Национальный
исследовательский
**Томский
государственный
университет**

Анализ новостных текстов (Covid-19)

Технологии автоматической обработки текста

Абакумова Мария, Зима Екатерина, Шведова София,
гр. 132004, 2023

Дизайн исследования

1. Предобработка: токенизация, лемматизация
2. Word2vec: обучающая модель
3. TF-IDF: выявление ключевых слов
4. Синонимическая близость
5. LDA и BERTopic: распределение близких по тематике слов в группы
6. Извлечение фактов: TOMITA

Описание данных

Новостной ресурс РИА Новости: https://ria.ru/

Рассматриваемая **категория**: Общество – Здоровье – Коронавирус

Исходные данные:

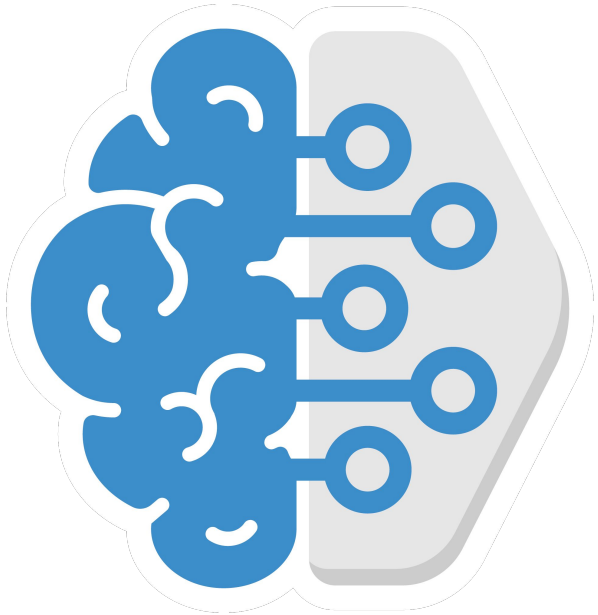
- 2020 год – 4592 новости
- 2021 год – 4737 новостей
- 2022 год – 2965 новостей

Гипотеза: Анализ новостей за разные годы позволяет отследить различную степень влияния пандемии на обстановку в мире

Препроцессинг данных

Включает в себя:

1. Приведение слов к нижнему регистру
2. Удаление цифр
3. Удаление одиночных букв
4. Удаление стоп-слов
5. Удаление ссылок



Word2vec

Для создания модели использованы следующие параметры:

- **min_count** — минимальное число вхождений слова
- **window** — расстояние между current и predicted слов
- **vector_size** — размер векторного представления слова (word embedding) — размер матрицы
- **negative** — сколько неконтекстных слов учитывать в обучении, используя negative sampling — число слов
- **alpha** — начальный learning_rate, используемый в алгоритме обратного распространения ошибки (Backpropogation). Задаем угол наклон распределения
- **min_alpha** — минимальное значение learning_rate, на которое может опуститься в процессе обучения.
- **sample** — пороговое значения для настройки того, какие высокочастотные слова подвергаются случайной понижающей дискретизации
- **sg** — если 1, то используется реализация Skip-gram; если 0, то CBOW.

TF-IDF

2020		2021		2022	
январь	0.271844	джонсон	0.297122	греция	0.255852
тип	0.215510	борис	0.278768	омикрон	0.255207
состояние	0.198511	штамм	0.222216	инфицировать	0.174546
эксперт	0.193178	великобританииотмечаться	0.205904	трудный	0.173862
возбудитель	0.181935	странеминистр	0.205904	поблагодарить	0.158591
метод	0.167650	новый	0.199481	полагать	0.142459
заражение	0.158415	великобритания	0.199222	стать	0.140846
заболевание	0.141961	вариантараный	0.189630	выразить	0.125858
известно	0.141808	анкара	0.161957	год	0.125359
диагностикипо	0.134121	мэтт	0.160217	мутация	0.113156

Синонимическая близость

2020	2021	2022
заражение	**штамм**	**омикрон**
[('зафиксировать', 0.9961381554603577), ('выявить', 0.9899018406867981), ('летальный', 0.9878172874450684), ('свыше', 0.9870333671569824), ('случай', 0.9864846467971802), ('исход', 0.9864648580551147), ('миллион', 0.9862198233604431), ('скончалисьва', 0.9846882224082947), ('умерлив', 0.9842034578323364), ('насчитываться', 0.9826470017433167)]	[('мутация', 0.9648011326789856), ('вариант', 0.9519167542457581), ('британский', 0.9437460899353027), ('дельта', 0.9411693215370178), ('омикрон', 0.9296159744262695), ('бразильский', 0.9270622730255127), ('ботсвана', 0.9208739995956421), ('южноафриканский', 0.9154641628265381), ('индийский', 0.9112163782119751), ('юар', 0.8990864157676697)]	[('вариант', 0.996515154838562), ('подвид', 0.9958195090293884), ('подвариант', 0.9951022863388062), ('дельта', 0.9948450326919556), ('кентавр', 0.9947988986968994), ('преобладать', 0.9947308301925659), ('особенность', 0.9945005178451538), ('геновариант', 0.9944378733634949), ('доминировать', 0.993434488773346), ('приходиться', 0.993298351764679)]

Синонимическая близость

2020	2021	2022
возбудитель, заражение, заболевание	**штамм, новый, великобритания**	**омикрон, инфицировать, греция**
[('тип', 0.9862245321273804), ('госкомитет', 0.9796692132949829), ('ncovпо', 0.9794099926948547), ('погибнуть', 0.9735077619552612), ('иностранец', 0.9728562235832214), ('предел', 0.9704283475875854), ('вызвать', 0.9702456593513489), ('материковый', 0.9688411951065063), ('заразиться', 0.9688306450843811), ('согласно', 0.9687695503234863)]	[('британский', 0.9784375429153442), ('дельта', 0.9638498425483704), ('ботсвана', 0.9592905044555664), ('десятка', 0.952200174331665), ('разновидность', 0.9511969685554504), ('бразильский', 0.9457120895385742), ('омикрон', 0.9383739233016968), ('впервые', 0.9376696348190308), ('индийский', 0.9364216923713684), ('обнаружить', 0.9362503886222839)]	[('приходиться', 0.9982255697250366), ('жертва', 0.99802166223526), ('максимум', 0.997991681098938), ('соответственно', 0.9979043006896973), ('завоз', 0.997746467590332), ('полтора', 0.9977391958236694), ('причём', 0.9976670145988464), ('цифра', 0.9976417422294617), ('аналогичный', 0.9976281523704529), ('значительно', 0.9976164102554321)]

Синонимическая близость - вычитание

2020	2021	2022
возбудитель, заражение – январь	**штамм, новый – великобритания**	**омикрон, инфицировать – греция**
***ncov =newCoronaVirus** [('ncov', 0.939436674118042), ('тип', 0.8930541276931763), ('пневмония', 0.8895509243011475), ('признать', 0.8850549459457397), ('предварительно', 0.8802609443664551), ('болезнь', 0.8768469095230103), ('хубэя', 0.8763914704322815), ('материковый', 0.8714083433151245), ('провинция', 0.8712812662124634), ('китай', 0.8672407865524292)]	[('covid', 0.7637748718261719), ('британский', 0.7581128478050232), ('обнаружить', 0.7576642036437988), ('дельта', 0.7576467990875244), ('тип', 0.7540638446807861), ('радио', 0.748855471611023), ('омикрон', 0.7486017346382141), ('сообщить', 0.7482305765151978), ('болезнираный', 0.7476321458816528), ('омикронштамм', 0.7463254928588867)]	[('штамм', 0.9948295950889587), ('максимум', 0.9906011819839478), ('целое', 0.9904603958129883), ('рекорд', 0.9902593493461609), ('прошлое', 0.9890199899673462), ('активно', 0.9890085458755493), ('локальный', 0.988959014415741), ('рекордный', 0.9888675212860107), ('коронавирусомть', 0.9888557195663452), ('свыше', 0.9886356592178345)]

Синонимическая близость

Наиболее близкое слово из списка к заданному

2020 – заболевание

↓

возбудитель,
заражение,
метод

2021 – штамм

↓

covid,
дельта,
обнаружить

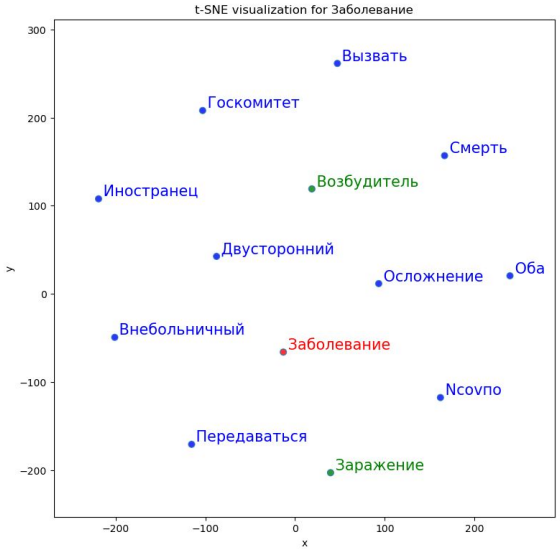
2022 – омикрон

↓

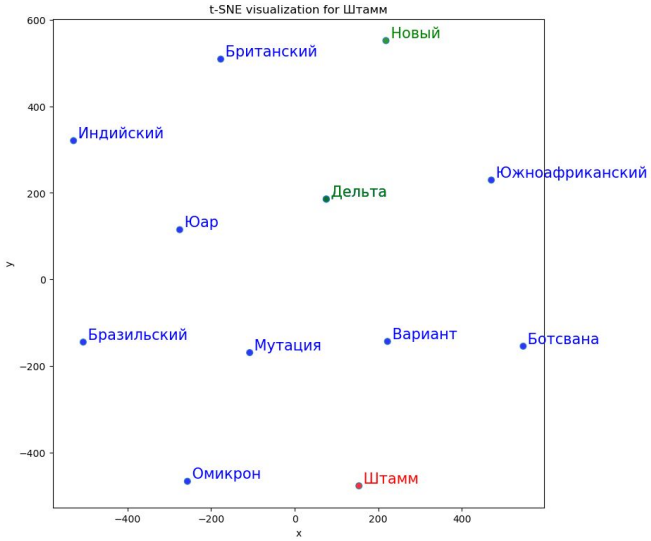
рекорд,
увеличиваться,
подвид

Синонимическая близость

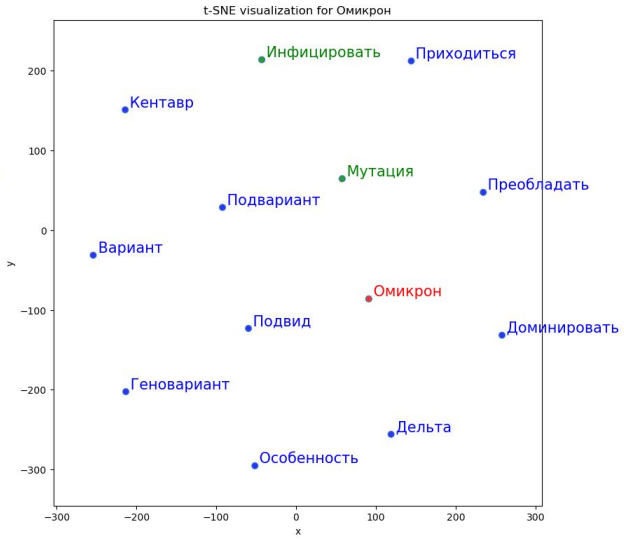
2020



2021



2022



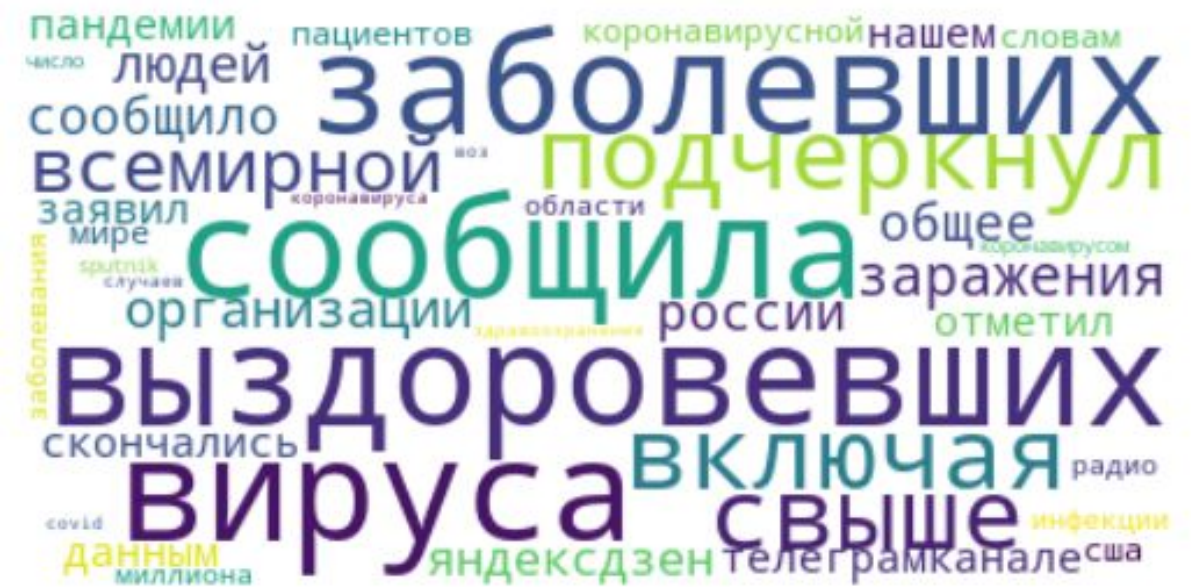
LDA – Latent Dirichlet allocation

Латентное размещение Дирихле (*LDA*, от англ. *Latent Dirichlet allocation*) — применяемый в машинном обучении метод тематического моделирования. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа.

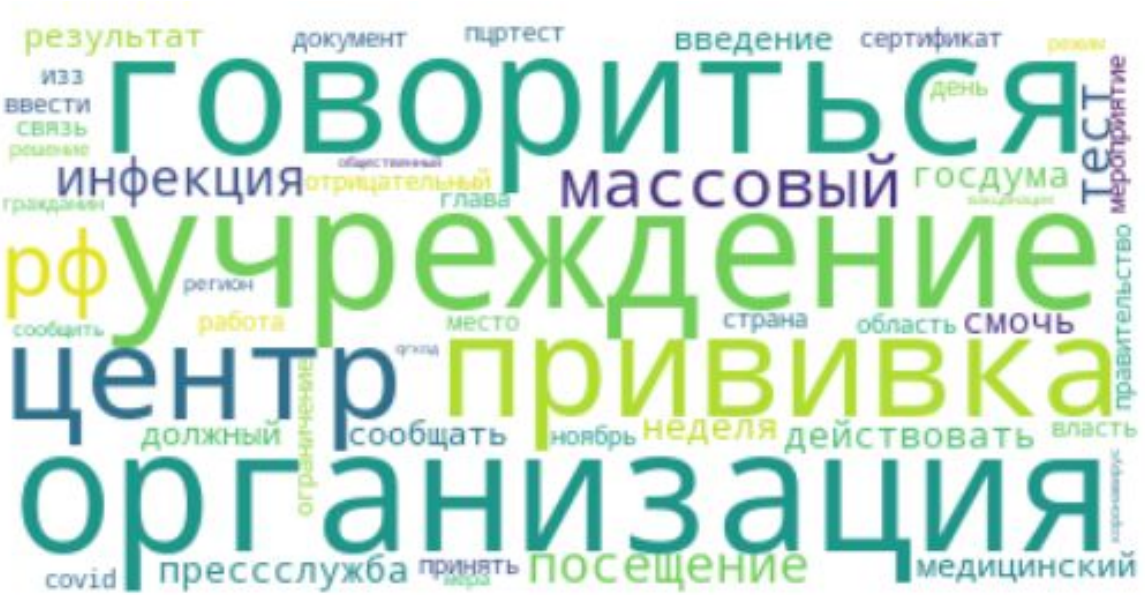
LDA - топ 10 слов

2020	2021	2022
регион	сообщить	covid
сообщить	режим	правительство
область	решение	вакцина
лечение	гражданин	прессслужба
коронавирус	регион	регион
помощь	мера	область
медицинский	общественный	режим
врач	qrкод	сообщить
пациент	коронавирус	коронавирус
больница	вакцинация	ограничение

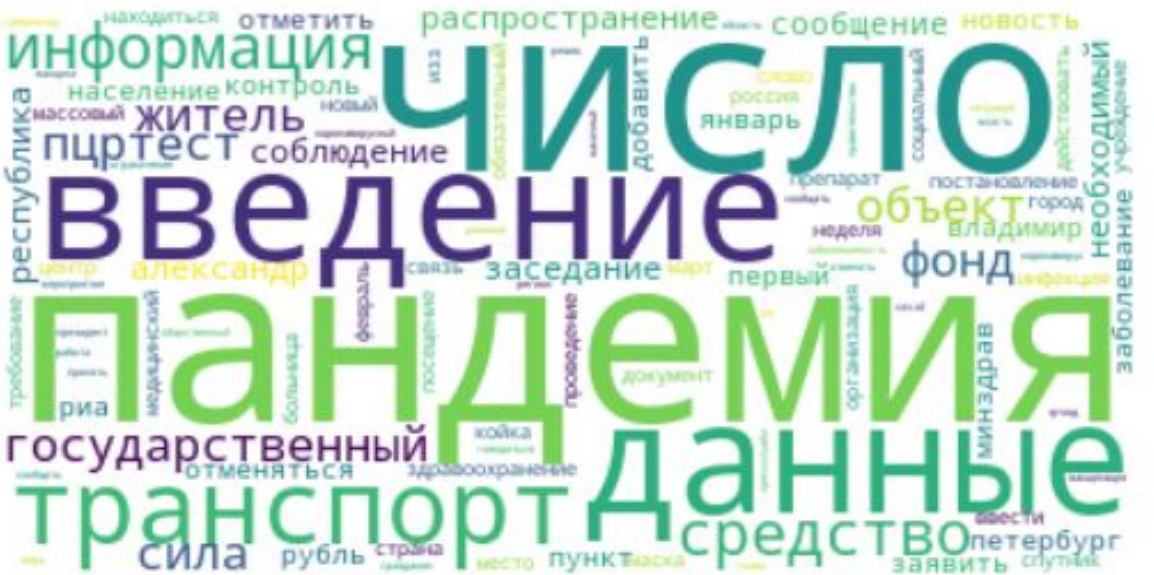
LDA – топ 100 слов (2020) – визуализация



LDA – топ 100 слов (2021) – визуализация



LDA – топ 100 слов (2022) – визуализация



Topic modeling: BERTopic

Кластеризация – см. Google Colab

BERTopic (Bidirectional Encoder Representations from Transformers) использует контекстуализированные эмбеддинги слов, основанные на Transformer-архитектуре.

Преимущества:

- обучается на больших объемах текста
- позволяет создавать высококачественные представления слов с учетом контекста и семантики

ГОД 2020	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6	TOPIC 7
LDA	**Случаи выявления ковида** (сообщить, пресслужба, инфекция, регион)	**Ковид в мире** (Великобритания, ООН, Путин, президент)	**Динамика развития инфекции** (умереть, сутки, выявить, март)	**Первые упоминания о вакцине** (разработать, спутник, клинический, испытание)	**Ковид в медиа** (сообщать, рассказать, телеграмканам, радио)	**Страна распространения** (провинция, Ухань, вспышка, Китай)	**Введение режима самоизоляции** (Москва, ввести, самоизоляция, правительство)
BERTopic	**Вакцина** (спутник, вакцина, россия, гамалеи*)	**Динамика развития инфекции** (случаев, число, сутки, заражения)	**Страна распространения** (Китай, Хубэя, коронавирус, пневмонии)	**Распространение covid** (новые, случаи, скончались, узбекистан)	**Летальность ковида** (смерти, скончались, умерших, достигло)	**Материальная поддержка** (рублей, выплаты, помощь, врачей)	***Футбол и ковид** (команды, игроков, матч, короновирус)

*Национальный исследовательский центр эпидемиологии и микробиологии имени Н. Ф. Гамалеи
*Хубэй – провинция на востоке центральной части Китая
*Чемпионат Европы 2020

ГОД 2021	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6	TOPIC 7
LDA	**Официально принятые меры по ковид** (пцртест, режим, решение, qrкод)	**Разработка российской вакцины** (разработать, препарат, спутник, российский)	**Распространение вакцинации** (иммунитет, вакцинировать, привить, население)	**Новый штамм** (заражение, мутация, омикрон, пациент, новый, вирус, заболевание)	**Обстановка в мире и государстве** (борьба, пандемия, мир, заявить)	**Ситуация в столице** (столичный, сообщить, мэр, сертификат)	**Новые случаи** (смерть, прирост, выявить, новый)
BERTopic	**Обстановка в столице** (пациент, коронавирус, ситуация, москва)	**Российская вакцина** (спутник, вакцинация, препарат, российский)	**Новые случаи** (случай, борьба, распространение, оперативный)	**Новый штамм** (штамм, новый, заражение, выявить, случай)	**Разработка вакцины** (вакцинация, препарат, страна, прививка)	**Информация по авиасообщению** (авиасообщение, турист, регулярный, рейс)	**Иммунитет к вирусу** (вакцина, иммунитет, антитело, прививка)

ГОД 2022	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6	TOPIC 7
LDA	**Ограничительные меры** (принять, мера, qrкод, масочный)	**Рост заболеваемости** (увеличиваться, заболевание, новый, пациент)	**Лечение** (лекарственный, медицинский, врач, препарат)	**Обстановка в Москве** (контроль, столица, сводка, москва)	**Омикрон** (новый, омикрон, штамм, заражение)	**Снижение заболеваемости** (снижение, ситуация, борьба, сравнение)	**Исследования** (работа, исследование, эпидемиология, здравоохранение)
BERTopic	**Оказание помощи** (помощь, пациент, больница, лечение)	**Рост заболеваемости** (госпитализация, сравнение, рост, сообщение)	**Иммунитет к вирусу** (коллективный, иммунитет, вакцинация, уровень)	**Вакцина** (вакцина, спутник, разработать, препарат)	**Состояние жителей** (раздражительность, житель, рф, нервозность)	**Изменение ситуации** (измениться, сутки, умереть, тысяча)	**Ограничительные меры** (qrкод, мероприятие, отменяться, ограничение)

Извлечение фактов: Томита

Томита-парсер — парсер, специализирующийся на извлечении фактов из текста на основе словарей и грамматик.
Извлекаемый факт – названия стран

	2020	**2021**	**2022**
Россия	34	1610	1870
Китай	66	91	52
США	763	318	190
Великобритания	28	143	76
Италия	34	119	32

Выводы

Гипотеза **подтвердилась** по следующим причинам:
1) Синонимическая близость: слова, выявленные в ходе анализа синонимической близости отражают события, происходящие в соответствующие года

2) LDA: на основе выявленных ключевых слов и топиков по каждому году

3) BERTopic: на основе выявленных топиков по каждому году

4) Tomita: новостные упоминания стран по годам соответствуют мировой обстановке

Спасибо за внимание!

Национальный исследовательский
Томский государственный университет

634050, г. Томск, пр. Ленина, 36
+7 (3822) 52-98-52, +7 (3822) 52-95-85 (факс)
rector@tsu.ru
www.tsu.ru