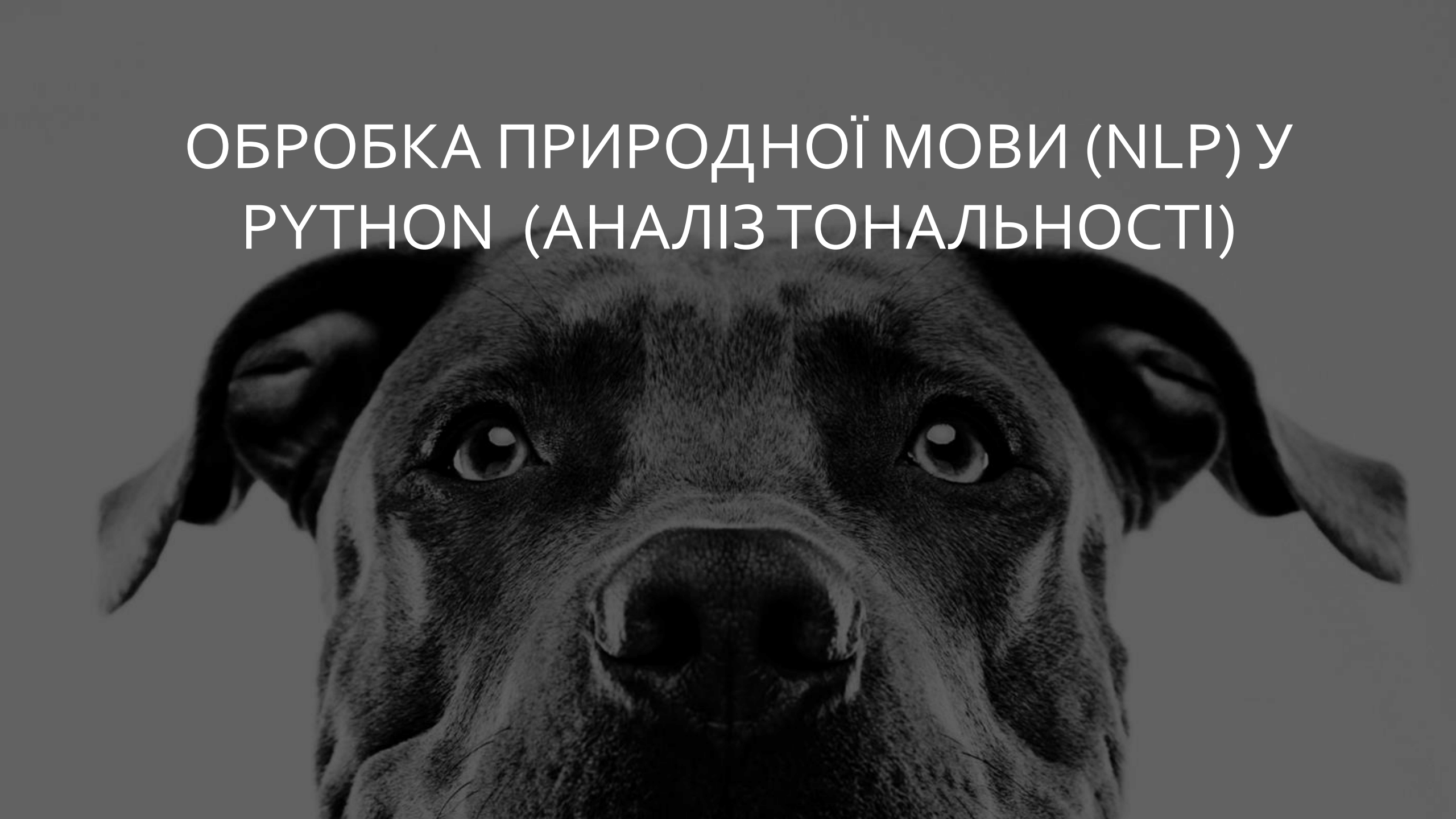
ОБРОБКА ПРИРОДНОЇ МОВИ (NLP) У PYTHON (АНАЛІЗ ТОНАЛЬНОСТІ)



ЩО ТАКЕ NLP (ОБРОБКА ПРИРОДНОЇ МОВИ)?

- Обробка природних мов (НЛП) це вивчення та застосування методів та інструментів, які дозволяють комп'ютерам обробляти, аналізувати, інтерпретувати та міркувати про людську мову. НЛП є міждисциплінарною сферою, яка поєднує в собі методи, розроблені в таких галузях, як лінгвістика та інформатика. Ці методи використовуються спільно з ШІ для створення чат-ботів і цифрових помічників, таких як Google Assistant і Alexa від Amazon.

ЧОМУ ОБРОБКА ПРИРОДНОЇ МОВИ (NLP) МАЄ ЗНАЧЕННЯ

- Обробка природної мови передбачає застосування різноманітних алгоритмів, здатних отримувати неструктуровані дані та перетворювати їх у структуровані дані. Якщо ці алгоритми застосовані неправильно, комп'ютер часто не зможе отримати правильне значення з тексту. Це часто можна побачити під час перекладу тексту між мовами, де часто втрачається точне значення речення. Хоча машинний переклад суттєво покращився за останні кілька років, помилки машинного перекладу все ще трапляються часто.

ТЕХНІКИ СИНТАКСИСУ НЛП

Приклади синтаксису:

- Лематизація
- Морфологічна сегментація
- Позначення частин мови
- Parsing
- Порушення речення
- Стерління
- Сегментація слів

ТЕХНІКИ СЕМАНТИЧНОГО НЛП

Техніки семантичного НЛП включають такі техніки, як:

- Визнання іменованої сутності
- Природне покоління мови
- Словосмислова неоднозначність

МОДЕЛІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ НЛП

- Повторювані нейронні мережі це типи нейронних мереж, які циклювати дані з попередніх часових кроків, враховуючи їх при розрахунку ваг поточного часового кроку. По суті, RNN мають три параметри, які використовуються під час прямого проходу навчання: матриця, заснована на попередньому прихованому стані, матриця, заснована на поточному вході, і матриця, яка знаходиться між прихованим станом і виходом. Оскільки RNN можуть враховувати інформацію з попередніх часових кроків, вони можуть витягувати відповідні шаблони з текстових даних, беручи до уваги попередні слова в реченні під час інтерпретації значення слова.

- Ще один тип архітектури глибокого навчання, який використовується для обробки текстових даних мережа довгострокової короткочасної пам'яті (LSTM).. Мережі LSTM подібні до RNN за структурою, але через деякі відмінності в їхній архітектурі вони, як правило, працюють краще, ніж RNN. Вони уникають специфічної проблеми, яка часто виникає під час використання RNN, яка називається проблема вибухового градієнта.

АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТІВ НА ПРИКЛАДІ НОВИН

Будемо використовувати, два інструменти NLTK:

- Інструмент «Аналіз настроїв **VADER**» (генерує позитивні, негативні і нейтральні оцінки настроїв для заданих вхідних даних).
- Інструмент токенізатора «**word_tokenize**» (розбиває великий текст на послідовність більш дрібних одиниць, таких як речення або слова).

ЩОБ ВИКОРИСТОВУВАТИ **VADER** І **WORD_TOKENIZE,** НАМ СПОЧАТКУ ПОТРІБНО ЗАВАНТАЖИТИ І ВСТАНОВИТИ ДОДАТКОВІ ДАННІ ДЛЯ **NLTK.**

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('vader_lexicon')
```

VADER ГОТОВИЙ ДО АНАЛІЗУ БЕЗ БУДЬ-ЯКОЇ СПЕЦІАЛЬНОГО НАЛАШТУВАННЯ. **VADER** УНІКАЛЬНИЙ ТИМ, ЩО ЧІТКО РОЗРІЗНЯЄ ПОЗИТИВНІСТЬ І НЕГАТИВНІСТЬ РІЗНОГО СТУПЕНЯ.

ОБЧИСЛИТИ ТОНАЛЬНІСТЬ ЗАГОЛОВКУ

```
!pip install feedparser
import feedparser

posts = []
rss_url='https://www.pravda.com.ua/ukr/rss/view_news/'
response = feedparser.parse(rss_url)
for each in response['entries']:
  if each['title'] in [x['title'] for x in posts]:
    pass
  else:
    posts.append({
        "title": each['title'],
        "link": each['links'][0]['href'],
        "tags": [x['term'] for x in each['tags']],
        "date": time.strftime('%Y-%m-%d', each['published_parsed'])
        })

for i, post in enumerate(p['title'] for p in posts):
  print(i, post)
```

```
0 В ЄС закликали Туреччину переглянути її вихід зі Стамбульської конвенції
1 У Києві виявили ще 434 зараження Covid, 28 хворих померли
2 За добу майже 4 тисячі українців отримали вакцину проти коронавірусу
3 Туреччина знову почала бомбити курдів у Сирії – ЗМІ
4 Завдяки вакцинації локдауни підуть в минуле – прогноз BioNTech
5 Початок тижня в Україні буде морозним та вітряним
6 Коронавірус: в Україні більше 11 тисяч нових заражень, 166 людей померли
7 Головні новини суботи і ночі: акція під ОП, Київ та Одещина "почервоніли"
8 Окупанти на Донбасі вели вогонь з забороненої "Мінськом" зброї
9 В школах ОРДЛО розповідають дітям про "майбутню військову агресію України" –
10 Україна опинилася між Туреччиною і Німеччиною за кількістю заражень COVID-19
11 Греція безкоштовно роздаватиме набори для тестування на COVID-19
12 Силовики побилися з протестувальниками біля відділку, де тримали учасника ак
13 Порушення "тиші" на Донбасі скоротилися майже втричі за час перемир'я – ТКГ
14 США попередили про санкції проти Індії – вона купила російські ЗРК
15 ЄС планує запровадити "паспорти вакцинації" влітку
16 У Зеленського вважають, що вандали під ОП хотіли спровокувати насильство
17 В Японії знову стався землетрус, є постраждалі
18 У МВС відреагували на події під Офісом президента
19 Другий за добу корабель ВМС США з ракетами "Томагавк" увійшов у Чорне море
```

ЗАВАНТАЖЕМО ТОНАЛЬНИЙ СЛОВНИК УКРАЇНСЬКОЇ МОВИ

```
import requests
import csv

url = 'https://raw.githubusercontent.com/lang-uk/tone-dict-
uk/master/tone-dict-uk.tsv'
r = requests.get(url)
with open(nltk.data.path[0]+'/tone-dict-uk.tsv', 'wb') as f:
    f.write(r.content)

d = {}
with open(nltk.data.path[0]+'/tone-dict-uk.tsv', 'r') as csv_file:
    for row in csv.reader(csv_file, delimiter='\t'):
        d[row[0]] = float(row[1])

from nltk.sentiment.vader import SentimentIntensityAnalyzer
SIA = SentimentIntensityAnalyzer()

SIA.lexicon.update(d)
```

АЛГОРИТМ VADER ВИВОДИТЬ ОЦІНКИ НАСТРОЇВ ДЛЯ 4 КЛАСІВ НАСТРОЇВ :

- neg: Негативний
- neu: Нейтральний
- pos: Позитивний
- compound: Складний (Сукупний бал)

ПОРАХУЄМО ОЦІНКУ НАСТРОЮ ДЛЯ НАДАНИХ ЗАГОЛОВКІВ ЗА ДОПОМОГОЮ **VADER**

```
for i, post in enumerate(p['title'] for p in posts):
  print(i, post, SIA.polarity_scores(post)["compound"])
```

```
0 В ЄС закликали Туреччину переглянути її вихід зі Стамбульської конвенції 0.0
1 У Києві виявили ще 434 зараження Covid, 28 хворих померли -0.25
2 За добу майже 4 тисячі українців отримали вакцину проти коронавірусу 0.0
3 Туреччина знову почала бомбити курдів у Сирії – ЗМІ -0.25
4 Завдяки вакцинації локдауни підуть в минуле – прогноз BioNTech 0.0
5 Початок тижня в Україні буде морозним та вітряним 0.0
6 Коронавірус: в Україні більше 11 тисяч нових заражень, 166 людей померли 0.0
7 Головні новини суботи і ночі: акція під ОП, Київ та Одещина "почервоніли" 0.0
8 Окупанти на Донбасі вели вогонь з забороненої "Мінськом" зброї 0.0
9 В школах ОРДЛО розповідають дітям про "майбутню військову агресію України" – активісти 0.0
10 Україна опинилася між Туреччиною і Німеччиною за кількістю заражень COVID-19 0.0
11 Греція безкоштовно роздаватиме набори для тестування на COVID-19 0.0
12 Силовики побилися з протестувальниками біля відділку, де тримали учасника акції під ОП 0.0
13 Порушення "тиші" на Донбасі скоротилися майже втричі за час перемир'я – ТКГ -0.4588
14 США попередили про санкції проти Індії – вона купила російські ЗРК 0.0
15 ЄС планує запровадити "паспорти вакцинації" влітку 0.0
16 У Зеленського вважають, що вандали під ОП хотіли спровокувати насильство -0.4588
17 В Японії знову стався землетрус, є постраждалі 0.0
18 У МВС відреагували на події під Офісом президента 0.0
19 Другий за добу корабель ВМС США з ракетами "Томагавк" увійшов у Чорне море 0.0
```

- Як видно з принту результату обробки — багато позицій визначено як нейтральна тональність (0.0), це через те, що ми не виконали попередньої підготовки тексту. Попередня обробка тексту використовується для поліпшення роботи алгоритмів. Тож, виконаємо очищення від стоп-слів та приведемо слова в нормальну форму.
- у NLTK, поки немає корпусу української мови, то для мормологічного аналізу скористаємось **pymorphy2** (наступний слайд)

```
url = 'https://raw.githubusercontent.com/olegdubetcky/Ukrainian-
Stopwords/main/ukrainian'
r = requests.get(url)
with open(nltk.data.path[0]+'/corpora/stopwords/ukrainian', 'wb') as
f:
    f.write(r.content)
# Retrieve HTTP meta-data
print(r.status_code)
print(r.headers['content-type'])
print(r.encoding)

import string
from nltk.corpus import stopwords
stopwords = stopwords.words("ukrainian")

!pip install git+https://github.com/kmike/pymorphy2.git
!pip install -U pymorphy2-dicts-uk

import pymorphy2
morph = pymorphy2.MorphAnalyzer(lang='uk')

stop_words = frozenset(stopwords+list(string.punctuation))
for i, post in enumerate(p['title'] for p in posts):
  sentences = nltk.sent_tokenize(post)
  for sentence in sentences:
    words = nltk.word_tokenize(sentence)
    without_stop_words = [word for word in words if not word in
stop_words]
    normal_words=[]
    for token in without_stop_words:
      p = morph.parse(token)[0]
      normal_words.append(p.normal_form)

    print(i, post, "RAW: ", SIA.polarity_scores(post)
["compound"],"NORM: ", SIA.polarity_scores(' '.join(normal_words))
["compound"])
```

РЕЗУЛЬТАТ

```
0 В ЄС закликали Туреччину переглянути її вихід зі Стамбульської конвенції RAW:  0.0 NORM:  0.25
1 У Києві виявили ще 434 зараження Covid, 28 хворих померли RAW:  -0.25 NORM:  -0.7184
2 За добу майже 4 тисячі українців отримали вакцину проти коронавірусу RAW:  0.0 NORM:  0.0
3 Туреччина знову почала бомбити курдів у Сирії – ЗМІ RAW:  -0.25 NORM:  -0.25
4 Завдяки вакцинації локдауни підуть в минуле – прогноз BioNTech RAW:  0.0 NORM:  0.0
5 Початок тижня в Україні буде морозним та вітряним RAW:  0.0 NORM:  0.0
6 Коронавірус: в Україні більше 11 тисяч нових заражень, 166 людей померли RAW:  0.0 NORM:  -0.4588
7 Головні новини суботи і ночі: акція під ОП, Київ та Одещина "почервоніли" RAW:  0.0 NORM:  0.0
8 Окупанти на Донбасі вели вогонь з забороненої "Мінськом" зброї RAW:  0.0 NORM:  -0.4588
9 В школах ОРДЛО розповідають дітям про "майбутню військову агресію України" – активісти RAW:  0.0 NORM:  -0.25
10 Україна опинилася між Туреччиною і Німеччиною за кількістю заражень COVID-19 RAW:  0.0 NORM:  -0.25
11 Греція безкоштовно роздаватиме набори для тестування на COVID-19 RAW:  0.0 NORM:  0.0
12 Силовики побилися з протестувальниками біля відділку, де тримали учасника акції під ОП RAW:  0.0 NORM:  -0.7184
13 Порушення "тиші" на Донбасі скоротилися майже втричі за час перемир'я – ТКГ RAW:  -0.4588 NORM:  -0.4588
14 США попередили про санкції проти Індії – вона купила російські ЗРК RAW:  0.0 NORM:  0.0
15 ЄС планує запровадити "паспорти вакцинації" влітку RAW:  0.0 NORM:  0.0
16 У Зеленського вважають, що вандали під ОП хотіли спровокувати насильство RAW:  -0.4588 NORM:  -0.7906
17 В Японії знову стався землетрус, є постраждалі RAW:  0.0 NORM:  0.0
18 У МВС відреагували на події під Офісом президента RAW:  0.0 NORM:  0.0
19 Другий за добу корабель ВМС США з ракетами "Томагавк" увійшов у Чорне море RAW:  0.0 NORM:  0.0
```