МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“Київський політехнічний інститут ім. Ігоря Сікорського”
ФАКУЛЬТЕТ Інформатики та Обчислювальної Техніки
КАФЕДРА Інформаційних систем та технологій

Звіт до лабораторної роботи №2

з предмету: Обробка та Аналіз текстових даних на мові Python

Перевірила:
Тимофєєва Ю.С.

Виконли:
студенти групи ІК-01
Філоненко І. Р.
Гацан С. Ю.

КИЇВ - 2023

Тема: Попередня обробка тексту за допомогою NLTK

Мета: Ознайомитись з представленням тексту Python в та регулярними виразами.

Варіант: 11

Завдання:

1. Зчитати файл text1.

- Порахувати кількість речень в тексті і вивести останнє.
- Позначити частини мови
- Знайти 10 слів, які зустрічаються найчастіше

2. Використайте корпус Brown, категорію science_fiction

- Порахувати загальну кількість слів в категорії
- Видалити всі дієслова в другому тексті.

Код програми

```
#!/usr/bin/env python3

from sty import fg, rs
import colorsys

import re

import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import brown

# nltk.download('all', download_dir="./nltk_data")
nltk.data.path.append("./nltk_data")


def h_to_color(h):
    (r, g, b) = colorsys.hsv_to_rgb(h, 1, 1)
    (r, g, b) = (int(r * 255), int(g * 255), int(b * 255))
    return fg(r, g, b)


if __name__ == "__main__":
    with open("text1.txt") as file:
        text = file.read()
        print(text, end="\n----------------------\n")

        print("Кількість речень в тексті: " + str(len(sent_tokenize(text))))

        print(sent_tokenize(text)[-1])

        pos_tags = dict()
        for item in nltk.pos_tag(word_tokenize(text)):
            try:
                pos_tags[item[1]] += 1
            except KeyError:
                pos_tags[item[1]] = 1
```

```
        pos_tags_len = len(pos_tags)
        list_pos_tags = list(pos_tags)

        print("Color guide:")
        for tag_index in range(len(list_pos_tags)):
            print(
                h_to_color(tag_index / pos_tags_len) + list_pos_tags[tag_index]
+ fg.rs,
                end='\t'
            )

        print('\n')
        for word, tag in nltk.pos_tag(word_tokenize(text)):
            index = list_pos_tags.index(tag)
            print(h_to_color(index / pos_tags_len) + word + fg.rs + " ", end="")

        print("\n")

        print("Частини мови: " + str(nltk.pos_tag(word_tokenize(text))), end="\
n\n")

        freqDist = nltk.FreqDist(word_tokenize(text))
        print("Частота зустрічі слів: " + str(freqDist.most_common(10)), end="\
n\n")

        print("---------------------\n")
        print(
            "Кількість слів в категорії science_fiction: "
            + str(len(brown.words(categories="science_fiction"))),
            end="\n\n",
        )

        # print(brown.fileids(categories='science_fiction'))
        sents = brown.tagged_words(fileids="cm02")
        # print(sents)
        regex = re.compile(r"^V.*")
        # print(len(' '.join(['' .join(sent_tokens) for sent_tokens, tags in
sents])))
        # print(len(' '.join(['' .join(sent_tokens) for sent_tokens, tags in
sents if not regex.search(tags)])))
        print("Другий текст без дієслів: ")
        print(
            " ".join(
                [
                    "".join(sent_tokens)
                    for sent_tokens, tags in sents
                    if not regex.search(tags)
                ]
            )
        )
```

Результат роботи програми

Isa Whitney, brother of the late Elias Whitney, D.D., Principal of the Theological College of St. George’s, was much addicted to opium. The habit grew upon him, as I understand, from some foolish freak when he was at college; for having read De Quincey’s description of his dreams and sensations, he had drenched his tobacco with laudanum in an attempt to produce the same effects. He found, as so many more have done, that the practice is easier to attain than to get rid of, and for many years he continued to be a slave to the drug, an object of mingled horror and pity to his friends and relatives. I can see him now, with yellow, pasty face, drooping lids, and pin-point pupils, all huddled in a chair, the wreck and ruin of a noble man.

One night—it was in June, ’89—there came a ring to my bell, about the hour when a man gives his first yawn and glances at the clock. I sat up in my chair, and my wife laid her needle-work down in her lap and made a little face of disappointment.

Кількість речень в тексті: 6

I sat up

in my chair, and my wife laid her needle-work down in her lap and made a little face of disappointment.

Color guide:

NNP , NN IN DT JJ VBD RB VBN TO
VB . PRP VBP WRB : VBG PRP$ NNS CC
JJR VBZ MD CD RP

Isa Whitney , brother of the late Elias Whitney , D.D. , Principal of the Theological College of St. George ’ s , was much addicted to opium . The habit grew upon him , as I understand , from some foolish freak when he was at college ; for having read De Quincey ’ s description of his dreams and sensations , he had drenched his tobacco with laudanum in an attempt to produce the same effects . He found , as so many more have done , that the practice is easier to attain than to get rid of , and for many years he continued to be a slave to the drug , an object of mingled horror and pity to his friends and relatives . I can see him now , with yellow , pasty face , drooping lids , and pin-point pupils , all huddled in a chair , the wreck and ruin of a noble man . One night—it was in June , ’ 89—there came a ring to my bell , about the hour when a man gives his first yawn and glances at the clock . I sat up in my chair , and my wife laid her needle-work down in her lap and made a little face of disappointment .

Частини мови: [('Isa', 'NNP'), ('Whitney', 'NNP'), (',', ','), ('brother',

'NN'), ('of', 'IN'), ('the', 'DT'), ('late', 'JJ'), ('E
lias', 'NNP'), ('Whitney', 'NNP'), (',', ','), ('D.D.', 'NNP'), (',', ','),
('Principal', 'NNP'), ('of', 'IN'), ('the', 'DT'), ('T
heological', 'NNP'), ('College', 'NNP'), ('of', 'IN'), ('St.', 'NNP'),
('George', 'NNP'), ('’', 'NNP'), ('s', 'NN'), (',', ','), (
'was', 'VBD'), ('much', 'RB'), ('addicted', 'VBN'), ('to', 'TO'), ('opium',
'VB'), ('.', '.'), ('The', 'DT'), ('habit', 'NN'), ('g
rew', 'VBD'), ('upon', 'IN'), ('him', 'PRP'), (',', ','), ('as', 'IN'), ('I',
'PRP'), ('understand', 'VBP'), (',', ','), ('from',
'IN'), ('some', 'DT'), ('foolish', 'JJ'), ('freak', 'NN'), ('when', 'WRB'),
('he', 'PRP'), ('was', 'VBD'), ('at', 'IN'), ('college
', 'NN'), (';', ':'), ('for', 'IN'), ('having', 'VBG'), ('read', 'VBN'), ('De',
'NNP'), ('Quincey', 'NNP'), ('’', 'NNP'), ('s', 'J
J'), ('description', 'NN'), ('of', 'IN'), ('his', 'PRP$'), ('dreams', 'NNS'),
('and', 'CC'), ('sensations', 'NNS'), (',', ','), ('
he', 'PRP'), ('had', 'VBD'), ('drenched', 'VBN'), ('his', 'PRP$'), ('tobacco',
'NN'), ('with', 'IN'), ('laudanum', 'NN'), ('in', '
IN'), ('an', 'DT'), ('attempt', 'NN'), ('to', 'TO'), ('produce', 'VB'), ('the',
'DT'), ('same', 'JJ'), ('effects', 'NNS'), ('.', '
.'), ('He', 'PRP'), ('found', 'VBD'), (',', ','), ('as', 'IN'), ('so', 'RB'),
('many', 'JJ'), ('more', 'JJR'), ('have', 'VBP'), ('
done', 'VBN'), (',', ','), ('that', 'IN'), ('the', 'DT'), ('practice', 'NN'),
('is', 'VBZ'), ('easier', 'JJR'), ('to', 'TO'), ('at
tain', 'VB'), ('than', 'IN'), ('to', 'TO'), ('get', 'VB'), ('rid', 'JJ'), ('of',
'IN'), (',', ','), ('and', 'CC'), ('for', 'IN'),
('many', 'JJ'), ('years', 'NNS'), ('he', 'PRP'), ('continued', 'VBD'), ('to',
'TO'), ('be', 'VB'), ('a', 'DT'), ('slave', 'NN'), (
'to', 'TO'), ('the', 'DT'), ('drug', 'NN'), (',', ','), ('an', 'DT'), ('object',
'NN'), ('of', 'IN'), ('mingled', 'JJ'), ('horror'
, 'NN'), ('and', 'CC'), ('pity', 'NN'), ('to', 'TO'), ('his', 'PRP$'),
('friends', 'NNS'), ('and', 'CC'), ('relatives', 'NNS'), ('
.', '.'), ('I', 'PRP'), ('can', 'MD'), ('see', 'VB'), ('him', 'PRP'), ('now',
'RB'), (',', ','), ('with', 'IN'), ('yellow', 'JJ'),
(',', ','), ('pasty', 'JJ'), ('face', 'NN'), (',', ','), ('drooping', 'VBG'),
('lids', 'NNS'), (',', ','), ('and', 'CC'), ('pin-p
oint', 'NN'), ('pupils', 'NNS'), (',', ','), ('all', 'DT'), ('huddled', 'VBN'),
('in', 'IN'), ('a', 'DT'), ('chair', 'NN'), (',',
',', ','), ('the', 'DT'), ('wreck', 'NN'), ('and', 'CC'), ('ruin', 'NN'), ('of',
'IN'), ('a', 'DT'), ('noble', 'JJ'), ('man', 'NN'), ('
.', '.'), ('One', 'CD'), ('night—it', 'NN'), ('was', 'VBD'), ('in', 'IN'),
('June', 'NNP'), (',', ','), ('’', 'VBD'), ('89—there',
'CD'), ('came', 'VBD'), ('a', 'DT'), ('ring', 'NN'), ('to', 'TO'), ('my',
'PRP$'), ('bell', 'NN'), (',', ','), ('about', 'IN'), (

'the', 'DT'), ('hour', 'NN'), ('when', 'WRB'), ('a', 'DT'), ('man', 'NN'), ('gives', 'VBZ'), ('his', 'PRP$'), ('first', 'JJ'), ('y
awn', 'NN'), ('and', 'CC'), ('glances', 'NNS'), ('at', 'IN'), ('the', 'DT'), ('clock', 'NN'), ('.', '.'), ('I', 'PRP'), ('sat', 'V
BD'), ('up', 'RP'), ('in', 'IN'), ('my', 'PRP$'), ('chair', 'NN'), (',', ','), ('and', 'CC'), ('my', 'PRP$'), ('wife', 'NN'), ('la
id', 'VBD'), ('her', 'PRP$'), ('needle-work', 'JJ'), ('down', 'NN'), ('in', 'IN'), ('her', 'PRP$'), ('lap', 'NN'), ('and', 'CC'),
('made', 'VBD'), ('a', 'DT'), ('little', 'JJ'), ('face', 'NN'), ('of', 'IN'), ('disappointment', 'NN'), ('.', '.')]

Частота зустрічі слів: [(',', 20), ('and', 9), ('of', 8), ('the', 8), ('to', 8), ('.', 6), ('a', 6), ('in', 5), ('his', 4), ('’',
3)]

Кількість слів в категорії science_fiction: 14470

Другий текст без дієслів:
The expense and time are astronomical . However , we a third vessel out , a much smaller and faster one than the first two . We ha
ve much about interstellar drives since a hundred years ago ; ; that is all I can you about them . `` But the third ship back seve
ral years ago and '' `` That it had a planet on which human beings could and which was already by sentient beings '' ! ! Hal , in
his enthusiasm that he had not been to . Macneff to at Hal with his pale blue eyes . `` How did you '' ? ? He sharply . `` me , Sa
ndalphon '' , Hal . `` But it was inevitable ! ! Did not the Forerunner in his Time and the World Line that such a planet would be
? ? I it was on page 573 '' ! ! Macneff and , `` I am glad that your scriptural lessons have such an impression '' . How could th
ey not ? ? Hal . Besides , they were not the only impressions . I still scars on my back where Pornsen , my gapt , me because I ha
d not my lessons well enough . He was a good impresser , that Pornsen . Was ? ? Is ! ! As I older and was , so was he , always whe
re I was . He was my gapt in the creche . He was the dormitory gapt when I to college and I was away from him . He is now my block
gapt . He is the one responsible for my such low M. R.'s . Swiftly , the revulsion , the protest . No , not he , for I , and I al
one , am responsible for whatever to me . If I a low M. R. , I do so because I it that way or my dark self does . If I , I because
I it so . So , me , Sigmen , for the contrary-to-reality thoughts ! ! `` me

again , Sandalphon '' , Hal . `` But did the expediti
on find any records of the Forerunner having been on this planet ? ? Perhaps , even , though this is too much to , the Forerunner
himself '' ? ? `` No '' , Macneff . `` Though that does not that there may not be such records there . The expedition was under or
ders to a swift survey of conditions and then to to Earth . I can't you now the distance in lightyears or what star this was , tho
ugh you can it with the naked eye at night in this hemisphere . If you , you will be where you're after the ship . And it very soo
n '' . `` You a linguist '' ? ? Hal . `` The ship is huge '' , Macneff , `` but the number of military men and specialists we are
the linguists to one . We have several of your professionals because they were lamechians and above suspicion . Unfortunately '' H
al : Macneff some more , . Then , he , `` Unfortunately , only one lamechian linguist , and he is too old for this expedition . Th
erefore '' `` A thousand pardons '' , Hal . `` But I have just of one thing . I am '' . `` No problem at all '' , Macneff . `` The
re will be no women aboard the Gabriel . And , if a man is , he will automatically be a divorce '' . Hal , and he , `` A divorce '
' ? ? Macneff his hands apologetically and , `` You are , of course . But , from our reading of the Western Talmud , we Urielites
that the Forerunner , this situation would , reference to and provision for divorce . It's inevitable in this case , for the coupl
e will be for , at the least , forty years . Naturally , he the provision in obscure language . In his great and glorious wisdom ,
he that our enemies the Israelites must not be able to therein what we '' . `` I '' , Hal . `` me more , Sandalphon '' . Six mont
hs later , Hal Yarrow in the observation dome of the Gabriel and the ball of Earth above him . It was night on this hemisphere , b
ut the light from the megalopolises of Australia , Japan , China , Southeast Asia , India , Siberia . Hal , the linguist , the dis
cs and necklaces in terms of the languages therein . Australia , the Philippine Islands , Japan , and northern China were by those
members of the Haijac Union that American . Southern China , all of southeast Asia , southern India and Ceylon , these states of
the Malay Federation Bazaar . Siberia Icelandic . His mind the globe swiftly for him , and he Africa , which Swahili south of the
Sahara Sea . All around the Mediterranean Sea , Asia Minor , northern India , and Tibet , Hebrew was the native tongue . In southe
rn Europe , between the Israeli Republics and the Icelandic-speaking peoples of northern Europe , was a thin but long stretch of t
erritory March . This was no man's land , by the Haijac Union and the Israeli Republic , a potential source of war for the last tw

o hundred years . Neither nation would up their claim on it , yet neither to any move that might to a second Apocalyptic War . So
, for all practical purposes , it was an independent nation and by now had its own government (unrecognized outside its own borde
rs) . Its citizens all of the world's tongues , plus a new one Lingo , a pidgin whose vocabulary was from the other six and whose
syntax was so simple it could be on half a sheet of paper . Hal in his mind the rest of Earth : Iceland , Greenland , the Caribbe
an Islands , and the eastern half of South America . Here the peoples the tongue of Iceland because that island had the jump on th
e Hawaiian-Americans who were busy North America and the western half of South America after the Apocalyptic War . Then there was
North America , where American was the native speech of all except the twenty descendants of French-Canadians on the Hudson Bay Pr
eserve . Hal that when that side of Earth into the night zone , Sigmen City would out into space . And , somewhere in that enormou
s light , was his apartment . But Mary would soon no longer be there , for she would be in a few days that her husband had in an a
ccident while on a flight to Tahiti . She would in private , he was sure , for she him in her frigid way , though in public she wo
uld be dry-eyed . Her friends and professional associates would with her , not because she had a beloved husband , but because she
had been to a man who unrealistically . If Hal Yarrow had been in a crash , he must have it that way . There was no such thing as
an `` accident '' . Somehow , all the other passengers (also to have in this web of elaborate frauds to up the disappearance of
the personnel of the Gabriel) had simultaneously `` '' to . And , therefore , being in disgrace , they would not be and their ash
es to the winds in public ceremony . No , the fish could their bodies for all the Sturch . Hal sorry for Mary ; ; he had a time th
e tears from to his own eyes as he in the crowd in the observation dome . Yet , he himself , this was the best way . He and Mary w
ould no longer have to and at each other ; ; their mutual torture would be over . Mary was free to again , not that the Sturch had
secretly her a divorce , that death had her marriage . She would have a year in which to up her mind , to a mate from a list by h
er gapt . Perhaps , the psychological barriers that had her from Hal's child would no longer be present . Perhaps . Hal if this ha
ppy event would . Mary was as below the navel as he . No matter who the candidate for marriage by the gapt The gapt . Pornsen . He
would no longer have to that fat face , that voice `` Hal Yarrow '' ! ! the voice . And , slowly , himself icy yet , Hal . There
was the squat loose-jowled man , lopsidedly up at him . `` My beloved ward , my

perennial gadfly '' , the voice . `` I had no idea
that you , too , would be on this glorious voyage . But I might have ! ! We to
be by love ; ; Sigmen himself must have foreseen i
t . Love to you , my ward '' . `` Sigmen love you , too , my guardian '' , Hal ,
. `` How wonderful to your self . I had we would
never again to each other '' . 5 the Gabriel towards her destination and , under
one-gee acceleration , to up towards her ultimate
velocity , 99.1 percent of the speed of light . Meanwhile , all the personnel
except those few to out the performance of the ship
, into the suspensor . Here they would in animation for many years . Some time
later , after a check had been of all automatic eq
uipment , the crew would the others . They would while the Gabriel's drive would
the acceleration to a point which the bodies of t
he personnel could not have . Upon the speed , the automatic equipment would off
the drive , and the silent but not empty vessel w
ould towards the star which was its journey's end . Many years later , the
photon-counting apparatus in the nose of the ship would
that the star was close enough to deceleration . Again , a force too strong for
bodies to would be . Then , after the vessel cons
iderably , the drive would to a one-gee deceleration . And the crew would be
automatically out of their animation . These members
would then the rest of the personnel . And , in the half-year before their
destination , the men would out whatever preparations w
ere . Hal Yarrow was among the last to into the suspensor and among the first to
out . He had to the recordings of the language of
the chief nation of Ozagen , Siddo . And , from the first , he a difficult
task . The expedition that had Ozagen had in two thous
and Siddo words with an equal number of American words . The description of the
Siddo syntax was very . And , as Hal out , obvious
ly in many cases . This discovery Hal anxiety . His duty was to a school text
and to the entire personnel of the Gabriel how to Oz
agen . Yet , if he all of the little means at his disposal , he would be his
students wrongly . Moreover , even this across would
be difficult . For one thing , the organs of speech of the Ozagen natives
somewhat from Earthmen's ; ; the sounds by these organs
were , therefore , dissimilar . It was true that they could be , but would the
Ozagenians these approximations ? ? Another obstacl
e was the grammatical construction of Siddo . the tense system . Instead of a
verb or an unattached particle to the past or future
, Siddo an entirely different word . Thus , the masculine animate infinitive
dabhumaksanigalu'ahai , to , was , in the perfect te
nse , ksu'u'peli'afo , and , in the future , mai'teipa . The same use of an
entirely different word for all the other tenses . Plu

s the fact that Siddo not only had the normal (to Earthmen) three genders of masculine , feminine , and neuter , but the two ext
ra of inanimate and spiritual . Fortunately , gender was , though the expression of it would be difficult for anybody not in Siddo
. The system of gender according to tense . All the other parts of speech : nouns , pronouns , adjectives , adverbs , and conjunc
tions under the same system as the verbs .