

به نام خدا



پروژه دوم درس بازیابی اطلاعات  
روش‌های مبتنی بر مدل زبانی و بردار معنایی عصبی

استاد درس: دکتر ممتازی

نام: زهرا اخلاقی

شماره دانشجویی: ۴۰۱۱۳۱۰۶۴

پاییز ۱۴۰۲

## فهرست مطالب

2	فصل اول - بازیابی با استفاده از مدل
2	۱-۱ مقدمه
2	۲-۱ پیش پردازش
2	۳-۱ پیاده سازی مدل زبانی Unigram
3	۱-۳-۱ نتیجه
3	۴-۱ پیاده سازی مدل زبانی Bigram
4	۱-۴-۱ نتیجه
5	فصل دوم - بازیابی با استفاده بردارهای معنایی عصبی
5	۱-۲ مقدمه
5	۲-۲ پیاده سازی الگوریتم word2vec (میانگین حسابی)
5	۱-۲-۲ نتیجه
5	۳-۲ پیاده سازی الگوریتم word2vec (میانگین وزنی)
5	۱-۳-۲ نتیجه
6	۴-۲ پیاده سازی با استفاده از مدل Bert
6	۱-۴-۲ نتیجه
7	فصل سوم - بازیابی با استفاده از مدل ترجمه
7	۱-۳ مقدمه
7	۲-۳ پیش پردازش
7	۳-۳ پیاده سازی الگوریتم
7	۴-۳ نتیجه
8	۵-۳ واژگان استخراج شده

## فصل اول - بازیابی با استفاده از مدل

### ۱-۱ مقدمه

در این بخش با استفاده از مدل‌های زبانی به استخراج سندهای مشابه می‌پردازیم و دو مدل زبانی Bigram و unigram را پیاده سازی می‌کنیم.

### ۲-۱ پیش پردازش

با توجه به بررسی داده‌های تمرین و و تنوع متون سوالات، تصمیم بر آن شد تا اولین پیش‌پردازش انجام شده تبدیل متن به حرف کوچک و سپس حذف تمامی کاراکتر علائم نگارشی و تگ‌های html از متن می‌باشد و سپس حذف تمامی کاراکترهای غیرحرفی از متن می‌باشد. بنابراین، تمامی علائم نگارشی و اعداد و حروف زبان‌های دیگر از سوالات حذف شده است.

در مرحله بعد پیش‌پردازش داده‌ها عبارات مخفف را به کامل آنها جایگزین کردم و با استفاده از کتابخانه nltk کلمات که شامل stop word ها هستند را از متن حذف کردم گه این کار به باقی‌ماندن کلمات تاثیرگذار کمک می‌کند. این مسئله باعث می‌شود کلماتی در شباهت دو جمله بررسی شوند که به احتمال بالاتری تاثیر کلیدی در معنی و مفهوم جمله داشته باشند.

گام بعدی در پیش‌پردازش متن استخراج ریشه کلمات است که باعث می‌شود شکل‌های مختلف از یک کلمه به یک صورت مشاهده شوند و اگر در دو جمله تکرار شدند، این تکرار در شباهت دو جمله تاثیرگذار باشد. این عمل با استفاده از کتابخانه nltk انجام شده است.

پیش‌پردازش داده‌ها در تابع preprocess پیاده سازی شده و روی ستون document در فایل hw2\_docs.csv و ستون query در فایل hw2\_queries.csv انجام شده و نتیجه به ترتیب در ستون‌های processed\_document و processed\_query ذخیره شده است.

## ۳-۱ پیاده سازی مدل زبانی Unigram

برای پیاده سازی این قسمت ابتدا با فراخوانی تابع `make_tf_cf_vocab_vector` برای هر کلمه در متن `tf`، `cf` محاسبه می‌شود و سپس تعداد کل کلمات متن (`all_corpus_tokens`) و `vocab_vector` که را به عنوان خروجی دارد.

در تابع `compute_jelinek_mercer_smoothed` احتمال هموارسازی یک کلمه را در سند محاسبه می‌کند و در تابع `compute_unigram` به ازای هر جفت کوثری و داکيومنت، به ازای هر کلمه از کوثری احتمال یونیگرم آن کلمه به شرط داکيومنت با روش `jelinek_mercer` محاسبه می‌شود.

### ۱-۳-۱ نتیجه

با توجه به توابع `mean_reciprocal_rank`، `mean_average_precision`، `precision_at_k` تمامی روش‌های پیاده‌سازی شده با استفاده از این توابع ارزیابی خواهند شد. این توابع نشانگر معیارهای ارزیابی `MRR`، `MAP`، `P@5`، `P@10` هستند.

توابع بالا به ازای مقادیر لاندای [0.7, 0.5, 0.3, 0.1, 0.01, 0.001] اجرا شده است و خروجی به صورت زیر است:

`landa=0.001 => MRR:0.56 --- MAP:0.50 --- P@5:0.42 --- P@10:0.36`

`landa=0.01 => MRR:0.55 --- MAP:0.51 --- P@5:0.42 --- P@10:0.37`

`landa=0.1 => MRR:0.63 --- MAP:0.56 --- P@5:0.44 --- P@10:0.40`

`landa=0.3 => MRR:0.70 --- MAP:0.59 --- P@5:0.47 --- P@10:0.42`

`landa=0.5 => MRR:0.69 --- MAP:0.59 --- P@5:0.48 --- P@10:0.42`

`landa=0.7 => MRR:0.69 --- MAP:0.60 --- P@5:0.49 --- P@10:0.43`

با توجه به اینکه هدف این تسک این است که برای هر سوال ۱۰ سند مرتبط را به عنوان خروجی داشته باشد، معیار `P@10` معیار مناسبی است (این معیار ارزیابی می‌کند که چه تعداد از اسناد استخراج شده از مدل واقعاً مرتبط با پرسش بوده‌اند و با انتخاب `K` برابر با ۱۰، می‌توانید دقت مدل در ارائه ۱۰ سند مرتبط را ارزیابی کنید) می‌توان مقدار مناسب لاندای ۰.۷ را برابر با ۰.۷ در نظر گرفت.

## ۴-۱ پیاده سازی مدل زبانی Bigram

برای پیاده سازی این مدل از دو تابع استفاده شده است، در تابع `compute_smoothed_bigram_p`، معادله داده شده پیاده سازی شده است و در تابع `compute_bigram_model` به ازای هر جفت کوئری و داکيومنت، به ازای هر کلمه از کوئری احتمال بایگرم آن کلمه به شرط داکيومنت محاسبه می‌شود.

### ۱-۴-۱ نتیجه

این مدل به ازای مقادیر لاندای  $[0.001, 0.01, 0.1, 0.3, 0.5, 0.7]$  و لاندای با توجه به بخش ۱ برابر ۰.۰ می‌باشد، اجرا شده است و خروجی به صورت زیر است:

landa1=0.7 landa2=0.001 => MRR:0.56 --- MAP:0.49 --- P@5:0.37 --- P@10:0.29

landa1=0.7 landa2=0.01 => MRR:0.58 --- MAP:0.51 --- P@5:0.38 --- P@10:0.31

landa1=0.7 landa2=0.1 => MRR:0.62 --- MAP:0.56 --- P@5:0.42 --- P@10:0.37

landa1=0.7 landa2=0.3 => MRR:0.73 --- MAP:0.64 --- P@5:0.46 --- P@10:0.39

landa1=0.7 landa2=0.5 => MRR:0.66 --- MAP:0.55 --- P@5:0.41 --- P@10:0.31

landa1=0.7 landa2=0.7 => MRR:0.61 --- MAP:0.53 --- P@5:0.39 --- P@10:0.30

همانطور که در بخش قبل توضیح داده شده در این تسک P@10 معیار مناسبی است و مقدار مناسب برای لاندای ۲ برابر با ۰.۳ در نظر گرفته می‌شود.

## فصل دوم - بازیابی با استفاده بردارهای معنایی عصبی

### ۱-۲ مقدمه

در این بخش برای به دست آوردن اسناد مرتبط از بردارهای معنایی عصبی استفاده می‌شود و از بردار معنایی word2vec (در حالت gram-skip) و Bert برای پیاده‌سازی این قسمت استفاده شده است.

### ۲-۲ پیاده سازی الگوریتم word2vec (میانگین حسابی)

برای پیاده سازی این بخش از بردار word2vec در google-news کتابخانه gensim استفاده شده است. تابع calculate\_doc\_vector به ازای کلمات ورودی از مدل استفاده شده بردار کلمات استخراج شده و میانگین آنها را به عنوان خروجی دارد و سپس به ازای هر جفت کوئری و داکيومنت شباهت cosine آنها محاسبه میشود.

#### ۱-۲-۲ نتیجه

نتایج معیارهای ارزیابی P@10 , P@5 , MAP, MRR به صورت زیر است:

P@10:0.29    P@5:0.37    MAP:0.53    MRR:0.60

### ۳-۲ پیاده سازی الگوریتم word2vec (میانگین وزنی)

برای پیاده سازی این بخش از بردار word2vec در google-news کتابخانه gensim استفاده شده است. تابع calculate\_weighted\_document\_vector به ازای کلمات ورودی از مدل استفاده شده بردار کلمات استخراج شده و میانگین وزن دار آن با استفاده از بردار tf-idf محاسبه می‌شود و سپس به ازای هر جفت کوئری و داکيومنت شباهت cosine آنها محاسبه میشود.

## ۱-۳-۲ نتیجه

نتایج معیارهای ارزیابی  $MRR, MAP, P@5, P@10$  به صورت زیر است:

$MRR:0.66 \text{ --- } MAP:0.57 \text{ --- } P@5:0.38 \text{ --- } P@10:0.31$

## ۴-۲ پیاده سازی با استفاده از مدل Bert

برای پیاده سازی این قسمت از مدل آمده bert-base-uncased استفاده شده است، سپس تابع `get_bert_embedding` برای هر جفت کوئری و داکيومنت اجرا می شود و سپس شباهت cosine آنها محاسبه می شود.

## ۱-۴-۲ نتیجه

نتایج معیارهای ارزیابی  $MRR, MAP, P@5, P@10$  به صورت زیر است:

$MRR:0.42 \text{ --- } MAP:0.35 \text{ --- } P@5:0.18 \text{ --- } P@10:0.18$

## فصل سوم - بازیابی با استفاده از مدل ترجمه

### ۱-۳ مقدمه

در این بخش از مدل ترجمه استفاده شده و با استفاده از آن واژگان مشابه هر واژه را یافته و سپس از آن برای یافتن داکيومنت‌های مرتبط به هر کوعری استفاده می‌شود.

### ۲-۳ پیش پردازش

پیش پردازش در این الگوریتم مشابه روش بیان شده در فصل اول، پیاده‌سازی شده است.

### ۳-۳ پیاده سازی الگوریتم

برای پیاده سازی این الگوریتم ابتدا بردار word2vec از کلمات موجود در داکيومنت ساخته و در تابع find\_similar\_words شباهت کسینوسی نرمال شده برای هر واژه محاسبه می‌شود (در سوال خواسته شده واژگانی که شباهت بزرگتر از ۰.۷ دارند در نظر گرفته شود ولی بالاترین میزان شباهت میان واژگان ۰.۰۰۰۱۹ بود) در این پیاده سازی حد آستانه میزان شباهت ۰.۰۰۰۱ در نظر گرفته شده و شبیه‌ترین ۱۰ واژه استخراج شده‌اند. در تابع conditional\_prob\_word\_given\_vector مقدار احتمال  $P(w|v)$  محاسبه شده و در تابع conditional\_prob\_vector\_given\_document مقدار احتمال  $P(v|D)$  محاسبه شده است. تابع calculate\_document\_vector به ازای کلمات داکيومنت از مدل استفاده شده بردار کلمات استخراج شده و میانگین آنها را به عنوان خروجی دارد و سپس به ازای هر جفت کوئری و داکيومنت شباهت cosine آنها محاسبه می‌شود.

### ۴-۳ نتیجه

نتایج معیارهای ارزیابی MRR, MAP, P@5, P@10 به صورت زیر است:

MRR:0.06 --- MAP:0.05 --- P@5:0.03 --- P@10:0.03



origin:['origin', 'major', 'pathogen', 'similar', 'first', 'new', 'releas', 'occur', 'agent', 'like']

covid:['covid', 'background', 'outbreak', 'pneumonia', 'ard', 'diseas', 'decemb', 'pandem', 'emerg', 'middl']

coronavirus:['coronavirus', 'caus', 'novel', 'respiratori', 'syndrom', 'acut', 'sar', 'east', 'distress', 'cov']

respond:['respond', 'take', 'make', 'la', 'general', 'learn', 'process', 'addit', 'set', 'propos']

chang:['chang', 'de', 'region', 'possibl', 'en', 'la', 'et', 'generat', 'contribut', 'larg']

weather:['weather', 'de', 'possibl', 'cluster', 'scale', 'quarantin', 'enhanc', 'whether', 'impact', 'process']

sar:['sar', 'cov', 'novel', 'coronavirus', 'caus', 'mer', 'respiratori', 'syndrom', 'east', 'acut']

cov:['cov', 'sar', 'novel', 'coronavirus', 'caus', 'mer', 'syndrom', 'respiratori', 'acut', 'east']

infect:['infect', 'middl', 'background', 'outbreak', 'tract', 'emerg', 'ard', 'human', 'decemb', 'worldwid']

peopl:['peopl', 'one', 'month', 'remain', 'appear', 'initi', 'individu', 'common', 'suscept', 'less']

develop:['develop', 'potenti', 'understand', 'pathogen', 'induc', 'anti', 'agent', 'evolut', 'similar', 'major']

immun:['immun', 'host', 'cell', 'vaccin', 'antibodi', 'entri', 'receptor', 'rbd', 'neutral', 'ace']

cross:['cross', 'determin', 'could', 'moreov', 'therefor', 'avail', 'via', 'e', 'contain', 'may']

protect:['protect', 'therapeut', 'drug', 'also', 'establish', 'activ', 'surfac', 'anim',  
'contain', 'function']

possibl:['possibl', 'chang', 'de', 'formul', 'various', 'allow', 'contribut', 'region',  
'generat', 'process']

caus:['caus', 'coronavirus', 'novel', 'respiratori', 'syndrom', 'acut', 'east', 'distress',  
'diseas', 'sever']

death:['death', 'case', 'mortal', 'risk', 'comorbid', 'rate', 'fatal', 'among', 'hypertens',  
'increas']

drug:['drug', 'therapi', 'therapeut', 'avail', 'protect', 'antivir', 'establish', 'contain',  
'direct', 'therefor']

activ:['activ', 'surfac', 'e', 'type', 'also', 'mutat', 'interact', 'gene', 'antigen', 'could']

anim:['anim', 'demonstr', 'also', 'contain', 'repres', 'may', 'determin', 'establish',  
'furthermor', 'function']

studi:['studi', 'data', 'databas', 'conduct', 'valid', 'find', 'singl', 'simul', 'practic',  
'analysi']

type:['type', 'e', 'demonstr', 'activ', 'may', 'reveal', 'could', 'thus', 'surfac', 'persist']

rapid:['rapid', 'sinc', 'ncov', 'worldwid', 'becom', 'china', 'wuhan', 'abstract', 'global',  
'name']

test:['test', 'negat', 'sampl', 'swab', 'rt', 'n', 'analys', 'collect', 'select', 'within']

serolog:['serolog', 'activ', 'surfac', 'acid', 'mutat', 'gene', 'avail', 'e', 'could', 'elisa']

detect:['detect', 'rna', 'viral', 'sequenc', 'strain', 'anti', 'envelop', 'similar', 'inactiv',  
'potenti']

antibodi:['antibodi', 'bind', 'rbd', 'neutral', 'tissu', 'recombin', 'replic', 'receptor',  
'sequenc', 'domain']

lack:['lack', 'also', 'anim', 'establish', 'contain', 'demonstr', 'repres', 'protect',  
'therapeut', 'context']

avail:['avail', 'contain', 'dynam', 'cross', 'determin', 'therefor', 'drug', 'direct', 'better', 'establish']

led:['led', 'object', 'continu', 'challeng', 'consequ', 'discuss', 'across', 'appear', 'contain', 'focus']

underreport:['underreport', 'liver', 'cvd', 'prognosi', 'citocina', 'accur', 'reason', 'toxic', 'constitut', 'respiratorio']

true:['true', 'better', 'start', 'near', 'date', 'variat', 'larg', 'most', 'chang', 'therefor']

incid:['incid', 'condit', 'follow', 'without', 'status', 'under', 'lower', 'proport', 'histori', 'increas']

affect:['affect', 'infecti', 'recent', 'known', 'failur', 'influenza', 'relat', 'like', 'serious', 'call']

canada:['canada', 'set', 'la', 'resourc', 'small', 'consider', 'general', 'combin', 'usa', 'u']

social:['social', 'measur', 'intervent', 'mask', 'counti', 'model', 'use', 'level', 'distan', 'adher']

distan:['distan', 'measur', 'social', 'mask', 'intervent', 'counti', 'lima', 'model', 'random', 'use']

impact:['impact', 'quarantin', 'govern', 'nation', 'epidemiolog', 'general', 'set', 'help', 'usa', 'long']

slow:['slow', 'epidem', 'strain', 'throughout', 'known', 'recent', 'call', 'current', 'pathogen', 'infecti']

spread:['spread', 'emerg', 'decemb', 'pandem', 'background', 'outbreak', 'infect', 'human', 'mer', 'global']

guidelin:['guidelin', 'practic', 'dataset', 'hand', 'recommend', 'lockdown', 'examin', 'need', 'optim', 'work']

triag:['triag', 'histori', 'accord', 'follow', 'status', 'abnorm', 'primari', 'decreas', 'condit', 'includ']

patient:['patient', 'critic', 'mortal', 'admit', 'diabet', 'among', 'hypertens', 'case', 'comorbid', 'death']

best:['best', 'practic', 'dataset', 'conduct', 'guidelin', 'work', 'paramet', 'il', 'school', 'databas']

practic:['practic', 'guidelin', 'best', 'assess', 'dataset', 'work', 'polici', 'hand', 'lockdown', 'conduct']

hospit:['hospit', 'admiss', 'icu', 'outcom', 'ventil', 'signific', 'male', 'higher', 'intens', 'comorbid']

home:['home', 'popul', 'exposur', 'find', 'work', 'april', 'probabl', 'paramet', 'dataset', 'conduct']

maintain:['maintain', 'hand', 'addit', 'dataset', 'peptid', 'distribut', 'select', 'concentr', 'work', 'search']

quarantin:['quarantin', 'impact', 'set', 'process', 'help', 'de', 'monitor', 'general', 'epidemiolog', 'physician']

transmiss:['transmiss', 'import', 'bat', 'understand', 'evolut', 'potenti', 'rout', 'pathogen', 'similar', 'origin']

rout:['rout', 'bat', 'role', 'import', 'evolut', 'agent', 'potenti', 'understand', 'describ', 'presenc']

evid:['evid', 'limit', 'approach', 'knowledg', 'propos', 'formul', 'allow', 'sourc', 'possibl', 'efficaci']

relat:['relat', 'complic', 'like', 'progress', 'unknown', 'affect', 'shock', 'asymptomat', 'neonat', 'releas']

super:['super', 'cluster', 'us', 'american', 'interest', 'question', 'larg', 'end', 'itali', 'creat']

spreader:['spreader', 'trend', 'would', 'especi', 'advers', 'stress', 'even', 'de', 'larg', 'us']

long:['long', 'de', 'possibl', 'epidemiolog', 'set', 'general', 'la', 'process', 'creat', 'help']

live:['live', 'itali', 'may', 'contain', 'investig', 'focus', 'us', 'appear', 'question', 'challeng']

outsid:['outsid', 'trend', 'end', 'would', 'near', 'januari', 'preval', 'even', 'relev', 'passeng']

bodi:['bodi', 'may', 'demonstr', 'identifi', 'contain', 'establish', 'howev', 'context', 'discuss', 'furthermor']

remain:['remain', 'one', 'individu', 'less', 'suscept', 'especi', 'prolong', 'relev', 'month', 'us']

stabl:['stabl', 'limit', 'allow', 'properti', 'site', 'ii', 'predict', 'effici', 'final', 'g']

surfac:['surfac', 'activ', 'e', 'also', 'type', 'mutat', 'gene', 'function', 'interact', 'determin']

clinic:['clinic', 'outcom', 'chest', 'enrol', 'without', 'laboratori', 'improv', 'record', 'requir', 'cohort']

trial:['trial', 'random', 'particip', 'randomis', 'demograph', 'placebo', 'regress', 'hydroxychloroquin', 'arm', 'data']

mask:['mask', 'counti', 'model', 'level', 'use', 'regress', 'surgic', 'random', 'intervent', 'measur']

prevent:['prevent', 'concern', 'therapeut', 'import', 'well', 'role', 'implic', 'agent', 'protect', 'futur']

hand:['hand', 'examin', 'work', 'access', 'assess', 'dataset', 'guidelin', 'order', 'consid', 'procedur']

sanit:['sanit', 'simul', 'practic', 'conduct', 'polici', 'base', 'databas', 'dataset', 'analysi', 'work']

need:['need', 'center', 'lockdown', 'guidelin', 'medic', 'practic', 'vulner', 'facil', 'dataset', 'home']

destroy:['destroy', 'warrant', 'citocina', 'safe', 'busi', 'fund', 'uncertainti', 'regardless', 'option', 'alcohol']

take:['take', 'set', 'make', 'la', 'general', 'violenc', 'propos', 'process', 'de', 'region']  
 angiotensin:['angiotensin', 'enzym', 'inhibitor', 'gene', 'convert', 'ii', 'molecular',  
 'specif', 'express', 'suggest']  
 convert:['convert', 'angiotensin', 'enzym', 'inhibitor', 'ace', 'express', 'ii', 'gene',  
 'aldosteron', 'interact']  
 enzym:['enzym', 'angiotensin', 'inhibitor', 'gene', 'specif', 'convert', 'express', 'ii', 'b',  
 'fusion']  
 inhibitor:['inhibitor', 'ii', 'product', 'predict', 'strong', 'allow', 'de', 'addit', 'chang',  
 'exist']  
 ace:['ace', 'receptor', 'target', 'bind', 'express', 'neutral', 'mediat', 'domain', 'inhibit',  
 'tissu']  
 increas:['increas', 'incid', 'reduc', 'under', 'rate', 'proport', 'factor', 'lower', 'excess',  
 'condit']  
 risk:['risk', 'hypertens', 'increas', 'factor', 'rate', 'comorbid', 'morbidity', 'among',  
 'corticosteroid', 'death']  
 mortal:['mortal', 'death', 'risk', 'patient', 'comorbid', 'case', 'hypertens', 'admiss',  
 'among', 'admit']  
 rate:['rate', 'increas', 'proport', 'fatal', 'lower', 'conclus', 'incid', 'higher', 'reduc',  
 'under']  
 overal:['overal', 'daili', 'estim', 'statist', 'least', 'ratio', 'valu', 'respect', 'dose', 'popul']  
 specif:['specif', 'assay', 'suggest', 'molecular', 'b', 'effici', 'adapt', 'site', 'epitop',  
 'structur']  
 popul:['popul', 'temperatur', 'home', 'find', 'exposur', 'area', 'april', 'work', 'highest',  
 'perform']  
 cardiac:['cardiac', 'treat', 'biomark', 'children', 'adult', 'pulmonari', 'chronic',  
 'manifest', 'diagnosi', 'event']

complic:['complic', 'relat', 'progress', 'shock', 'failur', 'injuri', 'upper', 'like', 'neonat',  
 'recov']  
 like:['like', 'releas', 'similar', 'major', 'origin', 'pathogen', 'relat', 'shock', 'call',  
 'known']  
 kind:['kind', 'inactiv', 'dysfunct', 'bcov', 'aspect', 'unlik', 'lung', 'cytokin', 'similar',  
 'corona']  
 associ:['associ', 'cardiovascular', 'complic', 'symptom', 'critic', 'relat', 'moder',  
 'progress', 'children', 'adult']  
 hypertens:['hypertens', 'among', 'comorbid', 'factor', 'onset', 'die', 'risk', 'condit',  
 'morbid', 'corticosteroid']  
 diabet:['diabet', 'children', 'adult', 'admit', 'among', 'liver', 'cardiac', 'prognosi',  
 'chronic', 'treat']  
 biomark:['biomark', 'cardiac', 'pulmonari', 'treat', 'diagnosi', 'chronic', 'children',  
 'event', 'adult', 'present']  
 predict:['predict', 'product', 'g', 'region', 'addit', 'chang', 'vitamin', 'exist', 'ii',  
 'distribut']  
 sever:['sever', 'distress', 'diseas', 'east', 'covid', 'tract', 'multiorgan', 'middl',  
 'background', 'caus']  
 cours:['cours', 'present', 'occur', 'previous', 'lead', 'aim', 'progress', 'pulmonari',  
 'event', 'first']  
 ncov:['ncov', 'rapid', 'sinc', 'abstract', 'becom', 'name', 'worldwid', 'trigger', 'china',  
 'beta']  
 initi:['initi', 'common', 'appear', 'one', 'month', 'still', 'peopl', 'secondari', 'main',  
 'due']  
 symptom:['symptom', 'associ', 'injuri', 'complic', 'upper', 'mild', 'failur', 'relat',  
 'critic', 'kidney']

known:['known', 'recent', 'infecti', 'affect', 'strain', 'like', 'call', 'cytokin', 'influenza', 'epidem']

asymptomat:['asymptomat', 'occur', 'like', 'similar', 'origin', 'first', 'relat', 'recov', 'major', 'progress']

valu:['valu', 'ratio', 'find', 'show', 'result', 'paramet', 'lymphocyt', 'differ', 'respect', 'least']

hydroxychloroquin:['hydroxychloroquin', 'placebo', 'receiv', 'mg', 'level', 'mean', 'match', 'particip', 'per', 'time']

treat:['treat', 'cardiac', 'diagnosi', 'pulmonari', 'chronic', 'biomark', 'initi', 'common', 'due', 'children']

protein:['protein', 'spike', 'viral', 'rna', 'genom', 'envelop', 'immun', 'host', 'rbd', 'cell']

human:['human', 'infect', 'spread', 'virus', 'enter', 'emerg', 'middl', 'mer', 'decemb', 'background']

interact:['interact', 'antigen', 'structur', 'mutat', 'express', 'gene', 'activ', 'fusion', 'adapt', 'surfac']

indic:['indic', 'observ', 'distribut', 'small', 'addit', 'la', 'five', 'within', 'surviv', 'violenc']

potenti:['potenti', 'understand', 'develop', 'induc', 'anti', 'pathogen', 'evolut', 'import', 'inactiv', 'similar']

target:['target', 'express', 'antigen', 'ace', 'inhibit', 'neutral', 'interact', 'structur', 'immunogen', 'mediat']

approv:['approv', 'set', 'efficaci', 'evalu', 'propos', 'help', 'process', 'support', 'action', 'general']

repurpos:['repurpos', 'various', 'knowledg', 'evid', 'process', 'approv', 'action', 'possibl', 'publish', 'therefor']

base:['base', 'provid', 'strategi', 'simul', 'assess', 'practic', 'polici', 'studi', 'hand', 'design']



inform:['inform', 'research', 'communiti', 'propos', 'help', 'local', 'set', 'limit', 'hand', 'approach']

remdesivir:['remdesivir', 'improv', 'period', 'curv', 'observ', 'low', 'area', 'benefit', 'standard', 'find']

effect:['effect', 'approach', 'strategi', 'provid', 'effici', 'research', 'limit', 'communiti', 'inform', 'evid']

treatment:['treatment', 'manag', 'primari', 'characterist', 'disord', 'featur', 'less', 'includ', 'defin', 'relev']

differ:['differ', 'observ', 'show', 'l', 'low', 'area', 'perform', 'blood', 'temperatur', 'score']

season:['season', 'de', 'chang', 'region', 'et', 'en', 'la', 'cluster', 'process', 'indic']

flu:['flu', 'great', 'success', 'domest', 'accur', 'character', 'discuss', 'still', 'one', 'stress']

subtyp:['subtyp', 'along', 'direct', 'effort', 'key', 'accur', 'knowledg', 'person', 'e', 'activ']

vaccin:['vaccin', 'target', 'antibodi', 'host', 'inhibit', 'immun', 'neutral', 'replic', 'bind', 'rbd']

candid:['candid', 'drug', 'key', 'avail', 'molecular', 'dynam', 'could', 'antivir', 'therefor', 'adapt']

longer:['longer', 'valu', 'ratio', 'result', 'statist', 'daili', 'compar', 'stay', 'exposur', 'respect']

term:['term', 'oxygen', 'medic', 'healthcar', 'resourc', 'optim', 'combin', 'examin', 'observ', 'point']

recov:['recov', 'occur', 'previous', 'present', 'first', 'progress', 'shock', 'asymptomat', 'dysfunct', 'diagnos']

new:['new', 'first', 'occur', 'origin', 'agent', 'import', 'major', 'aim', 'provinc', 'ongo']

public:['public', 'govern', 'nation', 'impact', 'challeng', 'mitig', 'research', 'intern', 'plan', 'offici']

dataset:['dataset', 'work', 'area', 'perform', 'hand', 'blood', 'assess', 'guidelin',  
'examin', 'procedur']

structur:['structur', 'gene', 'antigen', 'fusion', 'interact', 'mutat', 'express', 'adapt',  
'suggest', 'molecular']

spike:['spike', 'viral', 'protein', 'genom', 'rna', 'respons', 'envelop', 'anti', 'host',  
'pathogenesi']

result:['result', 'valu', 'compar', 'ct', 'conclus', 'statist', 'interv', 'receiv', 'cohort',  
'find']

phylogenet:['phylogenet', 'mutat', 'interact', 'gene', 'structur', 'avail', 'surfac', 'e',  
'depend', 'activ']

analysi:['analysi', 'show', 'perform', 'sensit', 'find', 'obtain', 'conduct', 'dataset',  
'lymphocyt', 'temperatur']

genom:['genom', 'viral', 'rna', 'strain', 'spike', 'detect', 'sequenc', 'anti', 'envelop',  
'respons']

sequenc:['sequenc', 'recombin', 'rbd', 'envelop', 'antibodi', 'host', 'bind', 'evolut',  
'replic', 'anti']

mechan:['mechan', 'de', 'multipl', 'especi', 'en', 'might', 'interest', 'involv', 'relev',  
'larg']

inflammatori:['inflammatori', 'function', 'also', 'form', 'demonstr', 'alter', 'type',  
'anim', 'repres', 'may']

respons:['respons', 'viral', 'host', 'strain', 'genom', 'anti', 'pathogenesi', 'induc',  
'spike', 'potenti']

pathogenesi:['pathogenesi', 'respons', 'genom', 'viral', 'strain', 'known', 'spike',  
'virus', 'anti', 'corona']

case:['case', 'death', 'among', 'confirm', 'admit', 'onset', 'fatal', 'comorbid',  
'hypertens', 'rate']

cytokin:['cytokin', 'like', 'releas', 'known', 'storm', 'affect', 'major', 'pathogen', 'shock', 'call']

storm:['storm', 'releas', 'like', 'shock', 'cytokin', 'origin', 'major', 'dysfunct', 'pathogen', 'lung']

syndrom:['syndrom', 'acut', 'respiratori', 'coronavirus', 'caus', 'novel', 'sar', 'cov', 'distress', 'sever']

observ:['observ', 'low', 'indic', 'l', 'oxygen', 'area', 'period', 'differ', 'non', 'score']

mutat:['mutat', 'gene', 'interact', 'structur', 'activ', 'surfac', 'antigen', 'fusion', 'phylogenet', 'suggest']

often:['often', 'especi', 'larg', 'relev', 'pattern', 'suscept', 'given', 'remain', 'natur', 'relationship']

occur:['occur', 'first', 'previous', 'present', 'new', 'origin', 'progress', 'diagnos', 'recov', 'aim']

among:['among', 'hypertens', 'factor', 'comorbid', 'older', 'fever', 'discharg', 'diabet', 'incid', 'liver']

african:['african', 'high', 'thicken', 'less', 'physician', 'poor', 'de', 'decreas', 'preval', 'consist']

american:['american', 'larg', 'de', 'even', 'scale', 'region', 'might', 'cluster', 'chang', 'various']

rest:['rest', 'incid', 'stratifi', 'accord', 'near', 'four', 'lower', 'follow', 'proport', 'period']

u:['u', 'la', 'scenario', 'region', 'chang', 'take', 'distribut', 'small', 'combin', 'predict']

vitamin:['vitamin', 'predict', 'addit', 'consid', 'g', 'within', 'exist', 'propos', 'combin', 'five']

pandem:['pandem', 'emerg', 'decemb', 'outbreak', 'background', 'spread', 'global', 'diseas', 'worldwid', 'infect']

violenc:['violenc', 'set', 'la', 'take', 'usa', 'indic', 'distribut', 'region', 'general', 'help']

societi:['societi', 'set', 'la', 'chang', 'sourc', 'region', 'take', 'process', 'help', 'de']

includ:['includ', 'less', 'primari', 'especi', 'disord', 'cough', 'abnorm', 'histori', 'follow',  
 'preval']  
 violent:['violent', 'assur', 'exercis', 'assumpt', 'outpati', 'synthesi', 'timelin', 'confin',  
 'drop', 'assign']  
 crime:['crime', 'polici', 'need', 'paramet', 'project', 'databas', 'conduct', 'school',  
 'practic', 'home']  
 much:['much', 'articl', 'involv', 'could', 'moreov', 'pattern', 'allow', 'cluster', 'multipl',  
 'larg']  
 mental:['mental', 'need', 'servic', 'practic', 'center', 'lockdown', 'polici', 'plan', 'facil',  
 'medic']  
 health:['health', 'public', 'plan', 'servic', 'offici', 'research', 'healthcar', 'impact',  
 'mental', 'communiti']  
 dexamethason:['dexamethason', 'indic', 'hour', 'benefit', 'small', 'materi', 'observ',  
 'la', 'subject', 'surviv']  
 outcom:['outcom', 'comorbid', 'admiss', 'icu', 'corticosteroid', 'hypertens', 'enrol',  
 'onset', 'clinic', 'die']  
 children:['children', 'adult', 'cardiac', 'manifest', 'chronic', 'treat', 'biomark',  
 'diagnosi', 'pulmonari', 'diabet']  
 contract:['contract', 'disord', 'factor', 'diabet', 'treatment', 'among', 'liver', 'preval',  
 'less', 'obes']  
 benefit:['benefit', 'l', 'low', 'combin', 'oxygen', 'observ', 'blood', 'non', 'addit', 'optim']  
 open:['open', 'analysi', 'exposur', 'conduct', 'sensit', 'singl', 'simul', 'paramet',  
 'school', 'il']  
 school:['school', 'polici', 'practic', 'exposur', 'face', 'paramet', 'conduct', 'find', 'home',  
 'best']  
 midst:['midst', 'induc', 'host', 'sl', 'strain', 'dead', 'rbd', 'oc', 'recombin', 'asthma']

individu:['individu', 'remain', 'one', 'month', 'suscept', 'six', 'secondari', 'may', 'earli',  
'appear']  
show:['show', 'perform', 'analysi', 'differ', 'find', 'temperatur', 'lymphocyt', 'area',  
'sensit', 'l']  
suffici:['suffici', 'standard', 'non', 'oper', 'surviv', 'hour', 'supplement', 'il', 'home',  
'resourc']  
level:['level', 'counti', 'time', 'respect', 'regress', 'hydroxychloroquin', 'dose', 'mean',  
'qualiti', 'averag']  
cell:['cell', 'immun', 'ace', 'target', 'antibodi', 'neutral', 'host', 'receptor', 'mediat',  
'rbd']  
mediat:['mediat', 'bind', 'receptor', 'inhibit', 'ace', 'also', 'neutral', 'replic', 'pro',  
'tissu']  
mrna:['mrna', 'avail', 'contain', 'activ', 'anim', 'surfac', 'e', 'interact', 'determin',  
'demonstr']  
virus:['virus', 'ncov', 'beta', 'trigger', 'rapid', 'abstract', 'pathogenesi', 'sinc', 'enter',  
'genom']