Аналіз постів користувачів Твіттер пов’язаних з темою «Коронавірус»:

Аналіз часових рядів згадування власних назв

Виконала:

Завальнюк Юлія

Зміст

Опис

У останні декілька місяців увесь світ захлиснула пандемія нового вірусу COVID-19. Зі закриттям кінотеатрів, магазинів, торгових центрів усе людське спілкування перемістилось у інтернет. Усі дискусії та обговорення, перемітились в соціальні мережі: Twitter, Facebook, Instagram та Вконтакті заполонили щоденні повідомлення про померлих чи нових заражених, теорії про походження вірусу чи надії на швидке закінчення карантину.

У цій роботі 5000 твітів пов'язаних з коронавірусом між 1 березня по 1 травня 2020 року були вилучені для аналізу.

Актуальність

Соціальні мережі були і залишаються індикаторами настрою користувачів. Для прикладу у Америці проводились дослідження, що знайшли кореляцію у аналізі твітів у певному регіоні та рівнем злочинності. Наше завдання не таке масштабне. Намагаючи в загальному проаналізувати твіти україномовного сегменту твіттеру за темою коронавірус, провести аналіз часових рядів згадування певних країн, організацій чи людей та зрозуміти їх динаміку.

Збір даних

Платформа з якої беруться дані

Твíтер — соціальна мережа мікроблогів, дає змогу користувачам надсилати короткі текстові повідомлення (до 280 символів), використовуючи SMS, служби миттєвих повідомлень і сторонні програми-клієнти. Літературний сегмент твíтера породив такий різновид короткотекстової літератури, як твітература (англ. Twitterature).

У Twitter є 330 мільйонів активних користувачів щомісяця та 145 мільйонів щоденних користувачів на Twitter. Тобто приблизно 42% користувачів Twitter щодня використовують платформу.

Інструмент для збору

Для збору даних з соціальної мережі Twitter була використана бібліотека GetOldTweets3.

GetOldTweets3 - це удосконалення гілка оригінального GetOldTweets-python Джефферсона Генріка. Офіційний API Twitter має обмежене обмеження у часі, та часто неможливо отримати твіти, старші чим тиждень. Деякі інструменти надають доступ до старих твітів, але в більшості з них використання платне.

Ця бібліотека працює напряму через імітацію прокрутки сторінки Twitter, збираючи дані в JSON. В результати отримується перевага в можливості знайти найглибші старі твіти. Також додаються такі функції, як підрахунок ретвітів, пошук у кількох облікових записах користувачів тощо, що дає можливість за необхідності конкретизувати запит.

Деякі твіти були занадто короткі (просто зображення без тексту, одне-два слова або просто один хештег). Щоб отримати більш релевантну інформацію, ми будемо використовувати лише довгі твіти (більше 40 символів). Деякі твіти були просто ретвітами чи копіями інших. Отже, з 5000 твітів, що ми зібрали, 4761 можуть бути актуальними, і лише 4537 - унікальними.

Попередня обробка

Чистий текст означає список слів або токенів, з якими ми можемо працювати в наших дослідженнях. Тобто необхідно перетворити необроблений текст в список слів.

Вхідне речення для прикладу:

«Я думаю, що надзвичайний стан це жахливо. От мої друзья ходили на роботу і все добре. А в новинах сказали, що «все меры не дадут никакого эфекта»!»

Неактуальні символи

Що зроблено? Проводиться очистка даних від будь-яких неактуальних символів, включаючи цифри, пунктуацію, посилання, згадки, російські літери тощо.

Для чого? Так як «,» вважається також словом, хоча для наших цілей не має жодного змісту, а знаходити слова з російськими літерами є найпростішим способом знаходити російські слова, що не входять до площини нашого завдання, та виключати їх з тесту.

Як? Бібліотека regex.

Регулярний вираз - це послідовність символів, що визначають шаблон пошуку. Зазвичай такі шаблони використовуються алгоритмами пошуку рядків для операцій "знайти" або "знайти і замінити" на рядках або для перевірки введення даних.

Приклад: Я думаю що надзвичайний стан це жахливо От мої друзья ходили на роботу і все добре А в новинах сказали що все не дадут никакого»

Нижній регістр

Що зроблено? Приведення до нижнього регістру.

Для чого? «Мова» та «мова» сприймаються як різні слова, а для хорошого аналізу нам необхідно, щоб вони рахувались як одне.

Як? str.lower()

Приклад: «я думаю що надзвичайний стан це жахливо от мої друзья ходили на роботу і все добре а в новинах сказали що все не дадут никакого»

Стоп-слова

Що зроблено? Стоп-слова з словника стоп-слів видалені з тексту.

Для чого? За законом Зіпфа (Ціпфа в інших перекладах, Zipf's law в оригіналі) що визначають закономірність розташування слів, де частота слова обернено пропорційна його місця в тексті. [1] У англійській мові таким словом є «the», при проведені оцінки на українських та російських корпусах на перших 10 місцях є переважно наступні слова:

1. і
2. в
3. не
4. на
5. що
6. з
7. то
8. я
9. він
10. як

Тобто приблизно 22% слів у тесті припадають на саме ці слова. Так як вони в більшості не несуть ніякого змістового навантаження (окрім можливо слова «не» що в поєднанні з іншими словами дає їм протилежний зміст). Для більшості мов присутні онлайн словники цих стоп-слів. Враховуючи специфіку розмовної української мови (ту, яку зазвичай використовують для твітів), також додаються російські стоп-слова.

Як? Російський та український список стоп-слів (враховуючи специфіку розмовної української мови).

Приклад: «думаю надзвичайний стан жахливо друзья ходили роботу добре новинах сказали не дадут никакого»

Іншомовні слова

Що зроблено? Застосована нейронна модель для пошуку і вилучення російських слів з тексту.

Для чого? Як виявилось після обробки тексту, не всі російські слова можуть бути вилучені через простий пошук російських літер. Так у реченні-прикладі залишились російські слова «дадут», «никакого» та «друзья».

Як? Класифікатор Naive Bayes - це імовірнісна модель машинного навчання, яка використовується для завдання класифікації. Суть класифікатора заснована на теоремі Байєса.

Найбільш часті слова у тренувальному корпусі (для цього набору 5000) розбиваються на підслова, що використовуються для розпізнавання мови.

Тобто склади (морфеми), що найчастіше використовуються в конкретній мові, отримують своє значення частоти для конкретної мови. І якщо в

подальшому модель зустрічає слово, якого не було в тренувальному корпусі, вона розбиває слово на морфеми і використовує це знання, щоб передбачити мову.

Точність моделі не ідеальна, так як українська та російська мова є дуже схожими. Але з точністю більше 90% російські слова будуть вилучені. Та майже з 100% всі слова з інших мов. Код можна знайти в репозиторії на github. [2]

Приклад: «думаю надзвичайний стан жахливо ходили роботу добре новинах сказали не дадут»

2 з 3 слів були виявленні. Слово «дадут» є дуже схожим з українським «дадуть», що пояснює помилку моделі.

Лематизація

Що зроблено? Приведення всіх слів до їх початкової форми.

Для чого? «Мова» та «Мовою» сприймаються як різні слова, а для хорошого аналізу нам необхідно, щоб вони рахувались як одне. Отже використовуючи відповідні бібліотеки можна добитись хороших результатів на перетворенні слів.

Як? Бібліотека pymorphy2.

pymorphy2 - морфологічний аналізатор для російської (та української) мови, написаний на мові Python і використовує словники з OpenCorpora.

Метод MorphAnalyzer.parse() приймає слово (обов'язково в нижньому регістрі, для цього, в тому числі. ми перевели весь текст в нижній регістр) і повертає всі можливі розбори слова. Ми використовуємо перше значення (саме поширене)

Деякі слова (найпоширеніший приклад в нашому тексті слово «влада») можуть бути розібрані і як іменник ("влада приймає рішення"), і як власна назва, тобто ім’я ("влад допомагав мені писати тези"). На основі однієї лише інформації про те, як слово пишеться, зрозуміти, який розбір правильний, не можна, тому аналізатор використовує в нашому прикладі перше значення.

Приклад: «думати надзвичайний стан жахливо ходити робота все добре новина сказати не дати»

Таксономія

Таксономія – це систематика класифікації речей. Таксономії є важливим інструментом як для обробки природних мов загалом, так і для конкретної аналітики тексту.

Запропонована таксономія на першому рівні поділяється на:

- вплив коронавірусу на життя людей
- статистичні повідомлення (нові випадки, статистика смертності)
- передбачення чи плани на майбутнє
- інформація про вірус
- карантин (загальна категорія, для зручного пошуку інформації саме про карантин)

Ці 5 категорії охоплюють всі твіти про минуле, майбутнє та теперішнє. Часто ці групи перетинаються, але це нормально для таких таксономій.

У категорії «Вплив коронавірусу» є підкатегорії другого рівня, що охоплюють більшість галузей життя людей, що на них повпливав вірус. А саме:

- Бізнес, індустрія
- Життя спільноти
- Економіка, фінанси
- Освіта
- Працевлаштування, робота
- Навколишнє середовище
- Галузь охорони здоров'я
- Міжнародні відносини
- Відпочинок, культура
- Закон та порядок
- Політика
- Релігія
- Спорт
- Транспорт, подорожі

У категорії «Статистичні повідомлення»:

- нові випадки
- люди, що видужали
- летальні випадки/кількість смертей
- хворі люди

У категорії «Передбачення чи плани» містити інформацію про передбачення щодо вірусу, плани влади та/чи інших організацій, та більш абстрактні речі.

У категорії «Інформація про вірус»:

- інформація про вакцини чи ліки
- симптоми
- превентивні міри

У категорії «Карантин» містяться всі твіти пов'язані з карантином.

Аналіз тексту

Аналіз частоти слів у загальному наборі речень

Один з найпростіших способів дослідження даних - це аналіз частоти. Хоча це не складно, в аналізі настроїв цей простий метод може бути напрочуд висвітлюючим.

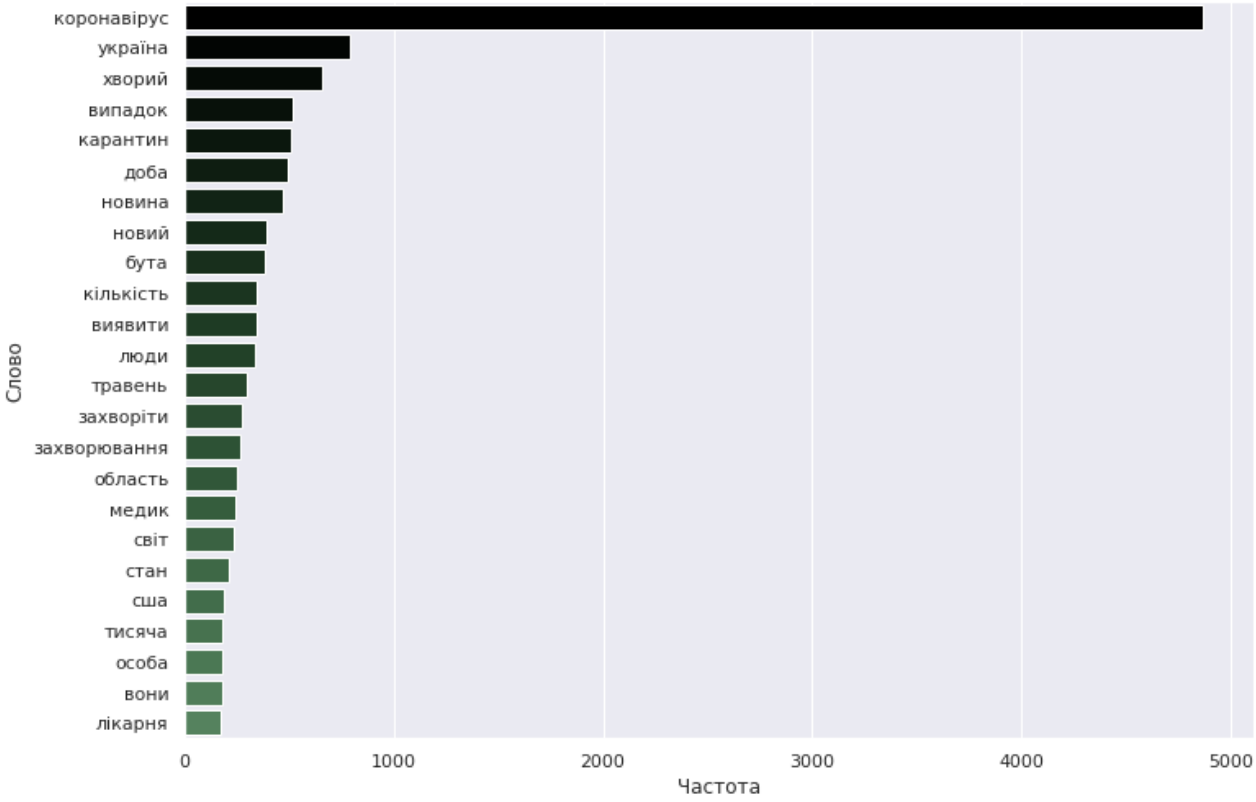
Два найчастіші слова це 'коронавірус' (4862) та 'україна' (791). Вони не представляють цінності для нас.

Наступні декілька слів з найвищою частотою:

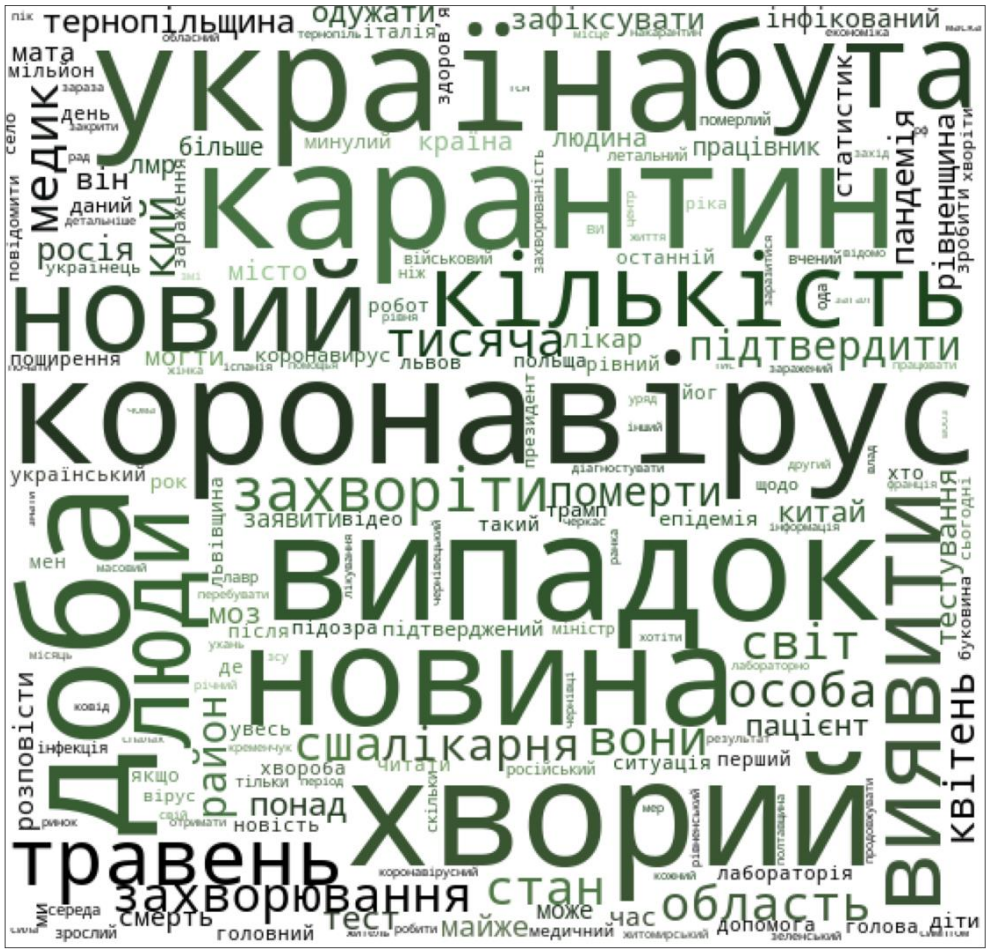
3. 'хворий' (660)
4. 'випадок' (518)
5. 'карантин' (510)
6. 'доба' (496)
7. 'новина' (473)
8. 'кількість' (348)
9. 'виявити' (346)

Познайомившись з даними стає зрозуміло, що уві ці слова використовуються для для підведення статистики. Тобто, переважно, твіти мають в собі кількість випадків за останній час, чи передбачення на травень (на момент збирання твітів велись дискусії на тему відміни карантину травні).

Візуалізація частоти слів на загальному наборі:

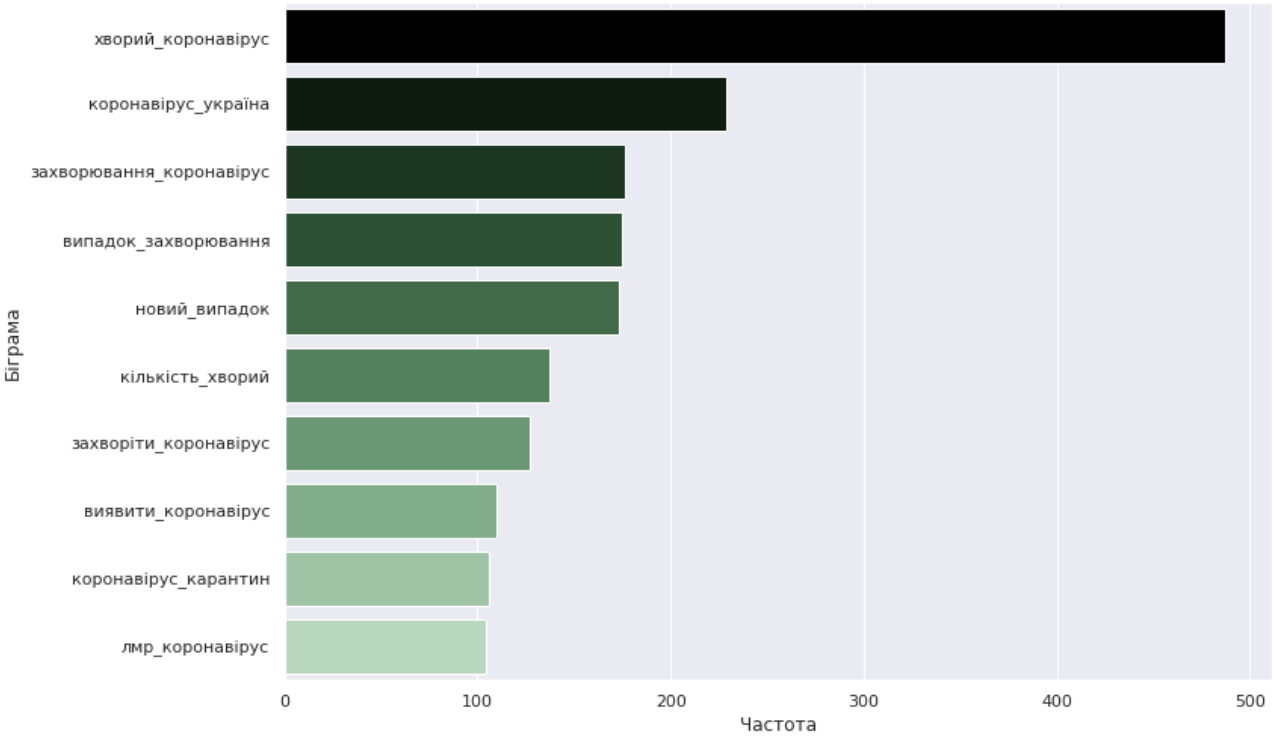


WordCloud для 100 найчастіших слів виглядає наступним чином:



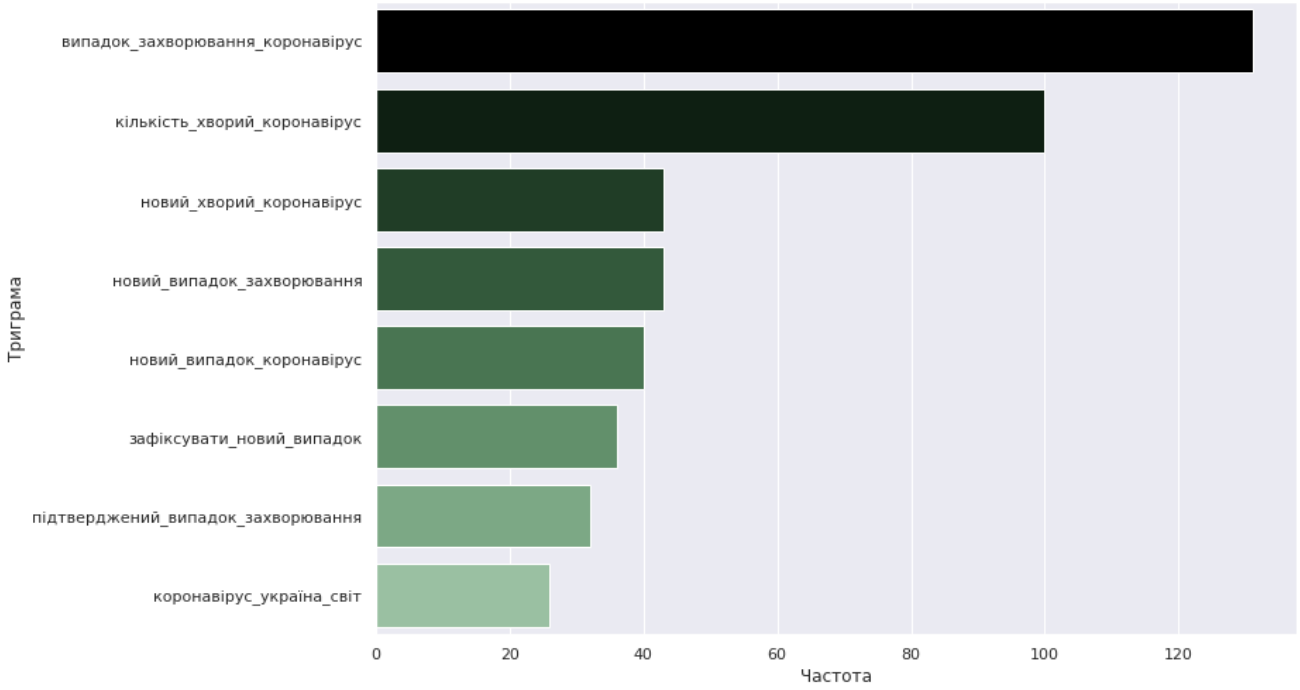
Використання біграм (2 слова) та триграм (3 слова) може дати нам кращі уявлення щодо даних, які ми аналізуємо.

Серед біграм найчастішими є:



Дані досить промовисті. Обговорення нових випадків та тестування. Також серед перших 50 біграм зустрічаються 'минулий_доба', 'карантин_коронавірус' та 'стан_травень'. Тобто люди обговорюють випадки за попередній день, карантин та перспективи на найближчий час.

Серед триграм:



Тематика така ж сама. Обговорення нових випадків, підбиття статистики.

Тобто навіть такий простий процес як підрахунок частоти слів у всьому наборі твітів уже дає зрозуміти загальну тематику текстів.

Розпізнавання іменованих утворень (NER) та аналіз в залежності від цих параметрів

Ця техніка є важливою для видобування інформації, адже намагається знайти і класифікувати іменовані сутності в неструктурованому тексті в заздалегідь визначені категорії, такі як імена людей, організації, місця, медичні коди, час, кількості, грошові значення, відсотки тощо. Тобто пов'язати елементи тексту з певними подіями, людьми чи установами.

А оцінивши емоційне забарвлення групи твітів пов'язаних з певною особою чи закладом можна зрозуміти настрої та відношення.

Отже, видобуток іменованих утворень з тексту був виконаний на не оброблених даних за допомогою Bidirectional LSTM Model (використовуючи keras для тренування) [3]. Двонаправлені LSTM - це розширення традиційних LSTM, які можуть покращити продуктивність моделі для проблем класифікації послідовностей.

У випадках, коли доступні всі часові кроки послідовності введення, двонаправлені LSTM вчать два замість одного LSTM на вхідній послідовності. Перша у вхідній послідовності як є, а друга - на перевернутій копії вхідної послідовності. Це забезпечує додатковий контекст для мережі та в більшості призводить до більш швидкого та більш повного вивчення проблеми. Для пошуку та класифікації таких іменних утворень для модель показує дуже хороші результати: 0.98 для англійської мови, >0.8 для української. Нижче значення для української мови є наслідком значно меншого тренувального набору даних та специфікою мови.

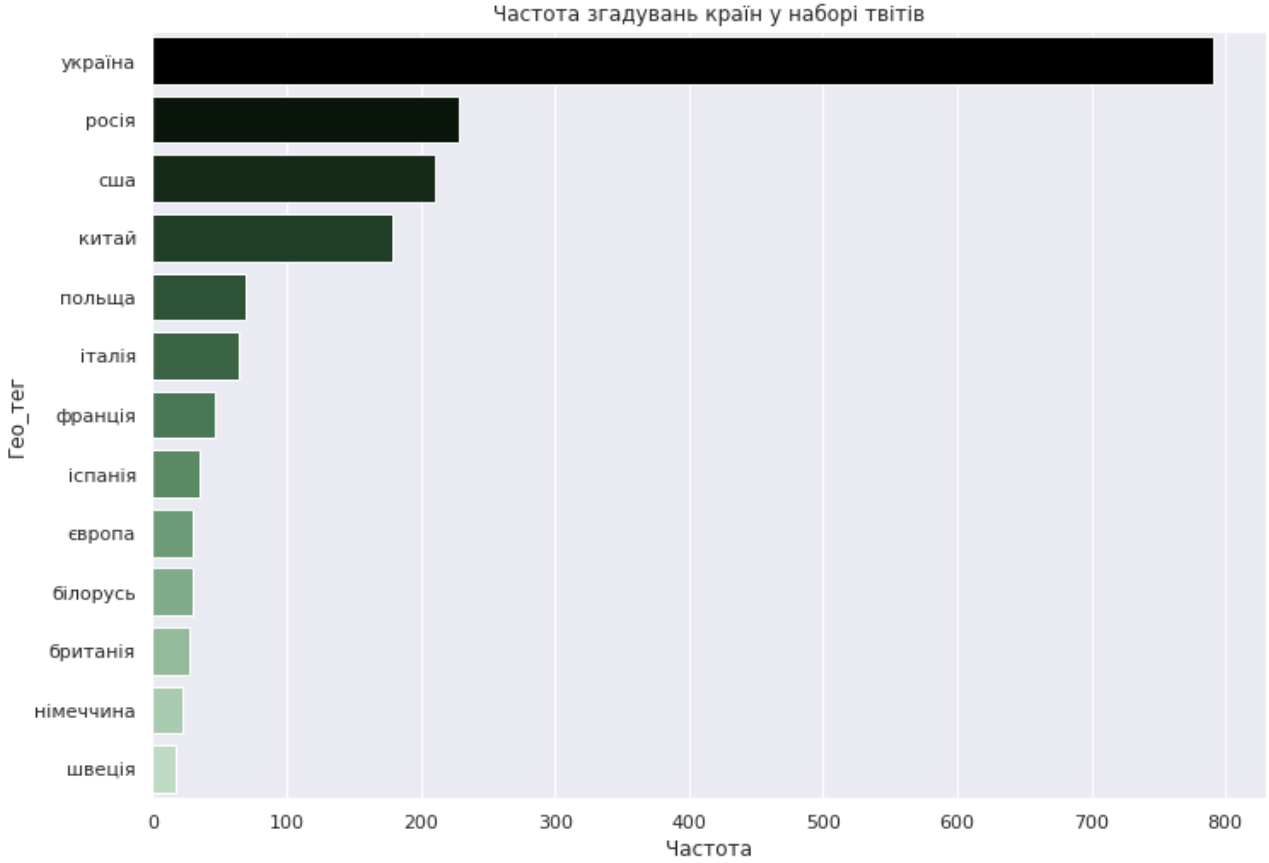
Аналіз проводиться на неочищених даних, щоб зберегти весь важливий контекст для мережі.

В результаті отрумується список слів з POS tagged і позначені Name Entity, а саме:

geo = Geographical Entity
org = Organization
per = Person
tim = Time indicator
eve = Event

Для аналізу використовуються дані geo, org та per. Для добування назв місяців з тексту немає необхідності використовувати машинне навчання.

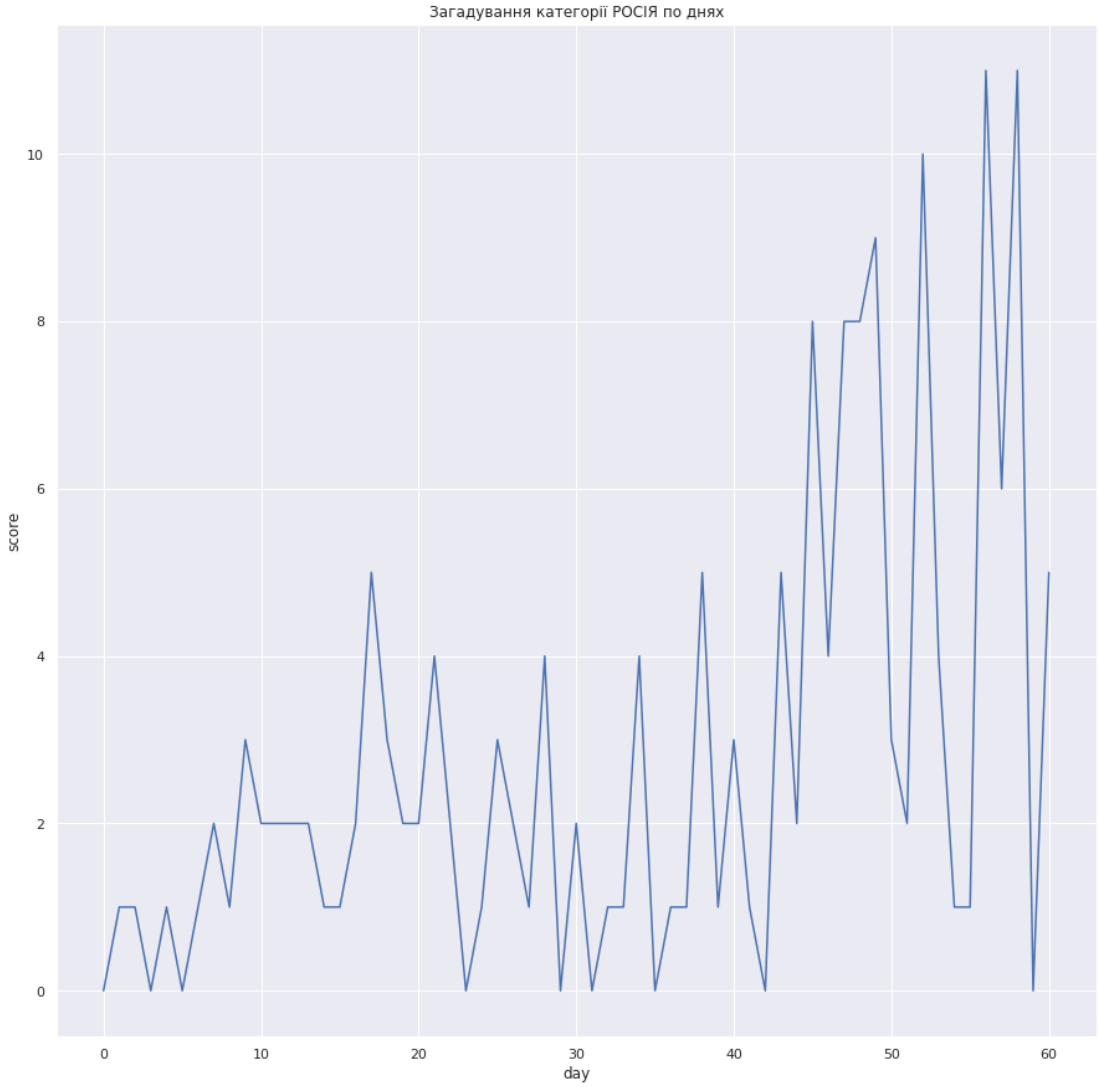
Частота згадувань країн



Частота тут залежить від 2 факторів: близькості країни до України та кількості заражень. Зрозуміло, що Україна згадується найчастіше.

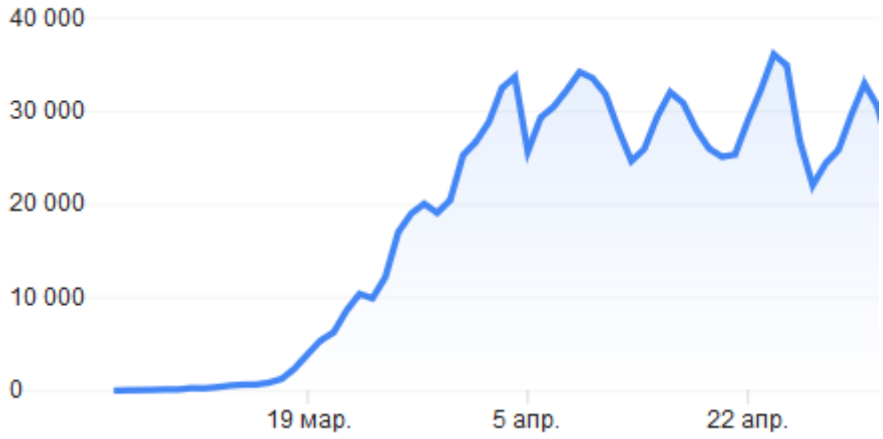
Частота згадувань певної країни має певну кореляцію з кривою нових випадків зараження. Звісно, говорити про сувору відповідність не має змісту, так як інформації недостатньо для таких рішучих висновків.

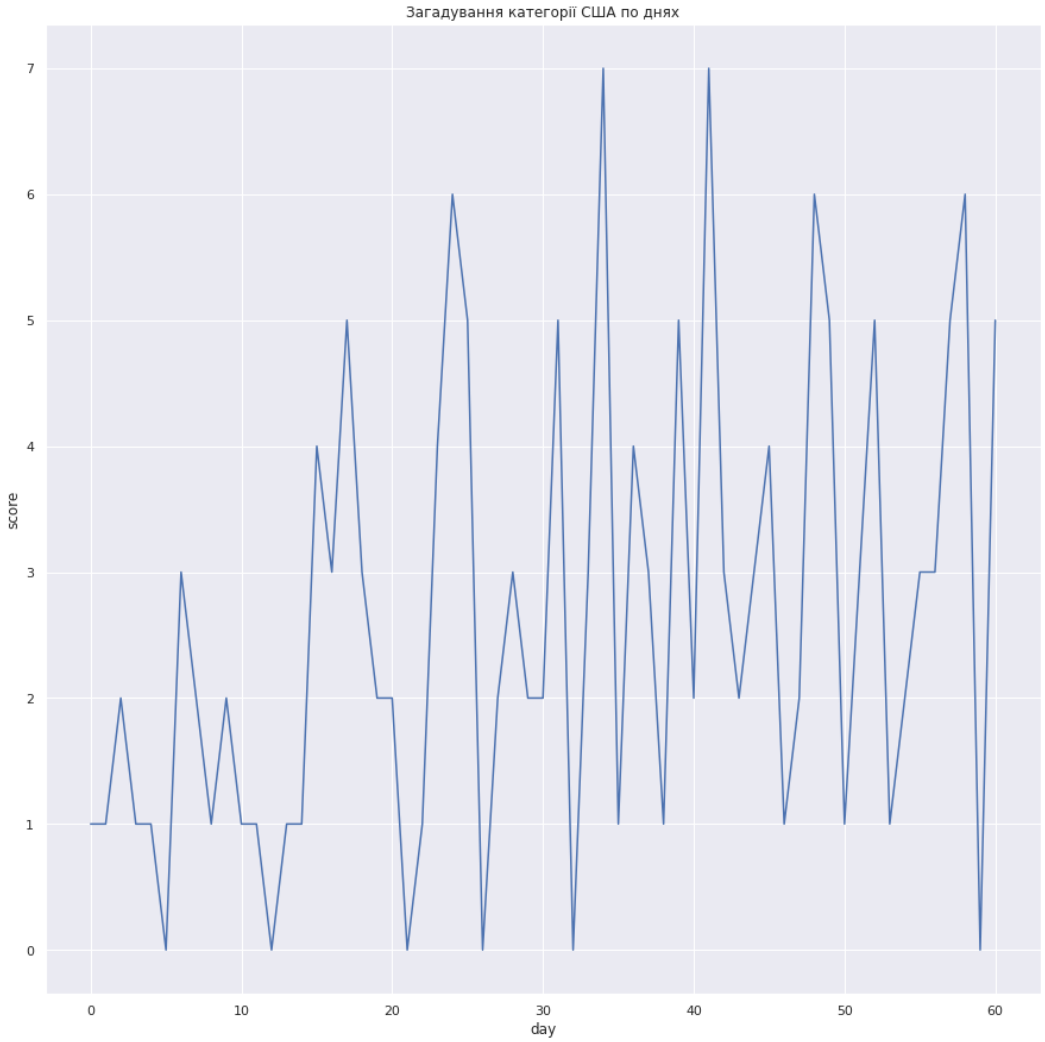
Росія



Кількість згадувань значно зросла починаючи з приблизно середини спостережень.

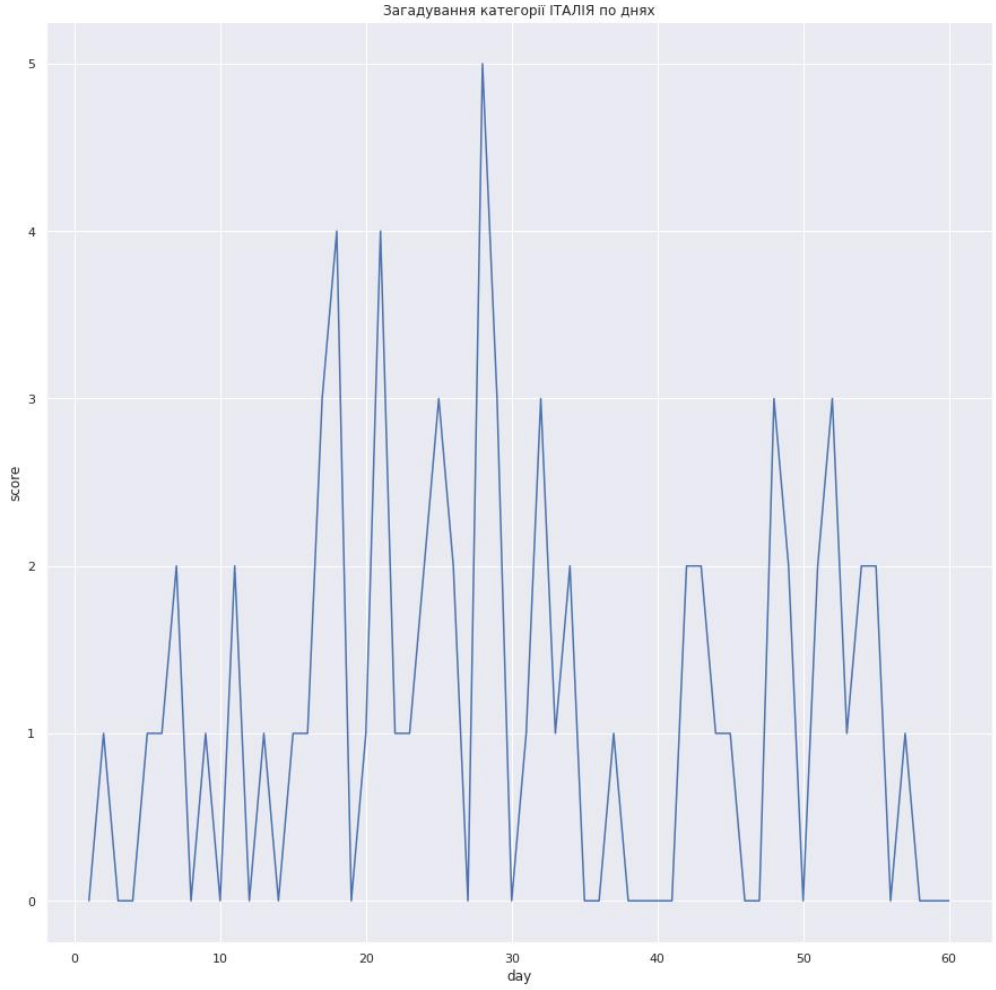
США





Тут кореляція менш помітна аніж для попереднього прикладу. Також навіть беручи до уваги масштаби епідемії у США українці приділяють менше уваги цій країні аніж сусідній Росії.

Італія



Пік захворюваності у країні приблизно сходиться з піком згадування.

Потрібно пам’ятати, що дані тут (та в інших випадках) є дуже зашумленими та не передають реальної ситуації, а лише наскільки люди обговорювали певну подію чи ідею.

Спільні згадування країн
Часто декілька з значень зустрічаються поруч і з цієї інформації можна добути трохи додаткового змісту.

В загальному, можна розділити на дві категорії.

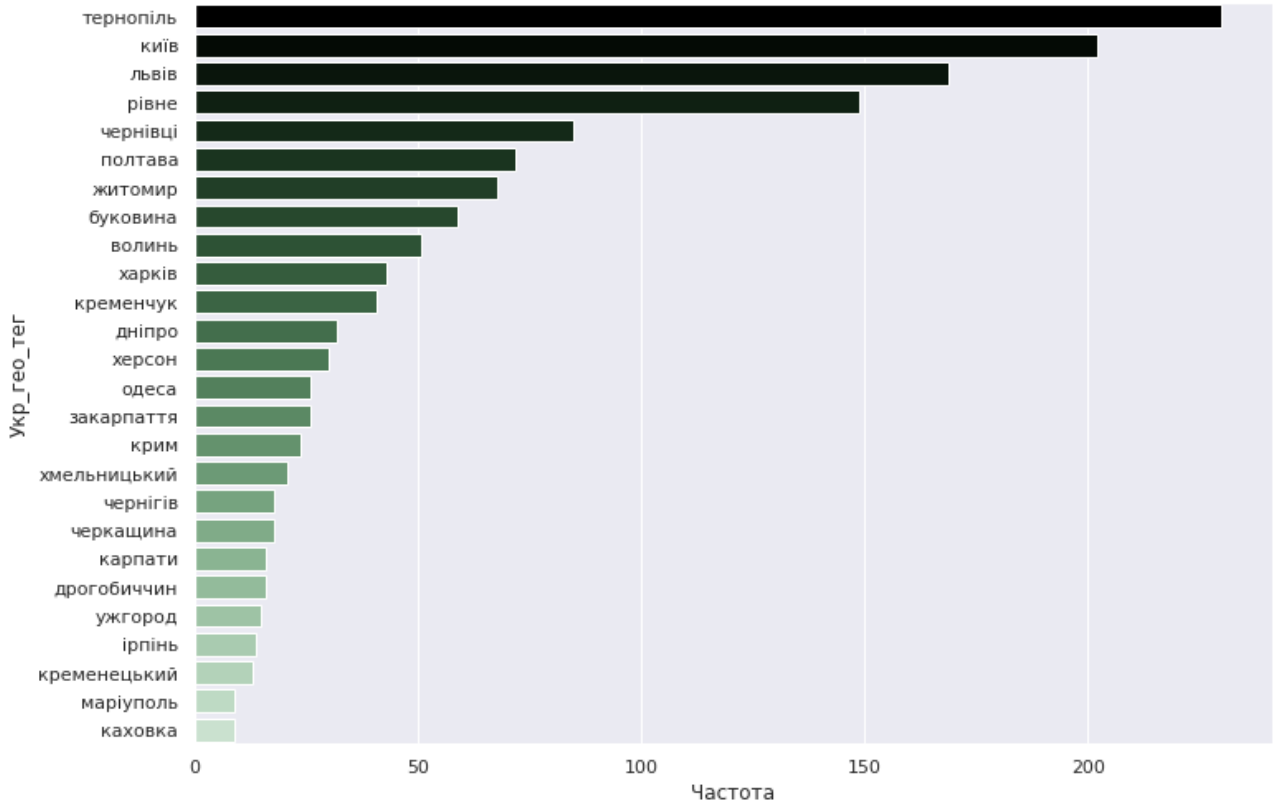
Порівняння:

- коронавірус росія доба майже стільки новий випадок україна весь час
- україна росія обігнати китай кількість хворий коронавірус
- коронавірус світ сша мітингувати рф кількість хворий зростати рекордно швидко

Взаємодія:

- тест коронавірус рф неякісний білорусь відмовитися
- медведчук поширювати україна російський фейк коронавірус держдеп сша

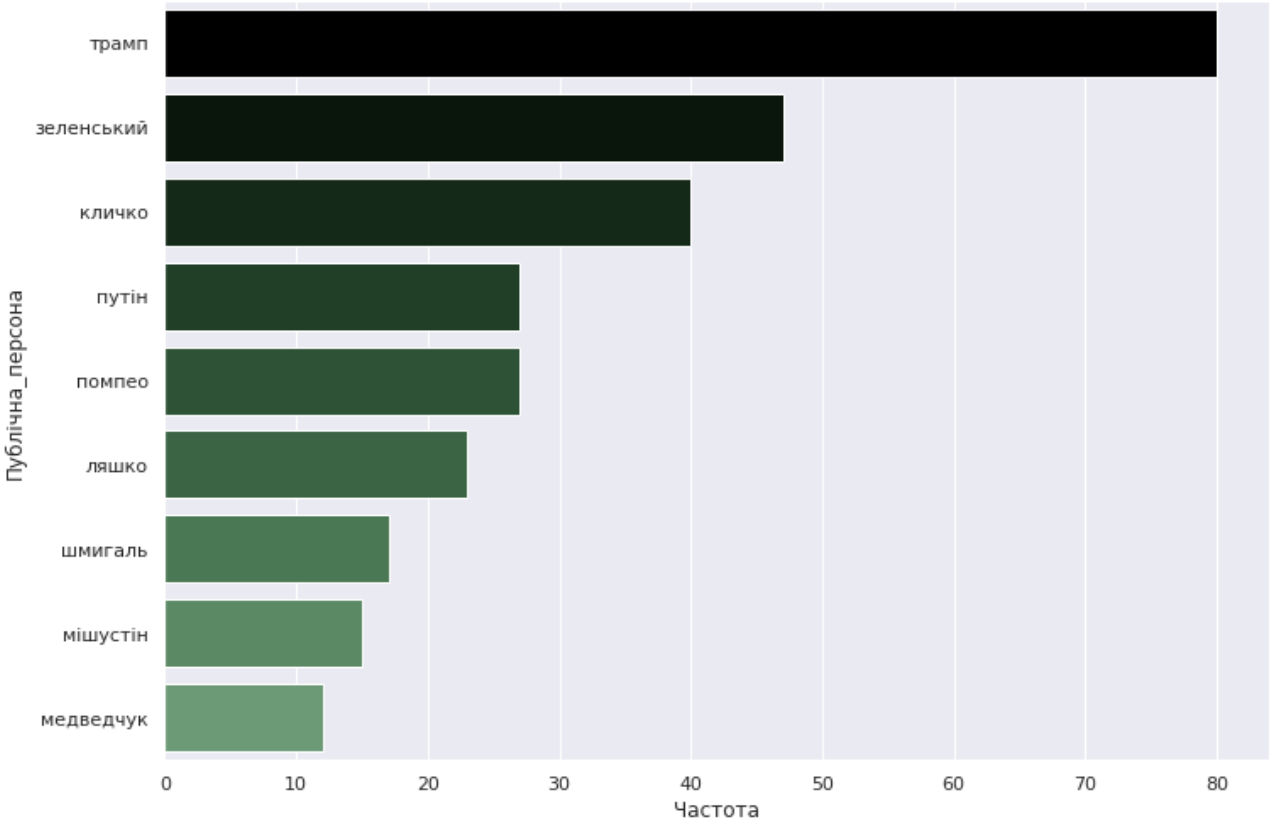
Частота згадувань регіонів України



Частота згадування різних міст має певну кореляцію з рейтингом по кількості захворювань у різних регіонах. Значення «Тернопіль» знаходиться так високо у рейтингу за частотою тому, що місцева служба новин щоранку викладає запис з статистикою.

Та аналіз часових рядів для цих значень не дають ніяких корисних результатів, що піддаються аналізу.

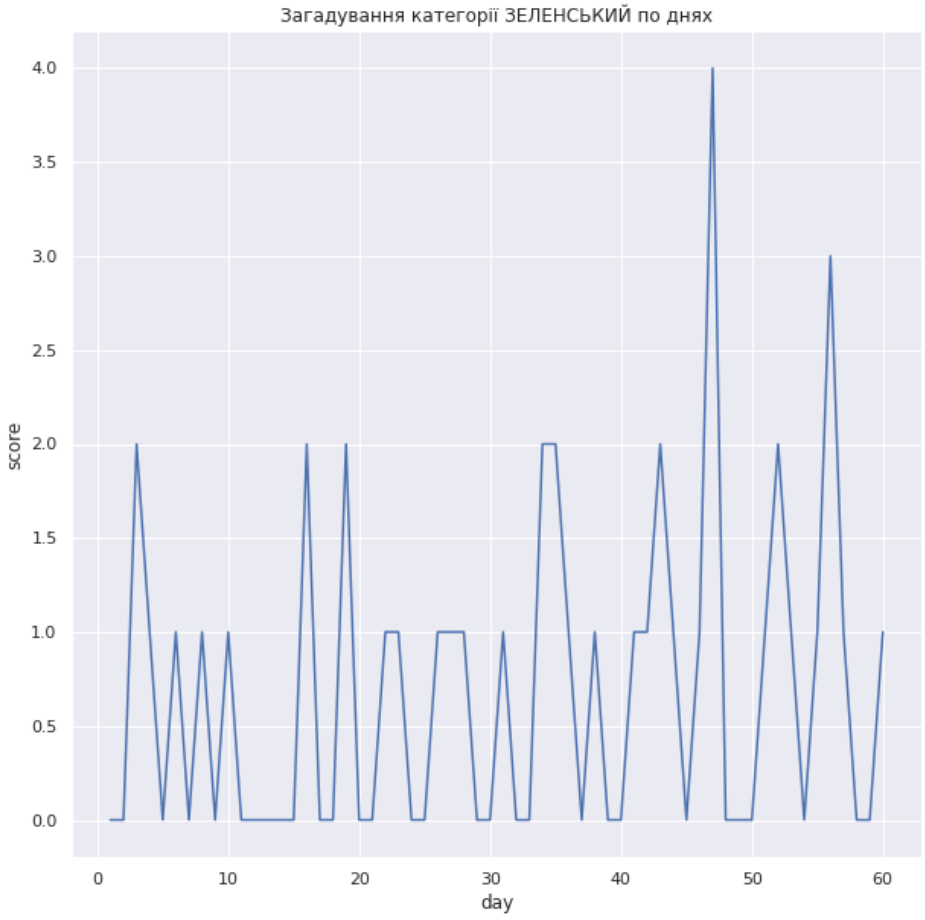
Частота згадувань громадських діячів України та світу

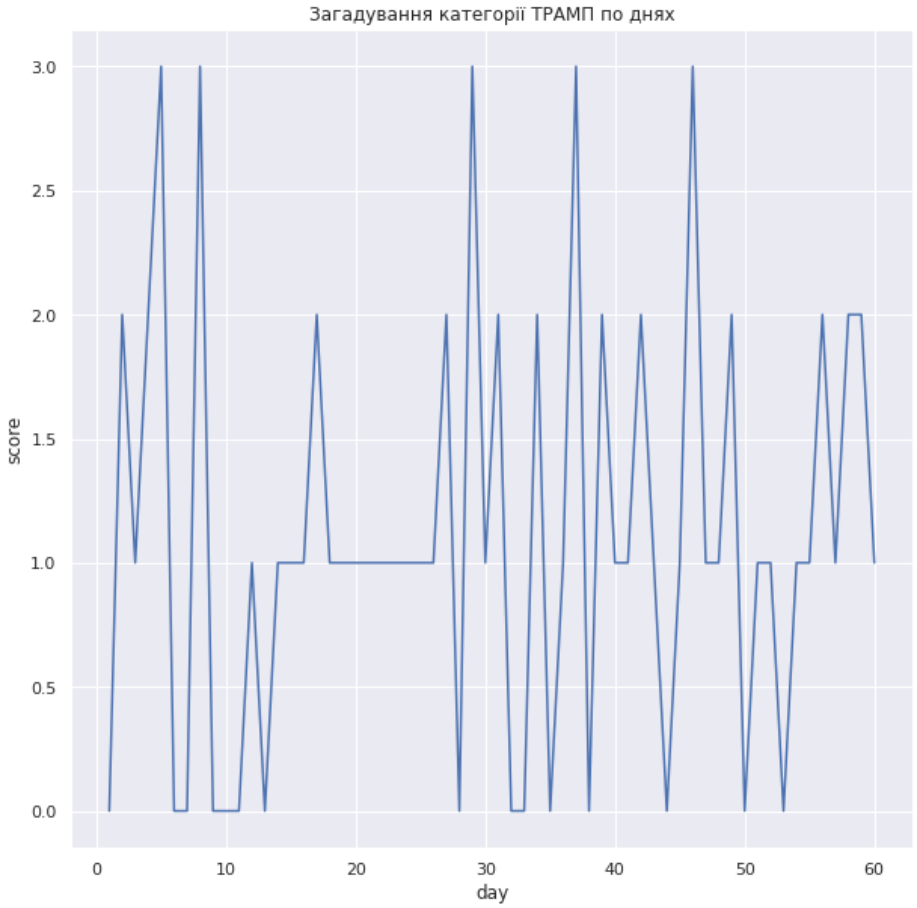


Досить цікавим є факт, що президент США зустрічається в твітах українців частіше аніж президент України.

Інші особи є визначними політичними діячами України та світу. Окрім Олега Ляшка, якому приписують ведення грамотної політики у твітері.

І Дональд Трамп і Володимир Зеленський зустрічаються в твітах з частотою, що не можна привязати до епідемії. Обоє діячів роблять досить багато заяв, та часто обговорюються чи критикуються користувачами.


Загадування категорії ЗЕЛЕНСЬКИЙ по днях
score
day
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
10
20
30
40
50
60


Загадування категорії ТРАМП по днях
score
day
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
10
20
30
40
50
60

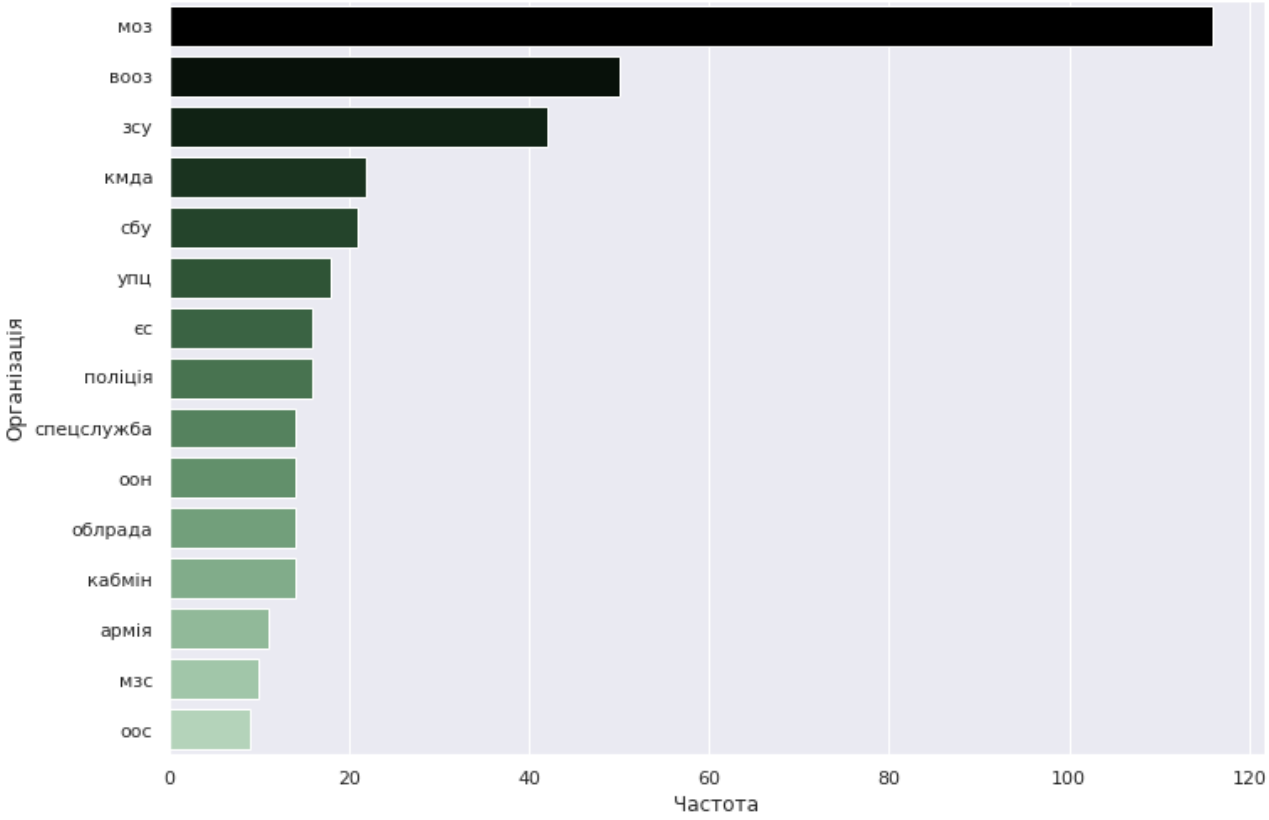
Спільні згадування діячів

Працювати з комбінаціями згадувань діячів складніше. В більшості якщо діячі згадуються поруч, то в цьому ж реченні є дуже багато нецензурної лексики та яскраво негативних фраз. Тобто в загальному такі твіти описують порівняння чи узагальнення певних осіб. Ці приклади тут приводитись не будуть.

Також мають місце твіти, що описують взаємодії певних діячів. Для прикладу:

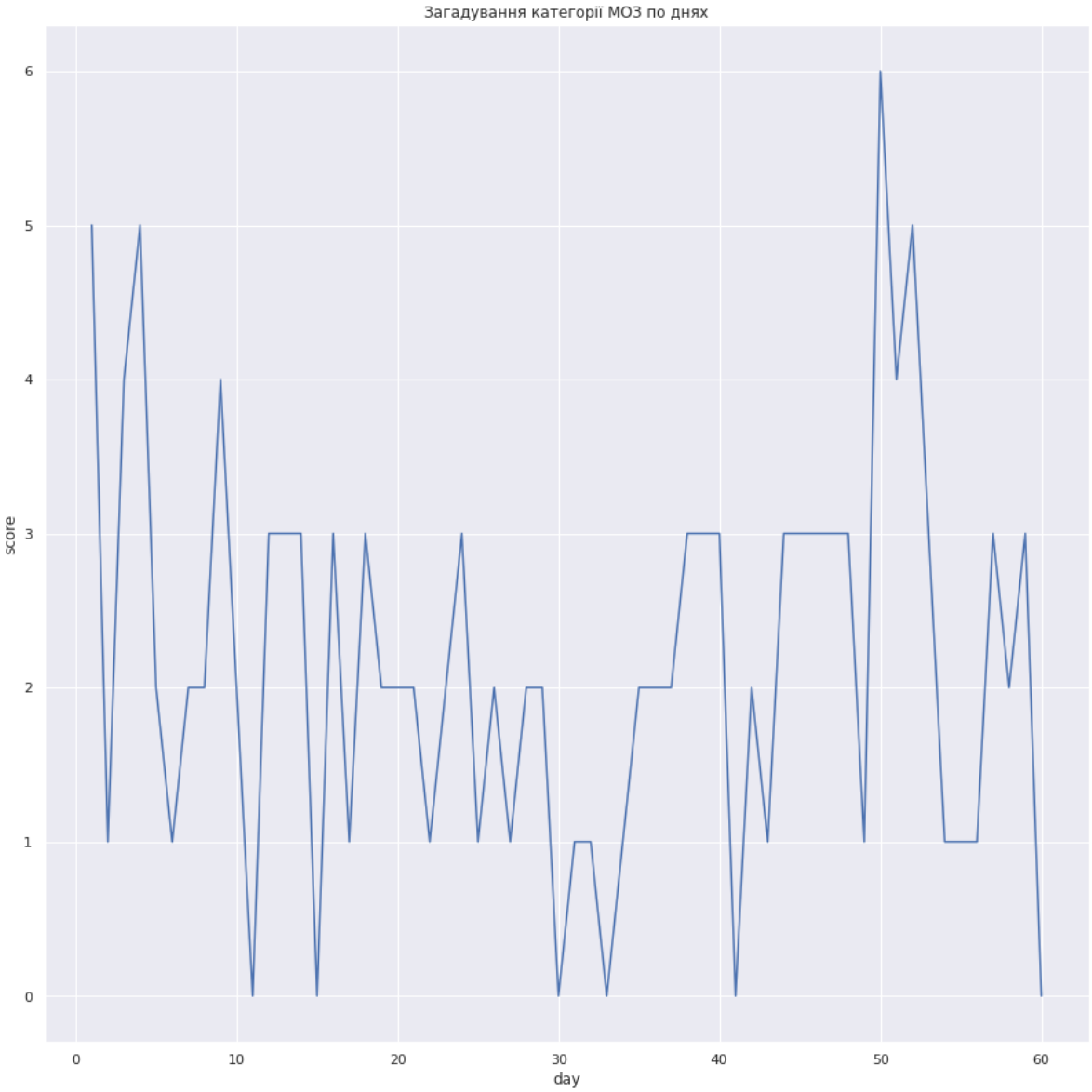
- трамп обговорити путін контроль озброєння коронавірус
- зеленський кличко обговорити карантин київ

Частота згадувань світових та українських організацій



Найвищі значення частоти мають українське міністерство охорони здоров’я та Всесвітня організація охорони здоров’я.

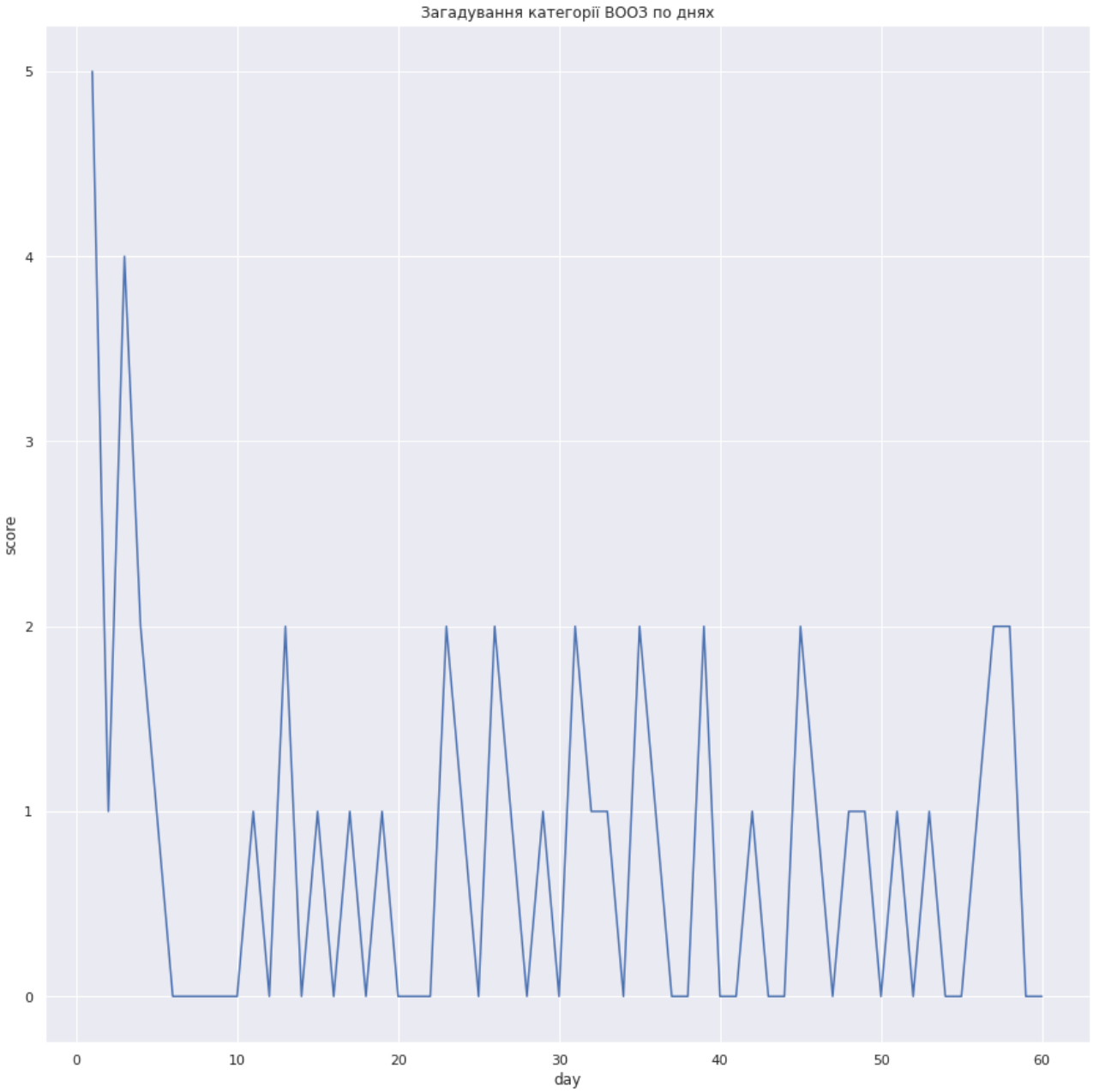
МОЗ



В кількості твітів у цій групі помітні два піки (особливо якщо спарсити дані за триваліший період часу). До 12 березня велись активні дебати і обговорення питання карантину та ситуації в Україні в загальному. На проміжку 15 лютого – 1 березня (що не є захоплений у цьому наборі даних) МОЗ зустрічається у 10 – 13 твітах на день.

Пізніше почали вестись бесіди про припинення карантину чи хоча б режиму надзвичацної ситуації. Та 24 квітня режим надзвичайної ситуації був знятий. На ці днів припадає другий пік у згадуваннях.

ВООЗ



Аналогічно для ВООЗ. Починаючи з 15 січня частота згадувань ВООЗ в контексті пандемії була вражаючи високою як для українського сегменту твіттеру. Та з часом (а саме приблизно з 11 березня, коли була оголошена пандемія) зацікавленість в ньому знизилась.

Цікавим є також факт, що був отриманий в результаті додатковоо проведеного сентимент аналізу на розширеному наборі даних (з 1 січня по 15 травня) для цієї категорії. Якщо до березня повідомлення були в більшості позитивні чи нейтральні, то пізніше вони стали рідші та більш негативні (в більшості критика діяльності організації).

Спільні згадування організацій

Здебільшого лише дві з організацій обговорюються разом. А саме, МОЗ та ВООЗ у контесті співпраці. Для прикладу:

- моз пропонувати пом’якшувати карантин тільки порада вооз україна коронавірус карантин

Також цікавою є комбінація МОЗ та УПЦ (чи просто «церква»), що зустрічається достатньо часто у контексті ігнорування другої вказівок першої.

Чому сентимент аналіз не може бути використаний в такій ситуації?

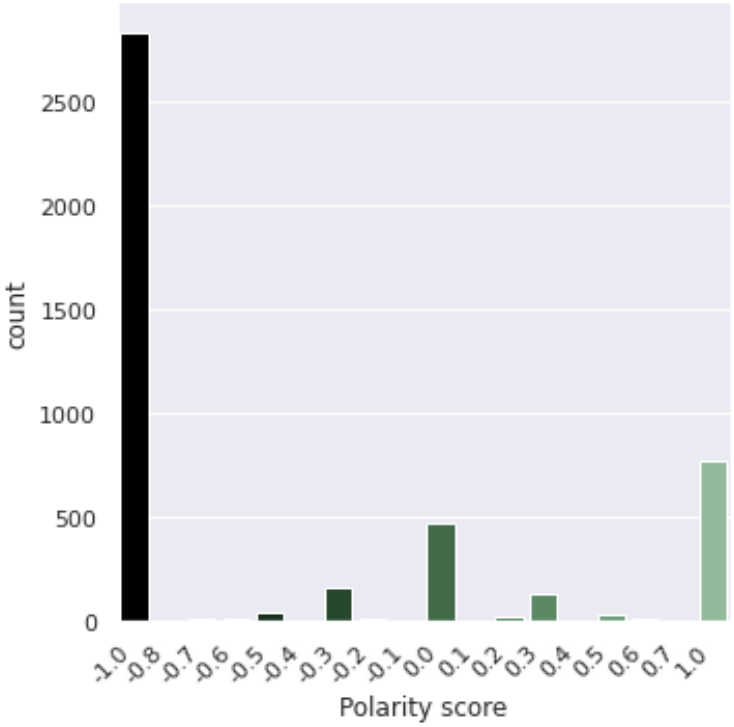
Оцінки полярності допомагають нам робити кількісні судження про почуття якогось тексту. Отже, ми класифікуємо слова з твітів на позитивні та негативні типи і даємо їм оцінку.

Оцінка полярності показує нам, якщо частина тексту має негативний, нейтральний чи позитивний тон. Використовується шкала від -1 до 1, де -1 це різко негативне значення, а 1 позитивне.

В результаті аналізу всіх текстів отримано наступні результати:

Медіана полярних значень твітів: -1.0
Мода полярних значень твітів: -1.0
Середнє полярних значень твітів: -0.4579
Стандартне відхилення полярних значень твітів: 0.7745

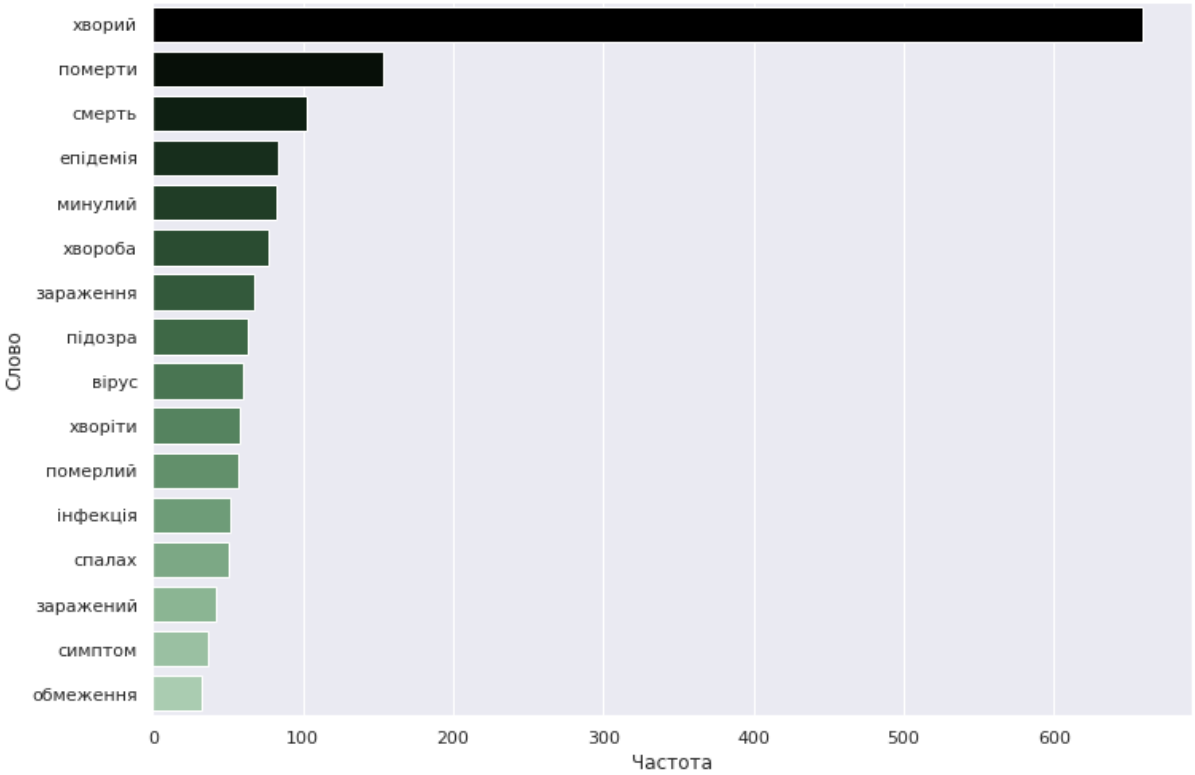
Розподіл кількості твітів виглядає наступним чином:



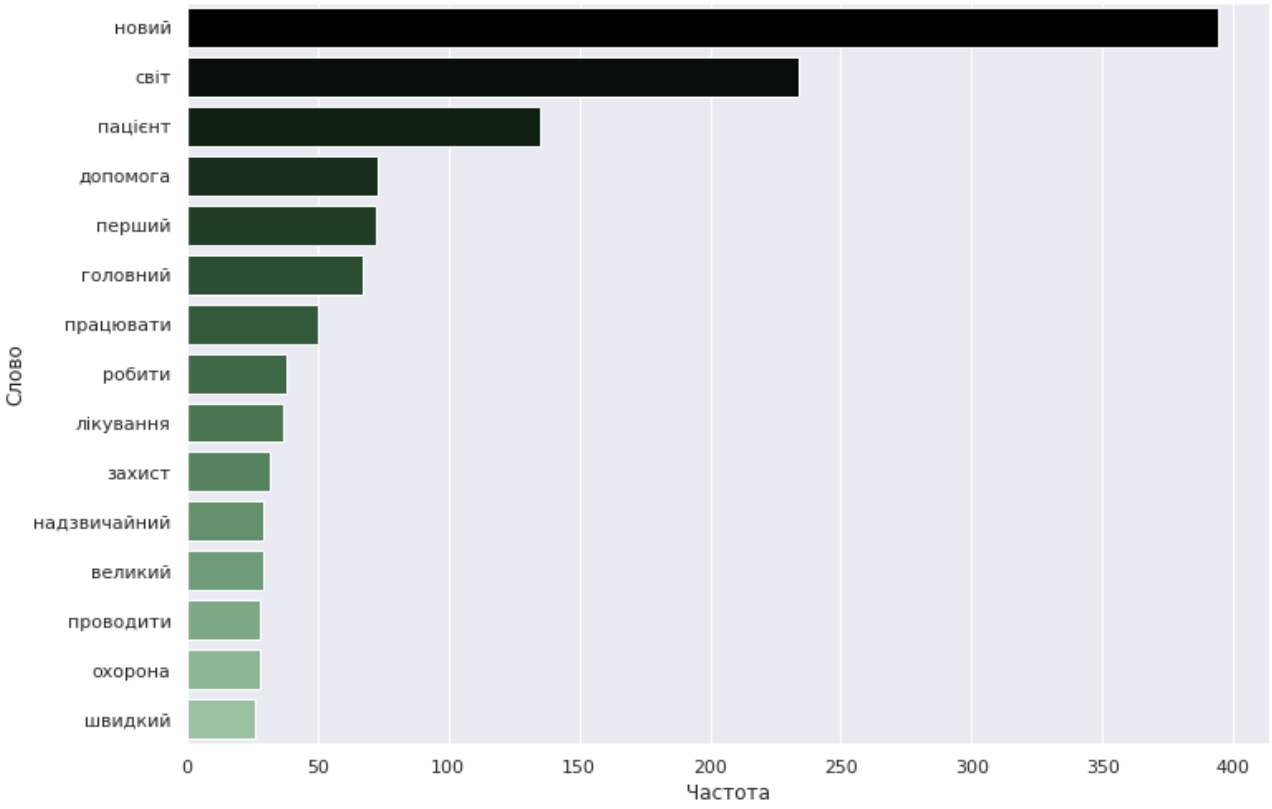
Проблема цього підходу у нашій ситуації в його «категоричності». Тобто маючи у реченні слова «вірус», «пацієнт» чи «хвороба» з більшою вірогідністю речення отримає яскраво негативний рейтинг. Хоча може в цілому йтись про те, що «пацієнт переміг хворобу» чи про винайдення якихось ліків.

Щоб проілюструвати це були створені рейтинги найчастіших негативно та позитивно забарвлених слів.

Найчастіше зустрічаються наступні слова з негативним значенням:



Найчастіше зустрічаються наступні слова з позитивним значенням:



І якщо для негативних слів класифікація ще має якийсь зміст, то з позитивних у контексті ситуації слово «новий» найчастіше використовується для повідомлення про нові випадки заражень. Інші слова мають ще менше змісту для нашої цілі.

Висновки

Що вдалось?

- Провести аналіз набору твітів.
- Використати класифікатор Naive Bayes та Bidirectional LSTM Model для аналізу тексту
- Скласти таксономію
- Виявити певну кореляцію між згадуваннями місць та статистикою захворюваності
- Виявити певну кореляцію між згадуваннями організацій та діями/заявами
- Дослідити комбінації згадування діячів, організацій та країн
- Проведені деякі додаткові дослідження для певних категорій, щоб краще зрозуміти динаміку

Що не вдалось?

- Ідея з сентимент аналізом була не дуже вдалою.
- Україномовний сегмент твіттеру є не настільки активним як хотілось би для цієї роботи.

Джерела

1. Zipf's law URL: https://en.wikipedia.org/wiki/Zipf%27s_law
2. Language Classification with Naive Bayes. URL: https://github.com/tiredwaffle/ML_projects_for_studying/blob/master/Language%20Classification%20with%20Naive%20Bayes/Language_Classification_with_Naive_Bayes.ipynb
3. Named Entity Recognition URL: https://github.com/tiredwaffle/ML_projects_for_studying/blob/master/Name%20entity%20recognition/Named_Entity_Recognition_with_Keras.ipynb