УДК 004.4

Роман Іващенко, Ольга Рєзіна

АНАЛІЗ ТОНАЛЬНОСТІ ТА ОБ'ЄКТИВНОСТІ ТЕКСТУ ЗАСОБАМИ МОВИ ПРОГРАМУВАННЯ PYTHON

(студент І курсу другого (магістерського) рівня вищої освіти факультету української філології, іноземних мов та соціальних комунікацій)
Науковий керівник – кандидат педагогічних наук, доцент Рєзіна О. В.

Стаття присвячена технології створення інтерактивного веб-додатка для аналізу тональності та об'єктивності тексту. Досліджені шляхи реалізації багатомовного сентимент-аналізу та визначені можливості візуалізації його результатів.

Ключові слова: *інтерактивний веб-додаток, бібліотека TextBlob, фреймворк Flask, хостинг-платформа Pythonanywhere, мультилінгвальний аналіз.*

Постановка проблеми. Думки відіграють критичну роль у нашому житті та мають великий вплив на людську поведінку. Наші переконання та сприйняття реальності, а також рішення, які ми приймаємо, починаючи від вибору нового телефону чи фільму для перегляду сьогодні ввечері до чогось важливішого, значною мірою залежать від того, як інші сприймають і оцінюють світ. Тому ми часто цікавимося думками інших, коли нам самим потрібно щось вирішити. Це стосується не лише окремих осіб, а й цілих організацій. Тому не дивно, що галузь обробки природної мови зараз набуває великого значення, зокрема й аналіз тональності текстів, який дає змогу визначити ставлення мовця до об'єкта висловлювання.

Аналіз тональності тексту (також *сентимент-аналіз*; англ. *sentiment analysis, opinion mining*) – це метод обробки природної мови (NLP) у комп'ютерній лінгвістиці, що використовується для автоматизованого виявлення емоційно забарвленої лексики та оцінювання ставлення автора тексту до об'єктів, про які йде мова [6]. Хоча сентимент-аналіз насамперед зосереджується на полярності тексту (позитивна чи негативна), він також має за мету виявити конкретні почуття та емоції (злість, радість, сум тощо), рівень суб'єктивності висловлювання, терміновість запитів (терміново або не дуже), рівень зацікавленості та навіть наміри користувачів [7].

Такий аналіз часто виконується на текстових даних, щоб допомогти компаніям відстежувати ставлення до бренду або продукту у відгуках клієнтів і дописах у соціальних мережах, а також краще розуміти їхні потреби. Залежно від того, як компанія хоче інтерпретувати відгуки та запити користувачів, вона можете самостійно визначати та налаштувати категорії відповідно до потреб аналізу настроїв [3].

Тому розробка програм, що дають змогу визначати тональність тексту, є актуальною. У запропонованій статті розглядається технологія створення інтерактивного веб-додатка для мультилінгвального аналізу тональності та об'єктивності тексту.

Аналіз досліджень і публікацій. Проблема аналізу тональності текстів привертає до себе увагу багатьох вітчизняних та зарубіжних дослідників, адже такий аналіз має безліч практичних застосувань, як-от визначення думки споживачів про продукт або бренд, що допомагає краще розуміти їхні бажання й приймати відповідні бізнес-рішення. Цій темі присвячені дослідження Немеш О., Теслюка В. [3], Лідді Е. [10], Шлера Д. [9] та інших. Зокрема Романишин М. та Романюк А. у своїй роботі розглядали процес створення тонального словника української мови на основі сентимент-анотованого корпусу [4].

Мета статті. Мета статті полягає у висвітленні особливостей створення інтерактивного веб-додатка для багатомовного сентимент-аналізу й візуалізації його результатів.

Виклад основного матеріалу. Сентимент-аналіз насамперед зосереджується на полярності тексту (позитивна чи негативна), але він також має за мету виявити конкретні почуття та емоції (наприклад, радість або сум), рівень суб'єктивності висловлювання, зацікавленості, терміновості запитів та навіть наміри користувачів [7]. За допомогою такого аналізу компанії можуть відстежувати ставлення клієнтів до бренду або продукту у відгуках і дописах у соціальних мережах, а отже, корегувати свою маркетингову стратегію. Інструменти для аналізу дають змогу самостійно визначати та налаштовувати його параметри й категорії відповідно до потреб [3].

Щоб розробити веб-додаток, необхідно дібрати потужні інструменти. Одним із таких засобів є Python-бібліотека TextBlob. Вона надає простий API для вирішення найпоширеніших завдань обробки природної мови, як-от розмічування частин мови, аналіз тональності та суб'єктивності текстів, їх класифікація, переклад тощо [20]. TextBlob базується на бібліотеках NLTK і Pattern та надає зручний і простий інтерфейс, зберігаючи потужний функціонал. Однією з переваг цієї бібліотеки є те, що її об'єкти схожі за структурою й використанням на звичайні рядки Python (string), тому користувач може працювати з ними за схожим принципом [12].

На першому етапі створення веб-додатка для мультилінгвального сентимент-аналізу було протестовано програму SENTIALIZER, яка дає змогу аналізувати тексти, наявні на інтернет-ресурсах. SENTIALIZER створено засобами Python, фреймворку Flask, бібліотеки Requests для отримання всієї інформації з веб-сторінки за запитом [17], BeautifulSoup для синтаксичного аналізу отриманих даних і виокремлення видимого тексту [31], а також TextBlob для безпосереднього обчислення тональності та суб'єктивності отриманого тексту [20].

На другому етапі була здійснена модифікація програми SENTIALIZER та розширено її функціонал. За допомогою SENTIALIZER тепер можна здійснювати аналіз не лише веб-сторінок, але й будь-якого довільного тексту, а також файлів Excel, вибраних користувачем. Також удосконалено її інтерфейс: числове відображення значень результатів аналізу замінено на словесне для тональності та відсоткове для рівня суб'єктивності. Додано 5 значень тональності тексту: *Very Negative, Negative, Neutral, Positive, Very Positive*. Для узгодження значень додамо такий програмний код до головного класу Analyze:

```
if self.overall.polarity > 0.6:
   self.very_positive_polarity = 'Very Positive'
elif self.overall.polarity > 0.10:
   self.positive_polarity = 'Positive'
elif self.overall.polarity < -0.10:
   self.negative_polarity = 'Negative'
elif self.overall.polarity < -0.6:
   self.very_negative_polarity = 'Very Negative'
else:
   self.neutral_polarity = 'Neutral'
```

Сучасні веб-додатки для аналізу тональності текстів здебільшого підтримують лише одну мову. Ми розглянули шляхи реалізації багатомовного аналізу та візуалізації отриманих результатів. Зрештою було обрано метод із залученням машинного перекладу за допомогою бібліотеки Googletrans, яка використовує програмний інтерфейс Google Перекладача. Хоча цей метод не такий точний, як попередньо укладені тональні словники для кожної мови окремо, він дає змогу охопити одразу 108 мов і позбавляє необхідності повністю переписувати програму та змінювати наявні інструменти.

Суть процесу полягає в тому, що програма перекладає отриманий від користувача текст англійською мовою, після чого й відбувається аналіз тональності. Для машинного

перекладу, як уже зазначалося вище, було залучено Googletrans – безкоштовну Python-бібліотеку, яка підтримує Google Translate API, що дає змогу виконувати як переклад, так і автоматичне виявлення мови тексту оригіналу [33].

Додамо визначення мови та переклад у разі необхідності до кожної з функцій додатка SENTIALIZER – аналізу веб-сторінок, тексту та файлів. Оскільки суть процесу не змінюється, розглянемо лише приклад із функцією аналізу тексту користувача.

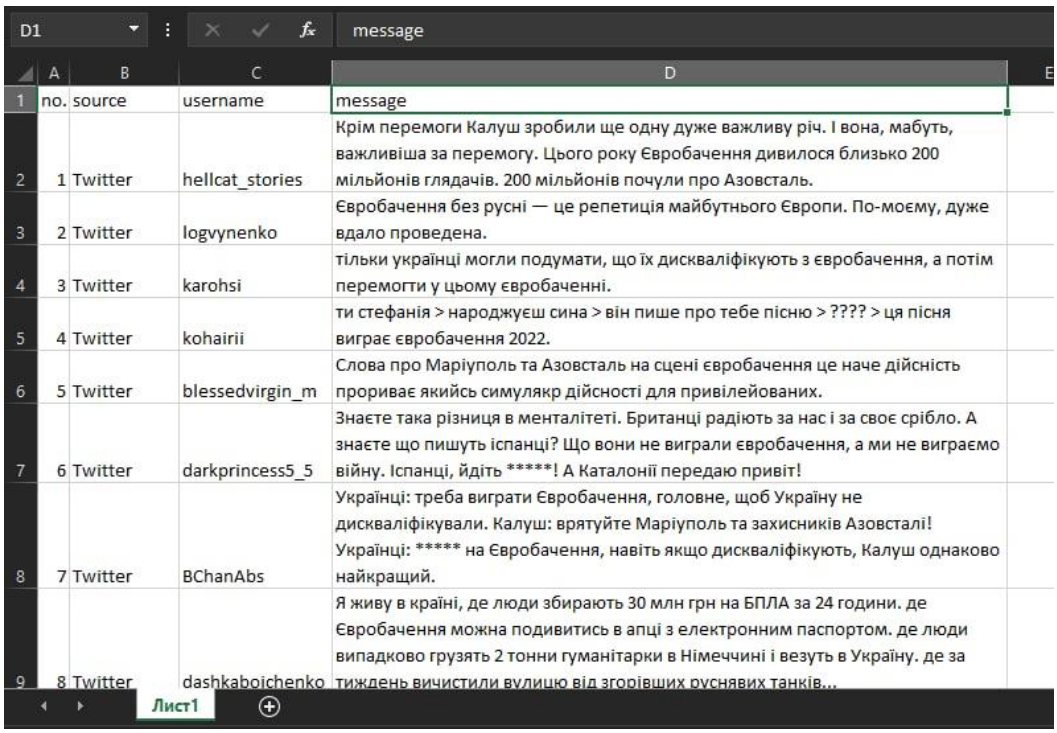
```
input_text = request.form['usertext']
detected_language = translator.detect(input_text)
detected_language = detected_language.lang
if detected_language == 'en':
   pass
else:
   inp_list = tokenize.sent_tokenize(input_text)
   inp_list_translated = []
   for sentence in inp_list:
      sentence = translator.translate(sentence, dest='en')
      sentence_translated = sentence.text
      inp_list_translated.append(sentence_translated)
```

Спочатку програма визначає мову тексту оригіналу. Якщо вона не англійська, то створюються два списки: з реченнями мовою оригіналу й мовою перекладу окремо. Аналіз виконується англійською, але на сторінці з результатами в полях вказуються речення мовою оригіналу.

Візуалізація результатів аналізу реалізована за допомогою діаграм. Для їх створення нам знадобляться бібліотеки Pandas і Matplotlib. Для роботи з діаграмами потрібно створити списки, у яких зберігатимуться значення полярності та суб’єктивності для кожного речення та код для їх отримання. Ці дані записуються у файл output.csv, що дає змогу Pandas і Matplotlib побудувати правильну діаграму.

Було проведено тестування створеного веб-додатка. Для аналізу ми вибрали матеріали з різних актуальних джерел, як-от дописи в сервісі мікроблогінгу Twitter, коментарі на відеохостингу YouTube, новини з телеграм-каналу «Суспільне Кропивницький», а також німецькомовні статі з веб-сайту Deutsche Welle.

Для початку було проаналізовано дописи україномовних користувачів у Twitter за запитом «Євробачення». Варто зазначити, що в рамках дослідження збережено оригінальний текст дописів. Для аналізу організуємо тексти в таблицю Excel, де також зазначено псевдоніми авторів (рис. 1.).



	A	B	C	D
1	no.	source	username	message
2	1	Twitter	hellcat_stories	Крім перемоги Калуш зробили ще одну дуже важливу річ. І вона, мабуть, важливіша за перемогу. Цього року Євробачення дивилося близько 200 мільйонів глядачів. 200 мільйонів почули про Азовсталь.
3	2	Twitter	logvynenko	Євробачення без русні — це репетиція майбутнього Європи. По-моєму, дуже вдало проведена.
4	3	Twitter	karohsi	тільки українці могли подумати, що їх дискваліфікують з євробачення, а потім перемогти у цьому євробаченні.
5	4	Twitter	kohairii	ти стефанія > народжуєш сина > він пише про тебе пісню > ???? > ця пісня виграє євробачення 2022.
6	5	Twitter	blessedvirgin_m	Слова про Маріуполь та Азовсталь на сцені євробачення це наче дійсність прориває якийсь симулякр дійсності для привілейованих.
7	6	Twitter	darkprincess5_5	Знаєте така різниця в менталітеті. Британці радіють за нас і за своє срібло. А знаєте що пишуть іспанці? Що вони не виграли євробачення, а ми не виграємо війну. Іспанці, йдіть *****! А Каталонії передаю привіт!
8	7	Twitter	BChanAbs	Українці: треба виграти Євробачення, головне, щоб Україну не дискваліфікували. Калуш: врятуйте Маріуполь та захисників Азовсталі! Українці: ***** на Євробачення, навіть якщо дискваліфікують, Калуш однаково найкращий.
9	8	Twitter	dashkaboichenko	Я живу в країні, де люди збирають 30 млн грн на БПЛА за 24 години. де Євробачення можна подивитись в апці з електронним паспортом. де люди випадково грузять 2 тонни гуманітарки в Німеччині і везуть в Україну. де за тиждень вичистили вулицю від згорівших руснявих танків...

Рис. 1. Структура файлу з дописами користувачів

Тепер можна скористатися функцією веб-додатка для аналізу файлів. Вона розроблена так, що одиницею аналізу вважається вміст клітинки з повідомленням, а не кожне речення окремо, як у випадку з аналізом довільних текстів та веб-сторінок. Для тестування візьмемо 10 дописів. Результат наведено нижче (рис. 2).

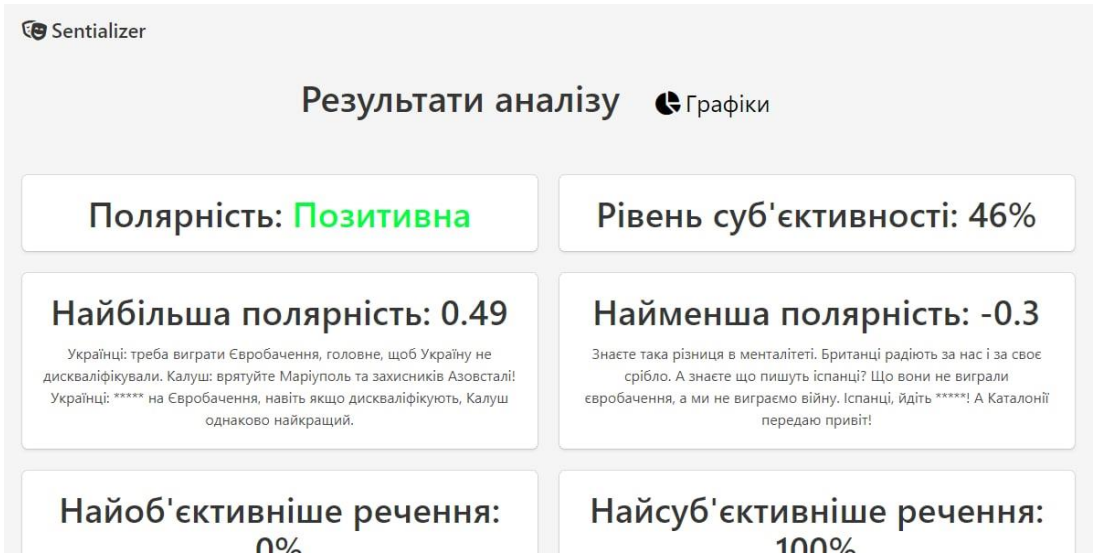


Рис. 2. Результати аналізу файлу з таблицею Excel

Більшість повідомлень мають позитивну чи нейтральну тональність, хоча є й дописи з негативним відтінком. Розглянемо створені діаграми за результатами аналізу (рис. 3).

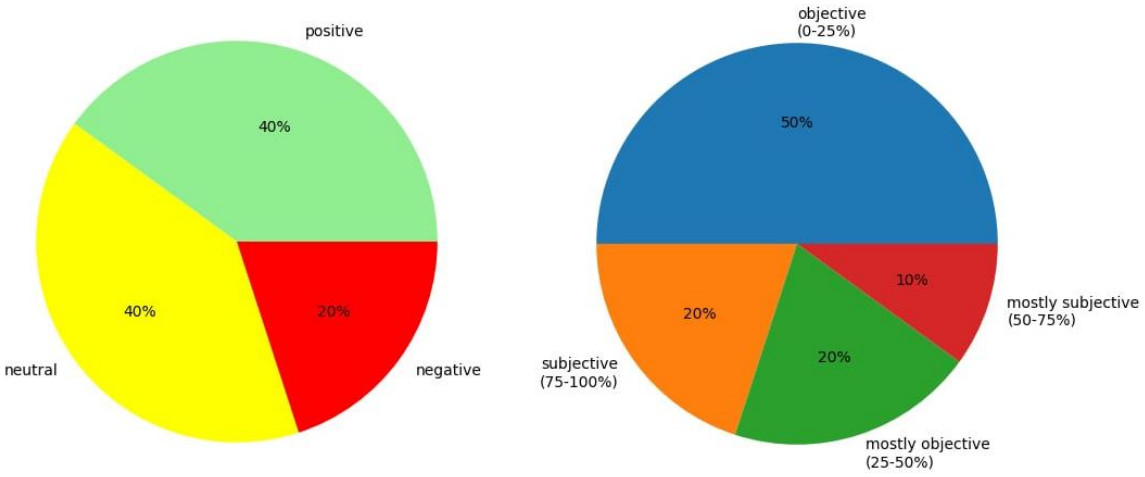


Рис. 3. Діаграми за результатами аналізу дописів користувачів

Діаграми тональності та суб'єктивності текстів будуються, як уже зазначалося, на основі файлу, куди записуються значення для кожної клітинки таблиці Excel окремо. Після автоматичного перекладу введеного тексту аналіз відбувається англійською. Для користувача це відбувається непомітно, адже завдяки словнику, що створюється в процесі аналізу, речення мовою оригіналу та перекладу знову зіставляються. Тож поряд зі значеннями найбільш та найменш полярних і суб'єктивних речень на сторінці з результатами відображатиметься текст мовою оригіналу.

Програма розміщена на хостингу Pythonanywhere, тож веб-додаток SENTIALIZER можна знайти за посиланням https://sentializer.pythonanywhere.com/.

Висновки та перспективи подальших пошуків у напрямі дослідження. Створений веб-додаток працює коректно і може бути використаний для потреб будь-якого користувача. Перспективи подальших досліджень вбачаються в удосконаленні роботи веб-додатка мультилінгвального сентимент-аналізу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Данилюк І. Г. Технологія автоматичного визначення тематики тексту [Текст] / І. Г. Данилюк // Лінгвістичні студії: зб. наук. пр. Вип. 17 / уклад.: Анатолій Загнітко (наук. ред.) та ін. – Донецьк : ДонНУ, 2008. – С. 290–293.
2. Іванов О. В. Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності / Іванов Олег Валерійович // Вісник Харківського національного університету імені В. Н. Каразіна, Харків: Видавничий центр ХНУ імені В. Н. Каразіна, 2013. — № 1045. — С.72
3. Немеш О., Романюк А., Теслюк В. Аналіз тональності тексту: основні поняття та приклади застосування. – 2015.
4. Романишин М., Романюк А. Тональний словник української мови на основі сентимент-анотованого корпусу / М. Романишин, А. Романюк, // Українське мовознавство . - 2013. - Вип. 43. - с. 63-74.
5. Adam McQuistan. Building a Text Analytics App in Python with Flask, Requests, BeautifulSoup, and TextBlob. : веб-сайт. URL: https://thecodinginterface.com/blog/text-analytics-app-with-flask-and-textblob/
6. Bing Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions – Cambridge University Press; 1 edition – 2015, 383 c.

7. Cambria, E., Das, D., Bandyopadhyay, A Practical Guide to Sentiment Analysis, Springer – 2017, 199 с.
8. Gaël Guibon, Magalie Ochs, Patrice Bellot. From Emojis to Sentiment Analysis. WACAI 2016, Lab-STICC; ENIB; LITIS, Jun 2016, Brest, France. – https://hal-amu.archives-ouvertes.fr/hal-01529708
9. Koppel, Moshe; Schler, Jonathan (2006). "The Importance of Neutral Examples for Learning Sentiment". Computational Intelligence 22. pp. 100–109.
10. Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. — P.1
11. Multilingual sentiment and subjectivity analysis – Rada Mihalcea /// C Banea, R Mihalcea, J Wiebe – Multilingual Natural Language Processing
12. Natural Language Processing for Beginners: Using TextBlob. : веб-сайт. URL: https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/
13. Natural Language Toolkit. NLTK Documentation. : веб-сайт. URL: https://www.nltk.org/
14. Pang, Bo; Lee, Lillian (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". Proceedings of the Association for Computational Linguistics (ACL).
15. Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
16. Python for NLP: Introduction to the TextBlob Library. Usman Malik. : веб-сайт. URL: https://stackabuse.com/python-for-nlp-introduction-to-the-textblob-library/
17. Requests: HTTP for Humans. Requests Documentation. : веб-сайт. URL: https://docs.python-requests.org/en/latest/
18. Snyder, Benjamin; Barzilay, Regina (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL).
19. Taboada, Maite; Brooke, Julian (2011). "Lexicon-based methods for sentiment analysis". Computational Linguistics. 37 (2): 272–274.
20. TextBlob: Simplified Text Processing. TextBlob Documentation. : веб-сайт. URL: https://textblob.readthedocs.io/en/dev/
21. The importance of neutral examples for learning sentiment [Електронний ресурс] : [Веб-сайт]. – Електронні дані. – Rita McCue, Jonathan Schler – 21.10.2005 – http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.9735
22. Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.
23. J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3):165–210, 2005.
24. R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In Proceedings of the Association for Computational Linguistics, Prague, Czech Republic, 2007.
25. C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii, 2008.
26. X. Wan. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008.
27. J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of the 6th International Conference on Intelligent Text

Processing and Computational Linguistics (CICLing-2005) (invited paper), Mexico City, Mexico, 2005.
28. C. Banea, R. Mihalcea, and J. Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In Proceedings of the Learning Resources Evaluation Conference (LREC 2008), Marrakech, Morocco, 2008.
29. S.-M. Kim and E. Hovy. Identifying and analyzing judgment opinions. In Proceedings of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics, New York City, NY, 2006.
30. Pandas documentation (2022). : веб-сайт. URL: https://pandas.pydata.org/pandas-docs/stable/index.html
31. Beautiful Soup 4.9.0 documentation. Beautiful Soup Documentation. : веб-сайт. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
32. Pang, Bo; Lee, Lillian (2008). "4.1.2 Subjectivity Detection and Opinion Identification". Opinion Mining and Sentiment Analysis. Now Publishers Inc.
33. Googletrans 3.0.0 documentation. : веб-сайт. URL: https://py-googletrans.readthedocs.io/en/latest/
34. Pattern 3.6 Documentation : веб-сайт. URL: https://pypi.org/project/Pattern/