Дипломна робота

НА ТЕМУ: “ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНОГО ТЕКСТУ.”

Виконав: ст. гр. 1АКІТ-18М Стовбчатий М. М.

Керівник: д.т.н., проф., зав.каф. АІІТ Квєтний Р.Н.

Консультант: д.т.н., проф. каф. АІВТ Бісікало О. В.

Актуальність:

Можливе використання для:

А) Виявлення плагіату;

Б) Встановлення авторства невідомого тексту;

В) Оперативного визначення недоброчесних або зловмисних дій користувачів інформаційних систем.

Мета роботи полягає в підвищенні якості визначення авторства україномовного тексту на основі методів і моделей комп’ютерної лінгвістики та машинного навчання, а також доступних програмних бібліотек і технологічних засобів.

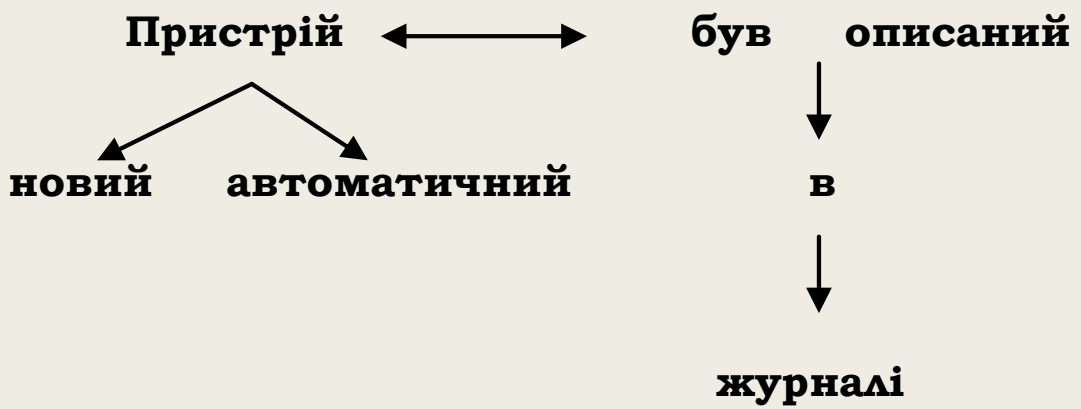
Об'єкт дослідження - процеси статистичного, синтаксичного та семантичного аналізу україномовних текстів.

Предмет дослідження – моделі, методи та засоби визначення авторства природно-мовних текстів.

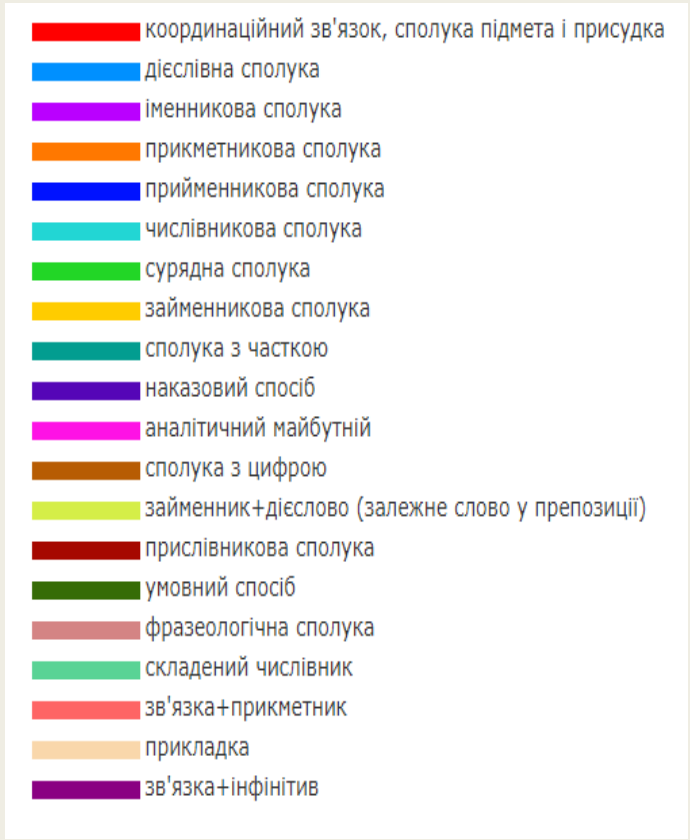
Речення у вигляді графа:

1 2 3 4 5 6 7

Новий автоматичний пристрій був описаний в журналі



Речення у вигляді графа:



Параметри, які формально описують дерево обрано:

- кількість вузлів у графі (словоформ) у реченні;
- кількість простих речень у складному;
- кількість рівнів у графі;
- максимальна кількість змін у шляху гілки графа;
- максимальна довжина дуги графа;
- загальна кількість вузлів у графі;
- середня кількість рівнів;
- середня кількість вузлів у рівні графа;
- співвідношення всіх вузлів речення, які не є термінальними (не є листями), до всіх вузлів цього речення;
- середня глибина гілки речення.

Для роботи було відібрано 3 автори твори яких були проаналізовані, створювалися графи по реченням , та сформувався файл який формально математичними параметрами описує стиль автора.

Приклад даних за якими вивчалася нейронна мережа

Автор	кількість вузлі	кількість прости	кількість рівнів у	ширина гілкуван	максимальна кіл	максимальна пр	загальна кількіс	середня кількіст	середня кількіст	співвідношенн	середня глибина гілки речення
Drach	6	1	4	2	1	1	0	0.5	1.5	0.5	3
Drach	17	2	6	2	2	10	3	0.5	3.3	0.5	4.8
Drach	10	1	4	4	1	3	1	0.5	2.5	0.5	3
Drach	15	1	10	2	1	3	2	2.66	1.5	0.73	5.5
Drach	17	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Drach	10	1	5	3	1	2	1	0.3	2	0.7	4
Drach	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Drach	20	2	6	2	2	10	3	0.5	3.3	0.5	4.8
Drach	10	1	4	4	1	3	1	0.5	2.5	0.5	3
Drach	15	1	10	2	1	3	2	2.66	1.5	0.73	5.5
Drach	9	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Drach	17	1	5	3	1	2	1	0.3	2	0.7	4
Drach	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Drach	10	1	4	4	1	3	1	0.5	2.5	0.5	3
Drach	15	1	10	2	1	3	2	2.66	1.5	0.73	5.5
Drach	13	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Drach	10	1	5	3	1	2	1	0.3	2	0.7	4
Drach	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Drach	13	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Drach	9	1	5	3	1	2	1	0.3	2	0.7	4
Drach	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Drach	10	1	4	4	1	3	1	0.5	2.5	0.5	3
Drach	15	1	10	2	1	3	2	2.66	1.5	0.73	5.5
Drach	13	3	5	5	2	6	0	0.46	2.6	0.53	3.33

Вирішення задачі за допомогою бібліотеки машинного навчання Scikit-learn

Для вирішення задачі була використана бібліотека машинного навчання «sklearn» а саме за допомогою її інструмента Multi-layer Perceptron (MLP).

Параметри були перетворені у csv-формат і розбиті на виборку навчання і тестову виборку у пропорції 60% (150) / 40% (111).

Результати класифікації:

Результати позитивні – нейронна мережа показала достовірність 95%

```
C:\Users\fkca2\Desktop\nn_classifier\venv\Scripts\python.exe C:/Users/fkca2/Desktop/nn_classifier/nn_classifier.py
[[50  0  0]
 [ 0 50  0]
 [ 0  0 50]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        50
         1.0       1.00      1.00      1.00        50
         2.0       1.00      1.00      1.00        50

    accuracy                           1.00       150
   macro avg       1.00      1.00      1.00       150
weighted avg       1.00      1.00      1.00       150

[[32  5  0]
 [ 0 37  0]
 [ 0  2 35]]
              precision    recall  f1-score   support

         0.0       1.00      0.86      0.93        37
         1.0       0.84      1.00      0.91        37
         2.0       1.00      0.95      0.97        37

    accuracy                           0.94       111
   macro avg       0.95      0.94      0.94       111
weighted avg       0.95      0.94      0.94       111


Process finished with exit code 0
```

Альтернативний експеримент з використанням алгоритму МГВА

Метод группового врахування аргументів - сімейство індуктивних алгоритмів для мате-матичного моделювання багатопараметричних даних. Метод заснований на рекурсивному селективному відборі моделей, на основі яких будуються складніші моделі. Точність моделювання на кожному наступному кроці рекурсії збільшується за рахунок ускладнення моделі.

Для тесту також було взято два масиви, один з вхідними даними та тестові дані, які були відібрані з загального масиву.

Результати класифікації МГВА

Мойсієнко	Результат	Вінграновський	Результат	Драч	Результат
-0,838841777	0	-9,452837995	1	-4,63644683	1
5,43485192	1	-9,215371344	1	-6,535968421	1
9,490017091	1	-8,965253295	1	-10,21520101	1
9,679178561	1	-7,173696854	1	-9,081437844	1
9,674322794	1	-7,465227715	1	-9,081437844	1
10,92877734	1	-9,912205063	1	-6,132787341	1
5,43485192	1	-9,215371344	1	-6,535968421	1
-0,838841777	0	-9,452837995	1	-4,63644683	1
5,43485192	1	-9,215371344	1	-6,535968421	1
9,490017091	1	-8,965253295	1	-10,21520101	1
-9,585205982	1	7,504215039	1	-9,77265416	1
-3,732990179	1	8,566666043	1	-5,916235136	1
2,602581617	0	6,380756829	1	-3,462892186	1
-7,71584839	1	4,776869905	1	-7,263811705	1
-7,320355897	1	9,86998084	1	-9,02613012	1
-5,362515209	1	-8,244848369	0	-7,092175545	1
-5,592998979	1	8,144605935	1	-7,092175545	1
-8,537733392	1	8,820832673	1	-7,48792009	1
-9,559387173	1	-8,82085707	0	-6,814692161	1
-6,291477883	1	9,770904524	1	-9,928450064	1
-9,506860578	1	-8,51349922	1	10,20464318	1
-8,337961264	1	-8,815593061	1	9,37328861	1
-9,816400328	1	-8,593943169	1	10,20725501	1
-9,816400328	1	-8,593943169	1	10,20725501	1
-9,602853	1	-9,773967609	1	10,20464318	1
-8,337961264	1	-8,815593061	1	9,37328861	1
-9,816400328	1	-8,593943169	1	10,20725501	1
0,333346283	1	-9,178381398	1	-5,787957436	0
-9,344067875	1	-9,181156558	1	9,886701053	1
-9,602853	1	-9,773967609	1	10,20464318	1
	90%		93%		97%

Висновки:

- Внаслідок дослідження запропоновано новий метод визначення авторства україномовного тексту, який, на відміну від існуючих, базується на лінгвістичній моделі побудови графу зв'язків між лексичними одиницями речення тексту та застосуванні методів машинного навчання за новими формальними ознаками множини речень тексту, що дозволяє підвищити якість визначення авторства україномовного тексту.
- Збіг позитивних результатів машинного навчання за допомогою нейронної мережі та методу МГВА демонструє інформативність обраних формальних ознак синтаксичної структури речення україномовного тексту для атрибуції авторства та підтверджує ефективність запропонованого методу визначення авторства україномовного тексту.