Лабораторна робота 5. Варіант 1.

Моделювання тем

Мета роботи: Ознайомитись з вирішенням задач пошуку ключових слів та моделювання тем.

1. Застосувати приховане семантичне індексування бібліотеки Gensim для моделювання тем. Вивести документи, що зробили найбільший вклад в теми. Обрати або створити три нових документи (які модель ще не бачила) та визначити їх теми.
2. Використати текст austen-persuasion.txt з корпусу gutenberg бібліотеки nltk та вивести ключові біграми.

```
import pandas as pd
data = pd.read_csv('news.csv')
data
```

```
                                                    text  label
0      Here are Thursday's biggest analyst calls: App...      0
1      Buy Las Vegas Sands as travel to Singapore bui...      0
2      Piper Sandler downgrades DocuSign to sell, cit...      0
3      Analysts react to Tesla's latest earnings, bre...      0
4      Netflix and its peers are set for a ‘return to...      0
...                                                  ...    ...
16985  KfW credit line for Uniper could be raised to ...      3
16986  KfW credit line for Uniper could be raised to ...      3
16987  Russian  https://t.co/R0iPhyo5p7 sells 1 bln r...      3
16988  Global ESG bond issuance posts H1 dip as supra...      3
16989  Brazil's Petrobras says it signed a $1.25 bill...      3

[16990 rows x 2 columns]
```

```
texts = data['text'].tolist()
```

```
from gensim.parsing.preprocessing import preprocess_string
from gensim import corpora
```

```
# Попередня обробка текстів
processed_texts = [preprocess_string(text) for text in texts]

# Створення словника та корпусу
dictionary = corpora.Dictionary(processed_texts)
corpus = [dictionary.doc2bow(text) for text in processed_texts]
```

Створити модель з модуля gensim

```
from gensim.models import LsiModel

# Побудова LSI моделі
lsi_model = LsiModel(corpus, id2word=dictionary, num_topics=10)
```

Вивести документи, що зробили найбільший вклад в теми.

```
topics = lsi_model.print_topics(num_topics=10, num_words=5)
for i, topic in topics:
    print(f"Topic #{i}: {topic}")

# Визначення внеску кожного документа в теми
corpus_lsi = lsi_model[corpus]

# Визначення топ документів для кожної теми
from collections import defaultdict

topic_contributions = defaultdict(list)
for doc_id, doc in enumerate(corpus_lsi):
    for topic_id, contribution in doc:
        topic_contributions[topic_id].append((contribution, doc_id))

for topic_id, contributions in topic_contributions.items():
    top_docs = sorted(contributions, reverse=True)[:3]
    print(f"\nTop documents for topic #{topic_id}:")
    for contribution, doc_id in top_docs:
        print(f"\nDocument #{doc_id} with contribution {contribution}:
{texts[doc_id]}")
```

```
Topic #0: 0.966*"http" + 0.060*"announc" + 0.058*"market" +
0.056*"new" + 0.055*"stock"
Topic #1: -0.508*"stock" + -0.469*"market" + -0.295*"trade" + -
0.264*"economi" + -0.250*"invest"
Topic #2: 0.377*"quarter" + 0.354*"second" + 0.319*"result" +
0.291*"earn" + 0.251*"announc"
Topic #3: -0.351*"rate" + -0.341*"inflat" + -0.304*"quarter" + -
0.283*"second" + -0.260*"year"
Topic #4: 0.629*"new" + 0.427*"est" + 0.290*"prev" + 0.263*"jun" + -
0.257*"market"
Topic #5: -0.730*"market" + 0.438*"stock" + -0.236*"est" + -
0.160*"jun" + -0.159*"prev"
Topic #6: -0.680*"new" + 0.392*"est" + 0.229*"prev" + 0.216*"jun" + -
0.182*"market"
Topic #7: -0.456*"price" + -0.419*"stock" + 0.396*"bank" +
0.246*"rate" + 0.178*"invest"
Topic #8: -0.473*"stock" + 0.294*"price" + 0.289*"trade" +
0.258*"economi" + -0.256*"market"
Topic #9: -0.410*"beat" + 0.378*"announc" + -0.338*"revenu" + -
0.300*"ep" + -0.294*"earn"
```

```
Top documents for topic #0:

Document #927 with contribution 3.8801262379041486: NovaBay
Pharmaceuticals’ DERMAdoctor Products Now Available at
https://t.co/tQjGxou7L8 and  https://t.co/9nezzpbAhf
https://t.co/XmjxG1Dxfq  https://t.co/6d6mdehcp4

Document #10118 with contribution 3.2152883108474914: South Korea Air
Treatment Systems Markets, 2021-2022 &amp; 2028: Total Markets, Air
Treatment Systems Markets, &amp; Filter Replacement Markets -
https://t.co/guyiBzPH8C  https://t.co/SNWojILr2B
https://t.co/kYjlxlrWE0

Document #10143 with contribution 3.1315457616203695: Global Lending
Market Report to 2031 - Featuring Citi Group, Bank of America and
State Bank of India Among Others -  https://t.co/guyiBzPH8C
https://t.co/CaK9s3yvWN  https://t.co/Q6l1W6mexs

Top documents for topic #1:

Document #1134 with contribution 0.5845055015481952: TotalEnergies SE
UK Regulatory Announcement: Papua New Guinea: TotalEnergies Announces
New Milestone towards Papua LNG Development  https://t.co/IZaoBm04WI
https://t.co/zsQ5lsjy3G

Document #943 with contribution 0.5332510955479566: TotalEnergies SE
UK Regulatory Announcement: United States: TotalEnergies Announces the
Start-up of New Ethane Cracker in Port Arthur  https://t.co/FCZzlm1FWz
https://t.co/6KfPlvGwzy

Document #12386 with contribution 0.5012628568440157:
https://t.co/GGf0VJaxvZ Announces Appointment of Karen Drexler to Its
Board of Directors and Jean-Olivier Racine to Its Advisory Board
https://t.co/oGjocDOMg9  https://t.co/5nkVakOc5Y

Top documents for topic #2:

Document #6890 with contribution 2.7706311737078253: Tesla reported
second-quarter earnings above Wall Street projections, defying
expectations.   $TSLA said it earned $2.3 billion, or $1.95 a share,
in the second quarter, compared with $1.1 billion, or $1.02 a share,
in the second quarter of 2021.  https://t.co/3z9kegGUiG
https://t.co/nBKoOf1J05

Document #4660 with contribution 2.355880077264461: USA Compression
Partners Announces Second Quarter 2022 Distribution; Second Quarter
2022 Earnings Release and Conference Call Scheduled for August 2
https://t.co/kDuvPiRHCU  https://t.co/u1IVYUQGD5

Document #6904 with contribution 2.066809659256738: Tesla ($TSLA) is
```

```
expected to report adjusted earnings of $1.86 a share in the second
quarter, which would compare with adjusted earnings of $1.45 a share
in the second quarter of 2021, according to analysts polled by
FactSet.   https://t.co/JQ7yfa5We9

Top documents for topic #3:

Document #10118 with contribution 0.6372971359487317: South Korea Air
Treatment Systems Markets, 2021-2022 &amp; 2028: Total Markets, Air
Treatment Systems Markets, &amp; Filter Replacement Markets -
https://t.co/guyiBzPH8C  https://t.co/SNWojILr2B
https://t.co/kYjlxlrWE0

Document #9422 with contribution 0.5965214248772098: uncertainty
https://t.co/rCq2sgwOew WATCH: Elon Musk's effort to delay Company's
trial against him flopped in court, after a Delaware judge ruled that
Company's lawsuit seeking to hold Elon Musk to his $44 billion
takeover will go to trial in October  https://t.co/rCq2sgwOew
https://t.co/3EDTIG0N83

Document #9595 with contribution 0.5488179314996751: Northleaf Capital
Partners to Acquire 40% Interest in New Zealand Mobile Tower
Infrastructure Business From Vodafone New Zealand Limited
https://t.co/ruFWvKrjWT  https://t.co/HFx3Ekz1gU

Top documents for topic #4:

Document #11112 with contribution 6.676393064391269: #OATT | Global
#WASDE Corn End Stocks New Jun: 313M (est 311M; prev 310M) - Soybean
End Stocks New: 100M (est 99M; prev 100M) - Wheat End Stocks New: 268M
(est 266M; prev 267M) - Cotton End Stocks New: 84M (est 82M; prev 83M)

Document #11113 with contribution 6.455319537916789: #OATT | US #WASDE
Corn End Stocks New Jul: 1470M (est 1450M; prev 1400M)     - Soybean
End Stocks New: 230M (est 203M; prev 280M)     - Wheat End Stocks New:
639M (est 641M; prev 627M)     - Cotton End Stocks New: 2.40M (est
2.79M; prev 2.90M)     https://t.co/twS4Y9iqnH

Document #10762 with contribution 4.753817550619759: U.S CPI (MOM)
(JUN) ACTUAL: 1.3% VS 1.0% PREVIOUS; EST 1.1%  U.S CPI (YOY) (JUN)
ACTUAL: 9.1% VS 8.6% PREVIOUS; EST 8.8%  U.S CORE CPI (MOM) (JUN)
ACTUAL: 0.7% VS 0.6% PREVIOUS; EST 0.5%  U.S CORE CPI (YOY) (JUN)
ACTUAL: 5.9% VS 6.0% PREVIOUS; EST 5.7%

Top documents for topic #5:

Document #14933 with contribution 1.327968131943603: Linear stocks
&gt; choppy stocks. Where a stock has come from &gt; relative
strength. Entries matter.
```

```
Document #4366 with contribution 1.3091793308907116: $C bucks the
banking trending and reports surprisingly big upside $JPM $XLF $MS
$WFC  https://t.co/z0XuABKmsY #earnings #economy #banks #financials
#inflation #Investment #banking #bank #stocks

Document #8653 with contribution 1.3043316607763182: So levered
single-stock ETFs are okay but volatility products using most liquid
option contracts (SPY) in the world are not? @HesterPeirce   Investors
in U.S. have a new way to supersize bets on high-profile stocks, with
the launch of single-stock ETFs   https://t.co/zgB5b3DsqD

Top documents for topic #6:

Document #11026 with contribution 4.157251779070796: Eurozone CPI
(Y/Y) Jun F: 8.6% (est 8.6%; prev 8.6%)  - Eurozone CPI (M/M) Jun F:
0.8% (est 0.8%; prev 0.8%)  - Eurozone CPI Core (M/M) Jun F: 0.2% (est
0.2%; prev 0.2%)  - Eurozone CPI Core (Y/Y) Jun F: 3.7% (est 3.7%;
prev 3.7%)

Document #10762 with contribution 4.140361563349039: U.S CPI (MOM)
(JUN) ACTUAL: 1.3% VS 1.0% PREVIOUS; EST 1.1%  U.S CPI (YOY) (JUN)
ACTUAL: 9.1% VS 8.6% PREVIOUS; EST 8.8%  U.S CORE CPI (MOM) (JUN)
ACTUAL: 0.7% VS 0.6% PREVIOUS; EST 0.5%  U.S CORE CPI (YOY) (JUN)
ACTUAL: 5.9% VS 6.0% PREVIOUS; EST 5.7%

Document #11013 with contribution 4.139541012334429: South African CPI
(M/M) Jun: 1.1% (est 0.9%; prev 0.7%)  - South African CPI (Y/Y) Jun:
7.4% (est 7.3%; prev 6.5%)  - South African CPI Core (M/M) Jun: 0.6%
(est 0.5%; prev 0.2%)  - South African CPI Core (Y/Y) Jun: 4.4% (est
4.3%; prev 4.1%)

Top documents for topic #7:

Document #5308 with contribution 1.6930081714065317: $BAC | Bank Of
America Q2 22 Earnings:  - EPS: 0.73$ (est $0.76)  - Revenue: $22.69B
(est $22.86B)  - Wealth &amp; Investment Rev: $5.43B (est $5.43B)  -
Trading Revenue EX DVA :$4B (est $4.01B)  - FICC Sales &amp; Trading
Rev: $2.34B (est $2.29B)

Document #4366 with contribution 1.6015048204900764: $C bucks the
banking trending and reports surprisingly big upside $JPM $XLF $MS
$WFC  https://t.co/z0XuABKmsY #earnings #economy #banks #financials
#inflation #Investment #banking #bank #stocks

Document #5969 with contribution 1.4221833967948219: UK Foreign
Secretary Liz Truss’s citing of the Bank of Japan’s inflation mandate
as a potential model for the Bank of England isn’t convincing central
bank watchers  https://t.co/bQMHhN5YaU

Top documents for topic #8:
```

```
Document #5579 with contribution 1.353332097960103: "I think you're
going to see supply upended at the end of the year, if not before,"
Truist's Neal Dingmann says, adding: "I do think... that if trade
flows and price caps happen... there's a very good chance for some
price spikes in oil."  https://t.co/Jb7k9tf4PH
https://t.co/zlGyhqqXLf

Document #15702 with contribution 1.2075195651548056: $TPG - TPG:
Short-Term Price Risk But Attractive Long-Term Business Fundamentals.
https://t.co/6JlcLtEySr #stockmarket #economy #investing

Document #11188 with contribution 1.1623802065821698: U.S.-China Trade
Rebounds As China Eases Its Zero-COVID Policy Lockdowns.
https://t.co/ZgQEdwk5IQ #trading #business #economy

Top documents for topic #9:

Document #10666 with contribution 1.7498802371380688: CANADA CPI (MOM)
(JUN) ACTUAL: 0.7% VS 1.4% PREVIOUS; EST 0.9%  CANADA CPI (YOY) (JUN)
ACTUAL: 8.1% VS 7.7% PREVIOUS; EST 8.4%  CANADA CORE CPI (MOM) (JUN)
ACTUAL: 0.3% VS 0.8% PREVIOUS  CANADA CORE CPI (YOY) (JUN) ACTUAL:
6.2% VS 6.1% PREVIOUS; EST 5.9%  @MtlExchange

Document #10762 with contribution 1.7327823286831745: U.S CPI (MOM)
(JUN) ACTUAL: 1.3% VS 1.0% PREVIOUS; EST 1.1%  U.S CPI (YOY) (JUN)
ACTUAL: 9.1% VS 8.6% PREVIOUS; EST 8.8%  U.S CORE CPI (MOM) (JUN)
ACTUAL: 0.7% VS 0.6% PREVIOUS; EST 0.5%  U.S CORE CPI (YOY) (JUN)
ACTUAL: 5.9% VS 6.0% PREVIOUS; EST 5.7%

Document #11026 with contribution 1.6504801282488137: Eurozone CPI
(Y/Y) Jun F: 8.6% (est 8.6%; prev 8.6%)  - Eurozone CPI (M/M) Jun F:
0.8% (est 0.8%; prev 0.8%)  - Eurozone CPI Core (M/M) Jun F: 0.2% (est
0.2%; prev 0.2%)  - Eurozone CPI Core (Y/Y) Jun F: 3.7% (est 3.7%;
prev 3.7%)
```

```python
new_texts = [
    "Analysts say Apple will continue to grow in the next quarter.",
    "Tesla's new model has impressed the market with its advanced
features.",
    "Amazon's stock prices are predicted to rise due to increased
sales."
]

new_processed_texts = [preprocess_string(text) for text in new_texts]
new_corpus = [dictionary.doc2bow(text) for text in
new_processed_texts]
new_corpus_lsi = lsi_model[new_corpus]

for i, doc in enumerate(new_corpus_lsi):
    print(f"New document #{i + 1}: {new_texts[i]}")
    for topic_id, contribution in doc:
        print(f" - Topic #{topic_id + 1} with contribution
{contribution}")

New document #1: Analysts say Apple will continue to grow in the next
quarter.
 - Topic #1 with contribution 0.0701707757892538
 - Topic #2 with contribution 0.02605186813960929
 - Topic #3 with contribution 0.33907283961691104
 - Topic #4 with contribution -0.3150724896901351
 - Topic #5 with contribution 0.009183449897806381
 - Topic #6 with contribution -0.08875207098864706
 - Topic #7 with contribution -0.11517570500442723
 - Topic #8 with contribution -0.13454558586140025
 - Topic #9 with contribution 0.03997909510178691
 - Topic #10 with contribution 0.008109876492719797
New document #2: Tesla's new model has impressed the market with its
advanced features.
 - Topic #1 with contribution 0.12970955198419243
 - Topic #2 with contribution -0.4276164854458416
 - Topic #3 with contribution -0.23161052124978343
 - Topic #4 with contribution 0.16704065259199743
 - Topic #5 with contribution 0.38628609795852714
 - Topic #6 with contribution -0.7702722717163895
 - Topic #7 with contribution -0.8674165923519286
 - Topic #8 with contribution -0.00011779822446590519
 - Topic #9 with contribution -0.2648436688181909
 - Topic #10 with contribution -0.28043369532303337
New document #3: Amazon's stock prices are predicted to rise due to
increased sales.
 - Topic #1 with contribution 0.1374158945219514
 - Topic #2 with contribution -0.565506307733751
 - Topic #3 with contribution -0.34045088648857896
 - Topic #4 with contribution -0.3248558149483241
 - Topic #5 with contribution 0.2076756959083238
 - Topic #6 with contribution 0.3597957302689959
 - Topic #7 with contribution 0.17682690435718748
 - Topic #8 with contribution -0.9868809722971388
 - Topic #9 with contribution -0.11910469732916602
 - Topic #10 with contribution 0.12563470079476136
```

1. Використати текст austen-persuasion.txt з корпусу gutenberg бібліотеки nltk та вивести ключові біграми.

```
import nltk
from nltk.corpus import gutenberg, stopwords
from nltk.collocations import BigramCollocationFinder
from nltk.metrics import BigramAssocMeasures
import string
```

```
# Завантажимо текст
nltk.download('gutenberg')
nltk.download('punkt')
persuasion_text = gutenberg.raw('austen-persuasion.txt')
stop_words = set(stopwords.words('english'))

# Токенізація
tokens = nltk.word_tokenize(persuasion_text)
cleaned_tokens = [token for token in tokens if token not in stop_words
and token not in string.punctuation and token not in ['``', "''"]]

# Знаходження біграм
bigram_finder = BigramCollocationFinder.from_words(cleaned_tokens)
bigrams = bigram_finder.nbest(BigramAssocMeasures.likelihood_ratio,
10)

print("Key Bigrams:")
for bigram in bigrams:
    print(bigram)
```

```
[nltk_data] Downloading package gutenberg to
[nltk_data]     C:\Users\murat\AppData\Roaming\nltk_data...
[nltk_data]   Package gutenberg is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\murat\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
Key Bigrams:
('Captain', 'Wentworth')
('Lady', 'Russell')
('Sir', 'Walter')
('Mr', 'Elliot')
('Mrs', 'Clay')
('Mrs', 'Smith')
('Captain', 'Benwick')
('Mrs', 'Musgrove')
('Camden', 'Place')
('great', 'deal')
```