بسم‌الله الرحمن الرحیم

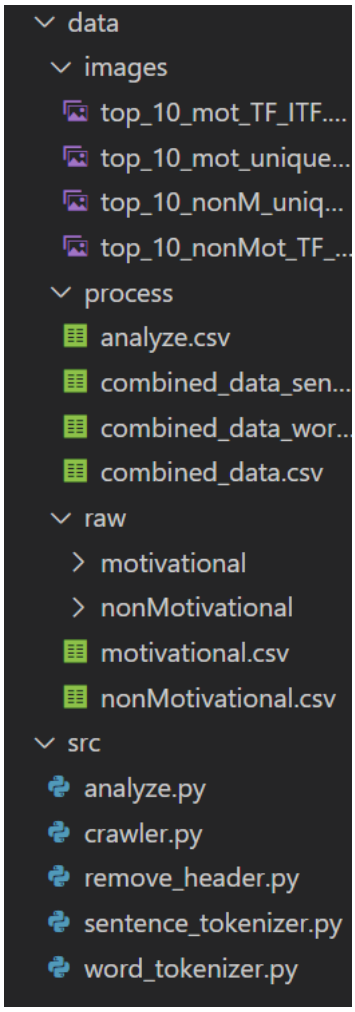
گزارش پروژه پایانی درس NLP

لینک دسترسی به پروژه : NLP Project

مقدمه

هدف اصلی این پروژه دسته‌بندی متن‌های motivational از non motivational می‌باشد.

ساختار پروژه



پروژه شامل دو پوشه اصلی src و data می‌باشد که کدهای در پوشه src و داده‌های و آنالیزهای در پوشه data می‌باشد.

پوشه src شامل فایل‌های زیر می‌باشد:

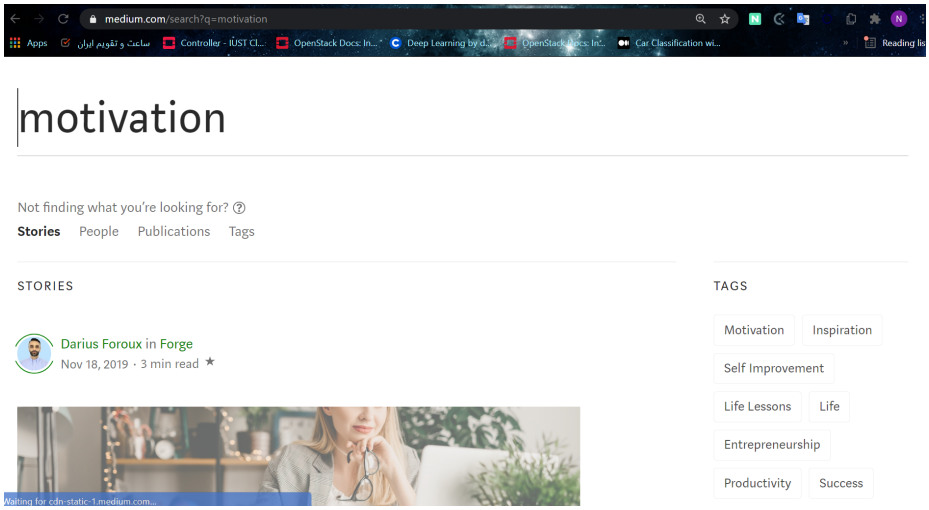
- analyze.py : وظیفه اصلی این فایل تهیه آمار و نمودارهای خواسته شده می‌باشد.
- crawler.py : برای جمع‌آوری داده خام و اولیه.
- remove_header.py : حذف داده اضافه در ابتدا وبلاگ‌ها(نام نویسنده و تاریخ).
- sentence_tokenizer.py : تبدیل متن به جملات تشکیل دهنده.
- word_tokenizer.py : تبدیل متن به کلمات و حذف کاراکترهای اضافی.

پوشه Data شامل :

- پوشه image برای ذخیره نمودارها.
- پوشه process برای ذخیره آمار و داده‌های پیش پردازش شده.
- پوشه raw برای ذخیره داده اولیه می‌باشد.

منبع جمع‌آوری داده

برای جمع‌آوری داده از وب‌سایت medium استفاده شده است به این‌صورت که یک سری کلمات کلیدی انتخاب شده و وبلاگ‌ها و پست‌های مرتبط با آن جمع آوری می‌شوند. به طور مثال برای کلمه کلیدی motivation از صفحه زیر استفاده شده است.



روش جمع‌آوری داده

برای جمع‌آوری داده از دو کتابخانه request و BeautifulSoup استفاده شده است به این‌صورت که ابتدا از صفحه جستجو لینک وبلاگ‌های مرتبط استخراج می‌شود و سپس با استفاده از request اطلاعات هر پست به دست می‌آید. با اجرای فایل crawler.py عملیات جمع‌آوری داده انجام می‌شود و داده به صورت دو فایل csv برای دو کلاس و هم چنین متن وبلاگ ها به صورت txt در دو فولدر جداگانه ذخیره می‌شوند. تمام این داده‌ها در مسیر data/raw ذخیره می‌شود. ساختار هر فایل csv به صورت زیر می‌باشد.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	link	subject	name	count	class	text							
2	https://for	motivation	0.txt	358	0	['Darius Foroux', 'Nov 18, 2019Â·3 min read', 'One of my greatest talents has always been c							
3	https://me	motivation	1.txt	1243	0	['Iâ€™m not highly motivated.', 'I donâ€™t have amazing willpower or self-control.', 'I donâ							
4	https://me	motivation	2.txt	639	0	['The most motivational statement comes down to three words: â€œYouâ€™re gonna die.â							
5	https://me	motivation	3.txt	884	0	['â€œDonâ€™t break the chain.â€', 'These four simple words helped Jerry Seinfeld become							
6	https://be	motivation	4.txt	980	0	['Benjamin Hardy, PhD', 'Jun 27, 2019Â·5 min read', 'When most people think of accountabil							
7	https://me	motivation	5.txt	906	0	['Neuroscience has discovered the remarkably simple source for endless self-motivation.', '							
8	https://me	motivation	6.txt	1070	0	['Daniel Vassallo', 'Feb 10, 2019Â·5 min read', 'Last week I left my cushy job at Amazon afte							
9	https://me	motivation	7.txt	919	0	['Itâ€™s the middle of the day and nothing has gone right.', 'Youâ€™ve had a late start to th							
10	https://me	motivation	8.txt	654	0	['Goals. Are. Awesome.', 'Used right, they are effective tools for improving yourself and ach							

همانطور که در تصویر مشخص می‌باشد این فایل دارای ۶ ستون می‌باشد. ستون اول لینک وبلاگ، ستون دوم کلمه کلیدی جستجو شده، ستون سوم نام فایل txt، ستون چهارم تعداد کلمات متن، ستون پنجم لیبل داده و ستون آخر متن وبلاگ می‌باشد.

پیش‌پردازش

1. در اولین مرحله با استفاده از remove_header.py می‌توان داده‌های اضافی همانند نام نویسنده و تاریخ انتشار را حذف کرد. همچنین این فایل دو فایل csv که برای هر کلاس به صورت جداگانه وجود داشت را ترکیب می‌کند و یک فایل جدید به نام combined_data.csv در مسیر data/process ذخیره می‌کند (تفکیک کلاس‌ها با استفاده از ستون class امکان پذیر می‌باشد).

	A	B	C	D	E	F	G	H	I	J	K	L
46	https://me	Success	44.txt	2515	0	stion has fascinated entire adult life: what causes some people become worl						
47	https://me	Success	45.txt	1050	0	define success lot different ways.', 'some think itâ€™s certain amount money						
48	https://me	Success	46.txt	1788	0	mfortable but powerful truth that took most 20â€™s internalize: thereâ€™s						
49	https://me	Success	47.txt	3038	0	read elon musk: tesla, spacex, and the quest for fantastic future over the sun						
50	https://me	Success	48.txt	478	0	farmer had become old and ready pass his farm down one his two sons. whe						
51	https://me	Success	49.txt	897	0	lly posted quora nicolas cole.', obsessive.', 'pull the covers off leg the same w						
52	https://me	Success	50.txt	782	0	'to listen this answer while you read, click here: quora: invest yourself nicolas						
53	https://ge	metoo	0.txt	1732	1	'sometimes wish could gather all the women iâ€™ve ever known, encountere						
54	https://me	metoo	1.txt	1679	1	'louis has just released statement the sexual misconduct accusations that hav						
55	https://me	metoo	2.txt	783	1	'sunday night, like many other women, created #metoo post. instead telling s						
56	https://me	metoo	3.txt	1514	1	cation leaves memories that last. everybody has story the eclectic curriculum						
57	https://hu	metoo	4.txt	2955	1	'â€œseparate the syllables the word gentleman, and you will see that the firs						
58	https://me	metoo	5.txt	1058	1	tâ€™s not because donâ€™t have them.', 'content warning: descriptions sexu						
59	https://ge	metoo	6.txt	658	1	'mark halperin wants his career back. the political pundit and author, fired 20						
60	https://me	metoo	7.txt	1339	1	ides what men are? decided decree? popular vote? decided you, individual m						
61	https://me	metoo	8.txt	1529	1	'one evening watched ron jeremy scavenge the leftover food few the table d						
62	https://me	metoo	9.txt	1100	1	past few days, iâ€™ve watched female friend after female friend post the sa						
63	https://me	terrorism	10.txt	665	1	'the purpose terrorism spread terror. this should statement the obvious, but						
64	https://ma	terrorism	11.txt	1601	1	went his high school when the year old mr. hacked into the largest internet c						

2. در مرحله بعد، قبل از حذف punctuation ها تعداد جمله‌ها و تفکیک جمله‌ها با استفاده از sentence_tokenizer.py صورت می گیرد. برای tokenize کردن از کتابخانه nltk.tokenize استفاده شده است. پس از تبدیل متن به جملات، تعداد جملات و تعداد کل وبلاگ‌ها در فایل analyze.csv برای اطلاعات آماری در مسیر data/process ذخیره می‌شود.

3. در مرحله بعدی هدف اصلی word tokenize می‌باشد. برای این کار در اولین قدم روی داده به دست آمده از مرحله ۱، پردازش اضافه انجام می‌شود. این پردازش شامل تبدیل متن به حروف کوچک، حذف punctuation، حذف stop word و حذف اعداد و ارقام و هر کاراکتری که جزو حروف الفبا نباشد، می‌باشد. مرحله آخر تبدیل متن به دست آمده از مراحل قبل به کلمات جداگانه می‌باشد. داده به دست آمده به صورت combined_data_word_broken.csv در مسیر data/process ذخیره می‌شود.

   a. برای پردازش داده در مرحله ۳ از کتابخانه nltk.tokenizer و re استفاده شده است.

   b. به طور مثال برای تشخیص stop words از داده nltk به نام nltk.corpus استفاده شده است.

4. در مرحله آخر با اجرای analyze.py اطلاعات و نمودارهای مفیدی در اختیار قرار می‌گیرد

   a. در اولین مرحله تعداد کلمات کل،تعداد کلمات منحصر به فرد، تعداد کلمات منحصر به فرد مشترک و غیر مشترک بین برچسب‌ها در فایل analyze.csv در مسیر data\process ذخیره می‌شود.

	0	1
blog_count		136
sen_count		11061
tot_word_count		713145
tot_unique_word_count		14141
tot_unique_word_count_1		2050
tot_unique_word_count_0		8352
tot_intersection_word_count		3738

   i. Blog_count : تعداد کل وبلاگ‌ها

   ii. Sen_count : تعداد جملات

   iii. Tot_word_count : تعداد کلمات

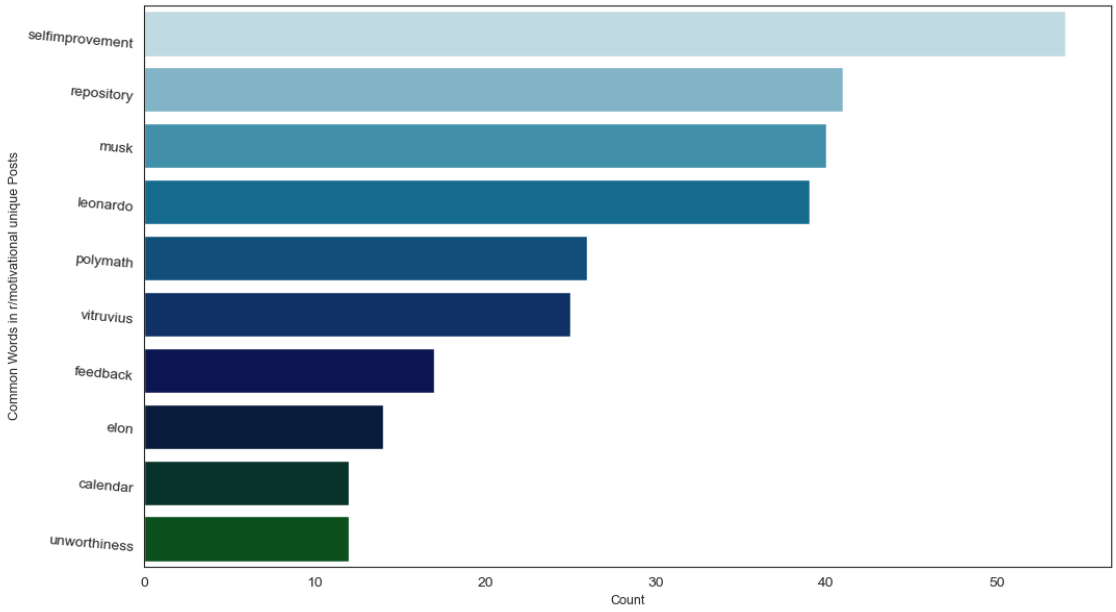
   iv. Tot_unique_word_count : تعداد کلمات منحصر به فرد

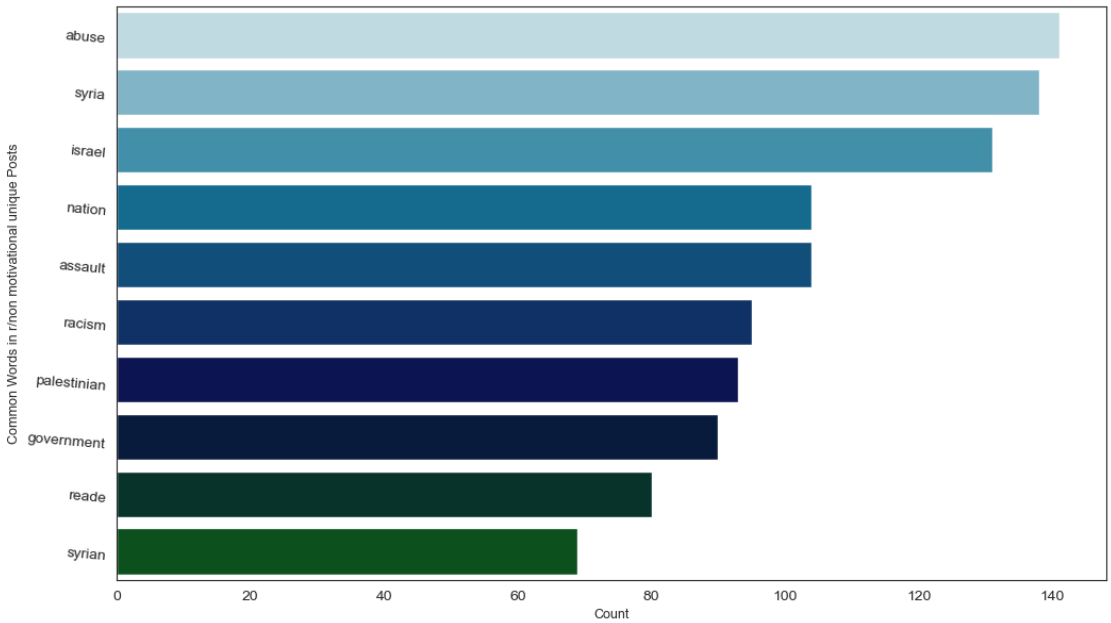
   v. Tot_unique_word_count_0: تعداد کلمات منحصر به فرد برچسب ۰

vi. Tot_unique_word_count_1: تعداد کلمات منحصر به فرد برچسب ۱

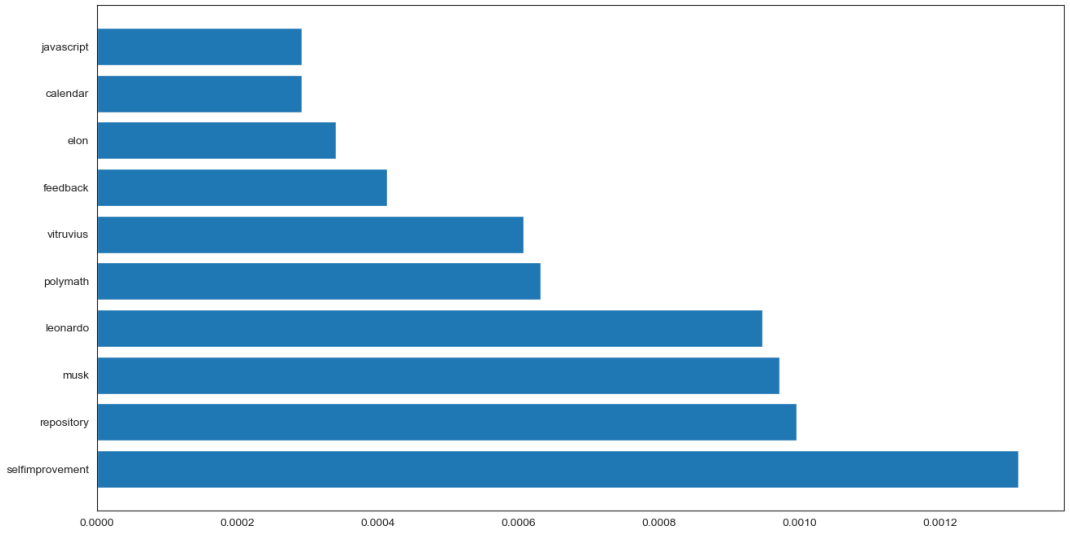
vii. Tot_intersection_word_count : تعداد کلمات منحصر به فرد میان دو برچسب

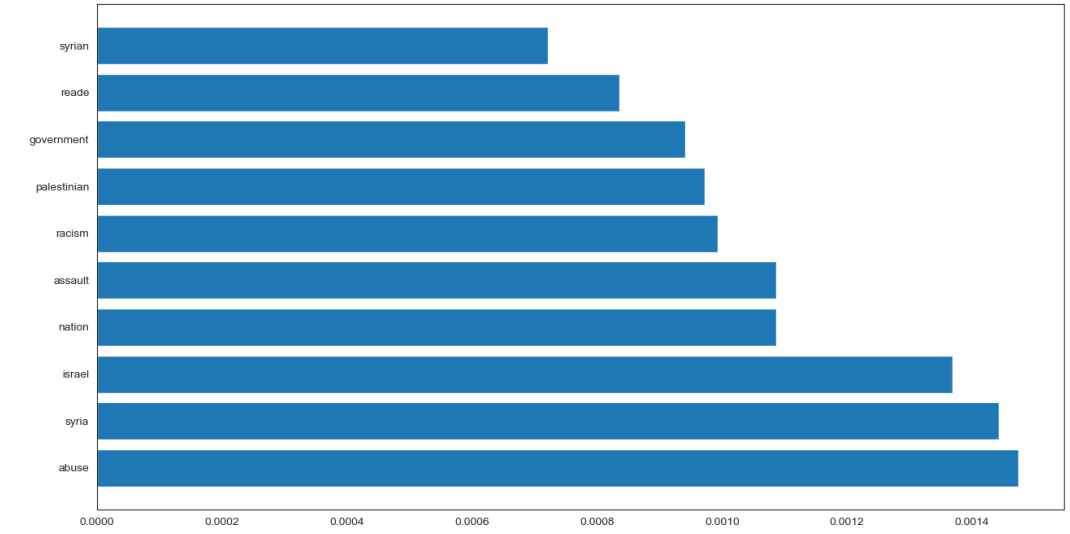
b. سپس نمودار ۱۰ کلمه برتر غیر مشترک هر برچسب رسم می‌شود.



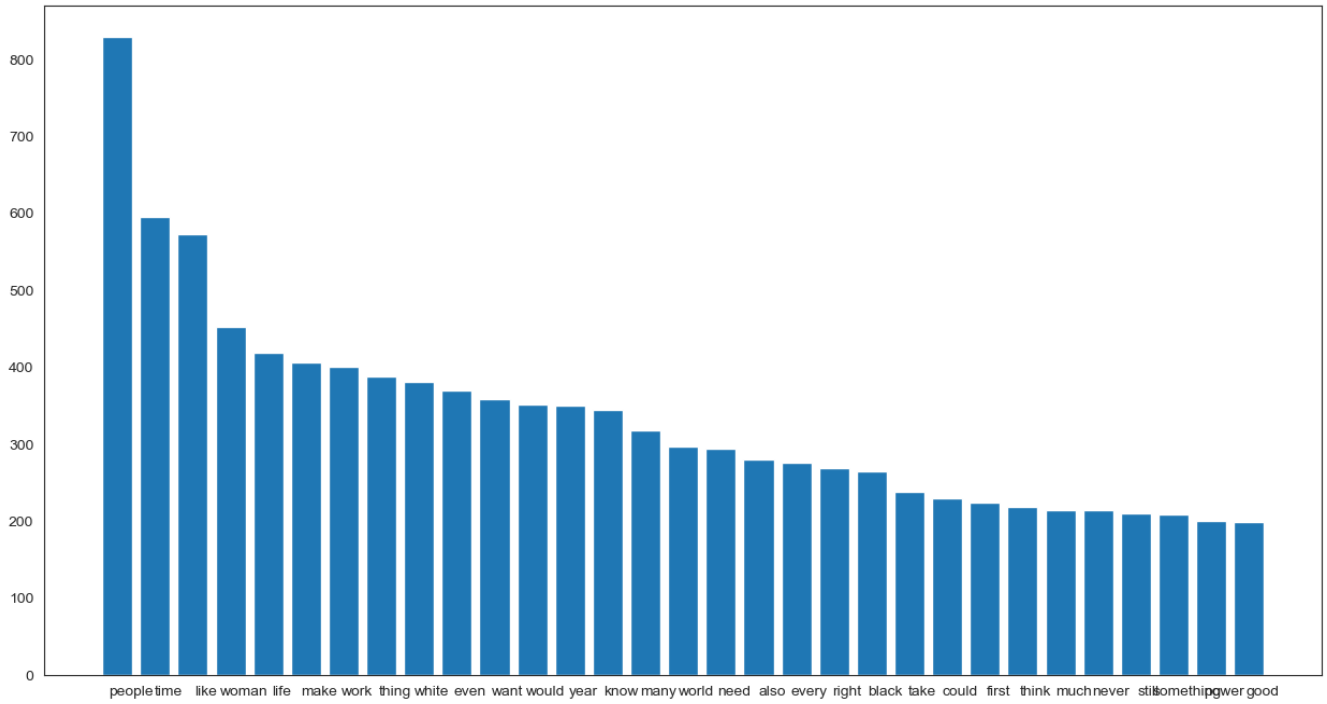


c. مرحله بعد رسم ۱۰ کلمه برتر هر برچسب بر اساس IF-IDF می‌باشد





d. در مرحله آخر نیز هیستوگرام ۳۰ کلمه منحصر به فرد با فرکانس بیشتر رسم شده است.



مراحل اجرای کد

1. python crawl.py
2. python remove_header.py
3. python sentence_tokenizer.py
4. python word_tokenizer.py
5. python analyze.py