

Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки
адекватності текстових даних засобами машинного навчання.
113 – Прикладна математика

Мета роботи: проаналізувати алгоритм визначення адекватності тестових даних з використанням підходу машинного навчання. Розробити відповідний програмний засіб.

ЗМІСТ

1. Теоретична довідка	1
1.1. Задача: відрізнити осмислений текст від нісенітниць	1
1.2. Алгоритм	2
1.3. Використання моделі	3
1.4. У реальному житті	3
2. Альтернатива машинному навчанню	4
2.1. Недоліки	4
2.2. Помилкові спрацьовування	4
2.3. Обхід моделі	5
3. Тестування розробленої системи	5
4. Індивідуальні завдання	14

1. Теоретична довідка

Останнім часом все технологічні компанії твердять про машинному навчанні. Багато завдань воно вирішує, які раніше тільки люди і могли вирішити.



Для машинного навчання використовують низку інформаційних технологій та суперкомп'ютери [1]. А оскільки використовуються технології, то нескладно все це пояснити людською мовою. Нижче буде описаний у словесній формі алгоритм, який підпадає під визначення машинного навчання. Складність цього алгоритму іграшкова – виле висновки він дозволяє

зробити суттєві.

1.1. Задача: відрізнити осмислений текст від нісенітниць

Текст, який пишуть справжні люди, виглядає так:

«Можу творити, можу і натворити! У мене два недоліки: погана пам'ять і щось ще. Ніхто не знає стільки, скільки не знаю я».

Нісенітниця виглядає так:

ОРПорывав аоьрОрпавор ОрОРАыдцуцзущгкгеуб
ыватьывдцулвдлоадузщ Йцхья ддваополц ыадолцлопиолым
бамдлотдламда .

Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки
адекватності текстових даних засобами машинного навчання.
113 – Прикладна математика

Наше завдання - розробити алгоритм машинного навчання, який би відрізняв одне від іншого. А оскільки говориться про це стосовно антивірусної тематики, то будемо називати осмислений текст «чистим», а нісенітницю – «шкідливим». Схоже завдання насправді вирішується при аналізі реальних файлів у реальній антивірусній системі.

Для людини завдання здається очевидним, адже відразу видно, де чистий, а де шкідливий текст, але ось формалізувати різницю або тим більше пояснити її комп'ютеру – це значно складніше. Тут використовується машинне навчання: спочатку системі потрібно дати зразки. Він на них «навчиться», а потім буде сам правильно відповідати, де що саме.

1.2. Алгоритм

Наш алгоритм буде вважати, як часто в нормальному тексті одна конкретна буква слідує за іншою конкретною літерою. І так для кожної пари букв. Наприклад, для першої чистої фрази — «Могу творить, могу и натворить!» — розподіл вийде такий:

ат — 1	мо — 2	ри — 2
во — 2	на — 1	тв — 2
гу — 2	ог — 2	ть — 2
ит — 2	ор — 2	

Що вийшло: за буквою «в» слід буква «о» – два рази, а за буквою «а» слід буква «т» – один раз. Для простоти ми не враховуємо знаки і пропуски.

На цьому етапі розуміється, що для навчання моделі однієї фрази мало: і поєднань недостатня кількість, і різниця між частотою появи різних сполучень не така велика. Тому треба взяти якийсь істотно більший обсяг даних. Наприклад, можна порахувати, які поєднання букв зустрічаються в першому томі «Война и мир»:

то — 8411	ал — 5637	тж — 1
ст — 6591	ра — 5273	ен — 4211
на — 6236	не — 5199	оу — 31
го — 5639	по — 5174	мб — 2

Зрозуміло, це не вся таблиця поєднань, а лише її мала частина. Виявляється, ймовірність зустріти «то» в два рази вища, аніж «ен». А щоб за буквою «т» слідувала «ж» – таке зустрічається лише один раз, в слові «отжившим».

«Модель» російської мови у нас тепер є, як же її використовувати? Аналогічно можна створити модель для навчання на українській мові або на іншій мові світового спілкування. Щоб визначити, наскільки ймовірно досліджуваний рядок чистий або шкідливий, порахуємо його «правдоподібність». Будемо брати кожну пару букв з цього рядка, визначати за «моделлю» її частоту (по суті, реалістичність поєднання букв) і перемножувати ці числа:

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

$F(\text{мо}) * F(\text{ог}) * F(\text{гу}) * F(\text{тв}) * \dots = 2131 * 2943 * 474 * 1344 * \dots = \text{правдоподібність}$

Також у фінальному значенні правдоподібності слід врахувати кількість символів в досліджуваному рядку – адже чим він буде довшим, тим більше чисел ми перемножуємо. Тому з добутку потрібно «взяти» корінь відповідного порядку (довжина рядка мінус один).

1.3. Використання моделі

Тепер ми можемо робити висновки: чим більше отримане число – тим правдоподібніший досліджуваний рядок лягає у нашу модель. Стало бути, тим більша ймовірність, що цей рядок писала людина, тобто він чистий.

Якщо ж досліджуваний рядок містить підозріло велику кількість вкрай рідкісних поєднань букв (наприклад, «ЇЇ», «тж», «'ь» і так далі), то, швидше за все, він штучний – шкідливий.

Для рядків вище правдоподібність вийшла наступна:

«Могу творить, могу и натворить!» — 1805 балів
«У меня два недостатка: плохая память и что-то еще.» — 1535 балів
«Никто не знает столько, сколько не знаю я.» — 2274 бали.
«ОРПорыв аоырОрпаяор ОрОРАыдцуцзущкгеуб ыватыивдцулвдлоадущц» — 44 бали.
«Йцхяь длавополц ыадоццлопиолым бамдлотдламда» — 149 балів.

Як видно, чисті рядки правдоподібні на 1000-2000 балів, а шкідливі не дотягують і до 150. Тобто все працює так, як задумано.

Щоб не ворожити, що таке «багато», а що є «мало», то краще довірити визначення порогового значення самій машині (нехай навчається). Для цього подаємо їй кілька чистих рядків і порахуємо їх правдоподібність, а потім згодуємо трохи шкідливих рядків – і теж порахуємо правдоподібність. І обчислимо деяке значення посередині, яке буде найкраще відокремлювати одні від інших. У нашому випадку вийде щось в районі 500.

1.4. У реальному житті

Давайте обдумаємо, що ж у нас вийшло.

1. Виділення ознак чистих рядків, а саме пар символів.

У реальному житті при розробці даного антивіруса теж виділяють ознаки з файлів або інших об'єктів. І це, до речі, найважливіший крок: від рівня експертизи та досвіду дослідників безпосередньо залежить якість виділюваних ознак. Зрозуміти, що ж насправді важливо, – це все ще завдання людини. Наприклад, хто сказав, що треба використовувати пари символів, а не трійки? Такі гіпотези якраз і перевіряють в антивірусній лабораторії. Само собою, у нас для відбору найкращих і взаємодоповнюючих ознак теж використовується машинне навчання.

2. На підставі виділених ознак побудова математичної моделі і її навчання на прикладах.

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

Само собою, в реальному житті використовуються моделі трохи складніше. Зараз найкращі результати показує ансамбль вирішальних дерев, побудований методом Gradient boosting, але прагнення до досконалості не дозволяє дослідникам заспокоїтися.

3. На основі математичної моделі підрахунок рейтингу «правдоподібності».

У реальному житті зазвичай вважається протилежний рейтинг – рейтинг шкідливості. Різниця, здавалося б, не суттєва, але наскільки неправдоподібною для нашої математичної моделі буде рядок на іншій мові або з іншим алфавітом. Антивірус не має права допустити помилкове спрацьовування на цілий клас файлів тільки з тієї причини, що «цього його не проходили».

2. Альтернатива машинному навчанню

20 років тому, коли шкідливий було мало, кожному «нісенітницю» можна було просто знайти за допомогою сигнатур – характерних уривків. Для прикладів вище «сигнатури» могли б бути такими:

*ОРПоряав аоырОрнаыор ОрОРАыдцууцзуцгкгеуб
ыватьывдцулвдлоадущц Йцхья длавополц ыадоцлопиолым бамдлотдламда*

Зараз без машинного навчання не працює жоден нормальний антивірус.

2.1. Недоліки

Переваги зрозумілі. Цей метод добре справляється, якщо описаний вище алгоритм буде працювати у хмарі або в інфраструктурі, постійно навчаючись на величезних кількостях як чистих, так і шкідливих об'єктах.

Також дуже добре, якщо за результатами навчання спостерігає команда експертів, які втручаються в тих випадках, коли без досвідченої людини не обійтися.

В цьому випадку недоліків дійсно небагато, а за великим рахунком тільки один – потрібна ця дорога інфраструктура і не менш дорога команда фахівців.

Інша справа, коли хтось намагається радикально заощадити і використовувати лише математичну модель і тільки на стороні продукту, прямо у клієнта. Тоді можуть розпочатися труднощі.

2.2. Помилкові спрацьовування

Детектування на базі машинного навчання – це завжди пошук балансу між рівнем детектування і рівнем помилкових спрацьовувань. І якщо захочеться детектувати побільше, то помилкові спрацьовування будуть. У разі машинного навчання вони будуть виникати в непередбачуваних і часто важко пояснюваних місцях. Наприклад, це чистий рядок – «Мцыри и Мкртчян» – розпізнається як неправдоподібний: 145 балів у моделі з нашого прикладу. Тому дуже важливо,

Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки
адекватності текстових даних засобами машинного навчання.

113 – Прикладна математика

щоб антивірусна лабораторія мала велику колекцію чистих файлів для навчання і тестування моделі.

2.3. Обхід моделі

Зловмисник може розібрати такий продукт і подивитися, як працює модель. Він людина і поки якщо не розумніший, то хоча б креативніший машини, тому він підлаштується під неє. Наприклад, наступний рядок вважається чистим (1200 балів), хоча його перша половина явно шкідлива: «лоыралоывврачигшуралорыловар Добавляем в конец много осмысленного текста, чтобы обмануть машину». Який би розумний алгоритм не використовувався, його завжди може обійти людина (досить розумна). Тому антивірусна лабораторія зобов'язана мати просунуту інфраструктуру для швидкої реакції на нові загрози.



Один із прикладі обходу описаного вище методу: усі слова виглядають правдоподібно, але насправді це нісенітниця.

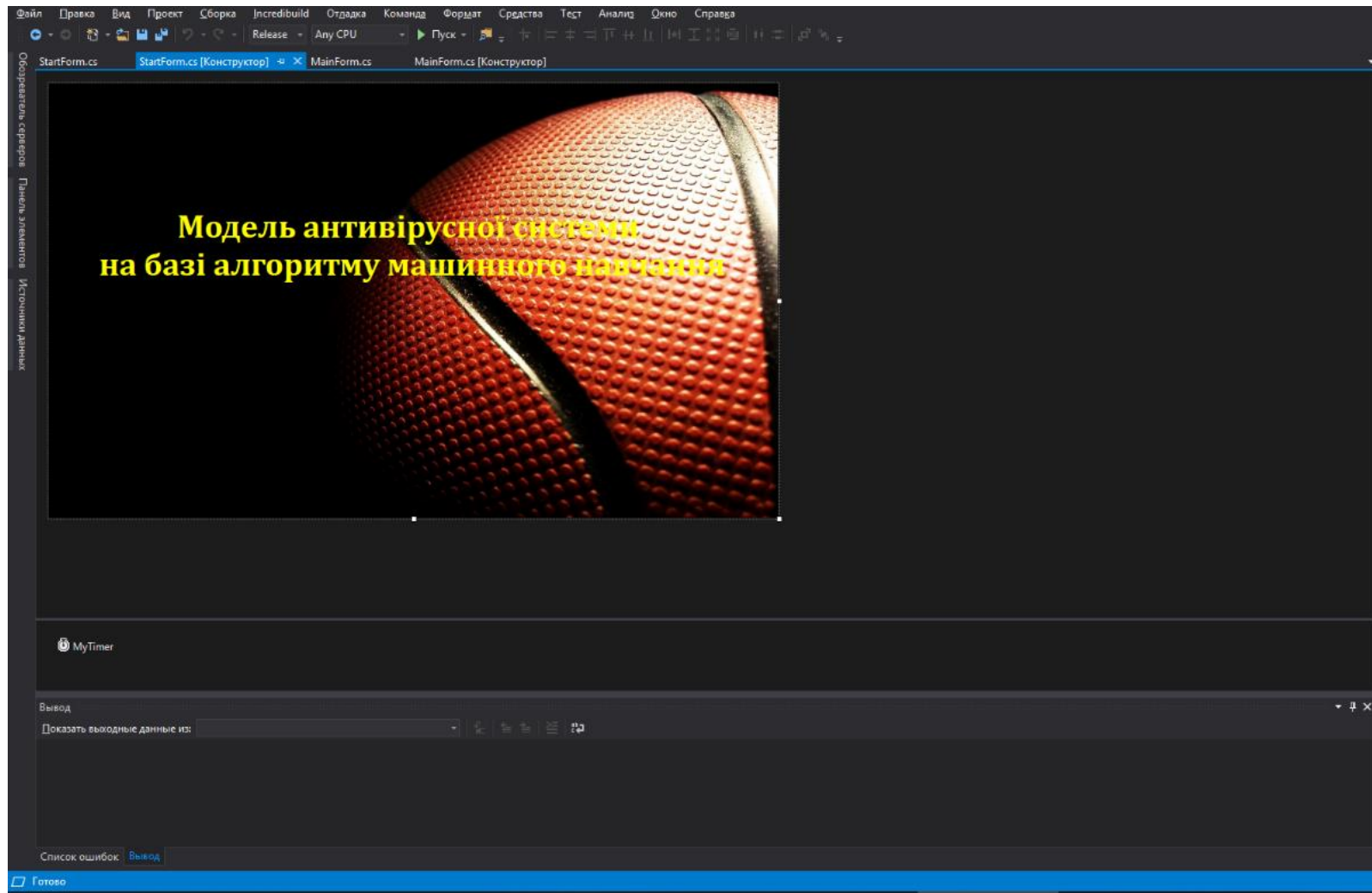
2.4. Оновлення моделі.

На прикладі описаного вище алгоритму згадувалося, що модель, навчена на російських текстах, буде непридатна для аналізу текстів з іншим алфавітом. А шкідливі файли, з урахуванням креативності зловмисників (див. попередній пункт), — це ніби поступово еволюціонуючий алфавіт. Ландшафт загроз змінюється досить швидко. За довгі роки досліджень розробили оптимальний підхід до поступового оновлення моделі прямо в антивірусних базах. Це дозволяє донавчати і навіть повністю перенавчати модель «без відриву від виробництва».

3. Тестування розробленої системи

Для навчання моделі було використано кілька словників (з різними мовами). Інтерфейс програмної частини зображено нижче.

Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки адекватності текстових даних засобами
машинного навчання.
113 – Прикладна математика



Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки адекватності текстових даних засобами
машинного навчання.

113 – Прикладна математика

Рис. 3.1 – Конструктор стартового вікна програми

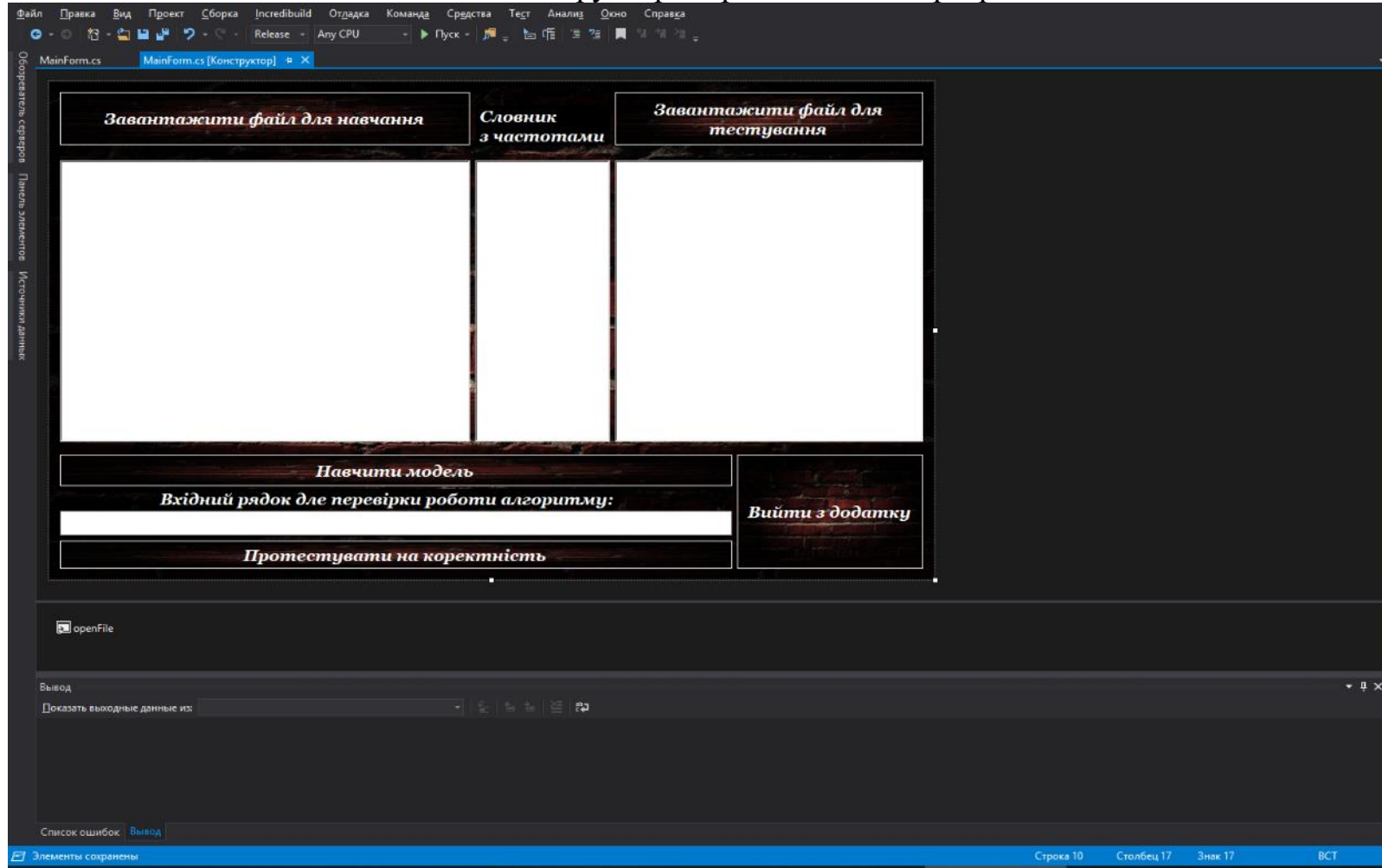


Рис. 3.2 – Конструктор головного вікна програми з функціональною частиною

Дисципліна «Структурні методи розпізнавання образів»

Комп'ютерний практикум № 11. Проєктування системи перевірки адекватності текстових даних засобами машинного навчання.

113 – Прикладна математика

Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проєктування системи перевірки
адекватності текстових даних засобами машинного навчання.

113 – Прикладна математика

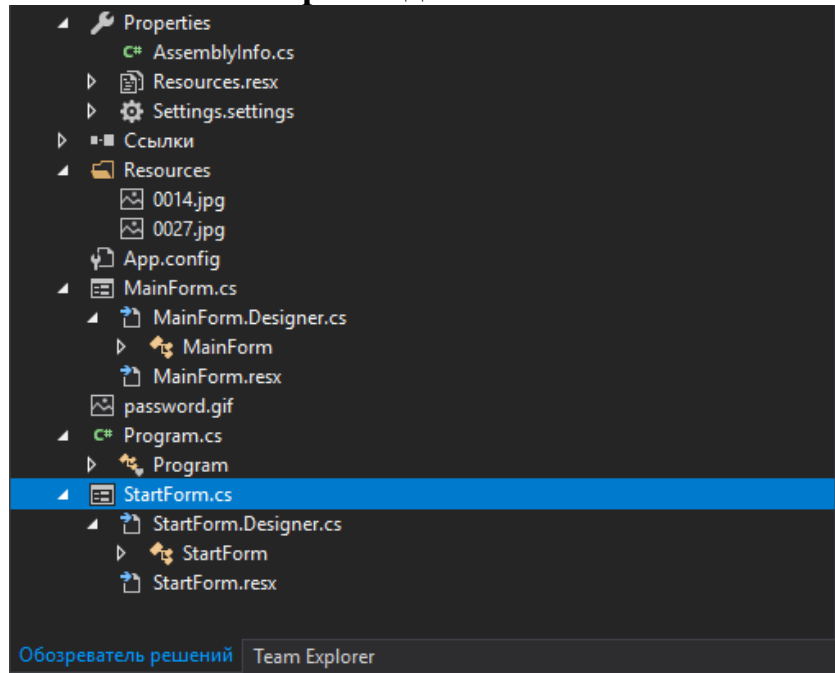


Рис. 3.3 – Загальна структура створеного проєкту мовою програмування C#

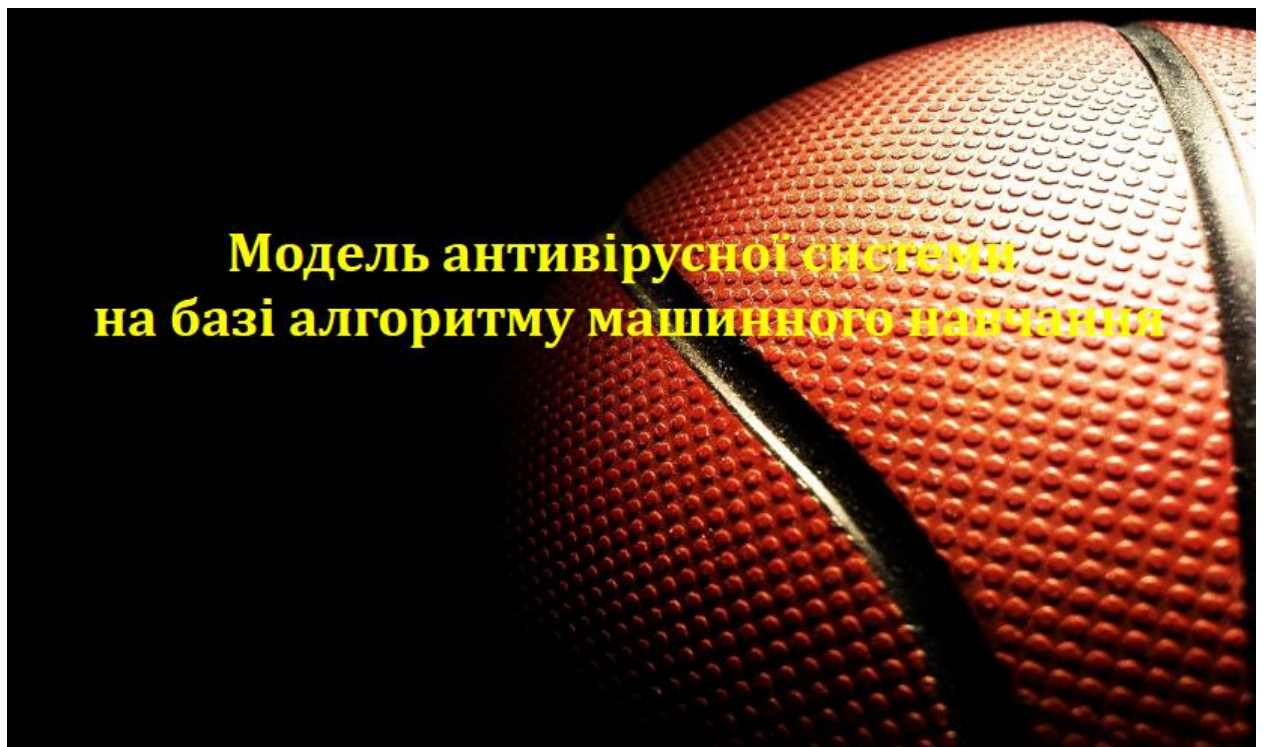


Рис. 3.4 – Запуск проєкту. Стартова форма програми

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика



Рис. 3.5 – Інтерфейс та робота у головній формі програми

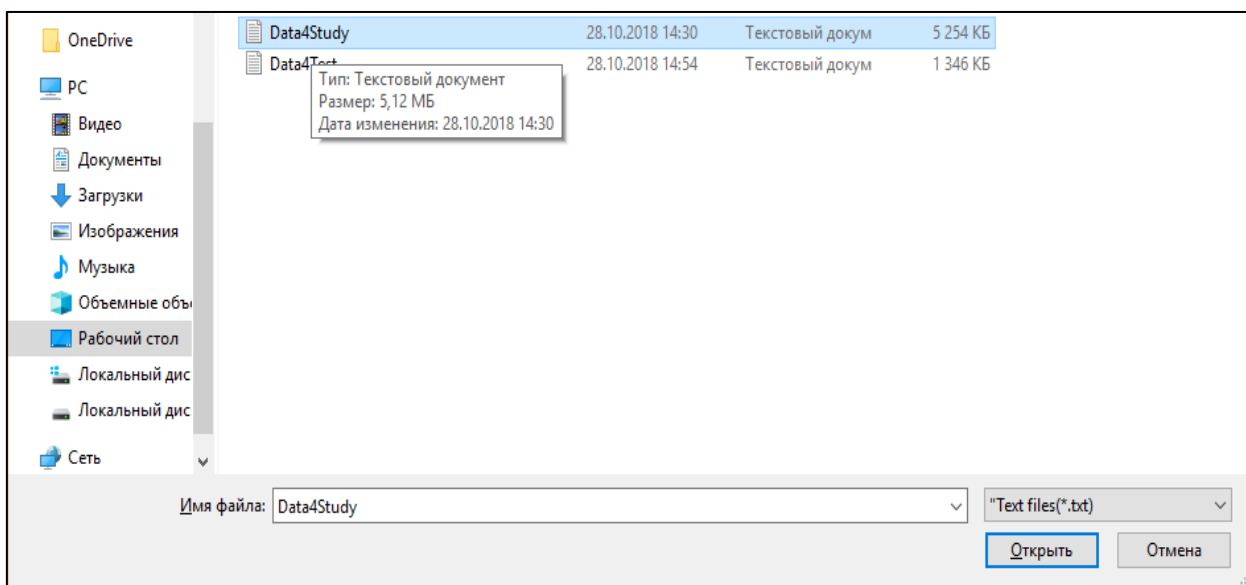


Рис. 3.6 – Завантаження даних з файлу для навчання згідно алгоритму

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

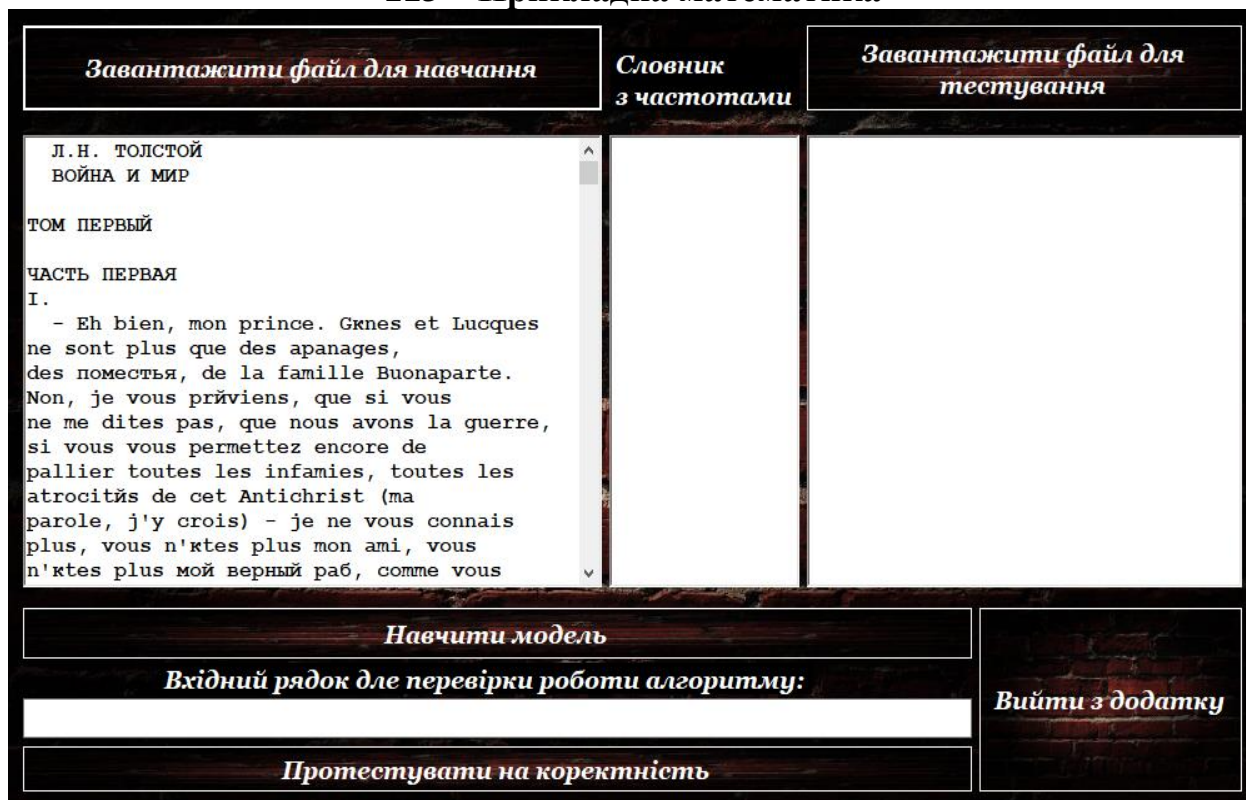


Рис. 3.7 – Результат завантаження даних з текстового файлу для навчання

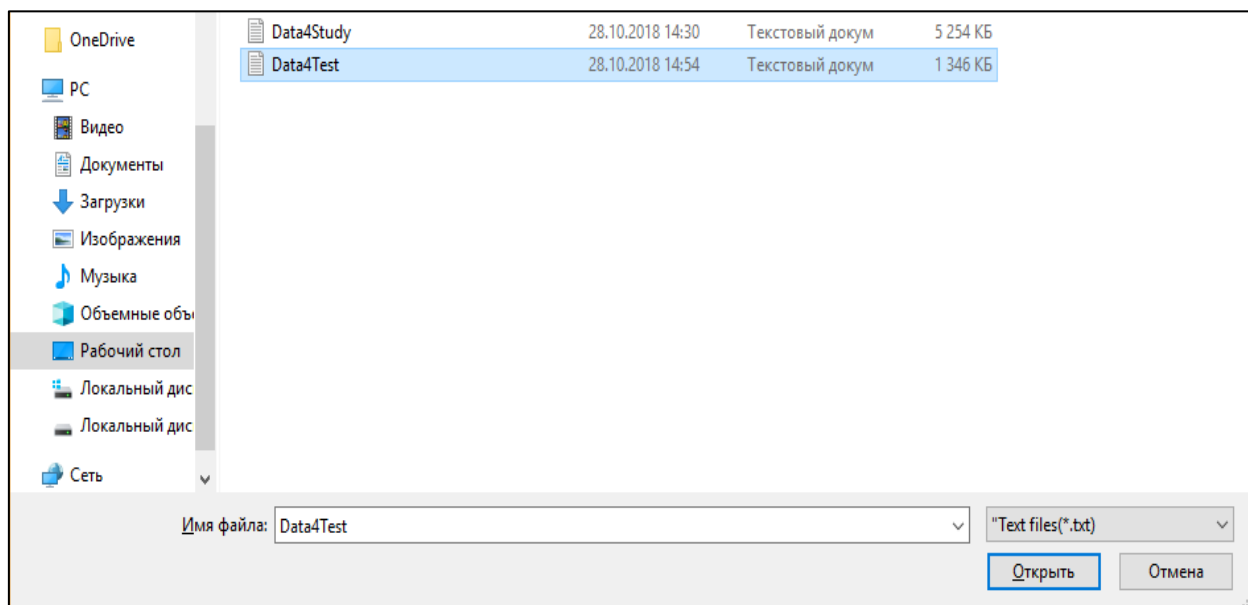


Рис. 3.8 – Завантаження даних з файлу для формування тестів та визначення порогового значення навчання згідно алгоритму

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

Завантажити файл для навчання	Словник з частотами	Завантажити файл для тестування
<p>Л.Н. ТОЛСТОЙ ВОЙНА И МИР</p> <p>ТОМ ПЕРВЫЙ</p> <p>ЧАСТЬ ПЕРВАЯ</p> <p>I.</p> <p>- Eh bien, mon prince. Gknes et Lucques ne sont plus que des apanages, des поместья, de la famille Buonaparte. Non, je vous prйviens, que si vous ne me dites pas, que nous avons la guerre, si vous vous permettez encore de pallier toutes les infamies, toutes les atrocitйs de cet Antichrist (ma parole, j'y crois) - je ne vous connais plus, vous n'кtes plus mon ami, vous n'кtes plus мой верный раб, comme vous</p>		<p>Михаил Афанасиевич Булгаков</p> <p>Мастер и Маргарита</p> <p>Русская классика Москва 1984</p> <p>Текст печатается в последней прижизненной редакции (рукописи хранятся в рукописном отделе Государственной библиотеки СССР имени В. И. Ленина), а также с исправлениями и дополнениями, сделанными под диктовку писателя его женой, Е. С. Булгаковой.</p> <p>Оглавление</p>
Навчити модель		Вийти з додатку
Вхідний рядок для перевірки роботи алгоритму:		
Протестувати на коректність		

Рис. 3.9 – Результат завантаження даних для перевірки та тестування алгоритму машинного навчання

Завантажити файл для навчання	Словник з частотами	Завантажити файл для тестування
<p>Л.Н. ТОЛСТОЙ ВОЙНА И МИР</p> <p>ТОМ ПЕРВЫЙ</p> <p>ЧАСТЬ ПЕРВАЯ</p> <p>I.</p> <p>- Eh bien, mon prince. Gknes et Lucques ne sont plus que des apanages, des поместья, de la famille Buonaparte. Non, je vous prйviens, que si vous ne me dites pas, que nous avons la guerre, si vous vous permettez encore de pallier toutes les infamies, toutes les atrocitйs de cet Antichrist (ma parole, j'y crois) - je ne vous connais plus, vous n'кtes plus mon ami, vous n'кtes plus мой верный раб, comme vous</p>	<p>14549</p> <p>Л 483</p> <p>Л. 5</p> <p>.Н 3</p> <p>Н. 7</p> <p>. 16937</p> <p>Т 1512</p> <p>ТО 11</p> <p>ОЛ 1</p> <p>ЛС 1</p> <p>СТ 18</p> <p>Ой 5</p> <p>Й 5</p> <p>47336</p> <p>10895</p>	<p>Михаил Афанасиевич Булгаков</p> <p>Мастер и Маргарита</p> <p>Русская классика Москва 1984</p> <p>Текст печатается в последней прижизненной редакции (рукописи хранятся в рукописном отделе Государственной библиотеки СССР имени В. И. Ленина), а также с исправлениями и дополнениями, сделанными под диктовку писателя его женой, Е. С. Булгаковой.</p> <p>Оглавление</p>
Навчити модель		Вийти з додатку
Вхідний рядок для перевірки роботи алгоритму:		
Протестувати на коректність		

Рис. 3.10 – Результат формування словника згідно алгоритму навчання – результат після натискання кнопки «Навчити модель»

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

Завантажити файл для навчання	Словник з частотами	Завантажити файл для тестування	
Л.Н. ТОЛСТОЙ ВОЙНА И МИР ТОМ ПЕРВЫЙ ЧАСТЬ ПЕРВАЯ I. - Eh bien, mon prince. Gknes et Lucques ne sont plus que des apanages, des поместья, de la famille Buonaparte. Non, je vous prйviens, que si vous ne me dites pas, que nous avons la guerre, si vous vous permettez encore de pallier toutes les infamies, toutes les atrocitйs de cet Antichrist (ma parole, j'y crois) - je ne vous connais plus, vous n'кtes plus mon ami, vous n'кtes plus мой верный раб, comme vous	14549 Л 483 Л. 5 .Н 3 Н. 7 . 16937 Т 1512 ТО 11 ОЛ 1 ЛС 1 СТ 18 ОЙ 5 Й 5 47336 10895	Михаил Афанасиевич Булгаков Мастер и Маргарита Русская классика Москва 1984 Текст печатается в последней прижизненной редакции (рукописи хранятся в рукописном отделе Государственной библиотеки СССР имени В. И. Ленина), а также с исправлениями и дополнениями, сделанными под диктовку писателя его женой, Е. С. Булгаковой. Оглавление	
<div align="center">Навчити модель</div> <div> Вхідний рядок для перевірки роботи алгоритму: <input type="text" value="Привет, моя любовь"/> </div> <div align="center">Протестувати на коректність</div>			<div align="center">Вийти з додатку</div>

Рис. 3.11 – Введення рядка для визначення його адекватності на основі навченої моделі на основі алгоритму машинного навчання

Завантажити файл для навчання	Словник з частотами	Завантажити файл для тестування	
Л.Н. ТОЛСТОЙ ВОЙНА И МИР ТОМ ПЕРВЫЙ ЧАСТЬ ПЕРВАЯ I. - Eh bien, mon prince. Gknes et Lucques ne sont plus que des apanages, des поместья, de la famille Buonaparte. Non, je vous prйviens, que si vous ne me dites pas, que nous avons la guerre, si vous vous permettez encore de pallier toutes les infamies, toutes les atrocitйs de cet Antichrist (ma parole, j'y crois) - je ne vous connais plus, vous n'кtes plus mon ami, vous n'кtes plus мой верный раб, comme vous	14549 Л 483 Л. 5 .Н 3 Н. 7 . 16937 Т 1512 ТО 11 ОЛ 1 ЛС 1 СТ 18 ОЙ 5 Й 5 47336 10895	Михаил Афанасиевич Булгаков Мастер и Маргарита Русская классика Москва 1984 Текст печатается в последней прижизненной редакции (рукописи хранятся в рукописном отделе Государственной библиотеки СССР имени В. И. Ленина), а также с исправлениями и дополнениями, сделанными под диктовку писателя его женой, Е. С. Булгаковой. Оглавление	
<div align="center">Навчити модель</div> <div> Вхідний рядок для перевірки роботи алгоритму: <input type="text" value="Привет, моя любовь"/> </div> <div align="center">Протестувати на коректність</div>			<div align="center">Вийти з додатку</div>

Рис. 3.12 – Результат розпізнавання введенного користувачем рядка

**Дисципліна «Структурні методи розпізнавання образів»
Комп'ютерний практикум № 11. Проектування системи перевірки
адекватності текстових даних засобами машинного навчання.**

113 – Прикладна математика

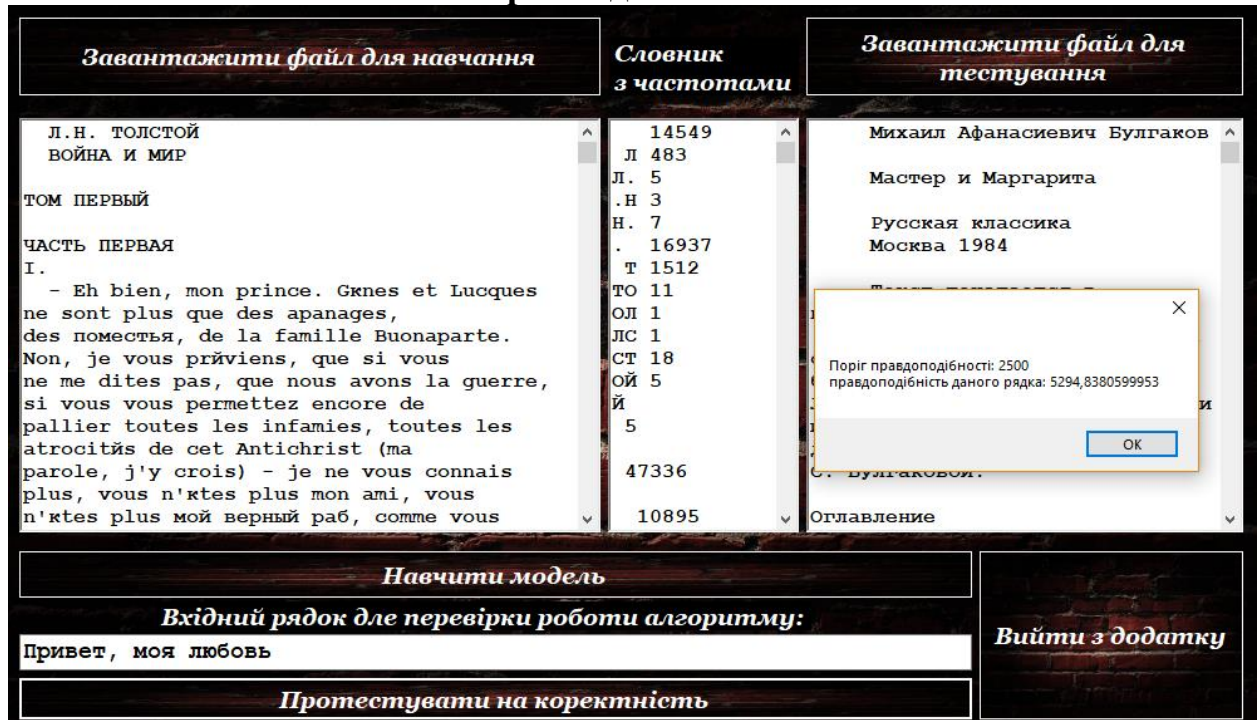


Рис. 3.13 – Визначення порогу адекватності моделі та рівня адекватності введеного рядка

4. Завдання до лабораторної роботи

1. Спроекувати та розробити програмний засіб, який навчається за вище описаним алгоритмом та має інтерфейс користувача, за допомогою якого користувач може завантажувати досить великий текст для навчання та може перевіряти адекватність уведеного користувачем рядка (для його перевірки).
2. Зразок інтерфейсу наведено нижче з відповідними результатами роботи програмного засобу.
3. Занести до звіту основні результати проектування. Серед таких результатів повинна бути діаграма класів, діаграма прецедентів, діаграма розгортання на апаратних засобах.
4. Програмну реалізацію засобу занести до додатків до лабораторної роботи. Основні фрагменти реалізація, які відображають алгоритмічну складову машинного навчання відокремити з поясненнями в основний текст звіту з результатами виконання лабораторної роботи.
5. Зробити висновки до отриманих результатів у лабораторній роботі.

Список використаних джерел

1. <https://www.kaspersky.ru/blog/machine-learning-explained/13605/>