Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет Комп’ютерних наук

Кафедра Програмної інженерії

**АТЕСТАЦІЙНА РОБОТА (ПРОЕКТ)**

**Пояснювальна записка**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_бакалавр\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Бібліотека визначення мови коротких повідомлень, що складаються з літер кириличного алфавіту

Виконала: студентка 4 курсу, групи ПІ-13-2

напряму підготовки (спеціальності)

6.050103 Програмна інженерія .

Сердюк Д.О. .

Керівник Бабій А.С.

Рецензент Сокорчук І.П.

Допускається до захисту

Зав. кафедри \_\_\_\_\_\_\_\_\_\_\_ Дудар. З.В.

(підпис) (прізвище, ініціали)

2017 р.

Харківський національний університет радіоелектроніки

Факультет Комп’ютерних наук

Кафедра Програмної інженерії

Освітньо-атестаційний рівень бакалавр

Напрям підготовки 6.050103 Програмна інженерія

Спеціальність \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ЗАТВЕРДЖУЮ:

Зав. кафедри \_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

«\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_ р.

**ЗАВДАННЯ**

НА АТЕСТАЦІЙНУ РОБОТУ

студентці\_\_\_\_\_\_\_\_\_\_\_\_Сердюк Дар’ї Олексіївні\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Тема роботи (проекту)\_\_\_\_\_\_\_\_\_ Бібліотека визначення мови коротких повідомлень, що складаються з літер кириличного алфавіту\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

затверджена наказом по університету від "\_\_\_"\_\_\_\_\_\_\_\_\_\_\_\_\_\_20 \_\_ р. № \_\_\_\_\_\_\_\_\_\_\_\_

2. Термін подання студентом роботи (проекту) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Вихідні дані до роботи (проекту)\_\_\_\_\_\_необхідно розробити бібліотеку розпізнавання мови коротких повідомлень; для розробки використовувати Python разом з бібліотекою Scikit-Learn\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Зміст пояснювальної записки (перелік питань, що потрібно розробити) \_вступ, аналіз предметної галузі, формування вимог до програмної системи, архітектура та проектування програмного забезпечення, опис прийнятих програмних рішень, тестування розробленого програмного забезпечення, впровадження програмного забезпечення, висновки за роботою, перелік джерел\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслеників, плакатів) діаграма пакетів, діаграма  розгортання,  діаграма послідовності, діаграма кооперацій, діаграма активності, структура бази даних, фрагменти коду програми, слайди презентації\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Консультанти розділів роботи

|  |  |  |  |
| --- | --- | --- | --- |
| Найменування  розділу | Консультант  (посада, прізвище, ім’я, по батькові) | Позначка консультанта  про виконання розділу | |
| підпис | дата |
| Спец. частина | ст.викл Бабій А.С. |  |  |
|  |  |  |  |
|  |  |  |  |

7. Дата видачі завдання «18» квітня 2017 р.

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Назва етапів роботи (проекту) | Термін  виконання етапів проекту | Примітка |
| 1 | Аналіз предметної галузі | 18.04.2017 - 22.04.2017 |  |
| 2 | Створення специфікації ПЗ | 23.04.2017 - 26.04.2017 |  |
| 3 | Проектування та розробка ПЗ | 27.04.2017 - 28.05.2017 |  |
| 4 | Тестування та дослідна експлуатація ПЗ | 26.05.2017 - 31.05.2017 |  |
| 5 | Написання пояснювальної записки та підготовка презентації | 01.06.2017 - 13.06.2017 |  |
| 6 | Перевірка пояснювальної записки керівником та рецензентом | 14.06.2017 |  |
| 7 | Проходження нормконтролю | 15.06.2017 |  |
| 8 | Здача роботи у електронний архів | 15.06.2017 |  |
| 9 | Підпис завідувача кафедри | 15.06.2017 |  |
| 10 | Захист випускної атестаційної роботи | 20.06.2017 |  |

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

Керівник роботи (проекту) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_ст.викл. Бабій А.С.\_\_\_\_

(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи бакалавра: 61 с., 18 рис., 7 табл., 3 дод., 16 джерел.

Головним об’єктом дослідження є алгоритми розпізнавання мови друкованого тексту, а саме підготовчі етапи та методи класифікація.

Метою роботи є створення бібліотеки на мові програмування Python.

Методи розробки використовують мову програмування Python та допоміжні бібліотеки для обчислень, такі як NumPy, SciPy та бібліотеку для машинного навчання (Machine Learning) і добування та аналізу даних (Data Mining, Data Analysis) Scikit-learn.

У результаті роботи були порівняні різноманітні методи розпізнавання мови: як тренування класифікатору, так і методи самої класифікації (Байесовська, метод опорних векторів (SVM) та інші); розроблена бібліотека розпізнавання мови коротких повідомлень (наприклад, повідомлень у соціальній мережі Twitter).

РОЗПІЗНАВАННЯ МОВИ, МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, N-ГРАМИ, НОРМАЛІЗАЦІЯ, БАЙЕСОВСЬКА КЛАСИФІКАЦІЯ, PYTHON, SCIKIT-LEARN, NUMPY, FLASK, SQLITE

The main objective of the research is the language identification algorithms of the printed text, preprocessing steps and classification in particular.

The goal of this piece of work is to create a Python library.

The developed methods are based on Python programming language and additional computing libraries, e.g. NumPy, SciPy, and a machine learning, data analysis and data mining library Scikit-learn.

As a result, a number of methods for language identification were studied and compared: both training the classifier and the classification methods (Bayesian, Support Vector Machine etc.); a language identification library for short messages (tweets in Twitter) was created.

LANGUAGE IDENTIFICATION, MACHINE LEARNING, CLASSIFICATION, N-GRAMS, NORMALIZATION, BAYESIAN CLASSIFICATION, PYTHON, SCIKIT-LEARN, NUMPY, FLASK, SQLITE