



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 ایندکس موجود در الستیک سرچ ذخیره شوند. با فرض وجود داده‌های مربوط به سخنرانی‌ها در الستیک سرچ (حاصل فاز اول پروژه)، تغییرات زیر در فاز دوم پروژه 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 carbon we release go? 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: 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	date: October 2020 labels: forest,nature,equality,climate,im climate activist from northwest region came 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 transcript: where does all carbon we release go carbon works natural cycle pi until new plants grow reabsorb carbon over millions years some carbon stored ancient trees sea life be
<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,insit 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 in 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

4,808 hits

Sort fields

Document

☒
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 emit

☒
author
Brent Loken
date
October 2020
labels
farmers,agricultural,food,methods,percent,future,gigafarm?
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

☒
author
Ernestine Leikeki Sevidzem
date
October 2020
labels
forest,nature,equality,kilumijim,trees,people,honey,bee,generation,community
likes
3,900
nature
transcript
my name sevidzem ernestine leikeki im climate activist from northwest region came

☒
author
Kristen Bell + Giant Ant
date
October 2020
labels
carbon,release,trees,natural,atmosphere,carbon we release go?
transcript
where does all carbon we release go carbon works natural cycle ph until new plants grow reabsorb carbon over millions years some carbon stored ancient trees sea life be

☒
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 wetter dry places drier increasing ferocity storm

☒
author
Aparna Nancherla
date
October 2020
labels
trash,recycling,stuff,recycled,lot,old,try,fact, 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

☒
author
Eif Shafat
date
October 2020
labels
time,stories,past,stop,future,tell tree,art,present,ink,sleep,smoke,picnic,secretly,kiss,our shade,pluck,our leaves,gorge,our fruits,break,our branches,care because think we obstruct view make cradles wine corks chewing gum rustic furniture produce most t

☒
author
Olafur Eliasson
date
October 2020
labels
future,message,speak,earth,listening,share,gift,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

☒
author
Prince Royce
date
October 2020
labels
music,change,spanish,singing,thank,lets,ends,ind "Carita de Inocente" / "Corazon Sin Cara" / "Darte un Beso"
transcript
whats up everyone im prince n true climate change defining issue our time histories defined moments when people rise up laws chang

☒
author
Climate Action Tracker
date
October 2020
labels
countries,targets,emissions,degrees,global

Rows per page: 100

Table

JSON

Q

Search field names

Actions

Field

Value

...

id

628

...

_index

ted-talk-index

...

_score

1

...

author

Ernestine Leikeki Sevidzem

...

date

October 2020

...

labels

forest,nature,equality,kilumijim,trees,people,honey,bee,generation,community

...

likes

3,900

...

link

https://ted.com/talks/ernestine_leikeki_sevidzem_a_forest_generation_living_in_harmony_with_nature

...

title

A "forest generation" living in harmony with nature

...

transcript

my name sevidzem ernestine leikeki im climate activist from northwest region cameroon i focus two things caring nature educating next generation do same we cant fight climate emergency we cannot protect regenerate our land thats why my colleagues i cameroon gender environment watch dedicated regenerating hectare kilumijim forest my community people depends nature livelihood we find daily solutions our forest our farmlands our natural resources face lot challenges deforestation overexploitation encroachment poor soil conservation made worse gender inequality cultural barriers little knowledge about good news nature worst then devastating bushfires make sure forest continues exist future we work educate our children protect forest turn support them we educate our children what means love nature treat care i call raising forest generation date my organization has provided environmental education more than people kilumijim forest area percent then women we have childfriendly sessions to ensure work our mission about trees found our forest how take care them

یک سخنرانی به همراه برچسب‌های جدید اضافه‌شده به آن

تابع save_transcript_with_labels

در این تابع یک آرایه شامل اطلاعات خروجی الستیک و به فرمت مخصوص آن و همچنین نام فایل خروجی دریافت می‌شود. برای استفاده از classifier کتابخانه FastText باید یک فایل شامل خطوط متن داشته باشیم. هر خط از این فایل، مربوط به متن یک سخنرانی بوده و قبل از متن سخنرانی در هر خط برچسب‌های مربوط به آن سخنرانی ذخیره می‌شوند. به ابتدای برچسب‌ها باید عبارت __label__ افزوده شود و باهم یک فاصله داشته باشند. در این تابع هر سخنرانی و برچسب‌های مربوط به آن که در مرحله پیشین استخراج شدند از الستیک سرچ استخراج می‌شوند و در یک خط با فرمت ذکرشده ذخیره می‌شوند. دو متن سخنرانی به همراه برچسب‌های مربوطه را می‌توان در شکل زیر مشاهده نمود. خروجی این تابع به شکل زیر خواهد بود. از این تابع برای ذخیره داده‌های test و train به صورت جداگانه استفاده می‌شود.

```
__label_meat __label_stories __label_film __label_muslim __label_halal __label_different __label_america __label_communities __label_community __label_laughter im blogger filmmaker butcher i
ll explain how identities come together started four years ago when friend i opened our first ramadan fast one busiest mosques new york city crowds men beards skullcaps were swarming streets fbi agent
s wet dream laughter being part community we knew how welcoming space years id seen photos space being documented lifeless cold monolith much like stereotypical image painted american muslim experienc
e frustrated myopic view my friend i had crazy idea lets break our fast different mosque different state each night ramadan share those stories blog we called mosques days we drove all states shared s
tories from over vastly different muslim communities ranging from cambodian refugees la projects black swiss living woods south carolina what emerged beautiful complicated portrait america media cover
age forced local journalists revisit muslim communities what really exciting seeing people from around world being inspired take own mosque journey were even two nfl athletes who took sabbatical from
league do so mosques blossoming around world i actually stuck pakistan working film my codirector omar i were breaking point many our friends how position film movie called birds walk about wayward st
reet kids who struggling find some semblance family we focus complexities youth family discord our friends kept nudging us comment drones target killings make film more relevant essentially reducing p
eople who have entrusted us stories sociopolitical symbols course we didnt listen then instead we championed tender gestures love headlong flashes youth ages behind our cinematic immersion only empha
thy emotion thats largely deficient from films come from our region world birds walk played film festivals theaters internationally i finally had my feet planted home new york all extra time still ran
l money my wife tasked me cook more us whenever id go local butcher purchase some halal meat something felt off those dont know halal term used meat raised slaughtered humanely following very strict i
slamic guidelines unfortunately majority halal meat america doesnt rise standard my faith calls more i learned about unethical practices more violated i felt particularly because businesses from my ow
n community were ones taking advantage my orthodoxy so emotions running high absolutely experience butcherery some friends i opened meat store heart east village fashion district laughter we call honest
chops were reclaiming halal sourcing organic humanely raised animals making accessible affordable workingclass families theres really nothing like america unbelievable part actually percent our insto
re customers even muslim many first time interacting islam intimate level so all disparate projects laughter result restlessness visceral response businesses curators who work hard oversimplify my bel
iefs my community only way beat machine play different rules we must fight inventive approach trust access love only we can bring we must unapologetically reclaim our beliefs every moving image every
cut meat because we whitewash our stories sake mass appeal only we fail we trumped those more money more resources tell our stories call creative courage novelty relevance simply because our communiti
es so damn unique so damn beautiful demand us find uncompromising ways acknowledged respected thank you applause
__label_forest __label_peoples __label_respect __label_people __label_indigenous __label_life __label_want __label_cowori __label_oil __label_love my name nemonte nenquimo i am waorani woman
mother leader my people when i young i went city missionary school so i could learn spanish i never had opportunity go university i have pikenis wise elders who have taught me how respect how love
forest from all i have grown leader forest our teacher pikenis wise both women men our scientists our teachers who have taught us value what we have knowledge love we have we say indigenous peoples
value has been lost outsiders forest our home forest gives us life food nourishment water spiritual connection arrival roads arrival colonization arrival evangelical missionaries arrival oil companies
has destroyed our forest i have met other indigenous peoples who live north who were first contacted colonization invasions roads very sad people watching listening i would like say our amazon forest
s continue burn oil continues spill miners continue enter our territory stealing our gold colonization continue invade cut down feed societies from abroad i want let known directly harming us risking
lives our amazonian peoples throughout whole country we indigenous peoples have been fighting our land thousands years because we have lot love respect why important you listen our voices our cries so
destruction our forest stops you stop harming our forest because you listening more than percent earth protected indigenous peoples includes nearly half world's forests i think people from outside wh
o we call cowori people who know less about forest therefore do care about life forest do care about spiritual life do care about life we have lived connected thousands years respecting mother earth's
why i think s important me woman young woman i have learned cowori think technology development better have awareness same time destroying planet continue acting blind we say people who know least ab
out forest blind we indigenous peoples ones whose eyes open we know what happening cowori us stranger who does value who has knowledge about forest what does forest mean us waorani people forest our h
ome our life full life full knowledge right now i can walk see plants around me we can eat leaves we can use heal vines make our baskets carry things wood build our homes good wood leaves cure headach
e i bring cowori here won't see way i n seeding cowori does have knowledge he thinks all gift place full resources he can keep extracting word cowori does have had you capable having same values knowle
dge we have amazonian peoples you can learn you can respect you can become our allies we indigenous peoples do need satellite images because we live forest we know what happening forest amazon burning
oil companies come our territory say develop our country support our communities say harm then come beautiful words say affect environment water surrounding forests lie we see our own eyes alongside
other peoples oil spills how months haven't been able clean them up how have contaminated our fish our rivers our spirits our people our people have been made sick indigenous peoples we say we do want
participate we do want exploitation government must respect our decision withdraw can't circumvent us although communities grassroots say do listen pretend blind deaf come anyway do respect our right
life nature s rights re killing us s why we ask we demand government listen our voices our decisions forest our home period we want them listen we want them longer consume oil longer consume our food
tear down trees thousands hectares because harming killing our spirits our life our forest our pharmacy so we ask where we going give water our children future generations our children your
children well what we do what we love what we respect only our people your lives well lives entire world we live we taking risks we want them listen wake up decide enough enough so longer enter our t
erritory exploit pollute i would ask you all reach understanding indigenous peoples general since we have had deep love true respect thousands years forest i also cannot teach all now i can't teach you
know what respect what connection land looks like connection spiritual all i ask you respect mother earth waiting us respect her mother earth waiting us save her we indigenous peoples expect same
```

بخشی از فایل‌های test و train ساخته‌شده شامل برچسب‌های هر transcript و متن آن

¹ Label

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. مقدار `K` در بخش `test` و `predict` بیانگر تعداد برچسب‌های درخواستی از مدل جهت پیشبینی است.

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`، `fossils`، `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
amido@win11 ~ • *.venv ~/Projects/ted-talk-classification > master
```

خروجی کار به همراه `precision` و `recall` مدل ساخته‌شده (`k=10`) برابر با ۷۳ درصد و `recall` برابر با ۱۰ درصد)

² Learning Rate

³ Epoch