Извлечение терминологии из корпуса специальных текстов

А.А. Матюшин

Москва, 2025

https://action-press.co.uk/wp-content/uploads/2018/02/terminology.jpg

Актуальность, цели, задачи, материалы, методы

- **Цель** – провести извлечение терминологии из корпуса специальных текстов
- **Задачи**
  - Скачать корпус
  - Провести его предобработку
  - Извлечь терминологию с использованием различных методов
  - На основе полученных данных определить оптимальный способ извлечения терминологии

FARMACIA
OFFICIAL JOURNAL OF THE ROMANIAN SOCIETY FOR PHARMACEUTICAL SCIENCES

Извлечение ТОП-30 биграмм (PMI)

$$\mathrm{PMI}(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$$P(w) = \frac{Freq(w)}{totalWordCount}$$

- $P(w_1, w_2)$ — вероятность совместного появления слов w1 и w2
- $P(w_1)$ — вероятность появления слова w_1
- $P(w_2)$ — вероятность появления слова w_2

Извлечение ТОП-30 биграмм (PMI)

acetylsalicilic acidacetaminophen
acordarea burse
ageu levai
aio krk
aleconsiliului stiintific
allopurinolxanthine oxidaselucigenin
amoutzias matakos
annulen acetylhydrazones
antiintestinal nematode
aparisthmium cordatum
approxi mately
arched recurved
arctostaphylos uvaursi
argic cambic
arylparachlorophenylsulfonylphenyl methyloxazoles
astra zeneca
augustynowiczkopec napiorkowska
aurobasidium pullulans
avco bsucrose
bacoside bacopaside
baumgartenii simonk
beitr tabakforsch
bergmeister podesser
bigovic roganovic
blankfassif peakfassif
blankfessif peakfessif
boberg taxvig
bonacucina cespi
boshy risha
bryophyllum pinnatum

Проверка полученных биграмм

('approxi', 'mately'): вероятно одно слово

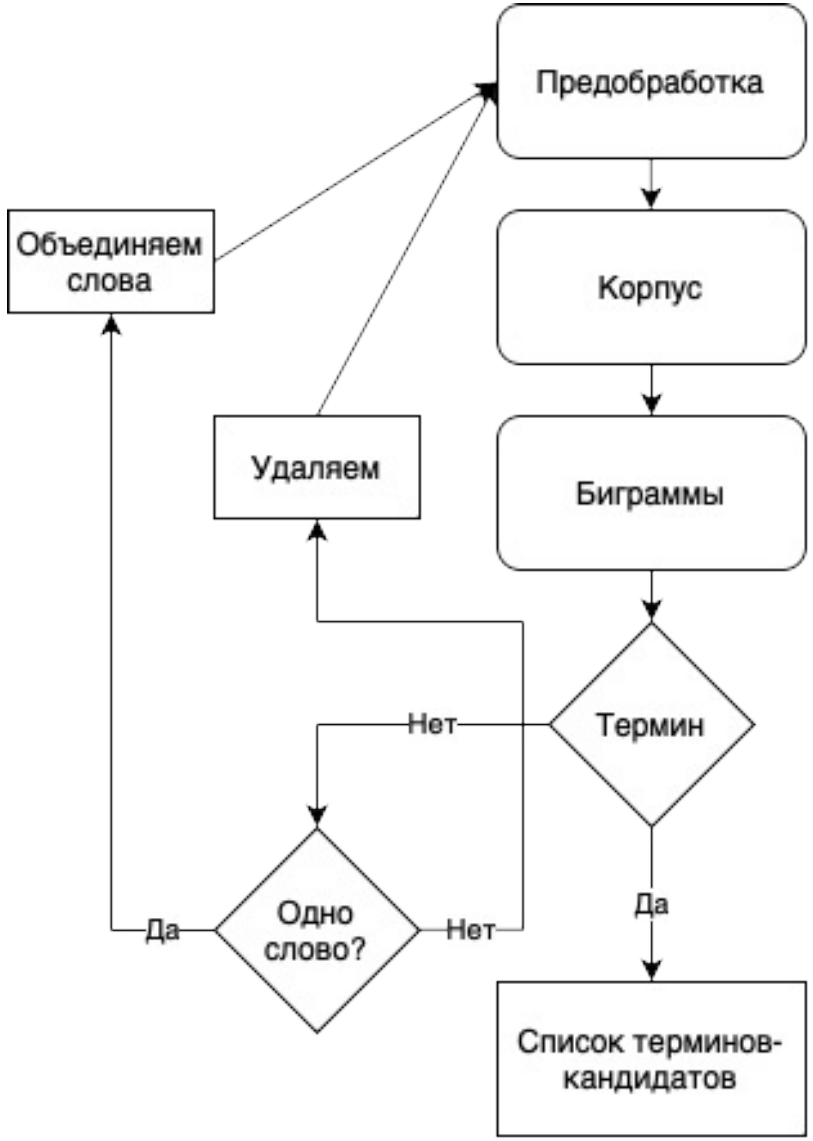
antiintestinal nematode

arched recurved
astra zeneca

arylparachlorophenylsulfonylphenyl methyloxazoles
acetylsalicilic acidacetaminophen annulen acetylhydrazones

aparisthmium cordatum arctostaphylos uvaursi
baumgartenii simonk

boberg taxvig
bacoside bacopaside aleconsiliului stiintific acordarea burse
bonacucina cespi



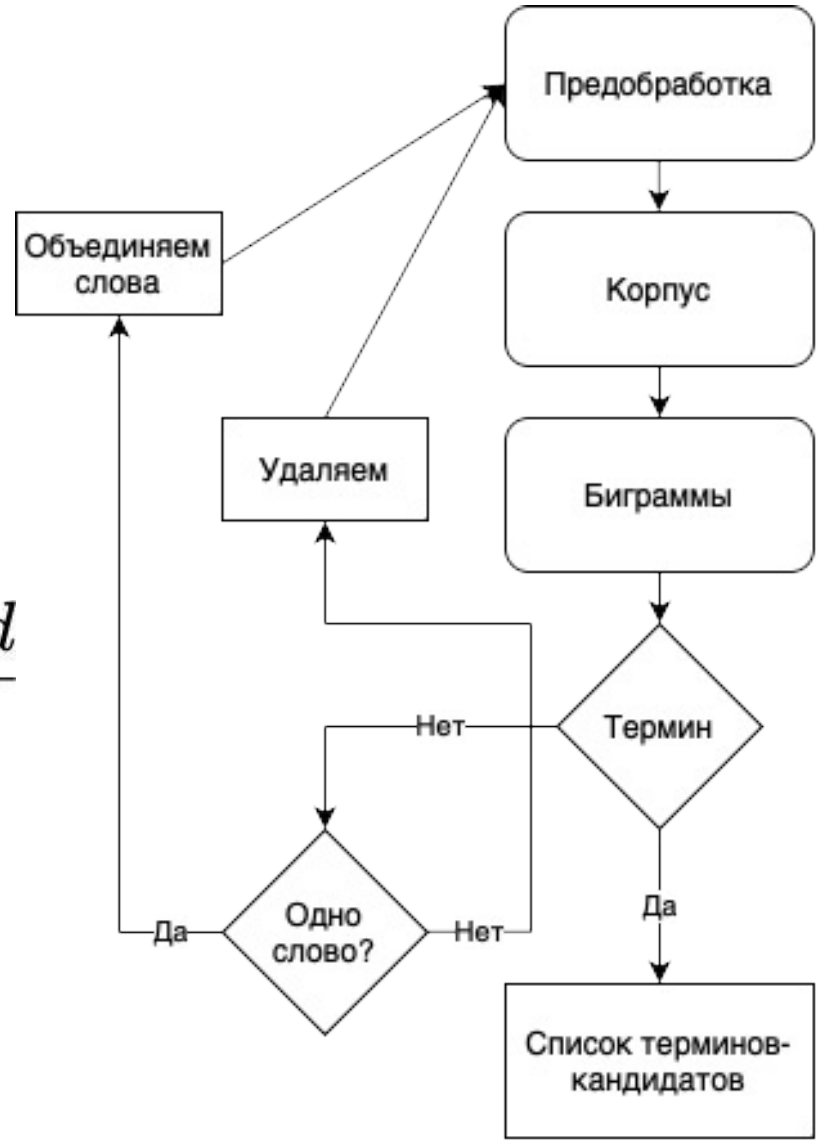
Альтернативный подход: TF-IDF

('approxi', 'mately'): вероятно одно слово
('centrif', 'uged'): вероятно одно слово
('choles', 'terol'): вероятно одно слово

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

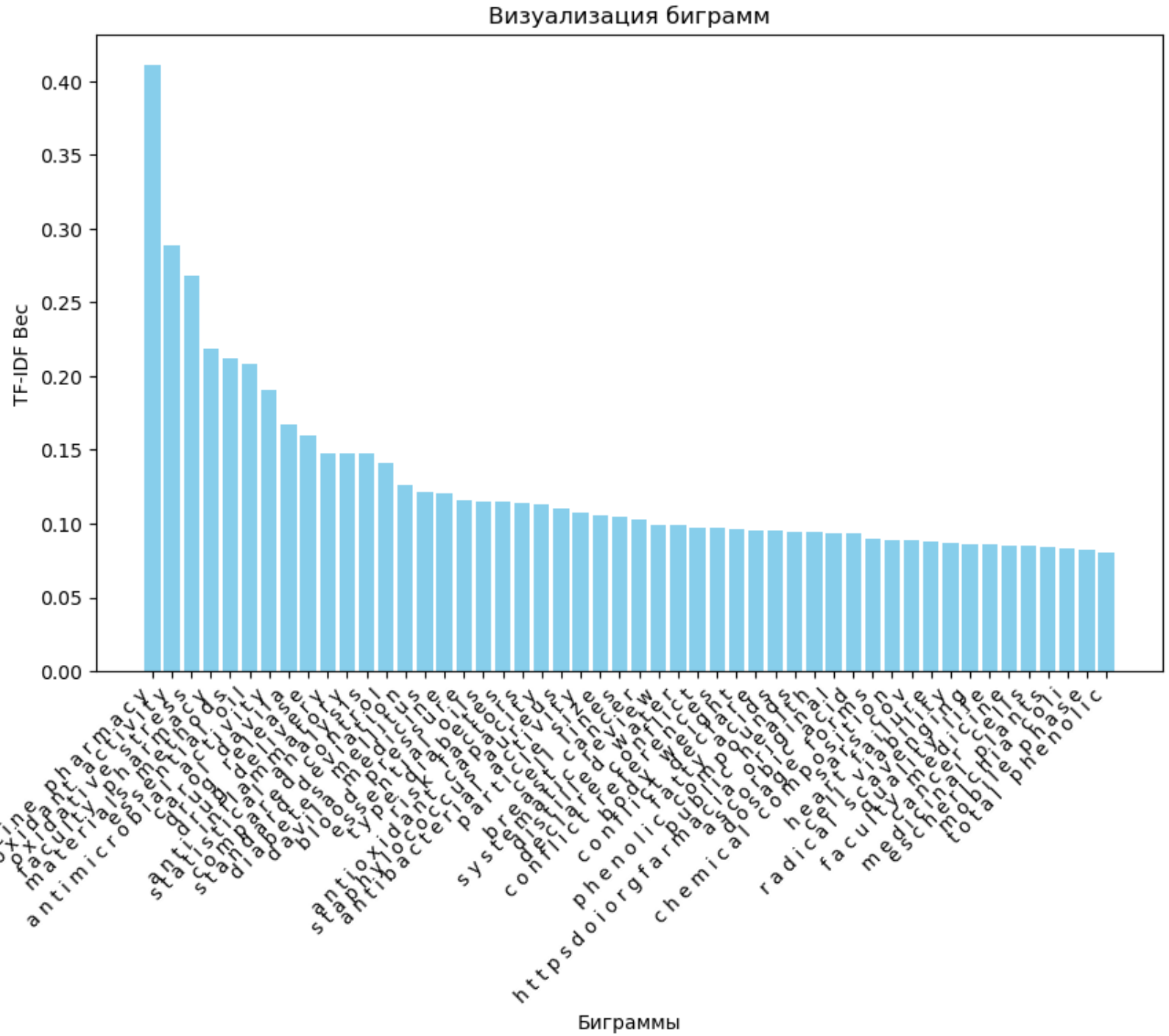


https://miro.medium.com/v2/resize:fit:1400/format:webp/1*1pTLnoOPJKKcKIcRi3q0WA.jpeg

TF-IDF (TOP-30 биграмм)

medicine pharmacy: 0.4109 (100.00%)
antioxidant activity: 0.2888 (70.29%)
oxidative stress: 0.2684 (65.32%)
faculty pharmacy: 0.2190 (53.29%)
materials methods: 0.2125 (51.72%)
essential oil: 0.2083 (50.70%)
antimicrobial activity: 0.1903 (46.31%)
carol davila: 0.1673 (40.71%)
drug release: 0.1594 (38.79%)
drug delivery: 0.1481 (36.04%)
anti inflammatory: 0.1478 (35.98%)
statistical analysis: 0.1477 (35.95%)
compared control: 0.1415 (34.44%)
standard deviation: 0.1259 (30.63%)
diabetes mellitus: 0.1210 (29.46%)
davila medicine: 0.1207 (29.37%)
blood pressure: 0.1163 (28.32%)
essential oils: 0.1148 (27.94%)
type diabetes: 0.1146 (27.88%)
risk factors: 0.1139 (27.73%)
antioxidant capacity: 0.1129 (27.48%)
staphylococcus aureus: 0.1107 (26.96%)
antibacterial activity: 0.1071 (26.06%)
particle size: 0.1054 (25.66%)
cell lines: 0.1050 (25.56%)
breast cancer: 0.1027 (25.01%)
systematic review: 0.0993 (24.17%)
distilled water: 0.0991 (24.11%)
declare conflict: 0.0972 (23.65%)
conflict references: 0.0970 (23.62%)

TF-IDF (TOP-30 биграмм)


Визуализация биграмм
TF-IDF Вес
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Биграммы

RAKE (YAKE)

Этот метод похож по смыслу на Rake, однако была убрана идея о выделении фраз на основе стоп-слов.
В данном методе используется стандартная для текстового анализа методика выделения слов и фраз с помощью *токенизации*.
Фактически такая методика позволяет проверить все сочетания слов на их важность, а не только разделенные стоп-словами.
YAKE! использует более сложную метрику, чем Rake – она собирается из 5 отдельных метрик.

antioxidant activity (Score: 3.0087148436266337e-09, 10.09%)
medicine pharmacy (Score: 3.059448082625984e-09, 10.26%)
drug release (Score: 5.437219680578785e-09, 18.24%)
patients patients (Score: 6.711015833803034e-09, 22.51%)
antimicrobial activity (Score: 7.559408321921399e-09, 25.36%)
materials methods (Score: 7.93816308479594e-09, 26.63%)
study patients (Score: 8.306468566469654e-09, 27.87%)
treatment patients (Score: 8.420550158860477e-09, 28.25%)
compared control (Score: 1.008944813237746e-08, 33.85%)
cancer cells (Score: 1.043746991979737e-08, 35.01%)
faculty pharmacy (Score: 1.1890855722811215e-08, 39.89%)
patients study (Score: 1.4687346692530432e-08, 49.27%)
drug delivery (Score: 1.5023315843638416e-08, 50.40%)
patients treated (Score: 1.569131537237309e-08, 52.64%)
statistical analysis (Score: 1.674248237879614 5e-08, 56.17%)
essential oil (Score: 1.722104681533698 8e-08, 57.77%)
patients treatment (Score: 1.820778646915904 4e-08, 61.08%)
risk factors (Score: 2.0108118969877545e-08, 67.46%)
plant extracts (Score: 2.017509224013513 4e-08, 67.68%)
cells cells (Score: 2.040598471801815 5e-08, 68.45%)
methods study (Score: 2.411255680213291 5e-08, 80.89%)
antibacterial activity (Score: 2.435621284777873e-08, 81.71%)
drug drug (Score: 2.454984292998198e-08, 82.36%)
activity extracts (Score: 2.478517167711605e-08, 83.15%)

Совпадающие термины (TOP-30 TF-IDF и YAKE)

antimicrobial activity
antibacterial activity
compared control
cancer cells
oxidative stress
drug delivery
materials methods
drug release
risk factors
essential oil
statistical analysis
medicine pharmacy
faculty pharmacy
antioxidant activity
phenolic compounds

Совпадающие термины (TOP-30 TF-IDF и YAKE)

antimicrobial **activity**
antibacterial **activity**
compared **control**
cancer **cells**
oxidative stress
drug delivery
materials methods
drug release
risk factors
essential oil
statistical **analysis**
medicine **pharmacy**
faculty **pharmacy**
antioxidant **activity**
phenolic **compounds**

3621337 слов в корпусе
patients: 22418 (0.62%)
study: 18277 (0.50%)
treatment: 15159 (0.42%)
activity: 14534 (0.40%)
drug: 13514 (0.37%)
acid: 12589 (0.35%)
cells: 10982 (0.30%)
analysis: 10380 (0.29%)
effects: 10034 (0.28%)
control: 8961 (0.25%)
compounds: 8835 (0.24%)
extract: 8822 (0.24%)
cell: 8752 (0.24%)
studies: 8306 (0.23%)
pharmacy: 8113 (0.22%)
concentration: 7874 (0.22%)
data: 7841 (0.22%)

Коллокации (TOP-30 слева/справа)

Термин: **patients**

Коллокации 'слева'
treatment patients: 412
study patients: 389
diabetic patients: 380
patients patients: 362
cancer patients: 209

Коллокации 'справа'
patients treated: 607
patients patients: 362
patients received: 334
patients receiving: 302
patients chronic: 289

Термин: **activity**

Коллокации 'слева'
antioxidant activity: 2274
antimicrobial activity: 1498
antibacterial activity: 843
scavenging activity: 504
antifungal activity: 462

Коллокации 'справа'
activity extracts: 187
activity tested: 172
activity essential: 142
activity compared: 136
activity dpph: 136

3621337 слов в корпусе
patients: 22418 (0.62%)
study: 18277 (0.50%)
treatment: 15159 (0.42%)
activity: 14534 (0.40%)
drug: 13514 (0.37%)
acid: 12589 (0.35%)
cells: 10982 (0.30%)
analysis: 10380 (0.29%)
effects: 10034 (0.28%)
control: 8961 (0.25%)
compounds: 8835 (0.24%)
extract: 8822 (0.24%)
cell: 8752 (0.24%)
studies: 8306 (0.23%)
pharmacy: 8113 (0.22%)
concentration: 7874 (0.22%)
data: 7841 (0.22%)

Альтернативный подход: сопоставление с корпусом

antioxidant: 7623.00 (0.210502%)
antimicrobial: 4477.00 (0.123628%)
res: 3903.00 (0.107778%)
anti: 3780.00 (0.104381%)
inflammatory: 3505.00 (0.096787%)
assay: 3097.00 (0.085521%)
oxidative: 2818.00 (0.077817%)
derivatives: 2804.00 (0.077430%)
pharmaceutical: 2709.00 (0.074807%)
pharmacists: 2692.00 (0.074337%)
tablets: 2646.00 (0.073067%)
clin: 2582.00 (0.071300%)
phenolic: 2427.00 (0.067019%)
usa: 2388.00 (0.065942%)
receptor: 2318.00 (0.064010%)
medicine pharmacy: 3235.00 (0.089332%)
antioxidant activity: 2274.00 (0.062794%)
oxidative stress: 2113.00 (0.058349%)
faculty pharmacy: 1724.00 (0.047607%)
materials methods: 1673.00 (0.046198%)
essential oil: 1640.00 (0.045287%)
antimicrobial activity: 1498.00 (0.041366%)
carol davila: 1317.00 (0.036368%)
drug release: 1255.00 (0.034656%)
drug delivery: 1166.00 (0.032198%)
anti inflammatory: 1164.00 (0.032143%)
compared control: 1114.00 (0.030762%)
diabetes mellitus: 953.00 (0.026316%)
davila medicine: 950.00 (0.026233%)
essential oils: 904.00 (0.024963%)

Общие биграммы (кандидаты в термины), полученные тремя методами