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 №2
з дисципліни
Обробка та аналіз текстових даних на Python

Виконав
студент групи ІП-11:
Сідак К. І.

Перевірила:
ст. викладач
Тимофєєва Ю. С.

Київ 2023

1.Зчитати файл text1. а) Порахувати кількість слів в тексті (не враховуючи знаки пунктуації та інші спеціальні символи); б) видалити стоп-слова; в) провести пошук кореня слів другого речення. 2.Використати корпус Reuters, перший текст категорії sugar. а) Вивести перші 5 речень; б) Вивести 10 іменників, що зустрічаються найчастіше.

In [1]:
```
import nltk
nltk.download('all')
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
[nltk_data]    | Downloading package abc to /Users/kyryl/nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_ru is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package basque_grammars to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package basque_grammars is already up-to-date!
[nltk_data]    | Downloading package bcp47 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package bcp47 is already up-to-date!
[nltk_data]    | Downloading package biocreative_ppi to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package biocreative_ppi is already up-to-date!
[nltk_data]    | Downloading package bllip_wsj_no_aux to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package bllip_wsj_no_aux is already up-to-date!
[nltk_data]    | Downloading package book_grammars to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package book_grammars is already up-to-date!
[nltk_data]    | Downloading package brown to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package brown_tei to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package brown_tei is already up-to-date!
[nltk_data]    | Downloading package cess_cat to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package cess_cat is already up-to-date!
[nltk_data]    | Downloading package cess_esp to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package cess_esp is already up-to-date!
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package city_database to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package city_database is already up-to-date!
```

```
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package comparative_sentences to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package comparative_sentences is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package comtrans to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package comtrans is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package conll2007 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package conll2007 is already up-to-date!
[nltk_data]    | Downloading package crubadan to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package crubadan is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package dolch is already up-to-date!
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package europarl_raw is already up-to-date!
[nltk_data]    | Downloading package extended_omw to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package extended_omw is already up-to-date!
[nltk_data]    | Downloading package floresta to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package floresta is already up-to-date!
[nltk_data]    | Downloading package framenet_v15 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package framenet_v15 is already up-to-date!
[nltk_data]    | Downloading package framenet_v17 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package framenet_v17 is already up-to-date!
[nltk_data]    | Downloading package gazetteers to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package ieer to /Users/kyryl/nltk_data...
[nltk_data]    |   Package ieer is already up-to-date!
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package indian to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package indian is already up-to-date!
```

```
[nltk_data]    | Downloading package jeita to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package jeita is already up-to-date!
[nltk_data]    | Downloading package kimmo to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package kimmo is already up-to-date!
[nltk_data]    | Downloading package knbc to /Users/kyryl/nltk_data...
[nltk_data]    |   Package knbc is already up-to-date!
[nltk_data]    | Downloading package large_grammars to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package lin_thesaurus to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package lin_thesaurus is already up-to-date!
[nltk_data]    | Downloading package mac_morpho to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package mac_morpho is already up-to-date!
[nltk_data]    | Downloading package machado to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package machado is already up-to-date!
[nltk_data]    | Downloading package masc_tagged to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package masc_tagged is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package maxent_treebank_pos_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package moses_sample to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package moses_sample is already up-to-date!
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package mte_teip5 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package mte_teip5 is already up-to-date!
[nltk_data]    | Downloading package mwa_ppdb to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package mwa_ppdb is already up-to-date!
[nltk_data]    | Downloading package names to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package names is already up-to-date!
[nltk_data]    | Downloading package nombank.1.0 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package nombank.1.0 is already up-to-date!
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data]    | Downloading package nps_chat to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package nps_chat is already up-to-date!
[nltk_data]    | Downloading package omw to /Users/kyryl/nltk_data...
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package omw-1.4 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package omw-1.4 is already up-to-date!
[nltk_data]    | Downloading package opinion_lexicon to
```

```
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package opinion_lexicon is already up-to-date!
[nltk_data]    | Downloading package panlex_swadesh to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package panlex_swadesh is already up-to-date!
[nltk_data]    | Downloading package paradigms to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package paradigms is already up-to-date!
[nltk_data]    | Downloading package pe08 to /Users/kyryl/nltk_data...
[nltk_data]    |   Package pe08 is already up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package perluniprops is already up-to-date!
[nltk_data]    | Downloading package pil to /Users/kyryl/nltk_data...
[nltk_data]    |   Package pil is already up-to-date!
[nltk_data]    | Downloading package pl196x to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package pl196x is already up-to-date!
[nltk_data]    | Downloading package porter_test to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package porter_test is already up-to-date!
[nltk_data]    | Downloading package ppattach to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package ppattach is already up-to-date!
[nltk_data]    | Downloading package problem_reports to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package problem_reports is already up-to-date!
[nltk_data]    | Downloading package product_reviews_1 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package product_reviews_1 is already up-to-date!
[nltk_data]    | Downloading package product_reviews_2 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package product_reviews_2 is already up-to-date!
[nltk_data]    | Downloading package propbank to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package propbank is already up-to-date!
[nltk_data]    | Downloading package pros_cons to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package pros_cons is already up-to-date!
[nltk_data]    | Downloading package ptb to /Users/kyryl/nltk_data...
[nltk_data]    |   Package ptb is already up-to-date!
[nltk_data]    | Downloading package punkt to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package qc to /Users/kyryl/nltk_data...
[nltk_data]    |   Package qc is already up-to-date!
[nltk_data]    | Downloading package reuters to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package reuters is already up-to-date!
[nltk_data]    | Downloading package rslp to /Users/kyryl/nltk_data...
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package rte to /Users/kyryl/nltk_data...
[nltk_data]    |   Package rte is already up-to-date!
[nltk_data]    | Downloading package sample_grammars to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package sample_grammars is already up-to-date!
[nltk_data]    | Downloading package semcor to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package semcor is already up-to-date!
[nltk_data]    | Downloading package senseval to
```

```
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package senseval is already up-to-date!
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package sentiwordnet to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package sentiwordnet is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package spanish_grammars to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package spanish_grammars is already up-to-date!
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package tagsets is already up-to-date!
[nltk_data]    | Downloading package timit to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package timit is already up-to-date!
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package toolbox is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package udhr to /Users/kyryl/nltk_data...
[nltk_data]    |   Package udhr is already up-to-date!
[nltk_data]    | Downloading package udhr2 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package udhr2 is already up-to-date!
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     /Users/kyryl/nltk_data...
```

```
[nltk_data]    |   Package unicode_samples is already up-to-date!
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package universal_tagset is already up-to-date!
[nltk_data]    | Downloading package universal_treebanks_v20 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package universal_treebanks_v20 is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package vader_lexicon is already up-to-date!
[nltk_data]    | Downloading package verbnet to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package verbnet is already up-to-date!
[nltk_data]    | Downloading package verbnet3 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package verbnet3 is already up-to-date!
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package webtext is already up-to-date!
[nltk_data]    | Downloading package wmt15_eval to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wmt15_eval is already up-to-date!
[nltk_data]    | Downloading package word2vec_sample to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package word2vec_sample is already up-to-date!
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet2021 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wordnet2021 is already up-to-date!
[nltk_data]    | Downloading package wordnet2022 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wordnet2022 is already up-to-date!
[nltk_data]    | Downloading package wordnet31 to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wordnet31 is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to
[nltk_data]    |     /Users/kyryl/nltk_data...
[nltk_data]    |   Package words is already up-to-date!
[nltk_data]    | Downloading package ycoe to /Users/kyryl/nltk_data...
[nltk_data]    |   Package ycoe is already up-to-date!
[nltk_data]    |
[nltk_data]  Done downloading collection all
```

Out[1]: True

Зчитування файлу text1.txt

In [2]:
```python
with open('text1.txt') as file:
    text = file.read()
```

Токенізація слів та підрахунок кількості слів у тексті

In [3]:
```
from nltk.tokenize.toktok import ToktokTokenizer
import re

tokenizer = ToktokTokenizer()
words = [word for word in tokenizer.tokenize(text) if re.search("\w", wor
for i in range(len(words) - 1, -1, -1):
    if words[i] == 's':
        if i > 0:
            words[i - 1] += ''s'
        words.pop(i)
words = [re.sub(r'[^\w\s]', '', word) if word not in ['D.D.', 'St.'] else
print(words)
```

```
['Isa', 'Whitney', 'brother', 'of', 'the', 'late', 'Elias', 'Whitney', 'D.
D.', 'Principal', 'of', 'the', 'Theological', 'College', 'of', 'St.', 'Geo
rges', 'was', 'much', 'addicted', 'to', 'opium', 'The', 'habit', 'grew',
'upon', 'him', 'as', 'I', 'understand', 'from', 'some', 'foolish', 'frea
k', 'when', 'he', 'was', 'at', 'college', 'for', 'having', 'read', 'De',
'Quinceys', 'description', 'of', 'his', 'dreams', 'and', 'sensations', 'h
e', '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', 'ea
sier', 'to', 'attain', 'than', 'to', 'get', 'rid', 'of', 'and', 'for', 'ma
ny', 'years', 'he', 'continued', 'to', 'be', 'a', 'slave', 'to', 'the', 'd
rug', 'an', 'object', 'of', 'mingled', 'horror', 'and', 'pity', 'to', 'hi
s', 'friends', 'and', 'relatives', 'I', 'can', 'see', 'him', 'now', 'wit
h', 'yellow', 'pasty', 'face', 'drooping', 'lids', 'and', 'pinpoint', 'pup
ils', 'all', 'huddled', 'in', 'a', 'chair', 'the', 'wreck', 'and', 'ruin',
'of', 'a', 'noble', 'man', 'One', 'night', 'it', 'was', 'in', 'June', '8
9', 'there', 'came', 'a', 'ring', 'to', 'my', 'bell', 'about', 'the', 'hou
r', 'when', 'a', 'man', 'gives', 'his', 'first', 'yawn', 'and', 'glances',
'at', 'the', 'clock', 'I', 'sat', 'up', 'in', 'my', 'chair', 'and', 'my',
'wife', 'laid', 'her', 'needlework', 'down', 'in', 'her', 'lap', 'and', 'm
ade', 'a', 'little', 'face', 'of', 'disappointment']
```

In [4]:
```
len(words)
```

Out[4]: 189

Видалення стоп-слів

In [5]:
```
from nltk.corpus import stopwords

stop_words = set(stopwords.words("english"))

filtered_words = [word for word in words if word.lower() not in stop_word
print(filtered_words)
```

```
['Isa', 'Whitney', 'brother', 'late', 'Elias', 'Whitney', 'D.D.', 'Princip
al', 'Theological', 'College', 'St.', 'Georges', 'much', 'addicted', 'opiu
m', 'habit', 'grew', 'upon', 'understand', 'foolish', 'freak', 'college',
'read', 'De', 'Quinceys', 'description', 'dreams', 'sensations', 'drenche
d', 'tobacco', 'laudanum', 'attempt', 'produce', 'effects', 'found', 'man
y', 'done', 'practice', 'easier', 'attain', 'get', 'rid', 'many', 'years',
```

```
'continued', 'slave', 'drug', 'object', 'mingled', 'horror', 'pity', 'frie
nds', 'relatives', 'see', 'yellow', 'pasty', 'face', 'drooping', 'lids',
'pinpoint', 'pupils', 'huddled', 'chair', 'wreck', 'ruin', 'noble', 'man',
'One', 'night', 'June', '89', 'came', 'ring', 'bell', 'hour', 'man', 'give
s', 'first', 'yawn', 'glances', 'clock', 'sat', 'chair', 'wife', 'laid',
'needlework', 'lap', 'made', 'little', 'face', 'disappointment']
```

Пошук кореня слів другого речення

In [6]:
```
from nltk.stem.snowball import EnglishStemmer
from nltk import sent_tokenize
import string

sentence = sent_tokenize(text)[1].translate(str.maketrans('', '', string.
words_sent = sentence.split()
stemmer = EnglishStemmer()
print([stemmer.stem(w) for w in words_sent])
```

```
['the', 'habit', 'grew', 'upon', 'him', 'as', 'i', 'understand', 'from',
'some', 'foolish', 'freak', 'when', 'he', 'was', 'at', 'colleg', 'for', 'h
ave', 'read', 'de', 'quincey', 'descript', 'of', 'his', 'dream', 'and', 's
ensat', 'he', 'had', 'drench', 'his', 'tobacco', 'with', 'laudanum', 'in',
'an', 'attempt', 'to', 'produc', 'the', 'same', 'effect']
```

Перші 5 речень тексту категорії sugar з корпусу Routers

In [7]:
```
from nltk.corpus import reuters

sentences = reuters.sents(categories='sugar')
sentences = [' '.join(sentence_tokens) for sentence_tokens in
             sentences[:][:5]]
print(sentences)
```

```
["THAI TRADE DEFICIT WIDENS IN FIRST QUARTER Thailand ' s trade deficit wi
dened to 4 . 5 billion baht in the first quarter of 1987 from 2 . 1 billio
n a year ago , the Business Economics Department said .", 'It said Janunar
y / March imports rose to 65 . 1 billion baht from 58 . 7 billion .', "Tha
iland ' s improved business climate this year resulted in a 27 pct increas
e in imports of raw materials and semi - finished products .", "The countr
y ' s oil import bill , however , fell 23 pct in the first quarter due to
lower oil prices .", 'The department said first quarter exports expanded t
o 60 . 6 billion baht from 56 . 6 billion .']
```

Визначення іменників у цьому тексті

In [8]:
```
from nltk import pos_tag
from nltk import word_tokenize
raw_text = reuters.raw(categories='sugar')
words = word_tokenize(raw_text)
tagged_words = pos_tag(words)
nouns = [word for word, tag in tagged_words if tag in ("NN", "NNS", "NNP"
print(nouns[:20])
```

```
['THAI', 'TRADE', 'DEFICIT', 'WIDENS', 'IN', 'FIRST', 'QUARTER', 'Thailan
d', 'trade', 'deficit', 'baht', 'quarter', 'year', 'Business', 'Economic
```

```
s', 'Department', 'Janunary/March', 'imports', 'baht', 'Thailand']
```

10 іменників, що зустрічаються найчастіше

In [9]:
```
from nltk import FreqDist
nouns_freq = FreqDist(nouns)
print(nouns_freq.most_common(10))
```

```
[('sugar', 496), ('tonnes', 353), ('mln', 189), ('year', 183), ('EC', 15
2), ('SUGAR', 129), ('U.S.', 100), ('production', 99), ('prices', 95), ('t
raders', 87)]
```