دانشکده مهندسی کامپیوتر

استاد درس: سید صالح اعتمادی

بهار ۱۴۰۲

تحلیل احساسات روی ویدئوهای نقد فیلم‌های سینمایی

درس پردازش زبان‌های طبیعی

حوریه سبزواری

شماره دانشجویی: ۹۸۴۱۲۰۰۴

گزارش فاز اول

موضوع پروژه تسک analysis sentiment روی ویدیوهای نقد و بررسی فیلم‌های سینمایی است. داده‌ها از منتقدهای مشغول به فعالیت در پلتفرم یوتیوب جمع‌آوری می‌شود. در انتها برای هر داده‌ی ورودی مشخص می‌گردد که نظر کلی ویدیو راجع به آن فیلم سینمایی خوب، متوسط یا بد بوده است.

لینک Repository: https://github.com/lhoorie/SentimentAnalysisOnReviewVideos

۱ منبع دقیق داده

منبع جمع‌آوری داده فیلم‌های ۲ کانال یوتیوب بنام‌های Jeremy Jahns و Chris Stuckmann بوده است. تلاش بر این بوده که داده‌ها به صورت متوازن از هر سه برچسب انتخاب و استخراج شوند.

۲ روش جمع‌آوری، مراحل و ابزارهای استفاده‌شده در جمع‌آوری داده

در مراحل این پروژه از کتابخانه‌های زیر برای عنوان‌های ذکرشده استفاده شده‌است:

- کتابخانهٔ YouTubeTranscriptApi برای گرفتن زیرنویس ویدئوهای داخل دیتاست
- کتابخانهٔ YouTube برای استخراج نام فیلم
- کتابخانهٔ extract برای استخراج id از لینک ویدئو
- کتابخانهٔ nltk برای tokenize کردن جملات
- کتابخانهٔ NNSplit برای جمله‌بندی زیرنویس‌ها
- کتابخانهٔ re برای تمیز کردن جملات
- کتابخانهٔ pandas برای ساختن دیتافریم‌ها
- و سایر کتابخانه‌های کمکی..

برای جمع‌آوری داده ابتدا لینک‌های ۹۰ فیلم از کانال‌های ذکرشده به صورت متوازن انتخاب شده است. سپس زیرنویس هر لینک استخراج شده و توابع موجود در کد روی آن اعمال می‌شوند. در انتها نیز اطلاعات مربوطه به صورت data frame درآمده و در فایل‌های csv ذخیره می‌شوند.

۳ فرمت داده‌ها

همانطور که در پروژه ذکر شده بود، فرمت داده‌های پروژه به صورت زیر است:

- پوشهٔ `data` شامل دو زیرپوشه به نام‌های `rawdata` و `cleandata` است که هرکدام شامل یک فایل `csv` از دادهٔ خام و دادهٔ تمیز است.
- پوشهٔ `src` شامل یک فایل `notebook` کدهای پروژه است.
- پوشهٔ `stats` شامل فایل‌های `csv` آمارهای خواسته شده است.

برچسب‌ها به صورت سه عدد ۰ و ۱ و ۲ هستند. برچسب ۰ به معنای نظر منفی، برچسب ۱ به معنای نظر خنثی و برچسب ۲ به معنای نظر مثبت منتقد است.

۴ پیش‌پردازش‌های انجام شده

۱.۴ روش/ابزار تفکیک جملات

همانطور که می‌دانیم در زیرنویس ویدئوها علائمی مانند نقطه وجود ندارد و واحد آن عباراتی است که برای چند ثانیه روی قسمتی از فیلم می‌آیند. لذا من با استفاده از ابزار `NNSplit` جملاتی حدودی به ازای هر زیرنویس استخراج کردم `NNSplit` یک کتابخانه برای تقسیم‌بندی متن است که از شبکه‌های عصبی برای انجام تشخیص مرز جمله استفاده می‌کند. با آموزش یک مدل یادگیری عمیق بر روی مجموعهٔ بزرگی از داده‌های متنی کار می‌کند تا مکان‌هایی را که جملات شروع و پایان می‌دهند را شناسایی کند.

۲.۴ روش/ابزار تفکیک توکن‌ها/کلمات

توکن‌های ورودی مجموعهٔ کلمات زیرنویس هر ویدئو هستند. برای بدست‌آوردن کلمات به راحتی می‌توانیم از ابزار `nltk.word-tokenize()` استفاده کنیم.

۳.۴ روش/معیارهای تمیزکردن داده

پس از مشاهدهٔ چند نمونه زیرنویس متوجه شدم که زیرنویس‌ها بدون علائم نگارشی، کلمات اضافی، لینک و .. هستند و به نوعی داده‌های استخراج‌شده از ابتدا تمیز بودند. اما برای اطمینان تابعی بنام clean data تعریف کردم که موارد زیر را حذف می‌کند:

- تمام عباراتی که با @ آغاز می‌شوند
- تمام علائم نگارشی موجود در متن
- تمام URL ها و ارجاعات

۴.۴ اندازهٔ داده قبل/بعد از تمیز کردن داده

همانطور که در موارد قبلی ذکر شد، داده‌ی خام موارد زیادی برای تمیز کردن نداشت و صرفا دارای کلمات بود. حتی برای پیدا کردن تعداد جملات، بدلیل عدم وجود نقطه و علائم نگارشی از کتابخانه‌های sentence detection استفاده شد. لذا اندازه‌ی داده قبل/بعد از اعمال عملیات تمیزکردن تغییری نداشته است.

۵ واحد و روش برچسب‌گذاری

واحد برچسب‌گذاری به صورت هر ویدئو و در واقع هر زیرنویس بوده است. همچنین برای روش برچسب‌گذاری ابتدا قصد داشتم تا تسک sentiment analysis را بر روی نظرات هر ویدئو اعمال کنم اما پس از بررسی متوجه شدم که نظرات هر ویدئو متفاوت با برچسب خود ویدئو است. برخی راجع به خود فیلم و برخی راجع به ویدئوی نقد نظر داده بودند. لذا استفاده از این روش دارای خطای بسیاری بود. سپس تلاش کردم که از روی imdb هر فیلم برچسب آن را تعیین کنم که باز هم دارای خطا بود. زیرا برخی از ویدئوهای نقد صرفا با امتیاز آن فیلم همخوانی نداشت. در انتها مجبور به برچسب‌گذاری دستی شدم که دارای خطای کمینه و دقت بیشینه بود. بدین منظور از سایت www.rottentomatoes.com نیز کمک گرفته شد.

۶ آمار داده

۱.۶ تعداد واحد داده

Count	Label
۳۰	۰
۳۷	۱
۳۳	۲

۲.۶ تعداد جملات

Sentences	Label
۲۳۷۵	۰
۲۴۶۵	۱
۲۱۹۷	۲

۳.۶ تعداد کلمات

Words	Label
۴۹۰۷۱	۰
۵۲۱۷۰	۱
۴۶۰۱۶	۲

۴.۶ تعداد کلمات منحصربه‌فرد

Unique	Label
۱۴۹۵۷	۰
۱۶۹۸۷	۱
۱۴۶۵۷	۲

۵.۶ تعداد کلمات منحصربه‌فرد مشترک و غیرمشترک بین برچسب‌ها

Uncommon	Common
۶۵۶۶	۱۷۸۲

۶.۶ ۱۰ کلمه پرتکرار غیرمشترک هر برچسب

۲ Label	۱ Label	۰ Label
elvis	blade	mcu
jennifer	mario	guardians
season	kong	uncharted
santa	snyder	holland
megan	dracula	rob
fox	godzilla	casino
creed	ted	royale
avatar	vampires	corn
ellie	sarah	diana
verse	snipes	warlock

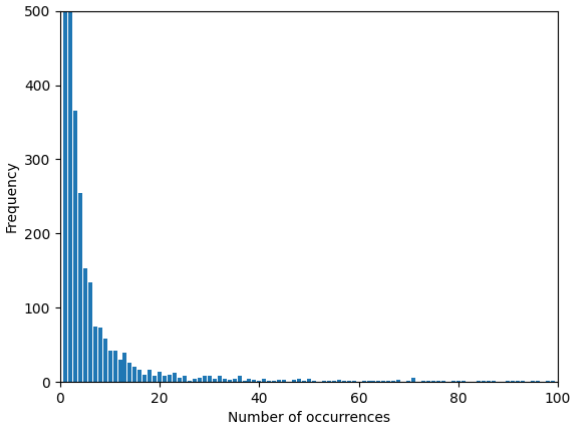
۷.۶ ۱۰ کلمه برتر هر برچسب بر اساس معیار `Relative-Normalized-Frequency`

۲ Label	۱ Label	۰ Label
the	the	the
and	and	and
it	that	it
that	it	to
to	to	that
of	of	of
you	you	you
this	this	this
in	in	in
is	is	like

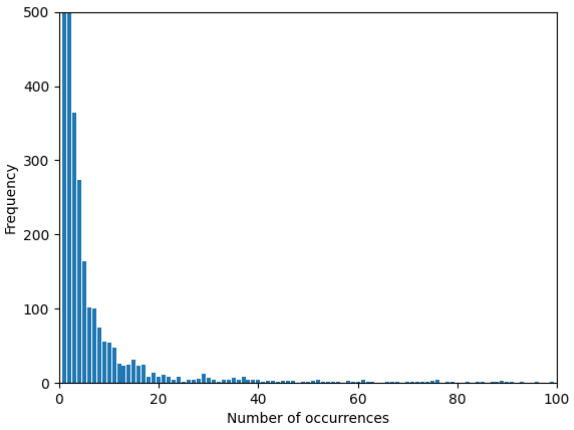
۸.۶ ۱۰ کلمه برتر هر برچسب بر اساس معیار TF-IDF

۲ Label	۱ Label	۰ Label
movie	like	like
like	movie	movie
film	film	just
really	just	film
just	really	really
lot	know	don
don	going	going
people	people	know
going	don	people
know	lot	lot

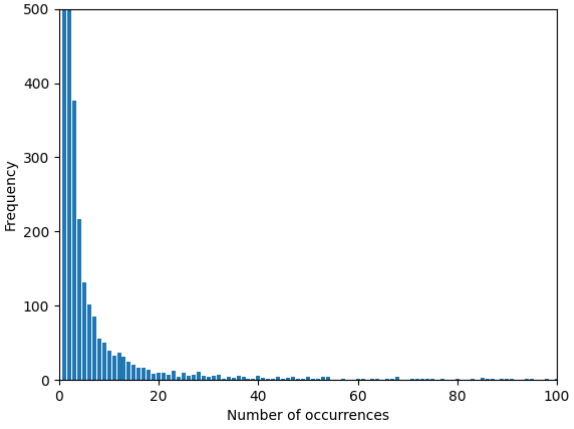
۹.۶ هیستوگرام تعداد تکرار هر کلمه منحصربه‌فرد به ترتیب از فرکانس بالا به پایین



شکل ۱: برچسب ۰



شکل ۲: برچسب ۱



شکل ۳: برچسب ۲