МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №1**

**Дисциплина: Основы компьютерной лингвистики**

**Тема: «Графематический анализатор»**

Работу выполнил:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Курбатский В.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Направленность (профиль): Компютерные технологии и прикладная математика.

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Вишняков Ю.М.

Краснодар

2022

**Подключение библиотек**

NLTK – это ведущая платформа для создания программ на Python для работы с данными на естественном языке. Он предоставляет простые в использовании интерфейсы для более чем 50 корпусов и лексических ресурсов, таких как WordNet, наряду с набором библиотек обработки текста для классификации, токенизации, обработки по меткам, разметки, анализа и семантического мышления, оболочек для промышленных библиотек NLP, и активный дискуссионный форум.

from nltk.corpus import stopwords

from nltk import sent\_tokenize

from nltk import woed\_tokenize

**Датасет**

«Уважаемые дамы и господа! Мне неизвестно, насколько каждый из вас из

литературы или понаслышке знаком с психоанализом. Однако само название моих

лекций "Элементарное введение в психоанализ" - предполагает, что вы

ничего не знаете об этом и готовы получить от меня первые сведения. Смею все

же предположить, что вам известно следующее: психоанализ является одним из

методов лечения нервнобольных; и тут я сразу могу привести вам пример,

показывающий, что в этой области кое-что делается по-иному или даже

наоборот, чем принято в медицине. Обычно, когда больного начинают лечить

новым для него методом, ему стараются внушить, что опасность не так велика,

и уверить его в успехе лечения. Я думаю, это совершенно оправданно, так как

тем самым мы повышаем шансы на успех. Когда же мы начинаем лечить невротика

методом психоанализа, мы действуем иначе. Мы говорим ему о трудностях

лечения, его продолжительности, усилиях и жертвах, связанных с ним. Что же

касается успеха, то мы говорим, что не можем его гарантировать, поскольку он

зависит от поведения больного, его понятливости, сговорчивости и выдержки.

Естественно, у нас есть веские основания для такого как будто бы

неправильного подхода к больному, в чем вы, видимо, позднее сможете

убедиться сами.

Не сердитесь, если я на первых порах буду обращаться с вами так же, как

с этими нервнобольными. Собственно говоря, я советую вам отказаться от мысли

прийти сюда во второй раз. Для этого сразу же хочу показать вам, какие

несовершенства неизбежно присущи обучению психоанализу и какие трудности

возникают в процессе выработки собственного суждения о нем. Я покажу вам,

как вся направленность вашего предыдущего образования и все привычное ваше

мышление будут неизбежно делать вас противниками психоанализа и сколько

нужно будет вам преодолеть, чтобы совладать с этим инстинктивным

сопротивлением. Что вы поймете в психоанализе из моих лекций, заранее

сказать, естественно, трудно, однако могу твердо обещать, что, прослушав их,

вы не научитесь проводить психоаналитическое исследование и лечение. Если же

среди вас найдется кто-то, кто не удовлетворится беглым знакомством с

психоанализом, а захочет прочно связать себя с ним, я не только не посоветую

это сделать, но всячески стану его предостерегать от этого шага.

Обстоятельства таковы, что подобный выбор профессии исключает для него

всякую возможность продвижения в университете. Если же такой врач займется

практикой, то окажется в обществе, не понимающем его устремлений,

относящемся к нему с недоверием и враждебностью и ополчившем против него все

скрытые темные силы. Возможно, кое-какие моменты, сопутствующие войне,

свирепствующей ныне в Европе, дадут вам некоторое представление о том, что

сил этих - легионы.

Правда, всегда найдутся люди, для которых новое в познании имеет свою

привлекательность, несмотря на все связанные с этим неудобства. И если

кто-то из вас из их числа и, несмотря на мои предостережения, придет сюда

снова, я буду рад приветствовать его. Однако вы все вправе знать, какие

трудности связаны с психоанализом.

Во-первых, следует указать на сложность преподавания психоанализа и

обучения ему. На занятиях по медицине вы привыкли к наглядности. Вы видите

анатомический препарат, осадок при химической реакции, сокращение мышцы при

раздражении нервов.»

**Графематический анализатор**

from nltk import \*

from nltk.corpus import stopwords

import docx

import string

def input\_doxc():

data\_docx = docx.Document('input.docx')

data = []

for paragraph in data\_docx.paragraphs:

data.append(paragraph.text)

data = ' '.join(data)

return data

def graphematic\_analiz(data):

list\_sent = sent\_tokenize(data) #Список предложений

data\_without\_znaks = data.lower()

for s in string.punctuation:

if s in data\_without\_znaks:

data\_without\_znaks = data\_without\_znaks.replace(s,'') #Строка без знаков препинания

list\_words = word\_tokenize(data\_without\_znaks)

download('stopwords')

list\_stop\_word = set(stopwords.words('russian'))

list\_words = [i for i in list\_words if i not in list\_stop\_word]

#list\_words - список без стоп слов и знаков + все в нижнем регистре

#Создаю словарь где ключем является частота встречаемости слова

dict\_frequency = {}

for i in list\_words:

if i in dict\_frequency:

dict\_frequency[i] = dict\_frequency[i]+1

else:

dict\_frequency[i] = 1

return [list\_sent, list\_words, dict(sorted(dict\_frequency.items(), key = lambda x: x[1], reverse = True))]

graph\_analiz = graphematic\_analiz(input\_doxc())

print(graph\_analiz[0])

print("\n")

print(graph\_analiz[1])

print('\n')

print(graph\_analiz[2])



Рисунок 1 – Результаты работы программы.