihu}_{h} = \log \frac{|\mathrm{Corpus}_{NH}|}{\sum_{t'} HFM\ [h]}$$

در محاسبه‌ی hf در صورت کسر $\mathrm{THFM}[t][h]$ همان خانه ی مربوط به تکرار کلمه‌ی t و هشتگ h در ساختمان داده است. در واقع در این معیار اگر کلمه‌ای ارتباط قوی و تکرار بیشتری فقط با هشتگ خاصی داشته باشد و رابطه‌ی ان با هشتگ‌های دیگر کم باشد احتمال این که آن هشتگ برای متنی با آن کلمه پیشنهاد شود بالا میرود. در محاسبه ihu رابطه‌ی هشتگ و دیگر کلمات بررسی می‌شود. به این ترتیب که اگر هشتگی با متن‌ها و کلمات زیادی از متن ضمینه ارتباط داشته باشد امتیاز آن کاهش می‌یابد. در این رابطه صورت کسر تعداد کل کلمات موجود در متن توییت‌های دیتای آموزش (به جز هشتگ ها) می باشد. در این الگوریتم ما ابتدا مهم‌ترین لغات از توییت ها را پیدا میکنیم و سپس از لیست مرتب شده‌ی هشتگ های مربوط به کلمات بنا به امتیازات انها مربوط‌ترین هشتگ‌ها را پیشنهاد می‌دهیم. ماکسیمم تعداد هشتگ‌های پیشنهادی پنج است.

ارزیابی

برای ارزیابی دو الگوریتم معیارهای precision ، recall ، f1-score و accuracy محاسبه شدند که مقادیر آن‌ها در جدول زیر آمده است.

برای محاسبه این معیارها به این شکل عمل می‌کنیم که برای هر توییت اگر هشتگ پیشنهادی در هشتگ‌های اصلی توییت بود، آن پیشنهاد true positive و در غیر این صورت false positive در نظر گرفته میشود. همچنین هر هشتگی که در هشتگ‌های اصلی توییت حضور داشت اما پیشنهاد نشده بود به عنوان false negative در نظر گرفته می‌شود. حال میتوانیم با روابطی که برای محاسبه این معیارها برای ۲ کلاس داشتیم، آن‌ها را محاسبه کنیم.

HF-IHU	
Precision	32.5802
Recall	92.1392
F1-Score	48.1378
Accuracy	81.0173

Multinomial Naïve Bayes	
Precision	35.8143
Recall	16.8166
F1-Score	22.8867
Accuracy	59.4694

سیستم نهایی

سیستم نهایی این پروژه در ابتدا قرار بود که به شکل تنها یک اپ iOS باشد که قابل اجرا بر روی iOS ۱۱ و ورژن‌های بعدی خواهد بود که. با دریافت متن توییت موردنظر و بهره‌گیری از یک Rest API پیشنهادات موردنظر را به شکل هشتگ‌های مختلف ارائه کند ولی بعد از اتمام پیاده سازی اپ تصمیم گرفتیم که با استفاده از امکانات Swift ۵ که به این شکل است که درصورت رعایت یک سری مسائل در پیاده سازی و رسیدگی به موارد موردنیاز میتوان با داشتن یک سورس‌کد مشترک هر سه اپ موردنظر برای سیستم‌عامل‌های IOS و MacOS وiPadOS را در خروجی نهایی داشته باشیم و از آن بهره ببریم که این موارد موردنیاز انجام شد و در قالب سه اپ با ظاهر و کاربری یکسان به خروجی نهایی رسیدیم.

برای دسترسی به اپ‌ها تا چندروز آینده و پس از تایی توسط اپل امکان دانلود اپ‌های نسخه iPadOS وiOS از اپ استور خواهد بود و نسخهMacOS فایل اپ در پیوست تمرین با نامHashtagica در سامانه کورسز بارگزاری شده است.

همچنین جهت ساده سازی مشاهده نتیجه نهایی فایل ویدئویی از عملکرد هر سه اپ با ضبط از صفحه نمایش لپتاپ آماده شده که در پیوست در سامانه کورسز بارگزاری شده‌است.

نحوه کارگروهی

در این پروژه مرحله تحقیق و بررسی نحوه پیاده‌سازی پروژه با روش‌های مختلف توسط همه اعضای گروه به شکل مشترک توسط بررسی و مطالعه مقالات مختلف و بحث درمورد نحوه اجرای پروژه به شگل گروهی انجام شد.

در مرحله بعدی توسط علیرضا طغیانی مرحله جمع‌اوری دیتاست با کمکAPI توئیتر برای استفاده در مراحل بعدی انجام شد.

در مرحله بعد تصمیم بر آن شد که پیاده سازی پروژه از دو روش منتخب مطالعه شده در مقالات باشد، یکی روش Naive Bayes classifier و دیگری روش HF-IHU که روش اولی توسط مهسا دیباجی و روش دوم توسط پرنیان تقی‌پور اجرا شد که در این فرایند در صورت نیاز به کمک در قسمت‌های کوچکی در کد ممووجود بر روی کولب اصلاحات انجام می‌شد.

در مرحله بعدی برای پیاده سازی اپ خروجی نهایی نیاز به یک Rest-API برای ارسال ریکوئست و دریافت هشتگ‌های پیشنهاد با توجه به متن وارد شده توسط کاربر بود درنتیجه مهسا دیباجی پیاده سازی اینAPI را با کمکPython وFlask انجام داد.

در این مرحله برای Deploy اینAPI طراحی شده با توجه به اینکه هیچکدام از اعضای گروه تجربه انجام اینکار را نداشتند به شکل گروهی درحال تلاش برای انجام اینکار از راه‌های مختلف هستیم.

در مرحله بعدی پیاده سازی یک نسخه اپ iOS و MacOS توسط علیرضا طغیانی انجام شد که نحوه انتشار اپ iOS در TestFlight خواهد بود که بتواند توسط افراد مختلف برای تست قابل دانلود باشد و فایل قابل اجرا اپ MacOS نیز به پیوست آپلود خواهد شد که قابل اجرا برروی سیستم عامل‌های حداقل ۱۰.۱۵ MacOSخواهد بود.

روش‌های بهبود دهنده

در این پروژه ما معیارهای مربوط به کلمات را با دو روش Naïve Bayes و HF-IHU بررسی کردیم. اما با توجه به ویژگی‌های منحصر به فرد توییت ها معیارهای متفاوتی میتوانند در انتخاب هشتگ مناسب برای یک توییت موثر باشند. یکی از مهمترین عامل‌ها ارتباط کلمات و موضوع کلی متن است؛ اگرچه کلمات نشانگر اصلی ویژگی‌ها هستند ولی استفاده از یک طبقه‌بند[۲] برای تشخیص رابطه کلی کلمات می‌تواند در دقت موثر باشد. علاوه بر این

---

[۲] Classifier

انتخاب هشتگ مناسب رابطه‌ی مستقیمی با کاربر و گراف اجتماعی او در حساب کاربریش دارد و احتمال زیادی برای نزدیکی موضوع بحث با مبحثی که در تایم لاین او بیشتر تکرار شده است وجود دارد. برای به روز بودن پیشنهادات با موصوعاتی که ترند هستند و با در نظر گرفتن سرعت تغییر موضوعات مورد بحث در توییتر می‌توان در بازه‌های زمانی مشخصی دیتاهای ترین را با وزن‌های مختلف تغییر داد به طوری که دیتاهای جدید اهمیت بیشتری داشته باشند. در اخر استفاده از دیگر اصلاعات متنی توییت مانند محتوای لینک موجود هم میتواند تاثیر به خصوصی در نتیجه‌ی این تسک داشته باشد.