Құрастырылған корпусты визуализациялау әдістерімен танысу

Болатбек М.А.

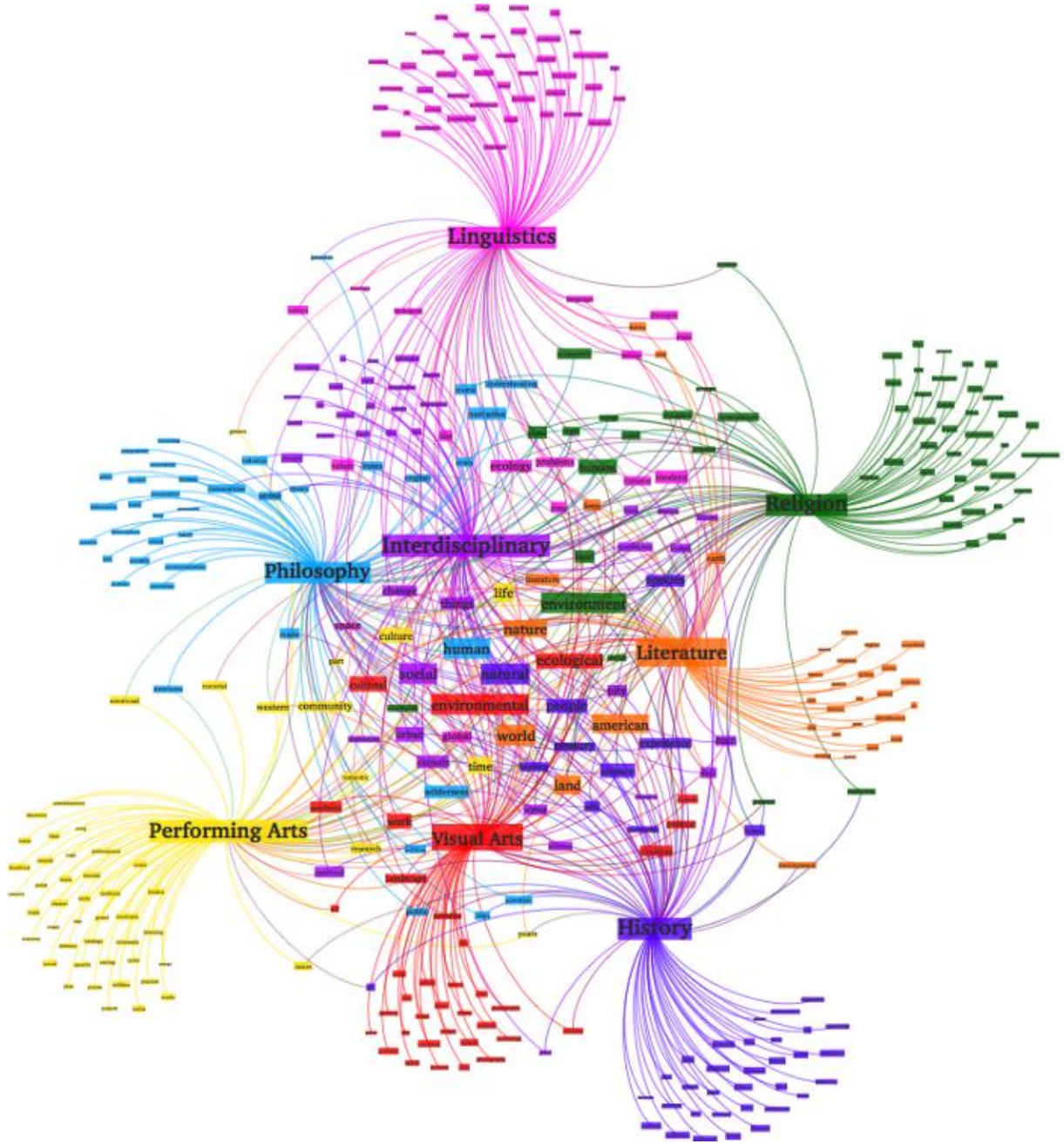
Неліктен мәтіндік деректерді визуализациялау маңызды?

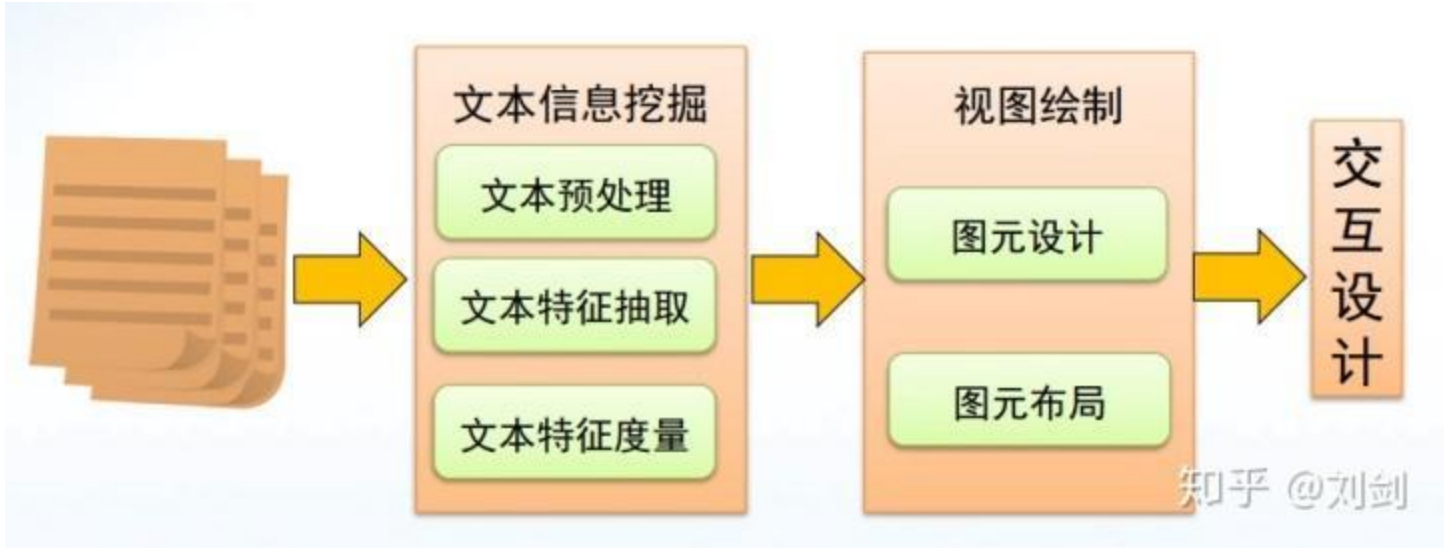
Мәтін-ақпарат беру үшін жиі қолданылатын медиа. Көлемді мәтіндердің пайда болуымен ақпараттың шамадан тыс жүктелуі және деректердің артық болуы сияқты проблемалар барған сайын жиі байқалуда. Бұрынғы уақытта мәтіннің үлкен абзацтары болған кезде, оларды адамдар шыдамдылықпен және мұқият оқыды. Ақпарат алудың неғұрлым тиімді әдісі қажет. Көрнекі тұрғыдан алғанда, мәтінді визуализациялау-бұл ең маңызды кезең.

Осылайша, мәтінді визуализациялау технологиясы мәтіндегі мазмұнды тез алу үшін адамдар визуалды қабылдауға тән параллельді өңдеу мүмкіндіктерін қолдана алатындай етіп, мәтінде айту қиын немесе қиын болатын мазмұн мен ережелерді көрнекі таңбалар түрінде білдіреді.

Мәтінді визуализациялау процессі

Мәтінді визуализациялау табиғи тілді өңдеуге негізделген, сондықтан мәтінді талдаудың кең таралған әдістері-сөз теру моделі, аталған нысандарды тану, кілт сөздерді шығару, тақырыпты талдау, тоналдылықты талдау және т. б. Мәтінді талдау процесі негізінен сөздерді сегментациялау, шығару және қалыпқа келтіру сияқты әрекеттерді қолдана отырып, сөздік деңгей мазмұнын шығаратын функцияларды шығаруды қамтиды, сонымен қатар векторлық кеңістік моделін құру үшін функцияларды қолданады және оны төмен өлшемді кеңістікте көрсету немесе тақырыптарды пайдалану үшін өлшемді азайтады. Модель сипаттамаларды өңдейді және соңында өңделген деректерді визуалды бейнелеу үшін икемді және тиімді түрде ұсынады. Келесі суретте мәтінді визуализациялаудың негізгі схемасы көрсетілген:


Linguistics
Religion
Interdisciplinary
Philosophy
Literature
Performing Arts
Visual Arts
History
ecology
life
environment
nature
culture
human
ecological
social
natural
environmental
people
community
american
world
time
land
wilderness
work


文本信息挖掘
文本预处理
文本特征抽取
文本特征度量
视图绘制
图元设计
图元布局
交互设计
知乎 @刘剑

Мәтіндік визуализация түрлері, гистограммалар, дөңгелек диаграммалар, сызықтық графиктер және т. б. сияқты әдеттегі диаграмма түрлерінен басқа, мәтін өрісінде жиі қолданылады:

(1) мәтіндік мазмұнға негізделген визуализация.

Мәтіндік мазмұнға негізделген визуализацияны зерттеу сөздердің жиілігіне негізделген визуализацияны және сөздік қордың таралуына негізделген визуализацияны қамтиды. Әдетте сөз бұлттары, тарату карталары және құжат карталары қолданылады.

(2) мәтіндік қатынастарға негізделген Визуализация.

Мәтіндік қатынастарға негізделген Визуализация адамдарға мәтіндік мазмұнды түсінуге және заңдарды ашуға көмектесетін мәтіндердің ішкі және сыртқы қатынастарын зерттейді. Көрнекілендірудің жиі қолданылатын формаларына ағаш тәрізді диаграммалар, түйіндерге қатысты желілік диаграммалар, күшке бағытталған диаграммалар, құрама диаграммалар және сөз ағашы жатады.

(3) көп деңгейлі ақпаратқа негізделген визуализация

Көп деңгейлі ақпаратқа негізделген визуализация негізінен пайдаланушыларға мәтіндік деректерді тереңірек түсінуге және оларға тән заңдарды анықтауға көмектесу үшін ақпараттың бірнеше аспектілерін қалай біріктіруге болатындығын зерттейді. Олардың ішінде уақыт пен географиялық координаттар туралы ақпаратты қамтитын мәтіндік визуализация соңғы жылдары көбірек назар аударуда. Әдетте географиялық жылу картасы, Темеривер, Спарклоудтар, Мәтін ағымы және матрицалық көріністерге негізделген эмоционалды талдау визуализациясы қолданылады.

Сөздер бұлты

Нақты процесс-сөздерді сегменттеу, тоқтату сөздерін алып тастау және сөздердің жиілігін санау, содан кейін Wordcloud сөздерінің бұлтын салу. Келесі екі әдісті қарастырайық.

Геотермалды карта

Географиялық жылу картасы пайдаланушының географиялық орнын арнайы бөлінген түрде көрсетеді.Жылу картасының көмегімен сіз пайдаланушының жалпы жағдайы мен қалауын интуитивті түрде байқай аласыз.

Негізгі қадамдар : * Folium Орнатыңыз;* Baidu көмегімен географиялық терминдерді ендік пен бойлыққа айналдырыңыз.
Сөздерді сегментациялау арқылы қала атауын алғаннан кейін, географиялық зат есімдер Baidu көмегімен ендік пен бойлыққа айналады.
Алдымен кілтті тіркеңіз, қажет мекен-жайдың ендік және бойлық координаттарын алу және оларды JSON құрылымының деректеріне айналдыру үшін Baidu веб-қызметінің API геокодтау интерфейсін қолданыңыз (жеке интерфейс, Baidu қоңыраулардың санын күніне 6000 рет шектейді), содан кейін ендік пен бойлықты анықтау функциясын анықтаңыз. :

```
 # Преобразование широты и долготы    def getlnglat(address):
 url = 'http://api.map.baidu.com/geocoder/v2/'        output = 'json'
ak = 'sqGDDvCDEZPSz24bt4b0BpKLnMk1dv6d'
add = quote (address) # Поскольку переменная города в этой статье китайская,
во избежание искажения символов используйте цитату для кодирования
 uri = url + '?' + 'address=' + add  + '&output=' + output + '&ak=' + ak
 req = urlopen(uri)
res = req.read (). decode () # Декодировать другие закодированные строки в unicode
   temp = json.loads (res) # Анализировать данные json        return temp
```


乌鲁木齐
阿拉木图
西宁
兰州
银川
呼和浩特
北京
石家庄
太原
济南
西安
郑州
沈阳
哈尔滨
上海
重庆
长沙
南昌
贵阳
昆明
福州
台北
南宁
澳门
香港
海口
拉萨
巴雷利
新德里
阿杰梅尔
坎普尔
安拉阿巴德
吉瓦哈蒂
达卡
兰契
加尔各答
印多尔
曼德勒
内比都
河内
万象
平壤
首尔
大田
光州
长崎
鹿儿岛
黄海
东海
渤海
知乎 @刘剑
地方方 & OpenStreetMap & HERE

Соңында, визуалды оқытуға арналған үш веб-сайтты ұсынамыз, олардың біріншісі - Baidu.Canvas негізіндегі Echarts деректерді визуализациялаудың классикалық парадигмаларына сүйене отырып, жаңадан бастаушыларға жарамды, егер деректер ұйымдастырылған болса, сіз өте әдемі диаграммаларды оңай ала аласыз;

Екінші ұсыныс d3.оңай орнату үшін SVG-ге негізделген js, D3 V4 Canvas + SVG, D3 қолдайды. js Echarts-ке қарағанда біршама күрделі, белгілі бір даму тәжірибесі бар адамдар үшін қолайлы; Үшіншісі

three.js-бұл WebGL негізіндегі үш өлшемді графикалық платформа, ол пайдаланушыларға JavaScript көмегімен WebGL жобаларын жасауға мүмкіндік береді.

TF-IDF-ті қалай есептеуге болады?

Интуитивті түрде, TF-IDF мәселесін Google-де сұрау жасаған кез-келген адам шешеді: сұрауды қай сөздер ең айқын сипаттайтынын біліп, олардан "сұрау" керек. Жақсы іздеу жүйесі, егер ол тиісті нәтиже бергісі келсе, мәтіндерде қандай сөздер ең мағыналы және сұрауларға сәйкес келетінін есептейді. Қандай сөздер басқаларға қарағанда маңызды және бұл нені білдіретіні туралы сұрақтарға сандық жауаптарды қалай алуға болады.

TF-IDF-екі фактордың көбейтіндісі. Жинақтың ішіндегі құжаттың тақырыбын анықтайтын сөз неғұрлым маңызды болса, соғұрлым көп жұмыс жасалады. Бірінші фактор — құжаттағы сөздің (терминнің) жиілігі. Бұл TF-term frequency. Екінші фактор — IDF-inverse document frequency-коллекциядағы барлық құжаттардың саны, қажетті терминмен құжаттар санына бөлінеді. IDF бөлшек болғандықтан, оның мәні неғұрлым аз болса, соғұрлым үлкен болады. Сондықтан, егер сөз корпуста "шоғырланған", аз мөлшерде мәтіндер болса, сөздің IDF өседі. Осылайша, IDF корпустың барлық мәтіндерінде жиі кездесетін сөздердің маңыздылығын азайтады. TF өте кішкентай болуы мүмкін, ал IDF өте үлкен болуы мүмкін және үлкен диапазондағы сандармен жұмыс істеу үшін TF — ке көбейту алдында IDF-тен логарифм алынады, бірақ қорытынды формулалардың әртүрлі нұсқалары бар. Біз мынаны ұсынамыз: TF-IDF = TF * log (IDF)

Мысал

- Бізде екі корпус болсын. Біріншісі-жазғы демалыс туралы, екіншісі-жемістер туралы. Жазғы демалыс туралы мәтіндердің корпусында алма сирек кездеседі (250-ден 5 мәтінде), негізінен бадминтон және серфинг туралы. Бірақ коттедж мен егін туралы мәтін бар, онда" алма " 20 рет кездеседі. "Егінжай" мәтіні үшін" алма " құжаттың тақырыбын сипаттайды.
- Жеміс корпусында алма жиі кездеседі: 250-ден 200 мәтінде. Бірақ біз бұл корпустан кездейсоқ алған Мәтін — цитрустық жемістер туралы, ал алма апельсиндермен салыстырылған кезде он рет кездеседі. Бұл мәтінде" алма " құжаттың тақырыбын сипаттамайды.
- Егінжай туралы мәтіндегі терминнің жиілігі-20/100 = 0.2 (әр мәтінде жүз сөз бар деп санаймыз)
- Цитрустық жемістер туралы мәтіндегі терминнің жиілігі-10/100 = 0.1
- Жазғы демалыс туралы корпуста алма-250-ден 5 мәтінде. Құжаттардың кері жиілігі (ADF) = 250/5 = 50. log(LDF) = log(50) = 1.69.
- Жеміс корпусында алма-250-ден 200 мәтінде. IDF = 250/200 = 1,25. log(IDF) = log(1,25) = 0,096
- Алма TF-IDF = 0,2 * 1,69 = 0,338 мәтіндерінде "топтастырылған" жағдайда
- Алма TF құжаттары бойынша "жағылған" корпуста-IDF = 0,1 * 0,096 = 0,0096
- Жоғары TF-IDF сөздің тақырыпты түсіну үшін маңызды екенін көрсетті.

Бұл әдіс қайда қолданылады?

TF-IDF-тің ең көп қолданылуы бұрын іздеуде болған: пайдаланушының сұранысына сәйкес қандай мақалалар бар екенін түсіну маңызды. Басқа идеялармен бірге TF-IDF мәтінді сандық векторларға аударуға мүмкіндік береді, олардың арасындағы қашықтықты өлшеуге болады (және мәтін біздің координаттар жүйесінде қайда орналасқанын түсінуге болады) — осылайша мәтіндерді мазмұны бойынша жіктеу және кластерлеу мәселесін шешуге болады.

```
de chocolate 0.0682629363327
leve uma companhia 0.0682629363327
e café mas 0.0682629363327
café 0.0434983219577
leve uma 0.0682629363327
chocolate e café 0.0682629363327
sozinho 0.0682629363327
não pense 0.0682629363327
em comer sozinho 0.0682629363327
torta especial 0.0682629363327
especial de 0.0682629363327
chocolate e 0.0682629363327
e café 0.0682629363327
torta 0.0682629363327
mas não pense 0.0682629363327
especial 0.0573044198973
pense em comer 0.0682629363327
```

$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

Қысқаша айтқанда, TF-IDF - бұл term frequency-inverse document frequency немесе, егер үлкен және күшті болса, терминдердің жиілігі құжаттардың кері жиілігі болып табылады.

Бұл кез-келген құжат үшін барлық басқа құжаттарға қатысты терминнің маңыздылығын бағалаудың қарапайым және ыңғайлы тәсілі. Принцип мынада: егер сөз кез-келген құжатта жиі кездесетін болса, ал басқа құжаттарда сирек кездесетін болса — бұл сөз сол құжат үшін өте маңызды.

Артықшылықтары

1. Барлық құжаттар үшін маңызды емес сөздер, мысалы, предлогтар немесе араласулар — TF-IDF салмағы өте төмен болады (өйткені олар барлық құжаттарда жиі кездеседі), ал маңыздылары жоғары болады.

2. Оны санау оңай

Бұл метриканы не үшін қолдануға болады?

1. Құжаттардағы маңызды сөздер мен сөздерді анықтау

2. Бұл формуланың кейбір кеңейтімдерін тональды классификаторлардың жұмысын жақсарту үшін қолдануға болады

Term Frequency

TF–бұл терминнің құжатта қаншалықты жиі кездесетінін өлшейтін терминнің жиілігі.
Ұзын құжаттарда бұл термин қысқа құжаттарға қарағанда көп мөлшерде кездеседі деп болжау қисынды, сондықтан абсолютті сандар мұнда оралмайды.
Сондықтан салыстырмалы түрде қолданыңыз
–олар мәтінде қажетті терминнің қанша рет кездескенін мәтіндегі сөздердің жалпы санына бөледі.
Яғни: A = терминінің TF (a термині мәтінде кездескендердің Саны / мәтіндегі барлық сөздердің саны)

```
import collectionsdef compute_tf(text):
#На вход берем текст в виде списка (list) слов
 #Считаем частотность всех терминов во входном массиве с помощью
#метода Counter библиотеки collections
 tf_text = collections.Counter(text)    for i in tf_text:
   #для каждого слова в tf_text считаем TF путём деления
   #встречаемости слова на общее количество слов в тексте
      tf_text[i] = tf_text[i]/float(len(text))    #возвращаем объект типа Counter
с TF всех слов текста
return tf_texttext = ['hasta', 'la', 'vista', 'baby', 'la', 'vista', 'la']print
compute_tf(text)
Out: Counter({'la': 0.4285714285714285, 'vista': 0.2857142857142857,
'hasta': 0.1428571428571428, 'baby': 0.1428571428571428})
```

Inverse Document Frequency

IDF-бұл құжаттардың кері жиілігі. Ол терминнің маңыздылығын тікелей өлшейді. Яғни, біз TF деп санаған кезде, барлық терминдер бір-біріне тең деп саналады. Бірақ, мысалы, предлогтар өте жиі кездесетінін бәрі біледі, бірақ олар мәтіннің мағынасына іс жүзінде әсер етпейді. Жауап қарапайым-IDF есептеу. Ол А термині кездесетін құжаттар санына бөлінген құжаттардың жалпы санынан логарифм ретінде қарастырылады.

Яғни:

А = логарифм терминінің IDF (а термині кездесетін құжаттардың жалпы саны / құжаттар саны)

Айтпақшы, логарифмді кез — келген қабылдауға болады-өйткені TF-IDF-салыстырмалы өлшем; яғни терминдердің салмағы кейбір бірліктерде көрсетілмейді.

```
import math

def compute_idf(word, corpus):
#на вход берется слово, для которого считаем IDF
#и корпус документов в виде списка списков слов
    #количество документов, где встречается искомый термин
    #считается как генератор списков
    return math.log10(len(corpus)/sum([1.0 for i in corpus if word in i]))

texts = [['pasta', 'la', 'vista', 'baby', 'la', 'vista'],
['hasta', 'siempre', 'comandante', 'baby', 'la', 'siempre'],
['siempre', 'comandante', 'baby', 'la', 'siempre']]
print compute_idf('pasta', texts)
Out: 0.47712125472
```


килобайт
мегабайт
терабайт
единицы
измерения
мегабайт
бит
байт
килобайт
единицы
терабайт
бит
бит
байт
мегабайт
бит
килобайт
терабайт
бит
измерения
бит
терабайт
байт
байт
мегабайт
байт
байт
измерения
единицы
терабайт

Сөздер немесе тегтер бұлтының көмегімен сіз Интернетте бірнеше рет кездескен шығарсыз. Әдетте, мұндай бұлт сайтта бүйірлік бағанға "ілінеді", ал жүгіргіні оған апарған кезде сөздер бізге көрінбейтін осьтің айналасында үлкейе бастайды немесе" айналады".

Сөздер немесе тегтер бұлты (ағылш. tag cloud, word cloud, wordle) - бұл белгілер, төте жолдар, кілт сөздер және т.б. деп аталатын санаттар немесе тегтер тізімінің көрнекі көрінісі. Әр сөз еренсілтеме болғандықтан, ол сайтта неғұрлым жиі кездессе, бұлтта соғұрлым үлкен өлшем болады. Бұлттар бар, онда сөздің маңыздылығы түспен ерекшеленеді. Осылайша, сөздер бұлты әрдайым қозғалмалы және сайтта жаңа материалдар жарияланған кезде мөлшері мен түсі өзгереді.

Пайдаланудың қарапайымдылығы мен сыртқы тартымдылығының арқасында сөз бұлттары блогтарда және тақырыптық сайттарда жиі қолданылады.

Бастапқыда бұлттар тек еренсілтемелерді ұйымдастырудың құралы ретінде әрекет етті. Бірте-бірте олардың функциялары өзгерді, ал бүгінде оларды қолдану аясы әлдеқайда кең.

Біріншіден, бұлт пайда болатын сөздер енді тек еренсілтемелер бола алмайды. Сіз кез-келген мәтінді алып, арнайы бағдарламалық құралдардың көмегімен оны сөз бұлтына айналдыра аласыз. Екіншіден, бұлттар сайт құрылысынан басқа көптеген басқа салаларда, соның ішінде білім беру саласында да қолданылды.

Аталған әдісті қолданудың әр түрлі жолдары бар:

* сабақтағы дидактикалық материал ретінде (электронды түрде немесе принтерде басылған);

* өзіңіз немесе қандай да бір адам туралы ақпаратты ұсыну үшін (портфолиода, тәжірибені жалпылау кезінде, презентацияларда, Сайтта және / немесе блогта);

* жарқын, есте қаларлық өнімдерді жасау үшін (ашық хаттар, ақпараттық-жарнамалық буклеттер, бюллетеньдер, презентациялар) ;

* маңызды күндерге, оқиғаларға, негізгі сәттерге назар аудару үшін (тәжірибені жалпылау кезінде, аналитикалық материалдарда, презентацияларда және т. б.);

* бір нәрсені бағалау критерийлерін қалай визуализациялау керек;

* сауалнама немесе талқылау нәтижелерін ұсыну үшін;

* сізге кәсіби тәжірибе мен шығармашылық қиял туралы айтатын көптеген басқа нұсқалар.


радо
слышишь
весна
запах
Талых
всё
Каком
крыш
теплый
запах
Каком
сада
Все
забыты
запах
весна
Как
стоишь
сада
свежый
теплый
радо
ей
И

Сөз бұлтын құру қызметтері

Wordcloud.pro сөздерден интерактивті бұлт жасауға мүмкіндік береді. Сөз бұлтының көмегімен сіз өзіңіздің сайтыңыздың іздеу бетін ұйымдастыра аласыз немесе графикалық файл түрінде сақтай отырып, әрі қарайғы жұмыста" бұлтты " қолдана аласыз. Кез-келген мәтін немесе жай сөздер жиынтығы тегтер бұлтына оңай айналады. Қызметтің сөзсіз артықшылығы-оның орыс тіліндегі интерфейсі.

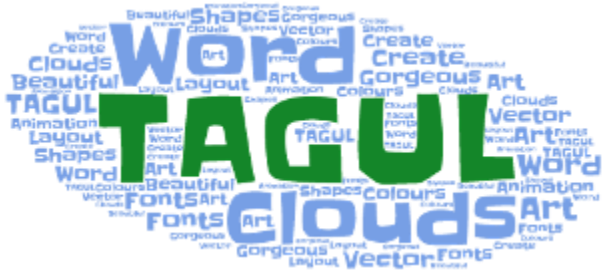
Тегтер бұлты екі жолмен қалыптасады:

1) Сіз берген сөздерден немесе мәтіннен,

2) сайттағы ұсынылған сөздер жиынтығынан. Бұл сервис сөз немесе кескін түрінде тегтер бұлтын жасауға мүмкіндік береді. Мүмкіндіктер тек сіздің қиялыңызбен шектеледі. Жұмысты бастау үшін тіркеу қажет емес.

Tagul.com. алдыңғы қызмет сияқты, пайдаланушы енгізген мәтіннен немесе көрсетілген мекен-жайдағы веб-беттен сөздер бұлтын жасауға мүмкіндік береді.

Бұлт әртүрлі пішіндер мен түстерге ие болуы мүмкін. Бұлттың әр сөзі Google-де іздеуге арналған еренсілтеме болып табылады.



Жұмысты бастау үшін Сіз қызметке тіркелуіңіз немесе әлеуметтік желілердегі аккаунтты пайдаланып кіруіңіз керек. Қызмет кириллицаны қолдайды.

Жасалған бұлтты сілтеме арқылы бөлісуге болады, сонымен қатар веб-сайттар мен блогтар беттеріне бұлтты енгізу кодын алуға болады.

Қызмет бұлт сөздерін тек нүктелік кескін (PNG кеңейтімі) ретінде ғана емес, сонымен қатар векторлық (SVG) ретінде де сақтауға мүмкіндік береді. Сондай-ақ, бұлтты принтерге басып шығаруға болады.

Wordclouds.com сіз берген мәтіннен тегін онлайн сөз немесе тег генераторы. Бұлтта бастапқы мәтінде жиі кездесетін үлкен сөздер ерекшеленеді. Бұлтты әртүрлі қаріптер, орналасулар, фон және түс схемалары арқылы реттеуге болады.

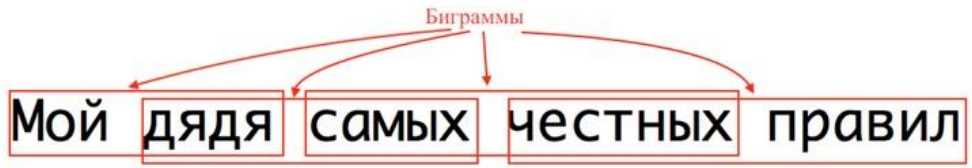
Білім беру саласында бұл қызметті сауалнамаларды, ойындарды, іс-шараларды қорытындылау үшін пайдалануға болады. Сіз мәтінді өңдеп, белгілі бір сөздің жиілігін анықтай аласыз.

Жасалған суреттерді галереяда сақтауға, Компьютердің қатты дискісіне сурет ретінде сақтауға, принтерге басып шығаруға болады. Сондай-ақ, веб-сайтыңызға, блогыңызға бұлт сілтемесін қосуға немесе достарыңызбен бөлісуге болады.

N-грамм

N-грамм-бұл мәтінде қатар жүретін n элементтерінің (дыбыстар, слогдар, сөздер немесе таңбалар) тізбегі. Іс жүзінде олар көбінесе бірқатар сөздерді білдіреді (сирек — таңбалар). Екі элементтің тізбегі биграмма деп аталады, үш элементтен-триграмма.


Униграммы
Мой
дядя
самых
честных
правил


Биграммы
Мой
дядя
самых
честных
правил


Триграммы
Мой
дядя
самых
честных
правил

Корпус мәтіндерінде N-граммның пайда болу жиілігін есептей отырып, сіз корпус немесе жеке мәтіндер туралы бір нәрсе біле аласыз. Мысалы, егер біз бір автордың мәтіндерін көптеген басқа мәтіндермен салыстыратын болсақ, онда автордың жиі қолданатын кейбір ерекшеліктерін, бұрылыстары мен идиомаларын анықтауға болады. Оның үстіне, ол тіпті оны бейсаналық түрде жасай алады.

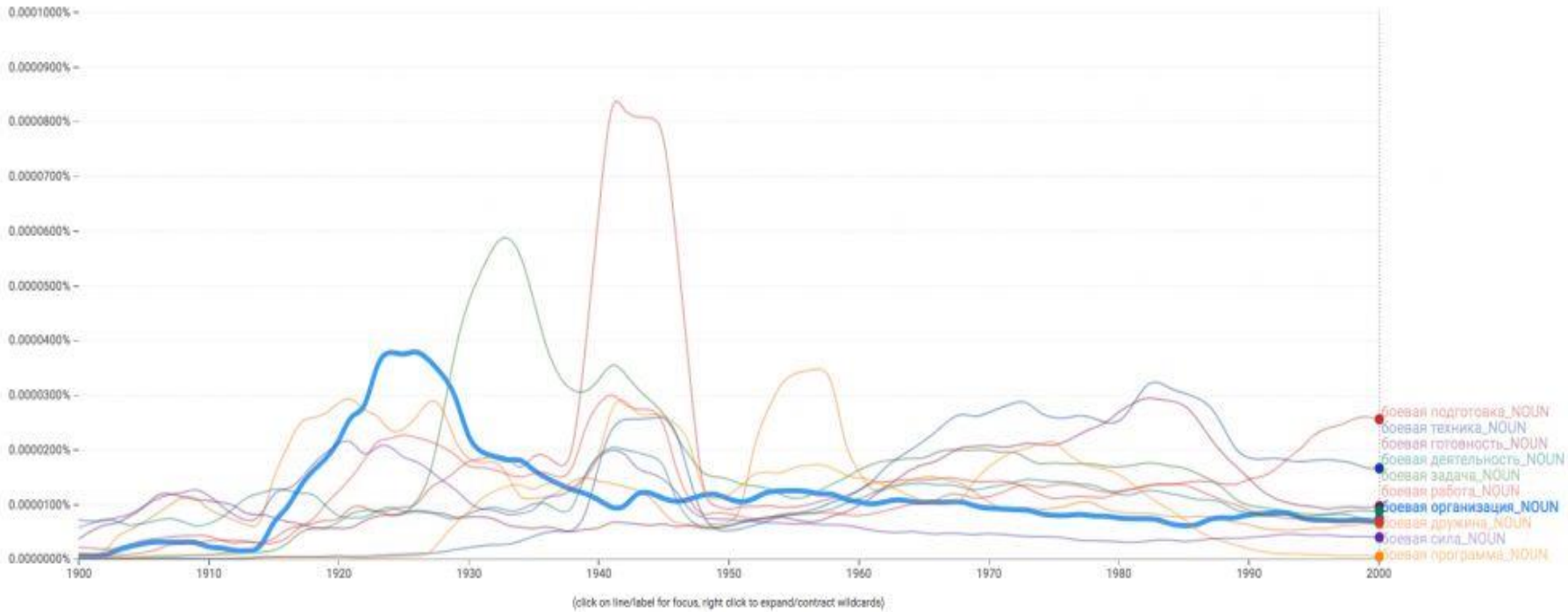
Егер сіз үлкен тілдік корпустарды зерттесеңіз (мысалы, Google Books немесе Википедия), онда кеңірек жоспардың заңдылықтарын анықтауға болады. Мысалы, тілдегі тұрақты тіркестерді немесе N-грамм жиілігінде көрінетін кейбір әлеуметтік тенденцияларды анықтауға болады.

XX ғасырдағы Google books-тегі "жауынгерлік" сын есімінен және кез-келген зат есімнен тұратын биграммалардың жиілігі. Барлық дерлік "жауынгерлік" 1940 жылдардың бірінші жартысында (ҰОС жылдары) жиіліктің жалпы өсуіне ие, мысалы, "жауынгерлік ұйым" ертерек пайда болады және 1910-1920 жылдары құлдырайды.және бұл табиғи: "жауынгерлік ұйым"-революциялық биграмма. Социалистік революционерлердің, большевиктердің және басқа да революциялық партиялардың әскери ұйымдары болды.

Google Books Ngram Viewer

боевая *_NOUN

1900 - 2000 | Russian (2012) | Case-insensitive | Smoothing



(click on line/label for focus, right click to expand/contract wildcards)

N-граммдар көбінесе келесі тапсырмаларда қолданылады:

* Келесі сөздің кеңестерін беру (мысалы, іздеу жолында). N-грамм моделі алдыңғы сөздер белгілі болса, келесі N-грамм сөзінің ықтималдығын есептеуге мүмкіндік береді.

* Авторлықты немесе плагиатты анықтау. Әр түрлі мәтіндер үшін N-граммдарды есептеп, ұқсастық дәрежесін салыстыруға болады.

* Қателерді іздеу және түзету

N-граммның қызықты қолданылуы Google-ді өзінің Google ngram Viewer құралында көрсетеді. Кітаптар жинағын (Google Books) цифрландыра отырып, Google бізге уақытты ескере отырып, мәтіндердегі өзгерістерді визуализациялау құралын берді. Мұнда кітаптарда маңызды әлемдік оқиғалар, тарихи тұлғалардың танымалдылығының шыңдары (Ленин мен Сталин туралы біздің зерттеуімізді қараңыз) және көркем кейіпкерлер қалай бейнеленгенін көруге болады.

N-грамм N элементтерінің тізбегі ретінде анықталады. С

емантикалық тұрғыдан алғанда, бұл дыбыстар, слогдар, сөздер немесе әріптер тізбегі болуы мүмкін. Іс жүзінде N-грамм бірқатар сөздер ретінде жиі кездеседі. Екі қатарлы элементтердің тізбегі жиі аталады биграммалар. Кемінде төрт және одан жоғары элементтер N-грамм ретінде белгіленеді, N тізбектелген элементтер санына ауыстырылады.

N граммдар, әдетте, ғылымның кең саласында қолданылады. Оларды, мысалы, теориялық математика, биология, картография, сонымен қатар музыка саласында қолдануға болады. Ең жиі қолданылатын N-грамм келесі бағыттарды қамтиды:

* ғарыштан жердің спутниктік суреттерін кластерлеу үшін мәліметтер алу, содан кейін суретте жердің қандай нақты бөліктері бар екенін шешу,

* генетикалық тізбекті іздеу,

* генетика саласында ДНҚ үлгілері қандай жануарлардың нақты түрлерінен жиналғанын анықтау үшін қолданылады,

* компьютерлік қысуда,

* N-граммды қолдана отырып, әдетте дыбыспен байланысты мәліметтер индекстеледі.

Сондай-ақ, N-граммдар табиғи тілді өңдеуде кеңінен қолданылады.

Табиғи тілді өңдеу қажеттіліктері үшін N-граммды қолдану

Табиғи тілдерді өңдеу саласында N-грамм негізінен ықтималды модельдер негізінде болжау үшін қолданылады. N-грамм моделі соңғы сөздің ықтималдығын есептейді N-грамм егер алдыңғы барлық белгілі болса. Тілді модельдеу үшін осы тәсілді қолданған кезде әр сөздің пайда болуы тек алдыңғы сөздерге байланысты болады деп болжанады.

N-граммның тағы бір қолданылуы-плагиатты анықтау. Егер сіз мәтінді n-граммммен ұсынылған бірнеше кішкене фрагменттерге бөлсеңіз, оларды бір-бірімен оңай салыстыруға болады және осылайша бақыланатын құжаттардың ұқсастық дәрежесін алуға болады. N-грамм, көбінесе мәтін мен тілді санаттау үшін сәтті қолданылады. Сонымен қатар, оларды мәтіндік деректерден білім алуға мүмкіндік беретін функцияларды құру үшін пайдалануға болады. N-граммды қолдана отырып, емле қателері бар сөздерді ауыстыру үшін үміткерлерді тиімді табуға болады.

Google ғылыми-зерттеу жобалары

Google зерттеу орталықтары көптеген зерттеулер мен әзірлемелер үшін N-грамм модельдерін қолданды. Оларға бір тілден екінші тілге статистикалық аударма, сөйлеуді тану, емле қателерін түзету, ақпарат алу және тағы басқалар сияқты жобалар кіреді. Осы жобалардың мақсаттары үшін бірнеше триллион сөзден тұратын корпустың мәтіндері қолданылды.

Google өзінің оқу корпусын құруға шешім қабылдады. Жоба Google teracorpus деп аталады және онда жалпыға қол жетімді веб-сайттардан жиналған 1 024 908 267 229 сөз бар.

N-грамм алу әдістері

Әр түрлі мәселелерді шешу үшін N-граммдарды жиі қолдануға байланысты оларды мәтіннен шығару үшін сенімді және жылдам алгоритм қажет. Тиісті N-грамм алу құралы Шексіз мәтін өлшемімен жұмыс істей алады, қол жетімді ресурстарды тез және тиімді қолдана алады. Мәтіннен N-грамм алудың бірнеше әдісі бар. Бұл әдістер әртүрлі принциптерге негізделген:

* Жапон тіліндегі мәтіндерге арналған Nagao 94 алгоритмі
* Lempel-Ziv-Welch Алгоритмі
* Жұрнақтар массиві
* Жұрнақтар ағашы
* инверттелген индекс

Назарларыңызға рақмет!