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

Лабораторна робота №5

Аналіз текстів на мові Python

Тема: Моделювання тем

Варіант: 1

Виконав
студент групи ІП-11:
Панченко С. В.

Перевірив:
Тимофєєва Ю. С

Київ 2023

ЗМІСТ

1 МЕТА ЛАБОРАТОРНОЇ РОБОТИ

Ознайомитись з вирішенням задач пошуку ключових слів та моделювання тем.

2 ЗАВДАННЯ

Застосувати приховане семантичне індексування бібліотеки Gensim для моделювання тем. Вивести документи, що зробили найбільший вклад в теми. Обрати три нових документи та визначити їх теми.

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

3 ВИКОНАННЯ

3.1 Використання тексту austen-persuasion.txt з корпусу gutenberg бібліотеки nltk та виведення ключових біграм.

Завантажимо текст austen-persuasion.txt.

In [13]:
```
import nltk
import re
from nltk.corpus import gutenberg
import numpy as np
name = 'austen-persuasion.txt'
text = [' '.join(sent) for sent in gutenberg.sents(name)]
text[:30]
```

Out[13]:
```
['[ Persuasion by Jane Austen 1818 ]',
 'Chapter 1',
 'Sir Walter Elliot , of Kellynch Hall , in Somersetshire , was a man who , for his own am
usement , never took up any book but the Baronetage ; there he found occupation for an idl
e hour , and consolation in a distressed one ; there his faculties were roused into admira
tion and respect , by contemplating the limited remnant of the earliest patents ; there an
y unwelcome sensations , arising from domestic affairs changed naturally into pity and con
tempt as he turned over the almost endless creations of the last century ; and there , if
every other leaf were powerless , he could read his own history with an interest which nev
er failed .',
 'This was the page at which the favourite volume always opened :',
 '" ELLIOT OF KELLYNCH HALL .',
 '" Walter Elliot , born March 1 , 1760 , married , July 15 , 1784 , Elizabeth , daughter
of James Stevenson , Esq .',
 'of South Park , in the county of Gloucester , by which lady ( who died 1800 ) he has iss
ue Elizabeth , born June 1 , 1785 ; Anne , born August 9 , 1787 ; a still - born son , Nov
ember 5 , 1789 ; Mary , born November 20 , 1791 ."',
 'Precisely such had the paragraph originally stood from the printer \' s hands ; but Sir
Walter had improved it by adding , for the information of himself and his family , these w
ords , after the date of Mary \' s birth -- " Married , December 16 , 1810 , Charles , son
and heir of Charles Musgrove , Esq .',
 'of Uppercross , in the county of Somerset ," and by inserting most accurately the day of
the month on which he had lost his wife .',
 'Then followed the history and rise of the ancient and respectable family , in the usual
terms ; how it had been first settled in Cheshire ; how mentioned in Dugdale , serving the
office of high sheriff , representing a borough in three successive parliaments , exertion
s of loyalty , and dignity of baronet , in the first year of Charles II , with all the Mar
ys and Elizabeths they had married ; forming altogether two handsome duodecimo pages , and
concluding with the arms and motto :--" Principal seat , Kellynch Hall , in the county of
Somerset ," and Sir Walter \' s handwriting again in this finale :--',
 '" Heir presumptive , William Walter Elliot , Esq ., great grandson of the second Sir Wal
ter ."',
 "Vanity was the beginning and the end of Sir Walter Elliot ' s character ; vanity of pers
on and of situation .",
 'He had been remarkably handsome in his youth ; and , at fifty - four , was still a very
fine man .',
 'Few women could think more of their personal appearance than he did , nor could the vale
t of any new made lord be more delighted with the place he held in society .',
 'He considered the blessing of beauty as inferior only to the blessing of a baronetcy ; a
nd the Sir Walter Elliot , who united these gifts , was the constant object of his warmest
respect and devotion .',
 'His good looks and his rank had one fair claim on his attachment ; since to them he must
have owed a wife of very superior character to any thing deserved by his own .',
 'Lady Elliot had been an excellent woman , sensible and amiable ; whose judgement and con
duct , if they might be pardoned the youthful infatuation which made her Lady Elliot , had
never required indulgence afterwards .-- She had humoured , or softened , or concealed his
failings , and promoted his real respectability for seventeen years ; and though not the v
ery happiest being in the world herself , had found enough in her duties , her friends , a
nd her children , to attach her to life , and make it no matter of indifference to her whe
n she was called on to quit them .',
 '-- Three girls , the two eldest sixteen and fourteen , was an awful legacy for a mother
to bequeath , an awful charge rather , to confide to the authority and guidance of a conce
ited , silly father .',
 'She had , however , one very intimate friend , a sensible , deserving woman , who had be
en brought , by strong attachment to herself , to settle close by her , in the village of
Kellynch ; and on her kindness and advice , Lady Elliot mainly relied for the best help an
d maintenance of the good principles and instruction which she had been anxiously giving h
er daughters .',
 'This friend , and Sir Walter , did not marry , whatever might have been anticipated on t
hat head by their acquaintance .',
 "Thirteen years had passed away since Lady Elliot ' s death , and they were still near ne
ighbours and intimate friends , and one remained a widower , the other a widow .",
 'Mrs Russell , of steady age and character , and extremely well provided for , shou
```

Рисунок 3.1.1 - Зчитування тексту

Визначимо стоп-слова англійської мови.

```
In [14]: wpt = nltk.WordPunctTokenizer()
         nltk.corpus.stopwords.words('english')
```


Рисунок 3.1.2 - Стоп-слова

Визначимо функцію, що виконує попередню обробку документу. Застосуємо декоратор np.vectorize для того, щоб функція могла працювати з корпусами.

```
In [15]: @np.vectorize
         def preproc_doc(doc):
             doc = re.sub(r'[^a-zA-Z\s]', '', doc, re.I | re.A)
             doc = doc.lower()
             doc = doc.strip()
             tokens = wpt.tokenize(doc)
             filtered_tokens = [token for token in tokens if token not in stop_words]
             doc = ' '.join(filtered_tokens)
             return doc


         text = preproc_doc(text)
         text

Out[15]: array(['persuasion jane austen', 'chapter',
                'sir walter elliot kellynch hall somersetshire man amusement never took book barone
         tage found occupation idle hour consolation distressed one faculties roused admiration res
         pect contemplating limited remnant earliest patents unwelcome sensations arising domestic
         affairs changed naturally pity contempt turned almost endless creations last century every
         leaf powerless could read history interest never failed',
                ...,
                'profession could ever make friends wish tenderness less dread future war could dim
         sunshine',
                'gloried sailor wife must pay tax quick alarm belonging profession possible disting
         uished domestic virtues national importance',
                '], dtype='<U662')
```


Рисунок 3.1.3 - Обробка документів

Завантажимо функції для пошуку сполучень та визначення тих, що зустрічаються найчастіше, або тих, що мають найвищі значення інших показників, наприклад, поточкової взаємної інформації.

```
In [16]: from nltk.collocations import BigramCollocationFinder
         from nltk.collocations import BigramAssocMeasures
         bigram_measures = BigramAssocMeasures()
         finder = BigramCollocationFinder.from_documents(
             [item.split() for item in text])
         finder.nbest(bigram_measures.raw_freq, 10)

Out[16]: [('captain', 'wentworth'),
          ('mr', 'elliot'),
          ('lady', 'russell'),
          ('sir', 'walter'),
          ('mrs', 'clay'),
          ('mrs', 'musgrove'),
          ('mrs', 'smith'),
          ('captain', 'benwick'),
          ('miss', 'elliot'),
```


Рисунок 3.1.4 - Ключові біграми

3.2Застосування прихованого семантичного індексування бібліотеки Gensim для моделювання тем.

Для початку імпортуємо модулі та зчитаємо файл.

In [17]:
```
import pandas as pd
df = pd.read_csv('bbc-news-data.csv', sep='\t')
df
```

Out[17]:

	category	filename	title	content
0	business	001.txt	Ad sales boost Time Warner profit	Quarterly profits at US media giant TimeWarne...
1	business	002.txt	Dollar gains on Greenspan speech	The dollar has hit its highest level against ...
2	business	003.txt	Yukos unit buyer faces loan claim	The owners of embattled Russian oil giant Yuk...
3	business	004.txt	High fuel prices hit BA's profits	British Airways has blamed high fuel prices f...
4	business	005.txt	Pernod takeover talk lifts Domecq	Shares in UK drinks and food firm Allied Dome...
...	...	...	...	...
2220	tech	397.txt	BT program to beat dialler scams	BT is introducing two initiatives to help bea...
2221	tech	398.txt	Spam e-mails tempt net shoppers	Computer users across the world continue to i...
2222	tech	399.txt	Be careful how you code	A new European directive could put software w...
2223	tech	400.txt	US cyber security chief resigns	The man making sure US computer networks are ...
2224	tech	401.txt	Losing yourself in online gaming	Online role playing games are time-consuming,...

2225 rows × 4 columns

Рисунок 3.2.1 - Зчитування файлу

Виділимо лише колонку "content"

In [18]:
```
text = df['content'].values
text
```

Out[18]:
```
array([' Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for
the three months to December, from $639m year-earlier.  The firm, which is now one of the
biggest investors in Google, benefited from sales of high-speed internet connections and h
igher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn.
Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and les
s users for AOL.  Time Warner said on Friday that it now owns 8% of search-engine Google.
But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in
the fourth quarter profits were lower than in the preceding three quarters. However, the c
ompany said AOL\'s underlying profit before exceptional items rose 8% on the back of stron
ger internet advertising revenues. It hopes to increase subscribers by offering the online
service free to TimeWarner internet customers and will try to sign up AOL\'s existing cust
omers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results follo
wing a probe by the US Securities Exchange Commission (SEC), which is close to concluding.
Time Warner\'s fourth quarter profits were slightly better than analysts\' expectations. B
ut its film division saw profits slump 27% to $284m, helped by box-office flops Alexander
and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord
of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.
36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financi
al performance was strong, meeting or exceeding all of our full-year objectives and greatl
y enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005,
TimeWarner is projecting operating earnings growth of around 5%, and also expects higher r
evenue and wider profit margins.  TimeWarner is to restate its accounts as part of efforts
to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300
m to settle charges, in a deal that is under review by the SEC. The company said it was un
able to estimate the amount it needed to set aside for legal reserves, which it previously
set at $500m. It intends to adjust the way it accounts for a deal with German music publis
her Bertelsmann\'s purchase of a stake in AOL Europe, which it had reported as advertising
revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of th
at stake. ',
       ' The dollar has hit its highest level against the euro in almost three months afte
r the Federal Reserve head said the US trade deficit is set to stabilise.  And Alan Greens
pan highlighted the US government\'s willingness to curb spending and rising household sav
ings as factors which may help to reduce it. In late trading in New York, the dollar reach
ed $1.2871 against the euro, from $1.2974 on Thursday. Market concerns about the deficit h
as hit the greenback in recent months. On Friday, Federal Reserve chairman Mr Greenspan\'s
speech in London ahead of the meeting of G7 finance ministers sent the dollar higher after
it had earlier tumbled on the back of worse-than-expected US jobs data. "I think the chair
man\'s taking a much more sanguine view on the current account deficit than he\'s taken fo
r some time," said Robert Sinche, head of currency strategy at Bank of America in New Yor
k. "He\'s taking a longer-term view, laying out a set of conditions under which the curren
t account deficit can improve this year and next."  Worries about the deficit concerns abo
ut China do, however, remain. China\'s currency remains pegged to the dollar and the US cu
rrency\'s sharp falls in recent months have therefore made Chinese export prices highly co
mpetitive. But calls for a shift in Beijing\'s policy have fallen on deaf ears, despite re
cent comments in a major Chinese newspaper that the "time is ripe" for a loosening of the
peg. The G7 meeting is thought unlikely to produce any meaningful movement in Chinese poli
cy. In the meantime, the US Federal Reserve\'s decision on 2 February to boost interest ra
tes by a quarter of a point - the sixth such move in as many months - has opened up a diff
erential with European rates. The half-point window, some believe, could be enough to keep
US assets looking more attractive, and could help prop up the dollar. The recent falls hav
e partly been the result of big budget deficits, as well as the US\'s yawning current acco
unt gap, both of which need to be funded by the buying of US bonds and assets by foreign f
irms and governments. The White House will announce its budget on Monday, and many comment
ators believe the deficit will remain at close to half a trillion dollars. ',
       ' The owners of embattled Russian oil giant Yukos are to ask the buyer of its forme
r production unit to pay back a $900m (£479m) loan.  State-owned Rosneft bought the Yugans
k unit for $9.3bn in a sale forced by Russia to part settle a $27.5bn tax claim against Yu
kos. Yukos\' owner Menatep Group says it will ask Rosneft to repay a loan that Yugansk had
secured on its assets. Rosneft already faces a similar $540m repayment demand from foreign
banks. Legal experts said Rosneft\'s purchase of Yugansk would include such obligations.
"The pledged assets are with Rosneft, so it will have to pay real money to the creditors t
o avoid seizure of Yugansk assets," said Moscow-based US lawyer Jamie Firestone, who is no
t connected to the case. Menatep Group\'s managing director Tim Osborne told the Reuters n
ews agency: "If they default, we will fight them where the rule of law exists under the in
ternational arbitration clauses of the credit."  Rosneft officials were unavailable for co
mment. But the company has said it intends to take action against Menatep to recover some
of the tax claims and debts owed by Yugansk. Yukos had filed for bankruptcy protection in
a US court in an attempt to prevent the forced sale of its main production arm. The sale w
ent ahead in December and Yugansk was sold to a little-known shell company which in turn w
as bought by Rosneft. Yukos claims its downfall was punishment for the political ambitions
of its founder Mikhail Khodorkovsky and has vowed to sue any participant in the sale. ',
       ...,
       ' A new European directive could put software writers at risk of legal action, warn
s former programmer and technology analyst Bill Thompson.  If it gets its way, the Dutch g
overnment will conclude its presidency of the European Union by pushing through a controve
rsial measure that has been rejected by the European Parliament, lacks majority support fr
om national governments and will leave millions of European citizens in legal limbo and fa
cing the possibility of court cases against them. If the new law was about border control
s, defence or even the new constitution, then our TV screens would be full of experts agon
ising over the impact on our daily lives. Sadly for those who will be directly affected, t
he controversy concerns the patenting of computer programs, a topic that may excite the bl
```

Рисунок 3.2.2 - Виділення тексту

Нормалізуємо текст.

In [19]:
```
text = preproc_doc(text)
text
```

Out[19]:
```
array(['quarterly profits us media giant timewarner jumped bn three months december yearea
rlier firm one biggest investors google benefited sales highspeed internet connections hig
her advert sales timewarner said fourth quarter sales rose bn bn profits buoyed oneoff gai
ns offset profit dip warner bros less users aol time warner said friday owns searchengine
google internet business aol mixed fortunes lost subscribers fourth quarter profits lower
preceding three quarters however company said aols underlying profit exceptional items ros
e back stronger internet advertising revenues hopes increase subscribers offering online s
ervice free timewarner internet customers try sign aols existing customers highspeed broad
band timewarner also restate results following probe us securities exchange commission sec
close concluding time warners fourth quarter profits slightly better analysts expectations
film division saw profits slump helped boxoffice flops alexander catwoman sharp contrast y
earearlier third final film lord rings trilogy boosted results fullyear timewarner posted
profit bn performance revenues grew bn financial performance strong meeting exceeding full
year objectives greatly enhancing flexibility chairman chief executive richard parsons sai
d timewarner projecting operating earnings growth around also expects higher revenue wider
profit margins timewarner restate accounts part efforts resolve inquiry aol us market regu
lators already offered pay settle charges deal review sec company said unable estimate amo
unt needed set aside legal reserves previously set intends adjust way accounts deal german
music publisher bertelsmanns purchase stake aol europe reported advertising revenue book s
ale stake aol europe loss value stake',
       'dollar hit highest level euro almost three months federal reserve head said us tra
de deficit set stabilise alan greenspan highlighted us governments willingness curb spendi
ng rising household savings factors may help reduce late trading new york dollar reached e
uro thursday market concerns deficit hit greenback recent months friday federal reserve ch
airman mr greenspans speech london ahead meeting g finance ministers sent dollar higher ea
rlier tumbled back worsethanexpected us jobs data think chairmans taking much sanguine vie
w current account deficit hes taken time said robert sinche head currency strategy bank am
erica new york hes taking longerterm view laying set conditions current account deficit im
prove year next worries deficit concerns china however remain chinas currency remains pegg
ed dollar us currencys sharp falls recent months therefore made chinese export prices high
ly competitive calls shift beijings policy fallen deaf ears despite recent comments major
chinese newspaper time ripe loosening peg g meeting thought unlikely produce meaningful mo
vement chinese policy meantime us federal reserves decision february boost interest rates
quarter point sixth move many months opened differential european rates halfpoint window b
elieve could enough keep us assets looking attractive could help prop dollar recent falls
partly result big budget deficits well uss yawning current account gap need funded buying
us bonds assets foreign firms governments white house announce budget monday many commenta
tors believe deficit remain close half trillion dollars',
       'owners embattled russian oil giant yukos ask buyer former production unit pay back
loan stateowned rosneft bought yugansk unit bn sale forced russia part settle bn tax claim
yukos yukos owner menatep group says ask rosneft repay loan yugansk secured assets rosneft
already faces similar repayment demand foreign banks legal experts said rosnefts purchase
yugansk would include obligations pledged assets rosneft pay real money creditors avoid se
izure yugansk assets said moscowbased us lawyer jamie firestone connected case menatep gro
ups managing director tim osborne told reuters news agency default fight rule law exists i
nternational arbitration clauses credit rosneft officials unavailable comment company said
intends take action menatep recover tax claims debts owed yugansk yukos filed bankruptcy p
rotection us court attempt prevent forced sale main production arm sale went ahead decembe
r yugansk sold littleknown shell company turn bought rosneft yukos claims downfall punishm
ent political ambitions founder mikhail khodorkovsky vowed sue participant sale',
       ...,
       'new european directive could put software writers risk legal action warns former p
rogrammer technology analyst bill thompson gets way dutch government conclude presidency e
uropean union pushing controversial measure rejected european parliament lacks majority su
pport national governments leave millions european citizens legal limbo facing possibility
court cases new law border controls defence even new constitution tv screens would full ex
perts agonising impact daily lives sadly directly affected controversy concerns patenting
computer programs topic may excite bloggers campaigning groups technical press obsess midd
le britain much fuss generate directive patentability computerimplemented inventions way a
mends article european patent convention yet new directive nodded next meeting one eus min
isterial councils seems likely allow programs patented europe us many observers computing
scene including think results disastrous small companies innovative programmers free open
source software movement let large companies patent sorts ideas give legal force want limi
t competitors use really obvious ideas us cannot build system stores customer credit card
details pay without reenter unless amazon lets hold patent oneclick online purchase small
invention amazon made patent office first owns relatively free sort thing perhaps long new
proposals go back although argument patentability software computerimplemented inventions
going since least mids come head year proposals made endorsed council ministers radically
modified european parliament represented original form national governments seem aware pro
blems poland rejected proposal germanys main political parties opposed enough opposition g
uarantee rejection early december british government held consultation meeting commented p
roposals science minister lord sainsbury went along listen outline uk position according p
resent embarrassing see little minister officials actually understood issues concerned dra
ft directive put council called item approved rejected discussion amendment allowed worrie
d first abuse democratic process involved disregarding views parliament abandoning careful
ly argued amendments goes heart european project even care software patents worried coders
treated like today say tomorrow directly software patents granted programmer worry code wr
iting infringing someone elses patent stealing software code already protected copyright p
atents copyright something much stronger patent gives owner right stop anyone else using i
nvention even person invented separately never shame managed read lord byrons childe harol
ds pilgrimage pointed one articles contained substantial chunk poem could defend court cla
```

Рисунок 3.2.3 - Нормалізація тексту

Розділимо текст на матрицю слів.

In [20]:
```
sentences = [sent.split() for sent in text]
sentences[:10]
```

Out[20]:
```
[['quarterly',
  'profits',
  'us',
  'media',
  'giant',
  'timewarner',
  'jumped',
  'bn',
  'three',
  'months',
  'december',
  'yearearlier',
  'firm',
  'one',
  'biggest',
  'investors',
  'google',
  'benefited',
  'sales',
  'highspeed',
  'internet',
  'connections',
  'higher',
  'advert',
  'sales',
  'timewarner',
  'said',
  'fourth',
  'quarter',
  'sales',
  'rose',
  'bn',
  'bn',
  'profits',
  'buoyed',
  'oneoff',
  'gains',
  'offset',
  'profit',
  'dip',
  'warner',
  'bros',
  'less',
  'users',
  'aol',
  'time',
  'warner',
  'said',
  'friday',
  'owns',
  'searchengine',
  'google',
  'internet',
  'business',
  'aol',
  'mixed',
  'fortunes',
  'lost',
  'subscribers',
  'fourth',
  'quarter',
  'profits',
  'lower',
  'preceding',
  'three',
  'quarters',
  'however',
  'company',
  'said',
  'aols',
  'underlying',
  'profit',
  'exceptional',
  'items',
  'rose',
  'back',
  'stronger',
  'internet',
  'advertising',
  'revenues',
  'hopes',
```

Рисунок 3.2.4 - Речення

Виділяємо біграми для всіх документів та створюємо словник.

```
In [21]: from gensim.models.phrases import Phrases, Phraser, ENGLISH_CONNECTOR_WORDS
         bigram = Phrases(sentences, min_count=20, threshold=20,
                          connector_words=ENGLISH_CONNECTOR_WORDS)
         bigram_model = Phraser(bigram)
```

Рисунок 3.2.5 - Модель

Виділяємо біграми для всіх документів та створюємо словник.

In [22]:
```
from gensim.corpora import Dictionary
norm_corpus_bigrams = [bigram_model[sent] for sent in sentences]
dictionary = Dictionary(norm_corpus_bigrams)
norm_corpus_bigrams[:20]
```

Out[22]:
```
[['quarterly',
  'profits',
  'us',
  'media',
  'giant',
  'timewarner',
  'jumped',
  'bn',
  'three_months',
  'december',
  'yearearlier',
  'firm',
  'one',
  'biggest',
  'investors',
  'google',
  'benefited',
  'sales',
  'highspeed',
  'internet',
  'connections',
  'higher',
  'advert',
  'sales',
  'timewarner',
  'said',
  'fourth_quarter',
  'sales',
  'rose',
  'bn_bn',
  'profits',
  'buoyed',
  'oneoff',
  'gains',
  'offset',
  'profit',
  'dip',
  'warner',
  'bros',
  'less',
  'users',
  'aol',
  'time',
  'warner',
  'said',
  'friday',
  'owns',
  'searchengine',
  'google',
  'internet',
  'business',
  'aol',
  'mixed',
  'fortunes',
  'lost',
  'subscribers',
  'fourth_quarter',
  'profits',
  'lower',
  'preceding',
  'three',
  'quarters',
  'however',
  'company',
  'said',
  'aols',
  'underlying',
  'profit',
  'exceptional',
  'items',
  'rose',
  'back',
  'stronger',
  'internet',
  'advertising',
  'revenues',
  'hopes',
  'increase',
  'subscribers',
```

Рисунок 3.2.6 - Біграми документів

Зменшимо об'єм словника через велику кількість унікальних рідкісних слів. та створюємо модель сумки слів.

In [23]:
```
dictionary.filter_extremes(no_below=20, no_above=0.6)
bow_corpus = [dictionary.doc2bow(text) for text in norm_corpus_bigrams]
bow_corpus[:20]
```

Out[23]:
```
[[(0, 2),
  (1, 2),
  (2, 1),
  (3, 1),
  (4, 2),
  (5, 1),
  (6, 1),
  (7, 1),
  (8, 1),
  (9, 1),
  (10, 1),
  (11, 1),
  (12, 3),
  (13, 1),
  (14, 1),
  (15, 1),
  (16, 1),
  (17, 1),
  (18, 1),
  (19, 1),
  (20, 1),
  (21, 1),
  (22, 1),
  (23, 2),
  (24, 1),
  (25, 1),
  (26, 2),
  (27, 2),
  (28, 1),
  (29, 1),
  (30, 1),
  (31, 1),
  (32, 1),
  (33, 2),
  (34, 1),
  (35, 1),
  (36, 1),
  (37, 2),
  (38, 1),
  (39, 1),
  (40, 1),
  (41, 1),
  (42, 1),
  (43, 3),
  (44, 1),
  (45, 1),
  (46, 1),
  (47, 1),
  (48, 1),
  (49, 2),
  (50, 1),
  (51, 1),
  (52, 1),
  (53, 2),
  (54, 2),
  (55, 1),
  (56, 1),
  (57, 1),
  (58, 1),
  (59, 4),
  (60, 1),
  (61, 1),
  (62, 1),
  (63, 1),
  (64, 1),
  (65, 1),
  (66, 1),
  (67, 1),
  (68, 1),
  (69, 1),
  (70, 1),
  (71, 1),
  (72, 1),
  (73, 1),
  (74, 1),
  (75, 1),
  (76, 1),
  (77, 1),
  (78, 1),
  (79, 1),
```

Рисунок 3.2.7 - Сумка слів

Застосуємо приховане семантичне індексування.

```
In [24]: from gensim.models import LsiModel
         total_topics = 10
         lsi_bow = LsiModel(bow_corpus, id2word=dictionary,
                            num_topics=total_topics,
                            onepass=True, chunksize=10000,
                            power_iters=1000)
```

Рисунок 3.2.8 - Приховане семантичне індексування

Переглянемо основні теми.

```
In [25]: for topic_id, topic in lsi_bow.print_topics(num_topics=10, num_words=20):
             print('Topic #'+str(topic_id+1)+':')
             print(topic)
```

```
Topic #1:
0.266*"would" + 0.254*"people" + 0.198*"mr" + 0.183*"also" + 0.172*"one" + 0.171*"new" +
0.161*"us" + 0.153*"could" + 0.125*"music" + 0.122*"government" + 0.117*"like" + 0.107*"ti
me" + 0.100*"get" + 0.096*"many" + 0.093*"first" + 0.090*"make" + 0.090*"year" + 0.085*"tw
o" + 0.083*"uk" + 0.082*"way"
Topic #2:
-0.447*"music" + 0.262*"mr" + 0.228*"would" + -0.216*"best" + 0.207*"government" + -0.160
*"game" + -0.137*"song" + 0.131*"labour" + -0.114*"awards" + -0.107*"games" + -0.092*"win"
+ -0.091*"award" + -0.090*"good" + -0.088*"like" + -0.088*"film" + -0.088*"think" + -0.087
*"last" + -0.086*"play" + 0.082*"bn" + 0.082*"plans"
Topic #3:
0.337*"music" + 0.303*"people" + -0.210*"game" + 0.158*"technology" + -0.134*"win" + -0.13
3*"england" + -0.125*"wales" + 0.120*"users" + 0.120*"mobile" + -0.116*"first" + -0.104*"t
wo" + 0.102*"services" + 0.102*"use" + 0.101*"digital" + -0.098*"back" + 0.094*"net" + -0.
093*"best" + 0.089*"tv" + 0.088*"software" + 0.087*"broadband"
Topic #4:
0.425*"music" + -0.247*"game" + -0.225*"games" + 0.186*"government" + 0.175*"best" + 0.170
*"would" + 0.140*"song" + -0.138*"technology" + 0.133*"british" + 0.116*"labour" + -0.109
*"mobile" + -0.104*"online" + 0.101*"awards" + 0.100*"think" + -0.099*"time" + -0.092*"gam
ing" + -0.092*"users" + -0.090*"play" + 0.089*"black" + -0.087*"net"
Topic #5:
0.431*"us" + -0.258*"people" + -0.176*"would" + 0.173*"year" + 0.154*"sales" + 0.148*"als
o" + -0.145*"game" + 0.145*"film" + 0.139*"bn" + -0.127*"labour" + 0.111*"company" + 0.109
*"best" + -0.103*"party" + 0.102*"market" + 0.097*"yukos" + -0.094*"games" + -0.093*"like"
+ 0.091*"growth" + 0.087*"oil" + -0.086*"government"
Topic #6:
0.484*"mr" + 0.375*"film" + -0.255*"music" + -0.211*"would" + 0.168*"best" + -0.114*"incre
ase" + 0.107*"british" + 0.100*"actor" + -0.098*"sales" + -0.096*"economy" + 0.095*"direct
or" + 0.094*"films" + -0.092*"year" + -0.089*"bn" + -0.086*"pay" + 0.083*"actress" + 0.082
*"awards" + -0.082*"tax" + 0.077*"aviator" + -0.075*"growth"
Topic #7:
0.382*"would" + 0.281*"film" + -0.275*"us" + -0.229*"government" + -0.223*"mr" + 0.179*"al
so" + -0.132*"threat" + -0.126*"game" + -0.114*"without" + -0.114*"people" + 0.113*"tv" +
-0.097*"agree" + 0.096*"party" + 0.091*"plans" + 0.088*"show" + 0.085*"mobile" + 0.081*"la
bour" + 0.076*"films" + 0.076*"actor" + -0.075*"state"
Topic #8:
0.467*"mr" + 0.351*"music" + -0.279*"people" + -0.172*"increase" + 0.162*"games" + -0.160
*"best" + -0.156*"film" + 0.148*"would" + 0.136*"game" + -0.114*"pay" + -0.105*"governmen
t" + 0.102*"players" + 0.101*"club" + -0.088*"many" + 0.087*"new" + -0.085*"british" + -0.
082*"last" + -0.077*"year" + -0.077*"jobs" + -0.076*"uk"
Topic #9:
0.370*"games" + 0.202*"film" + 0.184*"game" + 0.183*"gaming" + -0.172*"wales" + -0.165*"en
gland" + 0.125*"time" + 0.124*"online" + -0.124*"users" + -0.123*"ireland" + 0.117*"playin
g" + -0.103*"first" + 0.100*"play" + -0.100*"g" + 0.100*"would" + -0.099*"one" + 0.097*"ho
urs" + -0.096*"net" + -0.092*"france" + -0.090*"side"
Topic #10:
0.327*"would" + -0.202*"mr" + -0.196*"party" + -0.193*"labour" + 0.161*"software" + -0.154
*"mobile" + 0.145*"us" + -0.139*"uk" + -0.123*"election" + 0.122*"yukos" + -0.115*"year" +
-0.111*"sales" + 0.104*"could" + -0.103*"tv" + 0.103*"film" + 0.100*"users" + -0.099*"econ
omy" + 0.099*"club" + 0.099*"security" + 0.098*"liverpool"
```

Рисунок 3.2.9 - Основні теми

ДОДАТОК А ТЕКСТИ ПРОГРАМНОГО КОДУ

Тексти програмного коду

(Найменування програми (документа))

Жорсткий диск

(Вид носія даних)

(Обсяг програми (документа), арк.)

Студента групи ІП-113 курсу

Панченка С. В

```
import nltk
import re
from nltk.corpus import gutenberg
import numpy as np
name = 'austen-persuasion.txt'
text = [' '.join(sent) for sent in gutenberg.sents(name)]
text[:30]
wpt = nltk.WordPunctTokenizer()
stop_words = nltk.corpus.stopwords.words('english')
@np.vectorize
def preproc_doc(doc):
doc = re.sub(r'[^a-zA-Z\s]', '', doc, re.I | re.A)
doc = doc.lower()
doc = doc.strip()
tokens = wpt.tokenize(doc)
filtered_tokens = [token for token in tokens if token not in stop_words]
doc = ' '.join(filtered_tokens)
return doc
text = preproc_doc(text)
text
from nltk.collocations import BigramCollocationFinder
from nltk.collocations import BigramAssocMeasures
bigram_measures = BigramAssocMeasures()
finder = BigramCollocationFinder.from_documents(
[item.split() for item in text])
finder.nbest(bigram_measures.raw_freq, 10)
import pandas as pd
df = pd.read_csv('bbc-news-data.csv', sep='\t')
df
text = df['content'].values
text
text = preproc_doc(text)
text
sentences = [sent.split() for sent in text]
sentences[:10]
from gensim.models.phrases import Phrases, Phraser,
ENGLISH_CONNECTOR_WORDS
bigram = Phrases(sentences, min_count=20, threshold=20,
connector_words=ENGLISH_CONNECTOR_WORDS)
bigram_model = Phraser(bigram)
from gensim.corpora import Dictionary
norm_corpus_bigrams = [bigram_model[sent] for sent in sentences]
dictionary = Dictionary(norm_corpus_bigrams)
```

```
norm_corpus_bigrams[:20]
dictionary.filter_extremes(no_below=20, no_above=0.6)
bow_corpus = [dictionary.doc2bow(text) for text in norm_corpus_bigrams]
bow_corpus[:20]
from gensim.models import LsiModel
total_topics = 10
lsi_bow = LsiModel(bow_corpus, id2word=dictionary,
num_topics=total_topics,
onepass=True, chunksize=10000,
power_iters=1000)
for topic_id, topic in lsi_bow.print_topics(num_topics=10, num_words=20):
print('Topic #'+str(topic_id+1)+':')
print(topic)
```