Міністерство освіти і науки України
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"
Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Звіт

з лабораторної роботи №8 з дисципліни
«Аналіз даних в інформаційних системах»

„Аналіз текстів”

Виконав(ла) *ІП-11 Тарасьонок Дмитро Євгенович*
(шифр, прізвище, ім'я, по батькові)

Перевірив *Олійник Ю. А.*
(прізвище, ім'я, по батькові)

Київ 2023

ЗМІСТ

1 МЕТА ЛАБОРАТОРНОЇ РОБОТИ

Мета роботи – ознайомитись з методами аналізу текстів.

2 ЗАВДАННЯ

2.1 Основне завдання

Дані для виконання: текстові дані у форматі csv-файлів або дані з відкритих джерел (телеграм-канали, RSS-канали тощо). Приклад даних за посиланням

1. Нормалізація та попередня обробка даних.
2. провести очищення текстових даних від стоп-слів/тегів/розмітки;
3. виконати токенізацію текстових елементів;
4. провести лематизацію текстових елементів (можна використати бібліотеку Spacy - приклад роботи за посиланням). Зберегти результат в окремий файл.
5. Створити Bag of Words для всіх нормалізованих слів. Зберегти результат в окремий файл.
6. Порахувати метрику TF-IDF для 10 слів, що найчастіше зустрічаються в корпусі;

2.2 Додаткове завдання

2.2.1 Інтелектуальний аналіз текстів (+1 бал):

- провести сантимент аналіз (визначення емоційної тональності – позитивний / негативний) для даних ukr_text.csv.
- провести категоризацію (визначення категорій тексту) даних методом LSA.

2.2.2 Обробка даних оповідань А.К. Дойля та Е.По (+1 бал):

- Завантажити потрібні дані.
- Завантажити оповідання А.К. Дойля та Е.По з папки Texts/Task.
- Виконати попередню обробку текстів.
- Побудувати дві хмари слів, що використовують А.К. Дойль та Е.По.

- Який з письменників написав більш похмурі оповідання?

3 ВИКОНАННЯ ОСНОВНОГО ЗАВДАННЯ

Для виконання поставленого завдання було завантажено дані з Telegram-каналу «Адвокат Права». Для аналізу будуть використовуватися перша тисяча повідомлень. У першу чергу імпортуємо всі необхідні пакети.

```
import pandas as pd
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import nltk
from wordcloud import WordCloud
from functools import reduce
import json
import tqdm
import string
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from langdetect import detect
import numpy as np
from pymorphy2 import MorphAnalyzer
from nltk.stem import WordNetLemmatizer
import pickle

nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

Executed at 2023.05.26 02:19:01 in 3s 200ms

Рисунок 3.1 – Імпортування необхідних пакетів

Після цього визначимо функції, за допомогою яких будемо обробляти повідомлення. Це будуть: видалення пунктуації, емодзі, стоп-слів, токенізація речень на слова та їх лематизація в залежності від мови.

```python
morph_ru = MorphAnalyzer(lang='ru')
morph_uk = MorphAnalyzer(lang='uk')
lemmatizer_en = WordNetLemmatizer()

def clear_punctuation(text):
    text = text.translate(str.maketrans('', '', string.punctuation + '«»—–-’'))
    return text.replace('\n', ' ')


def clear_emojis(text):
    emojis = []
    with open('data/emoji.txt', encoding='utf-8') as file:
        for line in file.readlines():
            emojis.append(line.strip())
    return text.translate(str.maketrans('', '', ''.join(emojis)))


def get_stopwords_remover(stop_words):
    return lambda words: [word for word in words if not word.lower() in stop_words]


def preprocessing_pipeline(steps):
    return lambda raw_text: reduce(lambda data, func: func(data), steps, raw_text)


def prepare_words(text):
    lang = detect(text)
    match lang:
        case 'en':
            full_lang = 'english'
        case _:
            full_lang = 'russian'

    words = word_tokenize(text, language=full_lang)

    match lang:
        case 'uk':
            return [morph_uk.parse(word)[0].normal_form for word in words]
        case 'en':
            return [lemmatizer_en.lemmatize(word) for word in words]
        case _:
            return [morph_ru.parse(word)[0].normal_form for word in words]
```

Executed at 2023.05.26 02:19:02 in 463ms

Рисунок 3.2 – Визначення функцій для попередньої обробки текстів

Наступним кроком прочитаємо файл із повідомленнями, запишемо їх у DataFrame.

```python
with open('data/advocat_prava.json', encoding='utf-8', errors='ignore') as json_file:
    json_data = json.load(json_file)
json_messages = json_data['messages']
messages = []
for message in tqdm.tqdm(json_messages[:1000]):
    text = message['text']
    l = []
    for entity in text:
        if isinstance(entity, str):
            l.append(entity)
        elif isinstance(entity, dict):
            l.append(entity['text'])
    l = [s for s in l if s]
    if l:
        messages.append(''.join(l))

messages = pd.DataFrame(messages, columns=['text'])
messages.iloc[:50]
```

Executed at 2023.05.26 02:19:03 in 887ms

```
100%|██████████| 1000/1000 [00:00<00:00, 141265.16it/s]
```

1-11 50 rows × 1 columns pd.Dataframe

	text
0	⚡ Международный аэропорт Харькова закрыт на п...
1	❗ Центр противодействия дезинформации сообщает...
2	❗ Украина запросила срочное заседание Совбеза ...
3	❗ Воздушное пространство Украины полностью зак...
4	❗ ❗ Путин принял решение о специальной военной...
5	" ❗ Мною принято решение о проведение специально...
6	❗ Военная операция в Донбассе начинается – зая...
7	‼️ Вся линия фронта, очень громко. Бьют из Град...
8	Видео запуска ракет с территории России
9	Харьков, Мариуполь, Бердянск, Борисполь, Киев ...
10	CNN сообщает, что США и их союзники применят в...

Рисунок 3.3 – Завантаження повідомлень

Далі завантажимо список стоп-слів для української мови та створимо функцію, яка буде видаляти стоп-слова англійської, російської та української мов.

```
with open('data/stopwords_ua.txt', encoding='utf-8') as file:
    stopwords_ua = file.read().splitlines()
```
Executed at 2023.05.26 02:19:03 in 58ms

```
stop_words = set(stopwords.words('english') + stopwords.words('russian') + stopwords_ua)
remove_stopwords = get_stopwords_remover(stop_words)
stop_words
```
Executed at 2023.05.26 02:19:03 in 103ms

```
 'кожне',
 'усіляке',
 'чос',
 'is',
 'поперек',
 'відповідно',
 'навіщо',
 'жодне',
 'all',
 'зазвичай',
 'на знак',
 'сяка',
 'впрочем',
 'out',
 'поруч',
 ...}
```

Рисунок 3.4 – Визначення стоп-слів

Об’єднаємо всі дії в один пайплайн для простоти роботи та застосуємо ці функції до кожного повідомлення.

```
pipe = preprocessing_pipeline([
    clear_punctuation,
    clear_emojis,
    prepare_words,
    remove_stopwords
])
```
Executed at 2023.05.26 02:19:03 in 70ms

```
messages['words'] = messages.text.apply(pipe)
messages.iloc[:50]
```
Executed at 2023.05.26 02:19:11 in 8s 469ms

1-11 50 rows × 2 columns pd.Dataframe

	text	words
0	⚡ Международный аэропорт Харькова закрыт на п...	[международный, аэропорт, харьков, закрытый, п...
1	❗Центр противодействия дезинформации сообщает...	[центр, противодействие, дезинформация, сообща...
2	❗Украина запросила срочное заседание Совбеза ...	[украина, запросить, срочный, заседание, совбе...
3	❗Воздушное пространство Украины полностью зак...	[воздушный, пространство, украина, полностью, ...
4	❗ ❗Путин принял решение о специальной военной...	[путин, принять, решение, специальный, военный...
5	"❗Мною принято решение о проведение специально...	[принять, решение, проведение, специальный, во...
6	❗Военная операция в Донбассе начинается – зая...	[военный, операция, донбасс, начинаться, заявл...
7	‼️Вся линия фронта, очень громко. Бьют из Град...	[линия, фронт, очень, громко, бить, град, оста...
8	Видео запуска ракет с территории России	[видео, запуск, ракета, территория, россия]
9	Харьков, Мариуполь, Бердянск, Борисполь, Киев ...	[харьков, мариуполь, бердянск, борисполь, киев...
10	CNN сообщает, что США и их союзники применят в...	[cnn, сообщать, сша, союзник, применить, четве...

Рисунок 3.5 – Попередня обробка текстів

Після застосування функцій обробки текстів отримали список слів для кожного повідомлення.

Далі об’єднаємо ці слова в один масив.

```
1 advocat_prava_words = pd.Series(np.concatenate(messages.words))
2 advocat_prava_words.head()
```

Executed at 2023.05.26 02:19:11 in 72ms

5 rows Length: 5, dtype: object pd.Series

	<unnamed>
0	международный
1	аэропорт
2	харьков
3	закрытый
4	приём

Рисунок 3.6 – 5 перших слів, що зустрічаються в корпусі

Запишемо ці слова у файл.

```
1 with open('data/lemmatized_words.txt', 'w', encoding='utf-8') as file:
2     file.write('\n'.join(advocat_prava_words))
```

Executed at 2023.05.26 02:19:11 in 70ms

Рисунок 3.7 – Збереження слів у файл

Візуалізуємо частоту кожного зі слів.

```
_, ax = plt.subplots(figsize=(15, 10))
advocat_prava_words.value_counts().head(30).plot.bar();
```
Executed at 2023.05.26 02:19:12 in 382ms

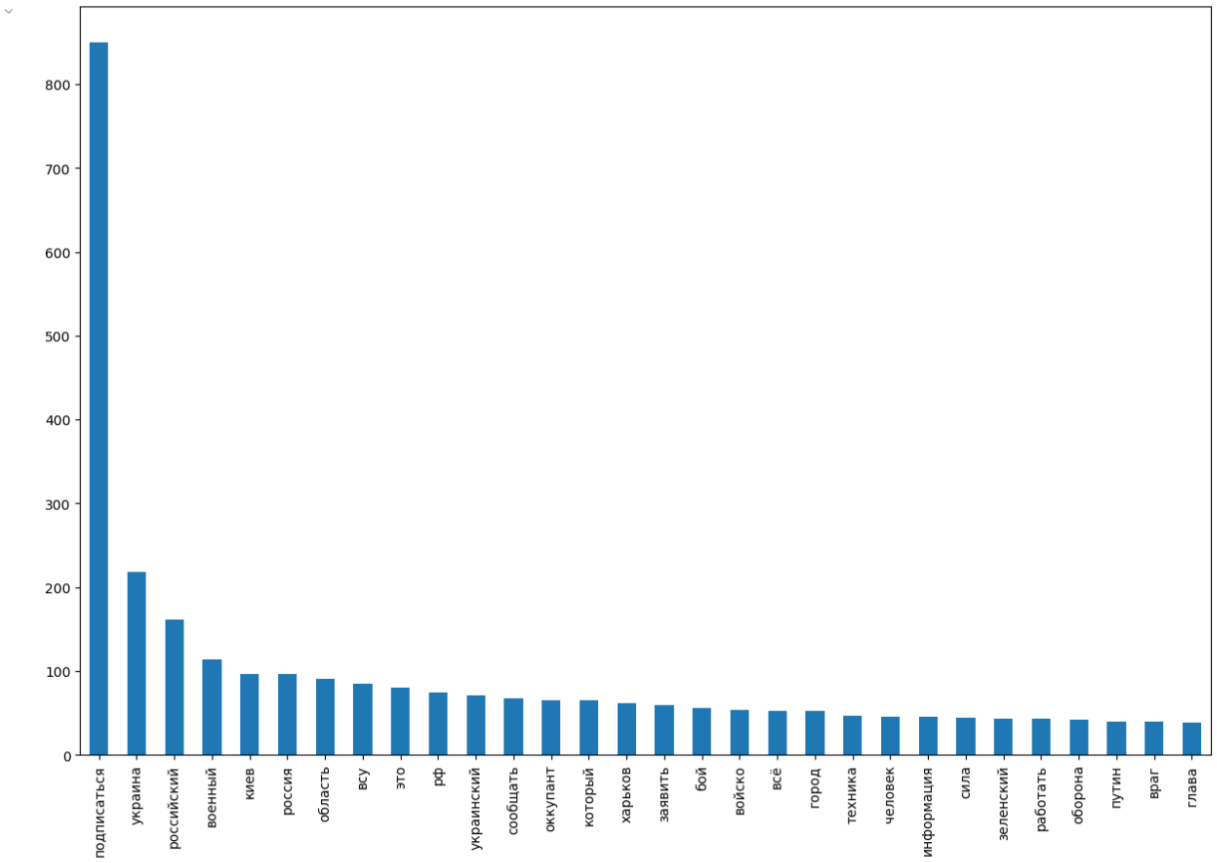

Рисунок 3.8 – Графік частоти кожного зі слів

Бачимо, що деякі стоп-слова не видалилися: «подписаться», «всё», «это». Видалимо і їх.

```
custom_stopwords = {'подписаться', 'это', 'всё'}
custom_stopwords_remover = get_stopwords_remover(custom_stopwords)
messages.words = messages.words.apply(custom_stopwords_remover)
advocat_prava_words = pd.Series(np.concatenate(messages.words))
```
Executed at 2023.05.26 02:19:12 in 35ms

Рисунок 3.9 – Видалення додаткових стоп-слів

Наступним кроком зобразимо хмару слів.

```
advocat_prava_cloud = wordcloud.generate_from_frequencies(advocat_prava_words.value_counts())

plt.figure(figsize=(12, 12))
plt.imshow(advocat_prava_cloud)
plt.axis(False);
```
Executed at 2023.05.26 02:19:13 in 1s 464ms

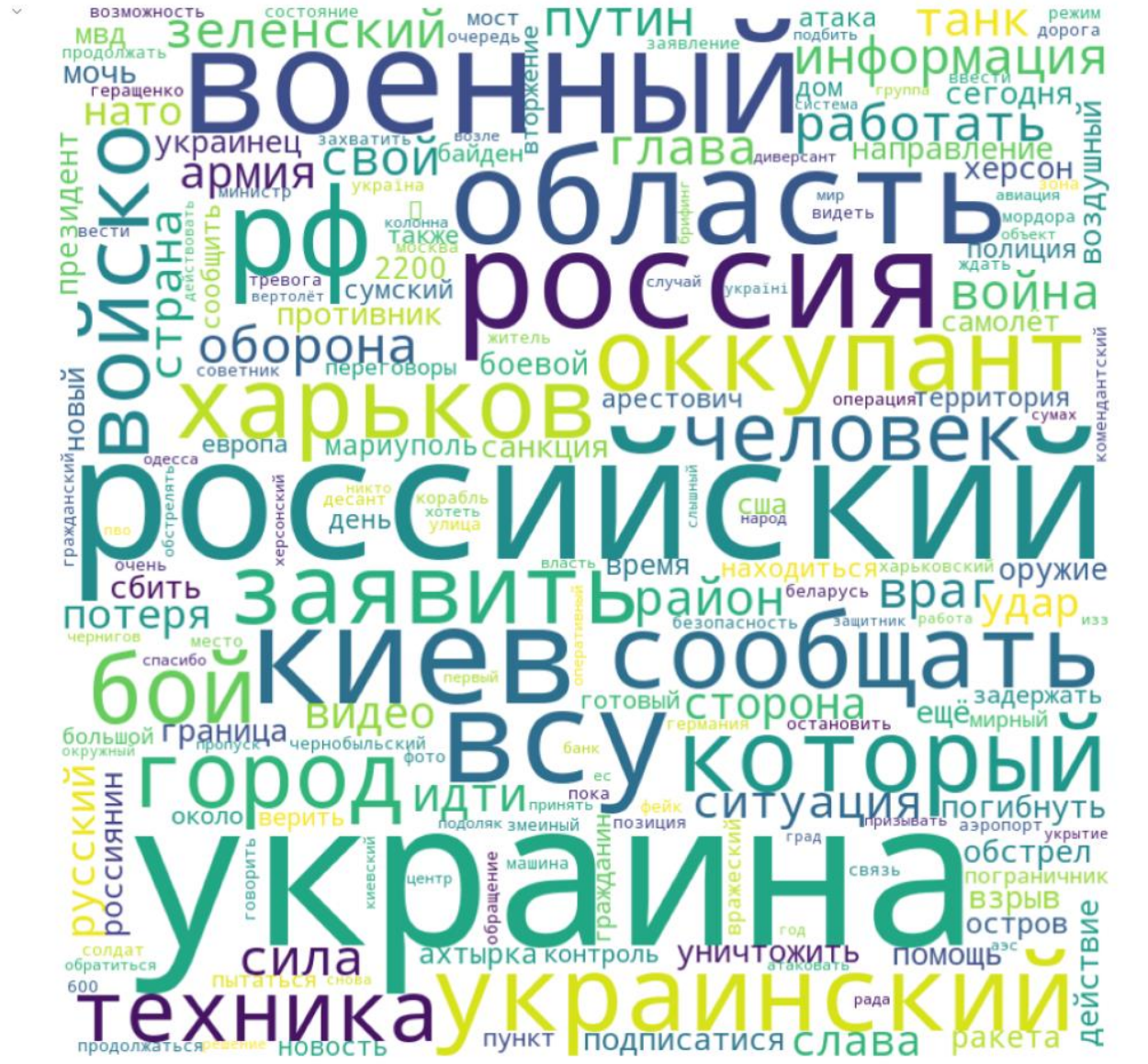

Рисунок 3.10 – Хмара слів для Telegram-каналу «Адвокат Права»

Тепер об’єднаємо очищені слова в речення просто додаючи між словами пробіли.

```
messages['clean_text'] = messages.words.str.join(' ')
messages.head()
```
Executed at 2023.05.26 02:19:13 in 28ms

5 rows × 3 columns pd.Dataframe

	text	words	clean_text
0	⚡ Международный аэропорт Харькова закрыт на п...	[международный, аэропорт, харьков, закрытый, п...	международный аэропорт харьков закрытый приём ...
1	❗ Центр противодействия дезинформации сообщает...	[центр, противодействие, дезинформация, сообща...	центр противодействие дезинформация сообщать г...
2	❗ Украина запросила срочное заседание Совбеза ...	[украина, запросить, срочный, заседание, совбе...	украина запросить срочный заседание совбез оон...
3	❗ Воздушное пространство Украины полностью зак...	[воздушный, пространство, украина, полностью, ...	воздушный пространство украина полностью закры...
4	❗ ❗ Путин принял решение о специальной военной...	[путин, принять, решение, специальный, военный...	путин принять решение специальный военный опер...

Рисунок 3.11 – Об’єднання слів у речення

Створимо Bag of Words для всіх нормалізованих речень, ознайомимося з тим, як виглядає результат роботи.

```
vectorizer = CountVectorizer()
bag_of_words_matrix = vectorizer.fit_transform(messages.clean_text)
```
Executed at 2023.05.26 02:19:13 in 78ms

```
bag_of_words_matrix
```
Executed at 2023.05.26 02:19:13 in 47ms

```
<896x3592 sparse matrix of type '<class 'numpy.int64'>'
    with 11198 stored elements in Compressed Sparse Row format>
```

```
df_bow_sklearn = pd.DataFrame(bag_of_words_matrix.toarray(), columns=vectorizer.get_feature_names_out())
df_bow_sklearn.head(10)
```
Executed at 2023.05.26 02:19:13 in 74ms

10 rows × 3592 columns pd.Dataframe

	000	0173	0200	0300	0600	0835	0836	0838	10	100	...	япония	ясно	яснопонятный	европа	єдиний	єпідтримка	інвалідність	інструкція
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Рисунок 3.12 – Створення та навчання Bag Of Words

Збережемо отримані результати у файл.

```
with open('data/bag_of_words_matrix.pkl', 'wb') as file:
    pickle.dump(bag_of_words_matrix, file)
with open('data/vocabulary.pkl', 'wb') as file:
    pickle.dump(vectorizer.vocabulary_, file)
```
Executed at 2023.05.26 02:19:13 in 146ms

Рисунок 3.13 – Збереження Bag of Words у файл

Далі створимо TF-IDF векторизатор та навчимо його на наших нормалізованих реченнях.

```
tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform(messages.clean_text)
```
Executed at 2023.05.26 02:19:13 in 197ms

Рисунок 3.14 – Створення векторизатора для обчислення TF-IDF метрики

Отримаємо всі слова з векторизатора та порахуємо для них TF-IDF метрики для кожного речення.

```
tfidf_feature_names = tfidf_vectorizer.get_feature_names_out()
```
Executed at 2023.05.26 02:19:13 in 197ms

```
tf_idf_metrics = messages[['text']].copy(deep=True)
for word in advocat_prava_words.value_counts().head(10).index:
    tf_idf_metrics[word] = 0
    word_index = np.where(tfidf_feature_names == word)[0][0]
    for document_index, _ in tf_idf_metrics.iterrows():
        tf_idf_metrics.loc[document_index, word] = tfidf_matrix[document_index, word_index]
tf_idf_metrics
```
Executed at 2023.05.26 02:19:15 in 1s 785ms

11 rows × 896 rows × 11 columns pd.Dataframe ↗ CSV

	text	украина	российский	военный	киев	россия	область	всу	рф	украинский	сообщать
0	ϟ Международный аэропорт Харькова закрыт на п...	0.000000	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.00000	0.000000
1	❗ Центр противодействия дезинформации сообщает...	0.000000	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.111134	0.11489	0.112490
2	❗ Украина запросила срочное заседание Совбеза ...	0.111351	0.0	0.13575	0.000000	0.145158	0.0	0.0	0.000000	0.00000	0.000000
3	❗ Воздушное пространство Украины полностью зак...	0.164756	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.00000	0.000000
4	❗ ❗ Путин принял решение о специальной военной...	0.000000	0.0	0.23892	0.000000	0.000000	0.0	0.0	0.000000	0.00000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...
891	❗ ϟ Audi, Jaguar Land Rover, BMW, Mercedes-Benz...	0.000000	0.0	0.00000	0.000000	0.133391	0.0	0.0	0.000000	0.00000	0.140245
892	❗ В разных частях Киева сообщают о сильных вз...	0.000000	0.0	0.00000	0.184283	0.000000	0.0	0.0	0.000000	0.00000	0.199265
893	‼ Взрывы на Троещине в Киеве\n\nПОДПИСАТЬСЯ	0.000000	0.0	0.00000	0.393919	0.000000	0.0	0.0	0.000000	0.00000	0.000000
894	❗ 🔥 Хорошо горит. Русские солдаты, не воюйте с ...	0.243146	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.00000	0.000000
895	Друзі! Ситуація зараз для Києва – без перебіль...	0.000000	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.00000	0.000000

Рисунок 3.15 – Обчислення TF-IDF метрики для 10 найбільш вживаних слів

4 ВИКОНАННЯ ПЕРШОГО ДОДАТКОВОГО ЗАВДАННЯ

Для початку буде повторено всі ті ж дії для підготовки даних, що й при виконанні основного завдання.

```
1  import numpy as np
2  import pandas as pd
3  from nltk.tokenize import word_tokenize
4  import nltk
5  from functools import reduce
6  import string
7  from pymorphy2 import MorphAnalyzer
8  from sklearn.feature_extraction.text import TfidfVectorizer
9  from sklearn.decomposition import TruncatedSVD
10
11 nltk.download('stopwords')
12 nltk.download('punkt')
```

Executed at 2023.05.26 02:28:37 in 2s 556ms

Рисунок 4.1 – Імпортування необхідних пакетів

```python
morph_uk = MorphAnalyzer(lang='uk')


def clear_punctuation(text):
    text = text.translate(str.maketrans('', '', string.punctuation + '«»—–-’'))
    return text.replace('\n', ' ')


def get_stopwords_remover(stop_words):
    return lambda words: [word for word in words if not word.lower() in stop_words]


def preprocessing_pipeline(steps):
    return lambda raw_text: reduce(lambda data, func: func(data), steps, raw_text)


def prepare_words(text):
    words = word_tokenize(text, language='russian')

    return [morph_uk.parse(word)[0].normal_form for word in words]

def clear_nbsp(text):
    return text.replace('NBSP', ' ')
```

Executed at 2023.05.26 02:28:38 in 271ms

Рисунок 4.2 – Визначення функцій для попередньої обробки текстів

```python
ukr_text = pd.read_csv('data/ukr_text.csv')
ukr_text
```

Executed at 2023.05.26 02:28:38 in 75ms

11 rows 1122 rows × 3 columns pd.Dataframe CSV

	Id	Title	Body
0	http://k.img.com.ua/rss/ua/4013798	Кличко покликав німецьких інвесторів до Києва	Київ - перспективний і відкритий ринок для біз...
1	http://k.img.com.ua/rss/ua/4001679	З'явилося відео, як байкер почав стріляти у во...	З'явилося відео конфлікту між мотоциклістом...
2	http://k.img.com.ua/rss/ua/4001390	У центрі Києва посеред вулиці помер чоловік	У Києві на Бессарабській площі вранці в четвер...
3	http://k.img.com.ua/rss/ua/4001239	Нічний ураган перетворив Хрещатик на смітник	Київ вночі 16 серпня пережив найсильнішу грозу...
4	http://k.img.com.ua/rss/ua/4001227	Потоп у Києві: столицю накрив ураган з градом	Уночі Київ вкотре накрила негода. Найсильніший...
...	...	...	...
1117	http://k.img.com.ua/rss/ua/3194862	Корреспондент: Діамантові руки. Історія успіху...	Київський офіс Класичного ювелірного дому Лобо...
1118	http://k.img.com.ua/rss/ua/3194633	Корреспондент: Роздача слонів. Янукович щедро ...	20 років тому орден За заслуги – тоді він нази...
1119	http://k.img.com.ua/rss/ua/3194587	Корреспондент: Рівняння з трьома відомими. Укр...	10 жовтня політичні важковаговики з табору опо...
1120	http://k.img.com.ua/rss/ua/3194570	Корреспондент: Точка зору. Мета обкрадає кошти...	Добре там, де нас немає. В Ізраїлі ми є, але т...
1121	http://k.img.com.ua/rss/ua/3194519	Корреспондент: Капітан Океанів. Інтерв'ю зі Св...	Початок розмови Корреспондента зі знаменитим у...

Рисунок 4.3 – Завантаження новин

```
with open('data/stopwords_ua.txt', encoding='utf-8') as file:
    stopwords_ua = file.read().splitlines()
```
Executed at 2023.05.26 02:28:38 in 20ms

```
stop_words = set(stopwords_ua)
remove_stopwords = get_stopwords_remover(stop_words)
stop_words
```
Executed at 2023.05.26 02:28:38 in 117ms

```
 'на базі',
 'чого',
 'би',
 'вподовж',
 'під знаком',
 'аніякісіньку',
 'геть',
 'такої',
 'тім',
 'нум',
 'більше',
 'звичайно',
 'кру',
 'ют',
 'пора',
 ...}
```

Рисунок 4.4 – Завантаження стоп-слів для української мови

```
pipe = preprocessing_pipeline([
    clear_nbsp,
    clear_punctuation,
    prepare_words,
    remove_stopwords
])
```
Executed at 2023.05.26 02:28:38 in 88ms

Рисунок 4.5 – Створення пайплайну для попередньої обробки

```
ukr_text['words'] = ukr_text.Body.apply(pipe)
ukr_text
```

Executed at 2023.05.26 02:28:59 in 20s 791ms

11 rows × 1122 rows × 4 columns pd.Dataframe

	Title	Body	words
mg.com.ua/rss/ua/4013798	Кличко покликав німецьких інвесторів до Києва	Київ - перспективний і відкритий ринок для біз...	[перспективний, відкритий, ринок, бізнес, інве...
mg.com.ua/rss/ua/4001679	З'явилося відео, як байкер почав стріляти у во...	З'явилося відео конфлікту між мотоциклістом...	[зявитися, відео, конфлікт, мотоцикліст, водій...
mg.com.ua/rss/ua/4001390	У центрі Києва посеред вулиці помер чоловік	У Києві на Бессарабській площі вранці в четвер...	[бессарабський, площа, вранці, четвер, 16, сер...
mg.com.ua/rss/ua/4001239	Нічний ураган перетворив Хрещатик на смітник	Київ вночі 16 серпня пережив найсильнішу грозу...	[вночі, 16, серпень, пережити, найсильніший, г...
mg.com.ua/rss/ua/4001227	Потоп у Києві: столицю накрив ураган з градом	Уночі Київ вкотре накрила негода. Найсильніший...	[уночі, вкотре, накрити, негода, найсильніший,...
	...	...	...
mg.com.ua/rss/ua/3194862	Корреспондент: Діамантові руки. Історія успіху...	Київський офіс Класичного ювелірного дому Лобо...	[київський, офіс, класичний, ювелірний, дома, ...
mg.com.ua/rss/ua/3194633	Корреспондент: Роздача слонів. Янукович щедро ...	20 років тому орден За заслуги – тоді він нази...	[20, том, орден, заслуга, називатися, почесний...
mg.com.ua/rss/ua/3194587	Корреспондент: Рівняння з трьома відомими. Укр...	10 жовтня політичні важковаговики з табору опо...	[10, жовтень, політичний, важковаговик, табір,...
mg.com.ua/rss/ua/3194570	Корреспондент: Точка зору. Мета обкрадає кошти...	Добре там, де нас немає. В Ізраїлі ми є, але т...	[ізраїль, країна, займати, місце, світ, надій,...
mg.com.ua/rss/ua/3194519	Корреспондент: Капітан Океанів. Інтерв'ю зі Св...	Початок розмови Корреспондента зі знаменитим у...	[початок, розмова, корреспондент, знаменитий, ...

Рисунок 4.6 – Результат виконання попередньої обробки

Після того, як ми отримали масиви слів для кожної новини, можемо завантажити словник тональності та створити функцію, яка буде цю тональність обраховувати. Функція працює наступним чином: додає всі значення тональності, після чого ділить на загальну кількість слів у новині для того, щоб врівноважити цю оцінку для новин із різною кількістю слів.

```
tone_dict = pd.read_csv('https://raw.githubusercontent.com/lang-uk/tone-dict-uk/master/tone-dict-uk.tsv', delimiter='\t', names=['word', 'tone'], index_col=0)
tone_dict
```

Executed at 2023.05.26 02:28:59 in 365ms

11 rows × 3442 rows × 1 columns pd.Dataframe

word	tone
Всевишній	1
Господь	1
Христовий	1
аборт	-1
абсурд	-1
...	...
янгол	1
яскравий	1
ясність	1
ясний	1
ясновельможний	1

Рисунок 4.7 – Завантаження словника тональності

Add Code Cell | Add Markdown Cell

```
def calculate_sentiment(words):
    tone_words = [word for word in words if word in tone_dict.index]
    if len(set(tone_words)) < 5:
        return 0
    return sum([tone_dict.loc[word].tone if word in tone_dict.index else 0 for word in words]) / len(words)
```

Executed at 2023.05.26 02:28:59 in 13ms

Рисунок 4.8 – Функція обрахунку тональності

Обрахуємо тональності.

```
ukr_text['tone'] = ukr_text.words.apply(calculate_sentiment)
ukr_text
```

Executed at 2023.05.26 02:29:00 in 820ms

11 rows × 1122 rows × 5 columns pd.Dataframe

	Title	Body	words	tone
rss/ua/4013798	Кличко покликав німецьких інвесторів до Києва	Київ - перспективний і відкритий ринок для біз...	[перспективний, відкритий, ринок, бізнес, інве...	0.026163
rss/ua/4001679	З'явилося відео, як байкер почав стріляти у во...	З'явилося відео конфлікту між мотоциклістом...	[зявитися, відео, конфлікт, мотоцикліст, водій...	-0.090909
rss/ua/4001390	У центрі Києва посеред вулиці помер чоловік	У Києві на Бессарабській площі вранці в четвер...	[бессарабський, площа, вранці, четвер, 16, сер...	-0.137255
rss/ua/4001239	Нічний ураган перетворив Хрещатик на смітник	Київ вночі 16 серпня пережив найсильнішу грозу...	[вночі, 16, серпень, пережити, найсильніший, г...	0.000000
rss/ua/4001227	Потоп у Києві: столицю накрив ураган з градом	Уночі Київ вкотре накрила негода. Найсильніший...	[уночі, вкотре, накрити, негода, найсильніший,...	0.000000
	...	...	...	...
rss/ua/3194862	Корреспондент: Діамантові руки. Історія успіху...	Київський офіс Класичного ювелірного дому Лобо...	[київський, офіс, класичний, ювелірний, дома, ...	0.042616
rss/ua/3194633	Корреспондент: Роздача слонів. Янукович щедро ...	20 років тому орден За заслуги – тоді він нази...	[20, том, орден, заслуга, називатися, почесний...	0.124682
rss/ua/3194587	Корреспондент: Рівняння з трьома відомими. Укр...	10 жовтня політичні важковаговики з табору опо...	[10, жовтень, політичний, важковаговик, табір,...	0.022042
rss/ua/3194570	Корреспондент: Точка зору. Мета обкрадає кошти...	Добре там, де нас немає. В Ізраїлі ми є, але т...	[ізраїль, країна, займати, місце, світ, надій,...	0.018657
rss/ua/3194519	Корреспондент: Капітан Океанів. Інтерв'ю зі Св...	Початок розмови Корреспондента зі знаменитим у...	[початок, розмова, корреспондент, знаменитий, ...	0.020802

Рисунок 4.9 – Результат обрахунку тональності

Виведемо найбільш негативну та позитивну новини.

```
print(ukr_text.sort_values(by='tone', ascending=True).iloc[0].Body)
```
Executed at 2023.05.26 02:29:00 in 47ms

У Державному департаменті США заявили, що США разом з усім світом згадують жертв Голодомору і вкотре підтвердили прихильність демократії, процвітанню і суверенітету України. Про це йдеться в заяві Держдепу, передає Голос Америки. Прес-секретар Державного департаменту США Морган Ортагюс заявила: "Ми об'єднуємося з усім світом, щоб згадати невинних жертв Голодомору і підтвердити нашу прихильність демократії, процвітанню і суверенітету України". У Держдепі заявили, що Голодомор - одна з найжорстокіших трагедій 20 століття. "Шляхом навмисного захоплення української землі, врожаю і примусової колективізації, Радянський Союз призвів до масштабного голоду, смертей і приніс надзвичайні людські страждання ... Хоча ця жахлива трагедія була однією з найжорстокіших в 20 столітті, Радянському Союзу не вдалося зломити дух українського народу". Раніше повідомлялося, що в Україні відзначають День пам'яті жертв Голодоморів. По всій території України приспущено державні прапори й обмежено проведення заходів розважального характеру.

Рисунок 4.10 – Найбільш негативна новина

```
print(ukr_text.sort_values(by='tone', ascending=False).iloc[0].Body)
```
Executed at 2023.05.26 02:29:00 in 96ms

груп планують навчальнии процес таким чином, щоб теоретична частина займала 20%, а практика - 80%. Подібний підхід до навчання вперше застосували навчальні заклади Великобританії, раніше з'ясувавши, що активне обговорення і постійна мовна практика підвищують загальну успішність студента і результативність всього курсу. Саме завдяки такому підходу, англійська засвоюється набагато швидше, словниковий запас студента збагачується новою лексикою і учень позбавляється від всіляких мовних перепон. Вивчення справжньої англійської мови Мало хто знає, але тільки на мовних курсах в Англії учень буде вивчати справжню "чисту" англійську, якою говорять в самому парламенті Великобританії без всіляких домішок або американських діалектів. На відміну від інших англомовних країн, де англійську мову можуть викладати представники інших національностей, в Великобританії - це корінні жителі країни, професора з багаторічним стажем роботи, бездоганною репутацією і безцінним досвідом. Для учнів, які бажають вивчити всі аспекти і нюанси правильної вимови, побудови речень, історію виникнення англійської, мовні курси в Великобританії - кращий варіант. Ніяких вікових обмежень Курси англійської мови в Англії не мають вікових рамок. Учень будь-якого віку має можливість вчитися в освітньому центрі країни на тих же умовах, як і інші студенти. Як показує практика, студенти пенсійного віку в середньому 65-70 років часті гості на курсах англійської мови в Англії. Для них освітні центри пропонують спеціалізовану програму, що передбачає спеціально розроблену методику навчання, завдяки чому, навчальний процес проходить легко і невимушено, а інформація засвоюється досить швидко і зрозуміло. Поєднання традиційних та інноваційних методів вивчення англійської Навчання на мовних курсах в Англії - з'єднання вікових традицій і суворих консервативних методів навчання з новаторськими технологіями і прогресивними поглядами. Британці дуже цінують традиції і методику навчання, по якій успішно навчаються студенти з усього світу вже багато поколінь. Однак, завдяки прийняттю сучасних освітніх нововведень і поваги поглядів молодших спеціалістів, навчання в Англії не стоїть на місці, постійно розвиваючись. Навчання на курсах англійської мови у Великобританії передбачає особливу методику навчання, яка передбачає захоплюючий навчальний процес у форматі інтерактивної гри, бесіди або активної дискусії. Так як 80% часу, на уроках студенти приділяють увагу практиці, а саме, розвитку розмовних навичок, британські педагоги розробили спеціальну програму навчання, завдяки якій інформація засвоюється легко і дуже доступно. Підготовлено спільно з освітнім агентством PFI

Рисунок 4.11 – Найбільш позитивна новина

Далі знову об’єднаємо слова кожної новини в новий текст, додавши між словами пробіли.

```
ukr_text['clean_text'] = ukr_text.words.str.join(' ')
ukr_text
```
Executed at 2023.05.26 02:29:00 in 93ms

11 rows × 1122 rows × 6 columns pd.Dataframe

	Body	words	tone	clean_text
в до Києва	Київ - перспективний і відкритий ринок для біз...	[перспективний, відкритий, ринок, бізнес, інве...	0.026163	перспективний відкритий ринок бізнес інвестиці...
ріляти у во...	З'явилося відео конфлікту між мотоциклістом...	[зявитися, відео, конфлікт, мотоцикліст, водій...	-0.090909	зявитися відео конфлікт мотоцикліст водій авто...
чоловік	У Києві на Бессарабській площі вранці в четвер...	[бессарабський, площа, вранці, четвер, 16, сер...	-0.137255	бессарабський площа вранці четвер 16 серпень в...
а смітник	Київ вночі 16 серпня пережив найсильнішу грозу...	[вночі, 16, серпень, пережити, найсильніший, г...	0.000000	вночі 16 серпень пережити найсильніший гроза з...
н з градом	Уночі Київ вкотре накрила негода. Найсильніший...	[уночі, вкотре, накрити, негода, найсильніший,...	0.000000	уночі вкотре накрити негода найсильніший дощ п...
	...	...	... ...	
орія успіху...	Київський офіс Класичного ювелірного дому Лобо...	[київський, офіс, класичний, ювелірний, дома, ...	0.042616	київський офіс класичний ювелірний дома лоборт...
ович щедро ...	20 років тому орден За заслуги – тоді він нази...	[20, том, орден, заслуга, називатися, почесний...	0.124682	20 том орден заслуга називатися почесний знак ...
домими. Укр...	10 жовтня політичні важковаговики з табору опо...	[10, жовтень, політичний, важковаговик, табір,...	0.022042	10 жовтень політичний важковаговик табір опози...
радає кошти...	Добре там, де нас немає. В Ізраїлі ми є, але т...	[ізраїль, країна, займати, місце, світ, надій,...	0.018657	ізраїль країна займати місце світ надій молоко...
ерв’ю зі Св...	Початок розмови Корреспондента зі знаменитим у...	[початок, розмова, кореспондент, знаменитий, ...	0.020802	початок розмова кореспондент знаменитий украї...

Рисунок 4.12 – Об’єднання слів у речення

Знову створимо векторизатор для обрахунку TF-IDF метрики.

Тепер створимо категоризатор. Встановимо кількість тем рівною 100, оскільки таке значення рекомендується для методу LSA в документації пакету scikit-learn. Навчимо цей перетворювач, виведемо матрицю для кожної новини.

```
tfidf_vectorizer = TfidfVectorizer()
X = tfidf_vectorizer.fit_transform(ukr_text.clean_text)
```
Executed at 2023.05.26 02:29:00 in 241ms

Рисунок 4.13 – Створення та навчання векторизатора для TF-IDF метрики

```
svd_vectorizer = TruncatedSVD(n_components=100, random_state=42)
X_lsa = svd_vectorizer.fit_transform(X)
```
Executed at 2023.05.26 02:29:02 in 1s 876ms

Рисунок 4.14 – Створення та навчання категоризатора

```
X_lsa
```
Executed at 2023.05.26 02:29:02 in 11ms

1-10 1,122 rows × 100 columns np.ndarray

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1
0	0.225543	-0.018412	0.070701	-0.083560	-0.039350	-0.018585	0.008417	-0.026353	-0.014447	0.034107	-0.032115	-0.006156	0.019872	-0.072587	0.035079	0.000957	-0.026
1	0.094064	-0.027904	-0.032069	0.001132	0.000290	0.041588	-0.017662	0.000309	0.023608	-0.017959	-0.008246	-0.010869	-0.017268	-0.010112	0.001489	-0.042816	-0.019
2	0.094172	-0.021887	-0.038988	-0.026011	-0.010463	0.043455	-0.052266	-0.037253	0.010384	-0.018007	-0.012249	-0.010664	-0.028943	-0.040078	-0.006671	-0.025068	-0.057
3	0.043385	-0.012315	-0.014534	-0.012375	-0.003493	0.033754	-0.028741	-0.009413	0.014932	-0.005994	-0.000252	-0.000585	-0.008453	0.000673	-0.000382	-0.022344	-0.031
4	0.059402	-0.021692	-0.029251	-0.016541	-0.018693	0.019230	-0.054412	-0.001225	0.029294	-0.029618	0.005188	-0.042832	-0.024581	0.047737	-0.013813	-0.017425	-0.034
5	0.113500	-0.038551	-0.024676	-0.001565	0.001591	0.069859	-0.034043	0.008222	0.019380	-0.025668	-0.014025	0.030187	-0.001441	-0.011670	-0.048556	-0.042937	-0.056
6	0.080400	-0.028267	-0.021085	-0.017458	-0.018908	0.027732	-0.049721	-0.017733	0.005627	-0.008596	-0.024920	0.007838	-0.036041	-0.045438	-0.008228	-0.030262	-0.050
7	0.071619	-0.015004	-0.004601	-0.001071	-0.001383	0.055652	-0.008666	-0.007712	0.019281	0.002495	-0.033959	0.023656	-0.039888	0.016596	-0.018764	0.004543	-0.016
8	0.110948	-0.037762	-0.033153	-0.011700	-0.003115	0.055305	-0.058385	-0.003451	0.029156	-0.016927	-0.034436	0.016167	-0.066749	-0.004806	-0.052763	-0.041763	-0.059
9	0.129260	-0.056900	-0.088444	-0.013063	-0.000221	0.047951	-0.022126	0.131658	-0.014638	-0.017116	-0.029350	-0.017313	-0.113649	-0.012735	-0.021992	-0.047299	-0.038

Рисунок 4.15 – Результат роботи категоризатора

Створимо функцію, яка буде показувати найважливіші слова для кожної категорії та виведемо їх для кожної категорії.

```
def get_topic_words(vectorizer, svd, n_top_words):
    words = vectorizer.get_feature_names_out()
    topics = []
    for component in svd.components_:
        top_words_idx = np.argsort(component)[::-1][:n_top_words]
        top_words = [words[i] for i in top_words_idx]
        topics.append(top_words)
    return topics
```
Executed at 2023.05.26 02:29:02 in 46ms

Рисунок 4.16 – Функція для виведення найважливіших слів для категорій

```
topics = get_topic_words(tfidf_vectorizer, svd_vectorizer, 10)
for i, topic in enumerate(topics):
    print(f'Topic {i}: {", ".join(topic)}')
```
Executed at 2023.05.26 02:44:12 in 184ms

```
Topic 0: україна, рок, долар, компанія, росія, новий, йога, країна, область, грудень
Topic 1: нафта, біржа, долар, індекс, барель, пункт, марка, торг, ньюйоркський, brent
Topic 2: область, регіон, інвестиція, газа, млрд, грн, населення, душити, транзит, україна
Topic 3: долар, газа, транзит, росія, україна, нафтогаз, газпром, нафта, російський, барель
Topic 4: індекс, пункт, jones, dow, nasdaq, фондовий, 500, виборчий, ринок, цвк
Topic 5: виборчий, округа, цвк, комісія, депутат, народний, серпень, обраний, зареєструвати, голос
Topic 6: гонка, область, кубок, україна, регіон, збірний, підручний, естафета, змішаний, інвестиція
Topic 7: матч, динамо, штрафний, мяч, шахтар, пробити, ворота, очко, команда, збірний
Topic 8: мерседес, феррарі, феттля, хемілтон, гранпрі, боттас, себастьян, булла, льюїс, хаас
Topic 9: шоу, учасник, випуск, мастершеф, сезон, санкція, північний, росія, конкурс, страва
Topic 10: шоу, випуск, учасник, мастершеф, сезон, млрд, страва, газа, долар, кулінар
Topic 11: гривня, курс, рок, автомобіль, зміцнення, електромобіль, доход, місяць, шоу, учасник
Topic 12: музичук, ганна, партія, шах, світ, фінал, чемпіонат, шахістка, танути, чжуня
Topic 13: музичук, instagram, санкція, ганна, північний, партія, газопровід, 2019, знімок, зіркий
Topic 14: гривня, курс, фільм, зміцнення, фото, instagram, бюджет, україна, зіркий, доход
```

Рисунок 4.17 – Виведення найважливіших слів для категорій

Тепер додамо стовпець topic у наш набір даних, який буде вказувати, власне, на категорію, до якої належить кожна з новин.

```
ukr_text['topic'] = X_lsa.argmax(axis=1)
ukr_text
```
Executed at 2023.05.26 02:29:02 in 65ms

1122 rows × 7 columns pd.Dataframe

	Body	words	tone	clean_text	topic
ıa	Київ - перспективний і відкритий ринок для біз...	[перспективний, відкритий, ринок, бізнес, інве...	0.026163	перспективний відкритий ринок бізнес інвестиці...	0
во...	З'явилося відео конфлікту між мотоциклістом...	[зявитися, відео, конфлікт, мотоцикліст, водій...	-0.090909	зявитися відео конфлікт мотоцикліст водій авто...	0
	У Києві на Бессарабській площі вранці в четвер...	[бессарабський, площа, вранці, четвер, 16, сер...	-0.137255	бессарабський площа вранці четвер 16 серпень в...	21
:	Київ вночі 16 серпня пережив найсильнішу грозу...	[вночі, 16, серпень, пережити, найсильніший, г...	0.000000	вночі 16 серпень пережити найсильніший гроза з...	57
ом	Уночі Київ вкотре накрила негода. Найсильніший...	[уночі, вкотре, накрити, негода, найсильніший,...	0.000000	уночі вкотре накрити негода найсильніший дощ п...	0
	...	...	... ...		...
ху...	Київський офіс Класичного ювелірного дому Лобо...	[київський, офіс, класичний, ювелірний, дома, ...	0.042616	київський офіс класичний ювелірний дома лобort...	0
ю ...	20 років тому орден За заслуги – тоді він нази...	[20, том, орден, заслуга, називатися, почесний...	0.124682	20 том орден заслуга називатися почесний знак ...	0
'кр...	10 жовтня політичні важковаговики з табору опо...	[10, жовтень, політичний, важковаговик, табір,...	0.022042	10 жовтень політичний важковаговик табір опози...	0
іти...	Добре там, де нас немає. В Ізраїлі ми є, але т...	[ізраїль, країна, займати, місце, світ, надій,...	0.018657	ізраїль країна займати місце світ надій молоко...	0
Св...	Початок розмови Корреспондента зі знаменитим у...	[початок, розмова, корреспондент, знаменитий, ...	0.020802	початок розмова корреспондент знаменитий украї...	0

Рисунок 4.18 – Додавання стовпця з категорією до набору даних

Виведемо новини з категорією №6, яка дуже схожа на новини про футбол.

```
ukr_text.query('topic == 7')
```
Executed at 2023.05.26 02:44:18 in 47ms

1-11 40 rows × 7 columns pd.Dataframe

	Id	Title	Body	words
9	http://k.img.com.ua/rss/ua/3999418	У Києві посилять заходи безпеки 10 і 12 серпня	У Києві у зв'язку з проведенням футбольних мат...	[звязка, проведення, футбольний, мат
62	http://k.img.com.ua/rss/ua/4170402	Шахтар - Аталанта 0:3. Онлайн матчу Ліги Чемпі...	Українці проведуть домашню гру на ОСК Металі...	[українка, провести, домашній, гру,
63	http://k.img.com.ua/rss/ua/4165628	Олександрія - Вольфсбург 0:1. Онлайн матчу ЛЄ	У четвер, 28-го листопада 2019 року, відбудеть...	[четвер, 28га, листопад, 2019, рок,
71	http://k.img.com.ua/rss/ua/4161290	Футболіст збірної Бельгії десять хвилин грав в...	У відбірковому матчі чемпіонату Європи з футбо...	[відбірковий, матч, чемпіонат, європ
72	http://k.img.com.ua/rss/ua/4160247	Україна - Естонія 1:0. Онлайн-трансляція матчу	Сьогодні, 14 листопада, на Korespondent.net - ...	[14, листопад, korespondentnet, онла
74	http://k.img.com.ua/rss/ua/4158548	Шахтар - Динамо 1-0. Онлайн матчу Прем'єр ліги	6:30 Попов підключився до атаки і заліз в офса...	[630, попов, підключитися, атака, за
76	http://k.img.com.ua/rss/ua/4157768	Олександрія - Сент-Етьєн 2:2. Онлайн матчу ЛЄ	Українці проведуть домашній поєдинок на стадіо...	[українка, провести, домашній, поєди
77	http://k.img.com.ua/rss/ua/4157749	Копенгаген - Динамо 1:1. Онлайн матчу Ліги Європи	Українці проведуть матч-відповідь на арені Пар...	[українка, провести, матчвідповідь,
82	http://k.img.com.ua/rss/ua/4170402	Шахтар - Аталанта 0:3. Онлайн матчу Ліги Чемпі...	Українці проведуть домашню гру на ОСК Металі...	[українка, провести, домашній, гру,
83	http://k.img.com.ua/rss/ua/4165628	Олександрія - Вольфсбург 0:1. Онлайн матчу ЛЄ	У четвер, 28-го листопада 2019 року, відбудеть...	[четвер, 28га, листопад, 2019, рок,
91	http://k.img.com.ua/rss/ua/4161290	Футболіст збірної Бельгії десять хвилин грав в...	У відбірковому матчі чемпіонату Європи з футбо...	[відбірковий, матч, чемпіонат, європ

Рисунок 4.19 – Новини з категорією про футбол

Як бачимо, із цією категорією категоризатор впорався чудово. Можемо також вивести новини категорії №5, у яких, схоже, йдеться про вибори.

```
ukr_text.query('topic == 51')
```
Executed at 2023.05.26 02:45:18 in 46ms

1-11 20 rows × 7 columns pd.Dataframe CSV

	Id	Title	Body	words
833	http://k.img.com.ua/rss/ua/4171554	ЦВК призначила довибори до Верховної Ради	У зв'язку з призначенням нардепа від Слуги нар...	[звязка, призначення, нардеп, слуга,
834	http://k.img.com.ua/rss/ua/4157103	Вибори в Раду: ЦВК оголосила останній результат	Центрвиборчком визнав обраним депутатом Верхов...	[центрвиборчком, визнати, обраний, д
835	http://k.img.com.ua/rss/ua/4156765	Вибори президента: члени однієї ОВК отримали у...	Мар'їнський районний суд Донецької області при...	[марїнський, районний, суд, донецьки
836	http://k.img.com.ua/rss/ua/4154741	Вибори в Раду: ЦВК завершила перерахунок голос...	Центральна виборча комісія (ЦВК) завершила пов...	[центральний, виборчий, комісія, цвк
837	http://k.img.com.ua/rss/ua/4146794	ЦВК відшкодувала парламентським партіям витрат...	Центральна виборча комісія (ЦВК) відшкодувала ...	[центральний, виборчий, комісія, цвк
838	http://k.img.com.ua/rss/ua/4133803	ЦВК завершила реєстрацію нардепів	Центральна виборча комісія (ЦВК) у середу, 28 ...	[центральний, виборчий, комісія, цвк
839	http://k.img.com.ua/rss/ua/4132428	ЦВК залишилося зареєструвати менше 30 нардепів	Центральна виборча комісія (ЦВК) на засіданні ...	[центральний, виборчий, комісія, цвк
840	http://k.img.com.ua/rss/ua/4131732	ЦВК зареєструвала три чверті нової Ради	Центральна виборча комісія вже зареєструвала 3...	[центральний, виборчий, комісія, зар
841	http://k.img.com.ua/rss/ua/4130907	ЦВК зареєструвала більше половини нардепів	Центральна виборча комісія (ЦВК) у понеділок, ...	[центральний, виборчий, комісія, цвк
842	http://k.img.com.ua/rss/ua/4130173	Вибори в Прилуках: ЦВК розпустила комісію	Центрвиборчком у п'ятницю, 16 серпня, прийняв ...	[центрвиборчком, пятниця, 16, серпен
843	http://k.img.com.ua/rss/ua/4130098	Округком №210 закидали димовими шашками	У Прилуках Чернігівської області під будівлею ...	[прилука, чернігівський, область, бу

Рисунок 4.20 – Новини з категорією про вибори

І справді, всі новини якраз про вибори.

5 ВИКОНАННЯ ДРУГОГО ДОДАТКОВОГО ЗАВДАННЯ

Для початку повторимо всі дії з попередніх розділів.

```
import pandas as pd
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.corpus import sentiwordnet as swn
import nltk
from wordcloud import WordCloud
import string
from functools import reduce
from nltk.stem import WordNetLemmatizer
import tqdm

nltk.download('stopwords')
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

Executed at 2023.05.26 02:48:32 in 2s 712ms

Рисунок 5.1 – Імпортування необхідних пакетів

```
lemmatizer = WordNetLemmatizer()

def clear_punctuation(text):
    return text.translate(str.maketrans('', '', string.punctuation))


def get_stopwords_remover(stop_words):
    return lambda words: words[~words.isin(stop_words)]


def lemmatize(words):
    return words.apply(lemmatizer.lemmatize)


def preprocessing_pipeline(steps):
    return lambda raw_text: reduce(lambda data, func: func(data), steps, raw_text)


def lowercase(words):
    return words.str.lower()
```

Executed at 2023.05.26 02:48:34 in 27ms

Рисунок 5.2 – Визначення функцій для попередньої обробки текстів

```
with open('data/doyle.txt') as file:
    doyle = file.read()
with open('data/doyle-2.txt') as file:
    doyle += file.read()

with open('data/poe.txt') as file:
    poe = file.read()
with open('data/poe-2.txt') as file:
    poe += file.read()
```

Executed at 2023.05.26 02:48:35 in 21ms

Рисунок 5.3 – Завантаження творів

```
stop_words = set(stopwords.words('english'))
remove_stopwords = get_stopwords_remover(stop_words)
stop_words
```
Executed at 2023.05.26 02:48:36 in 25ms

```
'will',
'with',
'won',
"won't",
'wouldn',
"wouldn't",
'y',
'you',
"you'd",
"you'll",
"you're",
"you've",
'your',
'yours',
'yourself',
'yourselves'}
```

Рисунок 5.4 – Визначення стоп-слів для англійської мови

```
pipe = preprocessing_pipeline([
    clear_punctuation,
    word_tokenize,
    pd.Series,
    lowercase,
    remove_stopwords,
    lemmatize
])
```
Executed at 2023.05.26 02:48:37 in 23ms

Рисунок 5.5 – Створення пайплайну для попередньої обробки текстів

```
1 doyle_words: pd.Series = pipe(doyle)
2 poe_words: pd.Series = pipe(poe)
Executed at 2023.05.26 02:48:42 in 1s 926ms
```

Рисунок 5.6 – Виконання попередньої обробки текстів

Виведемо графік частоти слів.

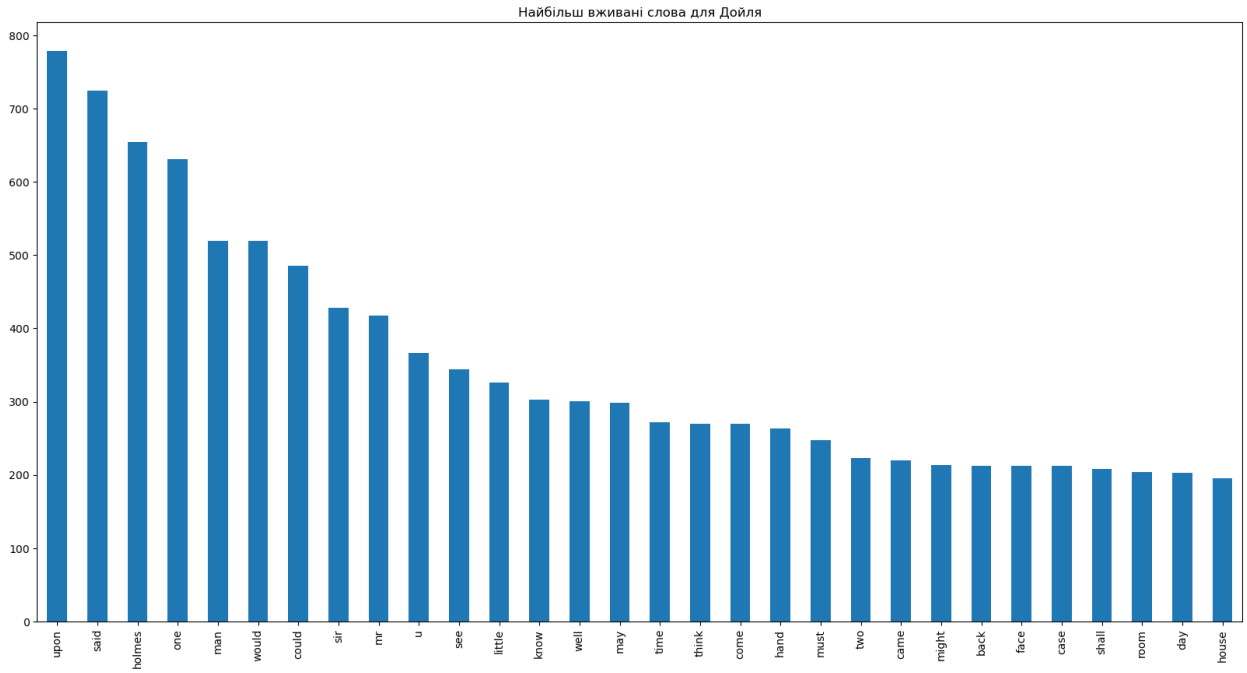

Рисунок 5.7 – Графік частот слів, що використовував Дойль

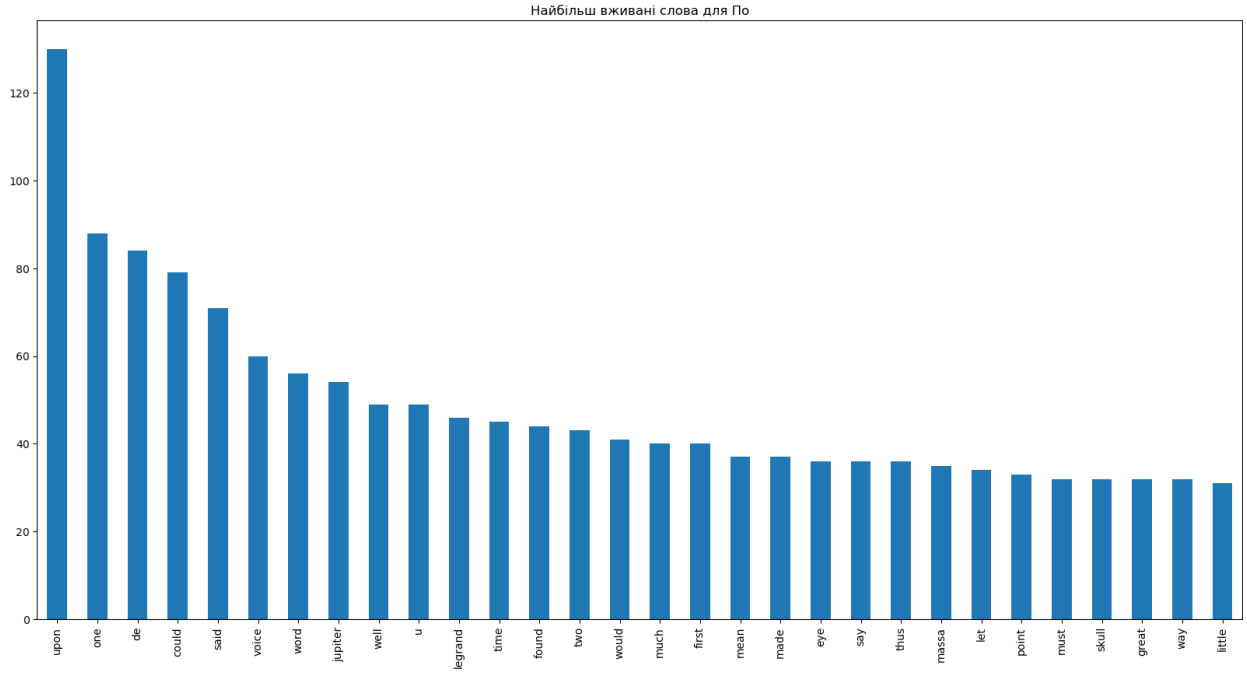

Рисунок 5.8 – Графік частот слів, що використовував По

Бачимо, на них велику кількість стоп-слів, що не були видалені. Видаляємо їх.

```
custom_stopwords = {'upon', 'could', 'would', 'us', 'u', 'man', 'mr', 'de', 'may', 'must', 'thus', 'say', 'much', 'two', }
remove_custom_stopwords = get_stopwords_remover(custom_stopwords)
```
Executed at 2023.05.26 02:49:23 in 43ms

```
doyle_words = remove_custom_stopwords(doyle_words)
poe_words = remove_custom_stopwords(poe_words)
```
Executed at 2023.05.26 02:49:25 in 59ms

```
doyle_words = pd.Series(doyle_words)
poe_words = pd.Series(poe_words)
```
Executed at 2023.05.26 02:49:31 in 31ms

Рисунок 5.9 – Видалення додаткових стоп-слів

Будуємо хмари слів.

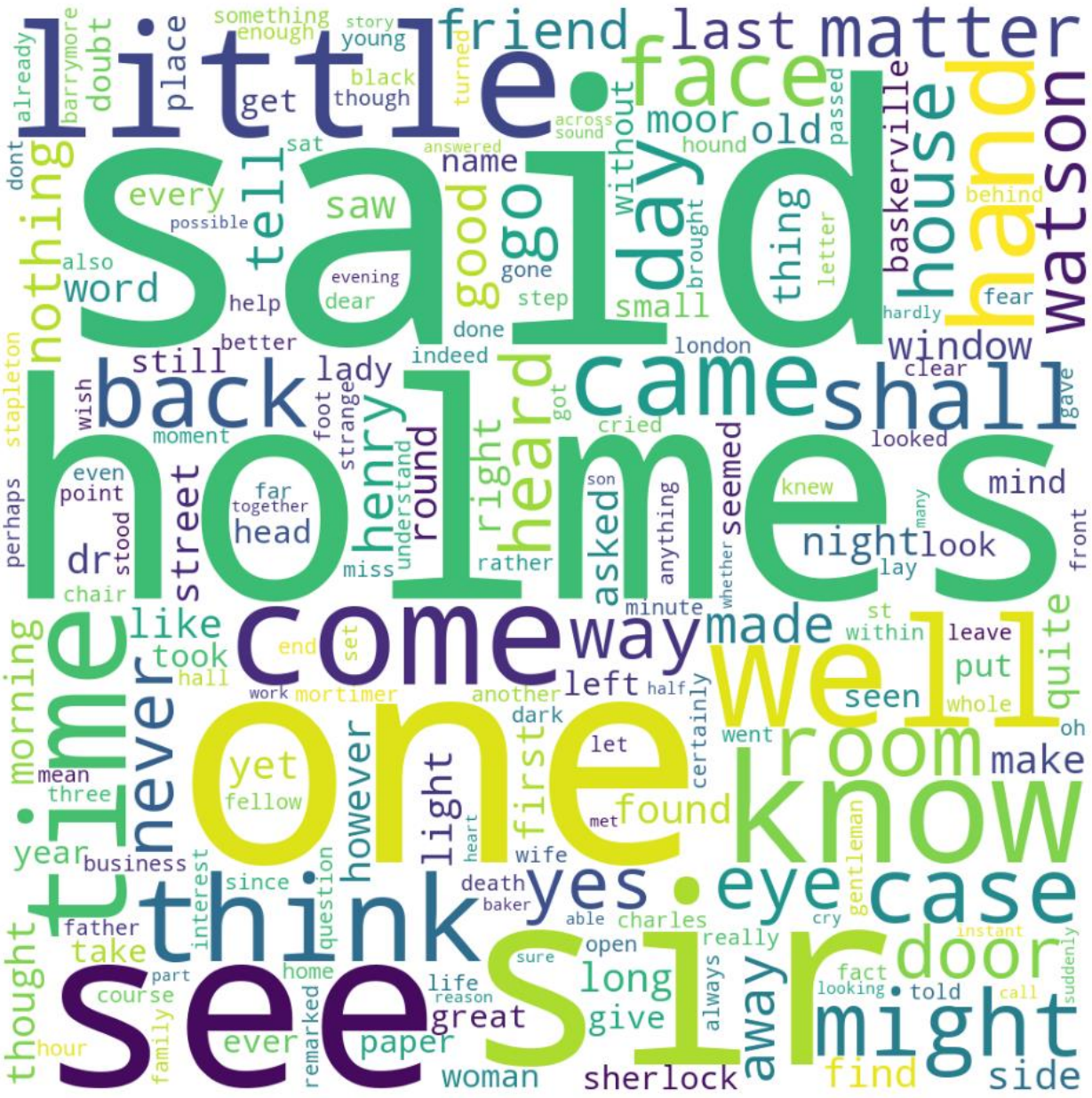

Рисунок 5.10 – Хмара слів творів Дойля

Рисунок 5.11 – Хмара слів творів По

Тепер обрахуємо «похмурість» для кожного з авторів. Для цього пройдемося кожним словом, що використовували автори, спочатку визначимо частину мови, до якої належить це слово, а потім за допомогою wordnet встановимо його тональність. Після обрахунку тональностей поділимо їх на загальну кількість слів, щоб збалансувати оцінку.

```python
def get_wordnet_pos(w: str) -> str:
    treebank_tag = nltk.pos_tag([w])[0][1]
    if treebank_tag.startswith('J'):
        return 'a'
    elif treebank_tag.startswith('V'):
        return 'v'
    elif treebank_tag.startswith('N'):
        return 'n'
    elif treebank_tag.startswith('R'):
        return 'r'
    else:
        return 'n'
```

Executed at 2023.05.26 02:50:26 in 12ms

```python
doyle_score = 0
poe_score = 0
doyle_count = 0
poe_count = 0

for word, count in tqdm.tqdm(list(doyle_words.value_counts().items())):
    synsets = list(swn.senti_synsets(word, get_wordnet_pos(word)))
    if synsets:
        score = synsets[0].pos_score() - synsets[0].neg_score()
        doyle_score += score * count
        doyle_count += count

for word, count in tqdm.tqdm(list(poe_words.value_counts().items())):
    synsets = list(swn.senti_synsets(word, get_wordnet_pos(word)))
    if synsets:
        score = synsets[0].pos_score() - synsets[0].neg_score()
        poe_score += score * count
        poe_count += count

doyle_score /= len(doyle_words)
poe_score /= len(poe_words)

print(f'Doyle score: {doyle_score}, Poe score: {poe_score}')
```

Executed at 2023.05.26 02:53:23 in 8s 188ms

```
100%|██████████| 9754/9754 [00:05<00:00, 1677.28it/s]
100%|██████████| 3910/3910 [00:02<00:00, 1669.54it/s]

Doyle score: 0.0028967595648804567, Poe score: 0.002898256992298338
```

Рисунок 5.12 – Обрахунок похмурості для творів Дойля та По

Бачимо, що оцінки дуже близькі одне до одного, тому тут не можна сказати, що хтось писав більш похмурі оповідання, але все одно виведемо підсумок.

```
1 print('По писав більш похмурі оповідання' if poe_score < doyle_score else 'Дойль писав більш похмурі оповідання')
```

Executed at 2023.05.26 02:53:30 in 25ms

```
Дойль писав більш похмурі оповідання
```

Рисунок 5.13 – Висновок про автора, який писав більш похмурі твори

6 ВИСНОВОК

У ході даної лабораторної роботи було отримано велику кількість практичних навичок із аналізу текстів. Я навчився виконувати попередню підготовку текстів: видалення розмітки, токенізацію слів, лематизацію. Навчився користуватися моделлю Bag of Words, обчислювати TF-IDF метрики. Також я провів обчислення тональності та категоризацію новин України за допомогою методу LSA. Для творів Дойля та По було намальовано хмару слів, обраховано тональність творів та отримано висновок, що Дойль писав більш похмурі твори, хоча різниця й невелика. При обчисленні тональності творів стикнувся з проблемою незбалансованості оцінок. Вирішив цю проблему шляхом ділення оцінки на загальну кількість слів у корпусі.