ДОДАТОК А
Фрагменти програмного коду

classification.py

```
# class of ML classificator based on sklearn
interfaces    from    sklearn.base    import
BaseEstimator, ClassifierMixin from sklearn
import cross_validation
from sklearn import ensemble
from sklearn.externals import joblib
from sklearn.linear_model import LogisticRegression,
LogisticRegressionCV from sklearn.svm import SVC


class MyClassifier(BaseEstimator,
  ClassifierMixin): def __init__(self,
  categories, directory):
    self.atomars = load_atomars(categories, directory)
    self.referee = joblib.load(directory +
    "referee.pkl") self.referee.probability = True # Some
    bug in joblib :( again self.categories = categories

  def
    dig(self,
    x):
    return
    int(x *
    10)

  def fit(self, X, y):
    if self.referie == None:
      self.referee =
      LogisticRegression()
      self.referee.fit(self._L_1_zip_prob
    a(X), y) return self

  def predict(self, X):
    return self.referee.predict(self._L_1_zip_proba(X))
  def predict_proba(self, X):
    return self.referee.predict_proba(self._L_1_zip_proba(X))

  def
    _L_1_proba(se
    lf, X): pred
    = None
    data = []
    for category in
      self.categories: res =
      []
      pred =
      self.atomars[category].predict_proba(X)
      pred_len = len(pred)
```

```
            for i in range(pred_len):
              res.append(self.dig(pred[i][1]))
            data.ap
          pend(res)
          return
          data

        def _L_1_zip_proba(self,
          X): return
          zip(*self._L_1_proba(X)
          )
        def load_atomars(categorie
          s, directory=''):
          atomars = { }
        for category in categories:
          atomars[category] = joblib.load(directory +

category+'_.pkl') return atomars
```

ДОДАТОК Б

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Інформаційних управляючих систем_____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

ГРАФІЧНИЙ МАТЕРІАЛ

_____«Дослідження методів класифікації веб-сторінок на основі технології інтелектуального аналізу даних»_____
(тема)

Студент гр. _ІУСТм-20-1_ _____ __Сотников К. В.__
(шифр групи) (підпис) (прізвище, ініціали)

Науковий керівник роботи _____ ___доц. Міхнова А. В.___
(підпис) (посада, прізвище, ініціали)

2021 р.

Слайд 2 - Загальна характеристика роботи

Таблиця Б.1 – Загальна характеристика роботи

Тема КР	«Дослідження методів класифікації веб-сторінок на основі технології інтелектуального аналізу даних»
Актуальність	Визначення тематики контенту веб-сторінок є однією з найважливіших задач багатьох інтернет-компаній. Наприклад, за умови коректної класифікації можна пропонувати користувачеві більш точну підбірку рекламних блоків, що в свою чергу дозволить підвищити продаж як місць розміщення рекламних банерів, так і рекламованого товару. Крім того, захист від небажаної інформації також є однією з основних можливих сфер застосування класифікації контенту.
Мета досліджень	Дослідження методів класифікації веб-сторінок та модифікація існуючих методів інтелектуального аналізу даних для підвищення точності класифікації веб-сторінок.
Задачі досліджень	- дослідження існуючих методів класифікації веб-контенту; - дослідження існуючих методів та алгоритмів інтелектуального аналізу даних; - вибір методології ведення проектів інтелектуального аналізу даних; - оцінка ефективності запропонованих методів
Методи досліджень	Теоретичні та емпіричні
Нові наукові результати	Наукова новизна полягає в модифікації методів класифікації на основі «сусідніх» веб-сторінок, що дозволило підвищити точність класифікації.
Практична значимість роботи	В ході проведення аналізу досліджуваної області було визначено, що існуючі методи в повній мірі не задовольняють сучасним вимогам точності і повноти класифікації веб-сторінок. Розроблені нові методи підвищення точності моделі класифікації веб-контенту на основі існуючих та розроблена модель, що дозволяє виконувати класифікацію веб-сторінок з точністю 96%

Слайд 3 - Use case диаграмма роботи системи фільтрації контенту

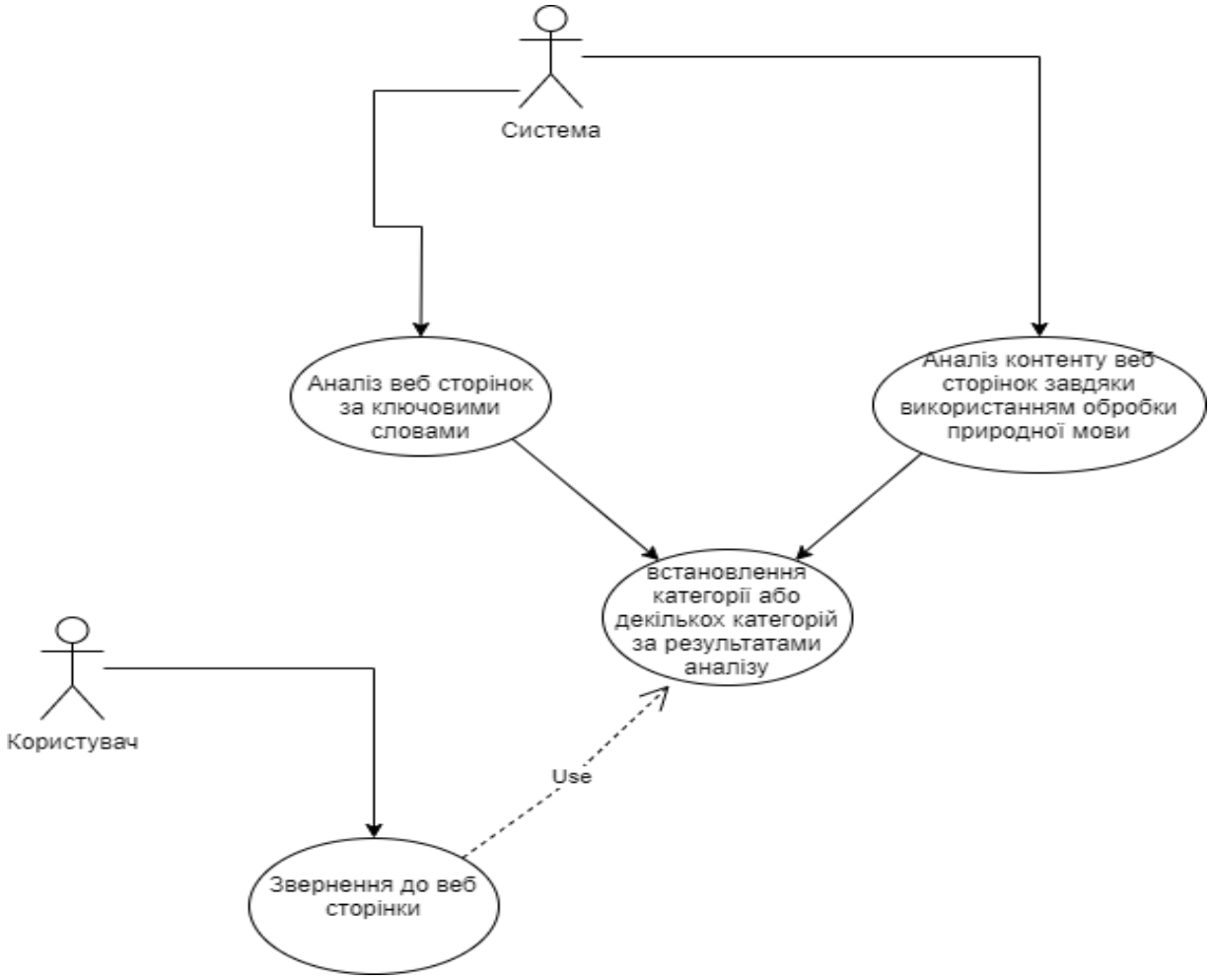

Рисунок Б.1 - Use case диаграмма роботи системи фільтрації контенту

Слайд 4 Основні завдання та відповідни методи інтелектуального аналізу даних

Таблиця Б.2 - Основні завдання та відповідни методи інтелектуального аналізу даних

Завдання ІАД	**Опис завдання**	**Методи ІАД**
Класифікація	Пошук спільних ознак в певному наборі даних відповідно до конкретної групи, класу	Нейроні мережи, дерево рішень, метод найближчого сусіда
Регресія	Отримання конкретних відомостей про те, яку форму і характер має залежність між змінними, що досліджуються	Метод найменших квадратів, статистичні методи
Кластерізація	Кластеризація представляє собою логічне продовженням ідеї класифікації. Це завдання більш складне, адже особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи	Нейроні мережі
Пошук асоціативних правил	Визначення закономірності між пов'язаними подіями в наборі даних, котрі відбуваються одночасно	Статистичні методи
Пошук послідовності	Завдання пошуку послідовності схоже з завданням асоціації, але її метою є встановлення закономірностей між подіями, що впорядковані в часі (тобто відбуваються в деякому порядку)..	Кореляційний аналіз
Прогнозування	Аналіз історичних даних та формування на їх основі майбутніх припущень	Статистичні методи, нейроні мережи, дерево рішень, лінійна регресія, метод опорних векторів

Слайд 5 - Порівняльна характеристика методів інтелектуального аналізу даних

Таблиця Б.3- Порівняльна характеристика методів інтелектуального аналізу даних

Алгоритм	Точність	Масштабованість	Інтерпретованість	Здатність до використання	Трудомісткість	Різнобічність	Швидкість	Популярність
Класичні методи	нейтральна	висока	висока/ нейтральна	висока	нейтральна	нейтральна	висока	низька
Нейронні мережі	висока	низька	низька	низька	нейтральна	низька	дуже низька	низька
Методи візуалізації	висока	дуже низька	висока	висока	дуже висока	низька	надзвичайно низька	висока/ нейтральна
Дерева рішень	низька	висока	висока	висока/ нейтральна	висока	висока	висока/ нейтральна	висока/ нейтральна
Поліноміальні нейронні мережі	висока	нейтральна	низька	висока/ нейтральна	нейтральна/низька	нейтральна	низька /нейтральна	нейтральна
k-найближчого сусіда	низька	дуже низька	висока/ нейтральна	нейтральна	нейтральна/низька	низька	висока	низька

Слайд 6 - Етапи інтелектуального аналізу даних

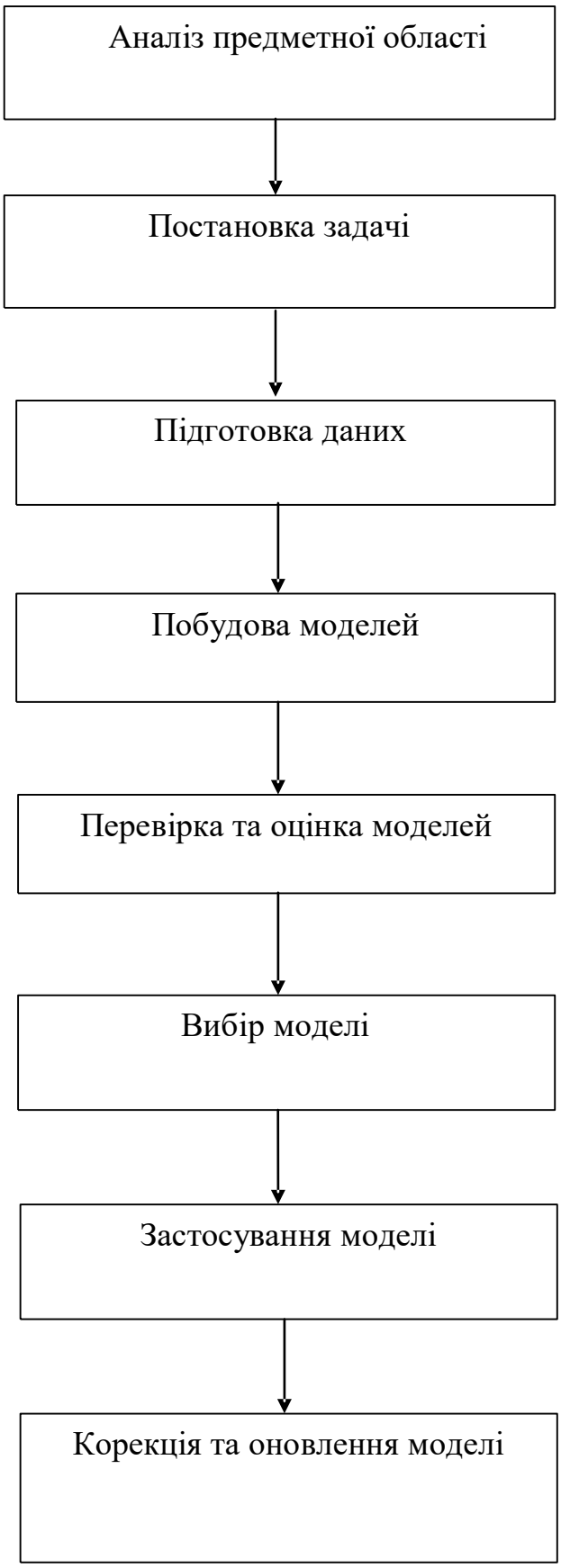

Рисунок Б.2 - Етапи інтелектуального аналізу даних

Слайд 7 - Життєвий цикл дослідження даних CRISP-DM

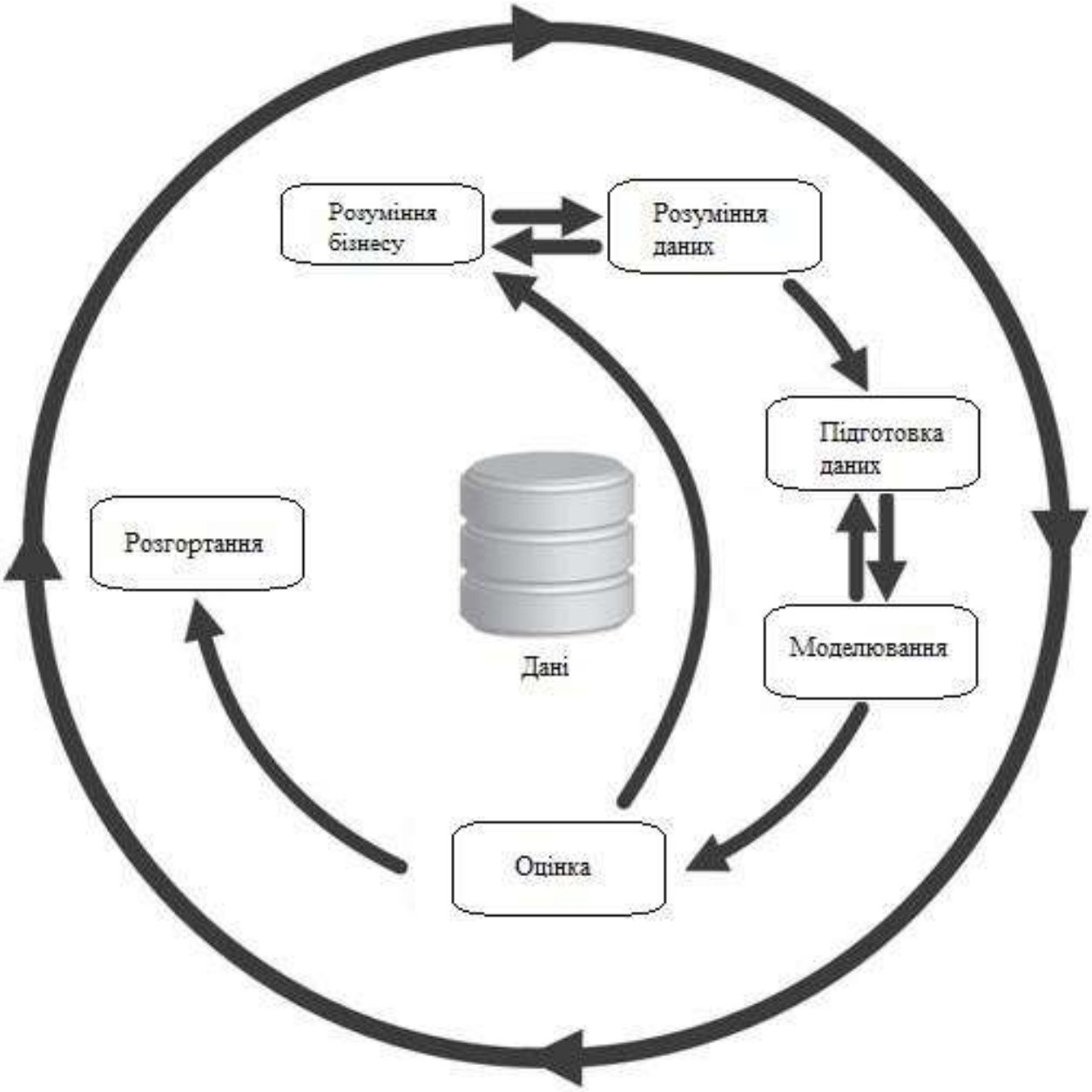

Рисунок Б.3 - Життєвий цикл дослідження даних CRISP-DM

Слайд 8 - Матриця неточностей та метрики пошуку інформації

Таблиця Б.4 - Матриця неточностей

		Вірні результати	
		1	0
Результат моделі	1	TP	FP
	0	FN	TN

Формула 1 - Метрика accuracy

$$A(accuracy) = \frac{tp+fp}{N}$$

Формула 2 - Метрика precision

$$P(precision) = \frac{tp}{tp+tn}$$

Формула 3 - Метрика повноти recall

$$R(recall) = \frac{tp}{tp+fn}$$

Формула 4 - Метрика F1 score

$$F_1 = 2 * \frac{P*R}{P+R}$$

Формула 5 - micro-averaging

$$P_{micro} = \frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k} (tp_i+fp_i)},$$

Формула 6 - macro-averaging

$$P_{macro} = \frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i+fp_i}}{k},$$

Слайд 9 – Заборонені та дозволені категорії

Таблиця Б.5 - Заборонені категорії

Категорія	Зміст
alcohol	Інформація про алкогольні продуктах
drugs	Інформація про наркотики
gambling	Інформація про азартні ігри
adults	Порнографічний контент
smoking	Інформація про тютюнову продукцію
violence	Інформація про насилля
weapons	Інформація про зброю
terrorism	Інформація про тероризм
suicide	Інформація про суїцид

Таблиця Б.6 - Дозволені категорії

Категорія	Зміст
news	Веб-сайти новин
sport	Спортивні веб-сайти
education	Веб-сайти навчальних установ
finance	Економіка і фінанси
shopping	Інтернет-магазини
whitePages	Контент для дітей

Слайд 10 - Схема структури бази даних

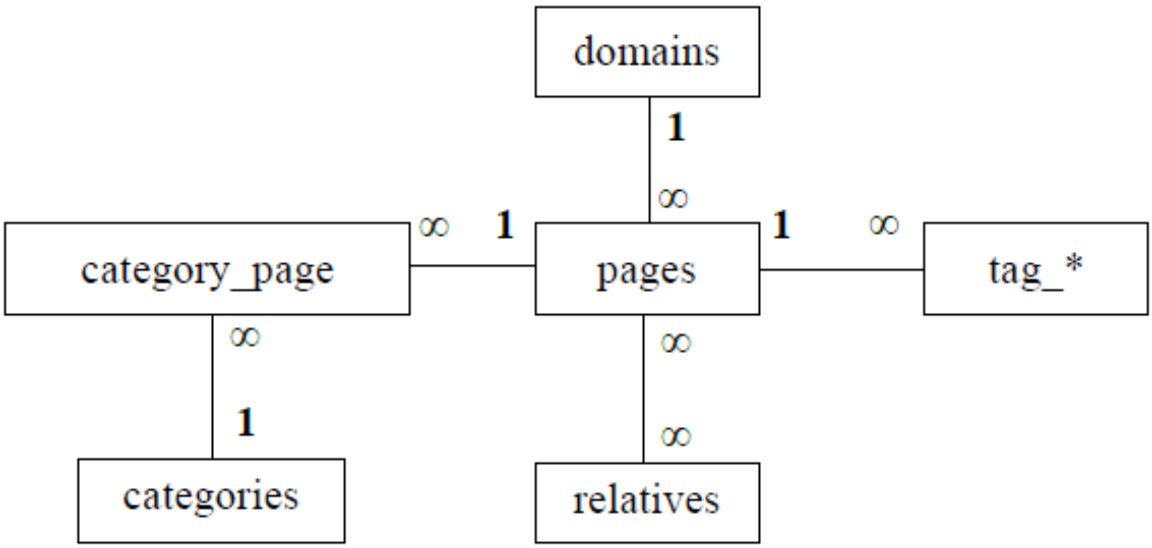

Рисунок Б.4 - Схема структури бази даних

Слайд 11 - Етапи роботи з адресами веб-сайтів

Рисунок Б.5 - Етапи роботи з адресами веб-сайтів

Слайд 12 - Етапи фільтрації тексту

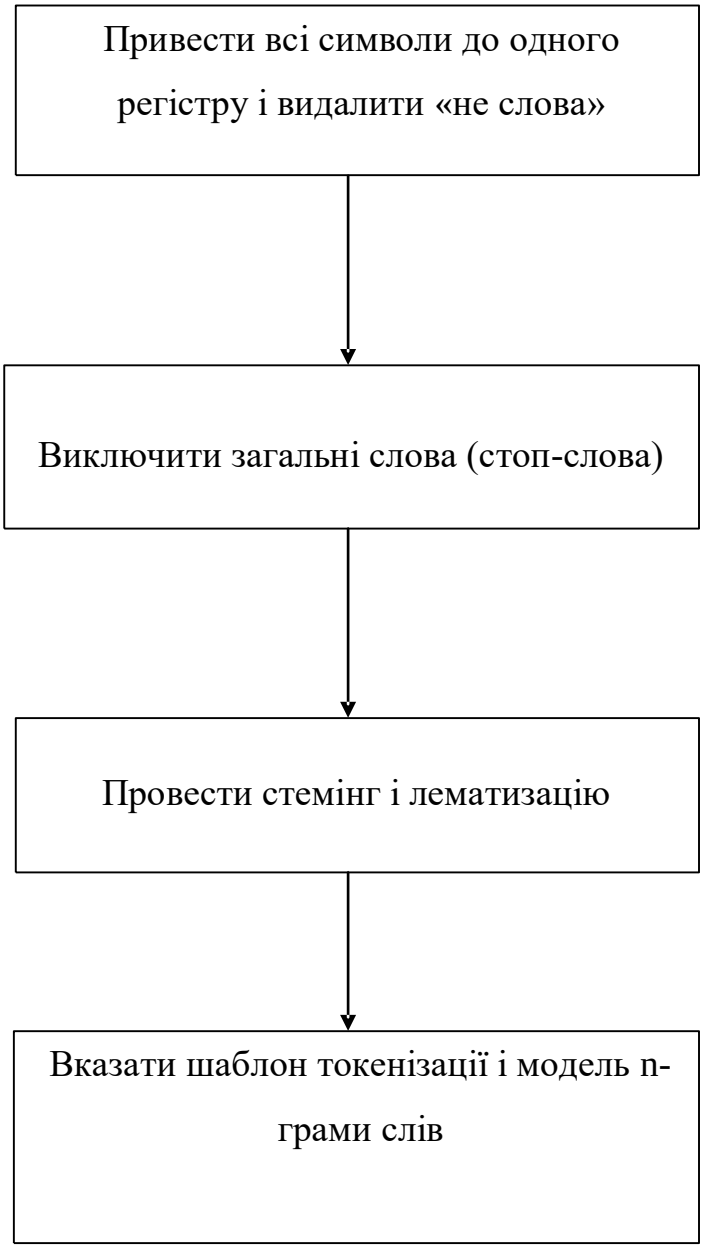

Рисунок Б.6 - Етапи фільтрації тексту

Слайд 13 - Результати векторізації

Таблиця Б.7 – Результати векторізації

Число атрибутів	HV	TF	TF+SB
n-gram(1, 2)			
5000	0,747	0,789	0,790
20000	0,767	0,778	0,781
max	0,767	0,772	0,778
n-gram (1, 3)			
5000	0,744	0,788	0,790
20000	0,757	0,783	0,780
50000	0,758	0,772	0,779
n-gram (2, 3)			
5000	0,529	0,687	0,699

Слайд 14 - Представлення зв'язків між веб-сторінками

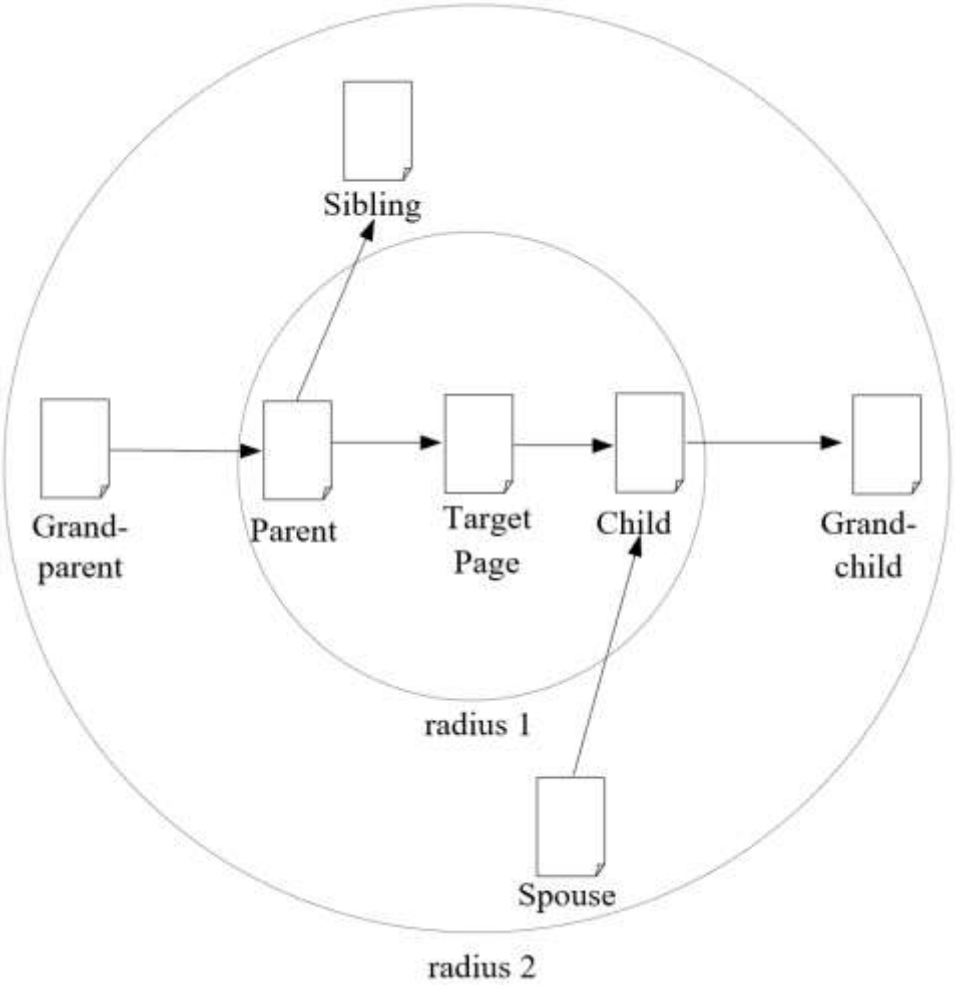

Рисунок Б.7 - Представлення зв'язків між веб-сторінками

Слайд 15 - Порівняльний аналіз методик класифікації

Таблиця Б.8 – Порівняльний аналіз методик класифікації

Застосована методика інтелектуального аналізу даних		Точність (precision micro-averaging)
На основі HTML-тегів		0,766
Навчання по «правильно передбаченим»		0,858
Метод головних компонент (PCA)		0,875
На основі заголовка веб-сторінки		0,669
На основі адреси веб-сторінки		0,631
На основі заголовка і адреси веб-сторінки		0,713
Word2vec	Average vector	0,860
	Bag of centroids	0,855
Ієрархічна класифікація	На основі голосування	0,879
	На основі голосування («правильно передбаченим»)	0,928
	Рефері	0,887
	Рефері («правильно передбаченим»)	0,960
Використання «сусідніх» веб-сторінок	На основі голосування за результатами дочірніх сторінок	0,690
	На основі рефері за результатами дочірніх веб-сторінок	0,850
	На основі об'єднаного тексту дочірніх Сторінок	0,770
	З урахуванням «впевненості» дочірніх веб-сторінок	0,851
	На основі зв'язків сусідніх веб-сторінок (Siblings, Spouse)	0,795
	Підрахунок посилань	0,347

Слайд 16 – Висновки

В ході роботи проведено аналіз існуючих методів класифікації веб-сторінок, який дозволив зробити висновок, що такі методи класифікації у повній мірі не задовольняють сучасним вимогам точності та повноти класифікації.

Методи, засновані на класифікації «сусідніх» веб-сторінок, не дозволяють отримати очікуване поліпшення точності. Можливо, це пов'язано з тим обмеженим набором «сусідніх» веб-сторінок, який використовувався в даному дослідженні. Також процес класифікації, при використанні «сусідніх» веб-сторінок займає чимало часу через необхідність їх скачування.

Проведенідослідження показують, що найбільш простим і ефективним способом класифікації, що досліджувалися у даній роботі, є класифікація на основі ієрархічної моделі з використанням бінарних моделей класифікації з рефері.

Також виявлено, що найбільш складно класифікувати веб-сторінки, які не містять тексту. Оскільки зазвичай людина оцінює вміст веб-сторінки на основі зображень, то таких сторінок досить багато. В майбутньому можливе додавання атрибутів такого типу, що допоможе поліпшити якість класифікації. Тому одним з можливих напрямків подальших досліджень може бути використання нових атрибутів, що базуються на зображеннях.

Слайд 17

Дякую за увагу