



استاد: دکتر احمد برآنی

دستیار: الهام اسماعیلی

شماره دانشجویی: ۴۰۰۳۶۱۴۰۰۴

تمرین اول: فاز سوم پروژه سخنرانی‌های TED

درس: بازیابی پیشرفته اطلاعات

نام و نام خانوادگی: سید عمید اسدالهی مجد

آدرس گیت: <https://github.com/amidmajd/ted-talk-classification>

یافتن برچسب‌های سخنرانی‌ها با استفاده از روش skip-gram

در فاز دوم پروژه با استفاده از روش skip-gram بردار مربوط به transcript هر سخنرانی ساخته شد. حال برای استفاده از این بردارها تغییری در کد skip-gram.py داده می‌شود تا با استفاده از این بردارها ده کلمه با بیشترین فرکانس به عنوان برچسب‌های هر سخنرانی در الاستیک سرچ ذخیره شوند. با فرض وجود داده‌های مربوط به سخنرانی‌ها در الاستیک سرچ (حاصل فاز اول پروژه)، تغییرات زیر برای اجرای فاز سوم پروژه انجام شدند.

در کد skip-gram.py در خط ۲۴ متغیر جدیدی با نام file_path_per_doc تعریف می‌شود که شامل محل ذخیره متن سخنرانی به ازای هر id سخنرانی است (برای ساخت بردار با استفاده از کتابخانه FastText نیاز است تا فایل متنی داشته باشیم). در ادامه در خط ۳۶، به ازای هر متن سخنرانی بردار مربوط به transcript ساخته می‌شود. سپس با فراخوانی model.words آرایه‌ای شامل کلمات سخنرانی به ترتیب فراوانی هر کلمه از مدل مربوط به بردارها دریافت می‌شود. پس از حذف کلمات stop-words، ده کلمه با بیشترین فرکانس انتخاب می‌شوند. سپس هریک از داده‌های ایندکس شامل متن سخنرانی‌ها با افزودن فیلد جدیدی با نام labels که شامل برچسب‌های بدست آمده است، آپدیت می‌شوند.

در شکل‌های زیر نمونه‌ای از خروجی حاصل از اجرای فایل skip-gram.py را مشاهده می‌نماییم.

4,808 hits		
Sort fields		
Document		
<input checked="" type="checkbox"/>	author: Kristen Bell + Giant Ant date: October 2020 labels: dioxide,carbon,today,advocate,act,emit,act yesterday second best time today every year we dont reduce emissions represents another gigaton forced move natural world transformed lives lost its much cheaper develop new technologies dont emi	
<input checked="" type="checkbox"/>	author: Brent Loken date: October 2020 labels: farmers,agricultural,food,methods,percent,future,g farm? transcript: about years ago humans began farm agricultural revolution turning point our history agricultural lands pieces global puzzle we all facing future how can we feed every member growing po	
<input checked="" type="checkbox"/>	author: Ernestine Leikei Seyidzem idate: October 2020 labels: forest,nature,equality,timeliness,people,honey,bee,generation,community likes: 3,96 nature transcript: my name sevidzem ernestine leikei im climate activist from northwest region came	
<input checked="" type="checkbox"/>	author: Kristen Bell + Giant Ant date: October 2020 labels: carbon,release,trees,natural,atmosphere extra heat also intensifies weather making wet places drier increasing ferocity storm	
<input checked="" type="checkbox"/>	author: Kristen Bell + Giant Ant date: October 2020 labels: degrees,climate,heat,places,extra,stable degrees such a big deal? transcript: why degrees big deal because warm our entire planet up degree extra heat also intensifies weather making wet places drier increasing ferocity storm	
<input checked="" type="checkbox"/>	author: Aparna Nancheria date: October 2020 labels: trash,recycling,stuff,recycled,jot,old,try,fact,h know just from looking me you might guess from smelling me one my favorite things do take out trash modern consumerist carbonpowered culture makes us buy endlessly often reason getting rid people n	
<input checked="" type="checkbox"/>	author: Elf Shafak date: October 2020 labels: time,stories,past,stop,future,tell,tree,art,present,insle sleep smoke picnic secretly kiss our shade pluck our leaves gorge our fruits break our branches carve because think we obstruct view make cradles wine corks chewing gum rustic furniture produce most t	
<input checked="" type="checkbox"/>	author: Olafur Eliasson date: October 2020 labels: future,message,speakr,earth,listening,share,girl, for the environment. Let's listen transcript: my name olafur eliasson im artist i work natural phenomena heard some time now i have fact collaborated young people artists well case make project earth speak	
<input checked="" type="checkbox"/>	author: Prince Royce date: October 2020 labels: music,change,spanish,singing,thank,lets,ends,indi "Carita de Inocente" / "Corazón Sin Cara" / "Darte un Beso" transcript: whats up everyone im prince r	

Table JSON		
Search field names		
Actions	Field	Value
...	_id	625
...	_index	ted-talk-index
...	_score	1
...	author	Kristen Bell + Giant Ant
...	date	October 2020
...	labels	dioxide,carbon,today,advocate,act,emit,cost,generations,action,time
...	likes	4,800
...	link	https://ted.com/talks/kristen_bell_giant_ant_why_act_now
...	title	Why act now?
...	transcript	> why act now best time act yesterday second best time today every year we dont reduce emissions represents another gigatons greenhouse gases future generati ons have scrub out air would cost trillion dollars year almost half entire us economy today price doesnt include cost people forced move natural world tran sformed lives lost its much cheaper develop new technologies dont emit carbon dioxide than capture carbon dioxide weve already emitted every ton carbon dio
...	views	162,000

یک سخنرانی به همراه برچسب‌های جدید اضافه شده به آن

توضیح کد مربوط به دسته‌بندی سخنرانی‌ها

در بدنه‌ی اصلی کد با استفاده از تابع `train_test_split` از کتابخانه `sklearn` با نسبت ۹۰ به ۱۰ درصد داده‌ها به `train` و `test` تقسیم شدند. سپس با استفاده از تابع `save_transcript_with_labels` این داده‌ها با فرمت مربوطه با تحت عناوین `train_data.txt` و `test_data.txt` ذخیره شدند.

سپس با استفاده از تابع `train_supervised` از کتابخانه `FastText` و با ورودی `train_data.txt` به‌عنوان فایل آموزش، نرخ یادگیری^۲ برابر با ۱.۲۵، `ngram`های ۳تایی و تعداد اجرای^۳ برابر با ۵۰۰۰ مدل آموزش دیده و ساخته می‌شود. سپس مدل را با نام `classifier_model.bin` ذخیره می‌نماییم تا در آینده بتوان آن را بارگیری نمود و استفاده کرد.

در ادامه با فراخوانی تابع `test` با ورودی `test_data.txt` مقادیر `precision` و `recall` بدست آمدند و نمایش داده می‌شوند که در شکل زیر قابل مشاهده است. برای نمونه متن زیر به مدل داده‌شده است و مدل برچسب‌هایی مثل `metals`، `fuels`، `fossil`، `electricity`، `bannana` و ... را با دقت‌های نمایش داده‌شده در شکل زیر پیشبینی کرده‌است.

متن ورودی:

We currently have enough fossil fuels to progressively transition off of them, says climate campaigner Tzeporah Berman, but the industry continues to expand oil, gas and coal production and exploration. With searing passion and unflinching nerve, Berman reveals the delusions keeping true progress from being made -- and offers a realistic path forward: the Fossil Fuel Non-Proliferation Treaty.

```
Read 0M words
Number of words: 28134
Number of labels: 1967
Progress: 100.0% words/sec/thread: 362276 lr: 0.000000 avg.loss: 3.085267 ETA: 0h 0m 0s
Example:
((('__label__fuels', '__label__metals', '__label__fossil', '__label__electricity', '__label__rocks', '__label__bananas', '__label__bel_banana', '__label__guatemala'), array([0.02530975, 0.02311448, 0.02227614, 0.02014352, 0.01846728, 0.01781257, 0.01735356, 0.01734817, 0.01705743, 0.01700926])))

Precision: 0.7308670520231214, Recall: 0.10112287661153588
amid@win11: ~/.venv ~/Projects/ted-talk-classification master ?
```

خروجی کار به همراه `precision` و `recall` مدل ساخته‌شده (`precision` برابر با ۷۳ درصد و `recall` برابر با ۱۰ درصد)

^۲ Learning Rate

^۳ Epoch