КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ
КАФЕДРА ОБЧИСЛЮВАЛЬНОЇ МАТЕМАТИКИ

ВИПУСКНА КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА
на тему:

Методи ідентифікації автора

Виконав
студент 2-го курсу магістратури
Михайлюк Владислав Юрійович

Науковий керівник:
доктор фізико-математичних наук
Клюшин Дмитро Анатолійович

Київ — 2020

1 Зміст

Зміст

2 Вступ

Література часто розглядається як предмет, який стосується лише читання та мислення. Але, як це відбувається майже у всіх областях інтелекту, воно перетинається з іншими. У своїй роботі "Принципи стильометрії"1890 року польський філософ Вінсентій Лутославський використовував статистичні підходи для побудови хронології діалогів Платона. Кажуть, що стилометрія почалася, коли в 1851 році Август де Морган сказав про тексти біблійних авторів, що «один з авторів використовує довші слова». Стильометрія - лінгвістична дисципліна, яка застосовує статистичний аналіз до літератури шляхом оцінки авторського стилю за допомогою різних кількісних критеріїв.

Розвиток копьютерів та їх можливостей для аналізу великох кільскоті данних створює можливості для дослудження більш складних статистичних методів. На тематику роботи з літературними творами та текстами існує багато досліджень, які використовують різні методи та підходи до вирішення задачі. У роботах були розглянуті такі підходи, як глибинне навчання Liuyu Zhou[?], Mike Kestemont[?], N. Smirnov[?], класичні методи машинного навчання Granichin[?], Grieve[?], та статистичні методи Stamatatos[?], Giacomo Inches[?], Shane Bergsma[?], Борисов Л.А[?].

Питяння статистичного аналізу знаходиться у сфері інтересів як літературознавців, так і математиків. Цікавими є задачі кластеризації текстів за автором, жанром, форматов, епохою написання, чи емоційним забарвленням. Математичний інтерес полягаж в дослудженні алгоритмики процессу творчї роботи мозку. Також самостійну цінність мають і статистичні методи аналізу таких многомірних по своїм атрибутам об'єктам, як літературні тексти, написані професійними письменниками.

Найбільш популярною є задача ідентифікації авторства тексту. Для її вирішення досить часто використовуютья статичтичні методи, що грунтуються на припущенні, що письменник в своїх творах притримується певної поведінки письма, що формує певні інварінти, які підкреслють стилистичну характеристику творів автора. Такими інваріантами можуть бути кількісні ознаки долі голосних та приголосних у тексті, частота використання певних комбінацій слів, частота слів-маркерів. Практичне застосування це питання може знайти в пошуку плагіату, лінвістичних, історичних та кримінілістих дослідженнях.

У данній роботі проводиться порівняльний аналіз двох підходів для ідентифікації автора. Первший грунтується на порівнянні частот буквосполучень. Другий метод полягає у висуванні гіпотези про належність тексту до певного та доведення її з використанням p-статистики[?]. Також так як автори можуть писати в різних стилях, то використання

одного еталону для автора може бути не ефективним. Тому розглядається також підхід з кластеризації текстів автора, і подальшому знаходженні еталону для кожного кластеру. Припускається, що такий елалон відповідає окремому стилю стилю автора. Кластеризація виконувалась ієрархічним методом кластерізації з "відстаню поділу авторів"$\hat{\rho}$, в якості параметру.
В алфавіті з 26 символами існує $26^3 = 17526$ різним n-грам довжиною 3, що накладає робичть обчислення відстані від автора до автора досить ресурсоємким, тому в данній роботі пропонується метод по знаходжненню найбільш ефективних n-грам для ідентифікації автора.

Головною метою данної роботи є порівняння методів з точки зору вікористания ресурсів, швидкості обчислень, складності реалізації та точності ідентифікації. Також в данній роботі достіджєються частоти n-грамм, як стилістичні характеристику тексту.

Так як автори можуть писати в різних стилях, то використання одного еталону для автора може бути не

3 Умовні позначення та терміни

A - кількість авторів
m - кількість символів з алфавіту
Алфавіт надалі вважається фіксованим
n-грама - послідовність з n символів алфавіту
K_α - кількість текстів автора α
$N_{i,\alpha}$ - кількість букв в i-ому тексті автора α
n-ЩФР - щільність функції розподілу n-грамм тексту відповідного алфавіту
$f_{i,\alpha}(j)$ - n-ЩФР i-го тексту автора α, де аргумент j відповідає деякому n-граму, та змінюється від 1 до m^n
$f^k_{i,\alpha}(j)$ - n-ЩФР k-ої частини i-го тексту автора α, де аргумент j відповідає деякому n-граму, та змінюється від 1 до $a(n) = m^n$
$a(n) = m^n$ - кількість n-грам довжини n
$pstatistic(x, y)$ - міра однорідності двох вибірок, побудована

4 Методи

4.1 Щільність функції розподілу та її застосування для ідентифікації автора

Нехай ми маємо бібліотеку, що містить тексти A авторів. K_α - кількість наявних текстів автора. $N_{i,\alpha}$ - кількість букв в i-ому тексті автора α. Вважається, що довжина кожного з текстів достатня для проведення статистичного аналізу. Для кожного тексту знайдемо його представлення у вигляді частот n-грамм та позначимо $f_{i,\alpha}(j)$ відповідну n-ЩФР(n-щільність функції розподілу) i-го тексту автора α, де аргумент j відповідає деякому n-граму, та змінюється від 1 до $a(n) = m^n$: Для кожного автора визначимо його сереньозважену ЩФР[?]:

$$F_\alpha(j) = \frac{1}{N_\alpha} \sum_{i=1}^{K_\alpha} f_{i,a}(j) N_{i,\alpha} \tag{1}$$

$$N_\alpha = \sum_{i=1}^{K_\alpha} N_{i,a} \tag{2}$$

Ці n-ЩФР далі будуть грати роль авторських еталонів.
В ??, ?? знехтувано тим фактом, що кількість різних n-грам на $n-1$ менше кількості символів в тексті, так як $N_\alpha >> n$.

Введемо бібліотечну норму ρ_{ik}, як відстань між ЩФР текстів i та k:

$$\rho_{ik} = \|f_i - f_k\| = \sum_{j=1}^{a(n)} \mid f_i(j) - f_k(j) \mid \tag{3}$$

Для кожного автора α побудуэмо щільність функцій розподілу $g_\alpha^+(\rho)$ відхилень $\rho_{i_\alpha,\alpha}$ "його текстів а також розподіл $g_\alpha^-(\rho)$ відхилень текстів інщих авторів від його середньої n-ЩФР F_α. Позначимо $G_\alpha^\pm(\rho)$ відповідні фуцнкії розподілу. Мінімальне значення ρ при якому $G_\alpha^-(\rho) = 1$, позначимо ρ_α^+, а максимальне значення ρ при якому $G_\alpha^(\rho) = 0$, позначимо ρ_α^-.
Смисл введених позначень в тому що, всі ЩФР текстів автора α знаходятся на відстані не більш ніж ρ_α^+ від його середньої ЩФР F_α, та аналогічно всі ЩФР текстів інших авторів знаходятся на відстані не менш ρ_α^-. Величина $1 - G_\alpha^+(\rho_\alpha^-)$ - ймовірність помилково ідентифікувати α як автора(помилка другого роду), а величина $G_\alpha^-(\rho_\alpha^+)$ - ймовірність помилково ідентифікувати текст автора α, як текст написаний іншим автором(помилка першого роду). Позначимо $G^+(\rho)$ - розподіл відхидень текстів від відповідних еталонів, та $G^-(\rho)$ - розподіл відхидень текстів від "чужих"еталонів:

$$G^+(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^+(\rho)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{4}$$

$$G^-(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^-(\rho)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{5}$$

Назвемо вілстаню поділу авторів таке значення $\hat{\rho}$, для якої помилка ідентифікації автора тексту мінімальна:

$$\hat{\rho_\alpha} = argmin(1 - G_\alpha^+(\rho) + G_\alpha^-(\rho)) = argmax(G_\alpha^+(\rho) - G_\alpha^-(\rho)) \tag{6}$$

$$\hat{\rho} = argmin(1 - G^+(\rho) + G^-(\rho)) = argmax(G^+(\rho) - G^-(\rho)) \tag{7}$$

Величина $\hat{\rho}$ може слугувати верхнім рівнем кластерізації текстів.

Нехай тепер ми маэмо текст "0 автора якого треба ідентифікувати. Тоді автором текста вважається той автор α, для якого норма $\rho_\alpha = \|f_0 - F_\alpha\|$ різноці між ЩФР $f_0(j)$ текста "0"та середньої авторської ЩФР $F_\alpha(j)$ мінімальна:

$$\rho_\alpha = \|f_0 - F_\alpha\|, \quad \alpha^0 = \underset{\alpha}{argmin} \rho_\alpha^0 \tag{8}$$

Правило ?? застосовується тільки в тому випадку, якщо $\min_\alpha \rho_\alpha^0 \leq \hat{\rho}$. Якщо $\min_\alpha \rho_\alpha^0 > \hat{\rho}$ - приймається рішення, що в данній бібліотеці немає відповідного автора.

4.2 Метод з використанням з p-статистики

4.2.1 P-статистика

P-значення(P-value),P-статистика - вірогідність заданої статистичної моделі для якої, при умові що нульова гіпотеза є істиною, статистичні суми будуть однакові або матимуть більш екстремальні значення ніж для фактично отриманих результатів.
Нехай маємо дві вибірки $x = (x_1, x_2, ..., x_n)$ $y = (y_1, y_2, ..., y_m)$ з генеральних сукупностей X та Y відповідно. Для вибірки z відомо, що вона належить X чи Y - задача полягає у ідентифікації до якох саме сукупності належить z. Покладемо гіпотезу H про однорідність двох вибірок з генеральних сукупностей з функціями розподілу $F_x(u)$ та $F_y(u)$ відповідно, $x = (x_1, x_2, ..., x_n) \in X$тТа $y = (y_1, y_2, ..., y_m) \in Y$ - тестові вибірки у яких виконуєтсья $x_1 \leq ... \leq x_n$ та $y_1 \leq ... \leq y_m$. Нехай гіпотеза H полягає в $F_x(u) = F_y(u)$, тоді відповідно до $A(n)$ припущення Хілла:

$$p(y_k \in (x_i, x_j)) = \frac{j-i}{n+1}, i < j \tag{9}$$

Використовуючи вибірку $y = (y_1, y_2, ..., y_m)$, ми можемо частоту h_{ij} випадкової події $y_k \in (x_i, x_j))$ та розрахувати довірчий інтервал I_{ij} для ймовірності $p(y_k \in (x_i, x_j))$ з заданим рівнем значущості β. Позначимо L кількість інтервалів для яких виконується $fracj - in + 1 \in I_{ij}$. Тоді, визначимо міру одноріності вибірок x та y, як пропорцію інтервалів для яких вірно $fracj - in + 1 \in I_{ij}$ серед усіх інтервалів:

$$h_{xy} = \frac{2L}{n * (n-1)} \tag{10}$$

Так як, h_{xy} - частота випадкової події $fracj - in + 1 \in I_{ij}$ з ймовірністю $1 - \beta$, ми можемо побудувати довірчий інтервал I_{xy} для події $fracj - in + 1 \in I_{ij}$ з рівнем значущості β. Якщо $1 - \beta \in I$ тоді гіпотеза H підтверджена, інакше відхилена. Показник h_{xy} є мірою однорідності вибірок x та y. Змінивши x та y місцями та знайшовши частоту h_{yx} та довірчий інтервал I_{yx} ми можемо побудувати ще один тест для перевірки гіпотези H. Так як міра h_{xy} не симетрична, ми можемо побудувати симетричну міру однорідності:

$$h = pstatistic(x, y) = \frac{1}{2}(h_{xy} + h_{yx}) \tag{11}$$

4.2.2 Адаптація P-статистики для ідентифікації автора тексту. Вариант 1

Нехай маємо бібліотеку з попереднього методу з K_α творами автора α, α змінюється від 1 до A. Кожен текст в бібліотеці розібьємо на K частин, та для кожноъ частини знайдемо $f^k_{i,\alpha}(j)$ - n-ЩФР k-ої частини тексту. Позначимо $g_{i,\alpha(j)}(j)$ множину множину частот j-ого n-граму j-ого тексту автора α:

$$g_{i,\alpha}(j) = \{f^1_{i,\alpha}(j), f^2_{i,\alpha}(j), ..., f^K_{i,\alpha}(j)\} \tag{12}$$

Тоді введемо відстань між текстами a та b, як частина n-грам, для яких відхиляється гіпотеза про однорідність вибірок $g_a(j) = \{f^1_a(j), f^2_a(j), ..., f^K_a(j)\}$ та $g_b(j) = \{f^1_b(j), f^2_b(j), ..., f^K_b(j)\}$:

$$\|f^{(.)}_a - f^{(.)}_b\| = 1 - \frac{\sum_{j=1}^{a(n)} pstatistic(g_a(j), g_b(j)}{a(n)} \tag{13}$$

5 Обчислювальні експерименти

5.1 Інструменти розробки

Бібліотека з текстами була завантажена з сайту https://www.gutenberg.org/, яка містить більш 60000 книг. Программа для завантаження бібліотеки написана на мові програмування NodeJS v.12.13.1.
Алгоритм для тестів розроблений на мові програмування C++ 17.Для підвищення швидкості обчислень використовувалися паралельні потоки. Для візуалізації даних було використано jupyter python.
Тестування проводилось на комп'ютері із наступними специфікаціями:

- ОС: 64bit Windows 10 Pro
- Центральний процесор: Processor Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, 2801 Mhz, 4 Core(s), 8 Logical Processor(s)
- Оперативна пам'ять: 32.0 GB

5.2 Вхідні дані

Тестування проводилося на сукпності текстів 16 авторів: George Manville Fenn, Sir Walter Scott, R.M. Ballantyne, U.S. Copyright Office, Robert Louis Stevenson, Jules Verne, W.H.G. Kingston, George Sand, Anthony Trollope, Charles Dickens, G. A. Henty, Mór Jókai, Fergus Hume, Alexandre Dumas, E. Phillips Oppenheim, William Le Queux. Кожен автор має щонаймеш 50 книг довжиною 200000 символів в бібліотеці. Тести проводилися з використанням 5000, 10000, 20000, 50000, 100000, 200000 перших символів текстів. Тексти кожного автора було розділено на тренувальну та тестову вибірки по 25 текстів в кожній.
Так як обчислення відстанней для триграм ресурсоємке, були відібрані найбільш "впливові"триграми, тобто такі які найбільше відрізняють одоного автора від іншого. У якості міри "впливовості"була обрана наступна статистика:

$$v(j) = \frac{\sum_{\alpha=1}^{A} D(f_{\cdot,\alpha}(j))}{D(f_{\cdot,\cdot}(j))} \tag{14}$$

$$D(f_{\cdot,\cdot}(j)) = \frac{\sum_{\alpha=1}^{A} \sum_{i=1}^{K_\alpha} (f_{i,\alpha}(j) - M(f_{\cdot,\cdot}(j)))^2}{\sum_{\alpha=1}^{A} K_\alpha} \tag{15}$$

$$M(f_{\cdot,\cdot}(j)) = \frac{\sum_{\alpha=1}^{A} \sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{16}$$

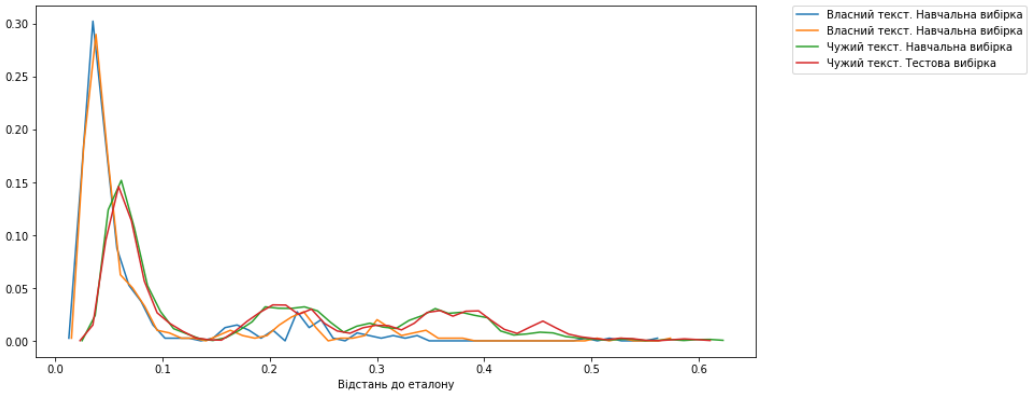

Рис. 1: Щільність розподілу відстаней. Монограми.

$$D(f_{\cdot\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha}(f_{i,\alpha}(j) - M(f_{\cdot\cdot}(j)))^2}{K_\alpha} \tag{17}$$

$$M(f_{\cdot\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{K_\alpha} \tag{18}$$

$D(\cdot)$ - це дисперсія, а $M(\cdot)$ - матсподівання. Отримуємо, що $v(j)$ - показує яку частину загальної варіації частот j-ого n-граму складає варіація цих частот локальна по авторам. Якщо $v(j)$ близьке до 0 - це означає, що при зміні автора тексту ця частота змінюється значно більше ніж при зміні текст на текст того ж автора. Серед усіх 17526 триграм було обрано 1802 з мірою "впливовості"менше 0.6.

5.3 Метод з використанням щільності функції розподілу

У таблицях ??, ??, ?? міститься інфирмація про середні відстані текстів до еталонів кожного автора для монограмів, біграмів, триграмів відповідно. А на рисунках ??, ??, ?? зображені щільності розподілів відстаней до еталонів n-грам. З отриманих таблиц бачимо, що Jules Verne, Charles Dickens та Alexandre Dumas - мають досить великі середні відстані до власного еталону, це пояснюється тим, що ці автори пишуть в досить різних стилях.

Автор	Середня відстань власних текстів до еталону		Середня відстань чужих текстів до еталону	
	Навчальна вибірка	Тестова вибірка	Навчальна вибірка	Тестова вибірка
George Manville Fenn	0.0338465	0.0340254	0.148186	0.160377
Sir Walter Scott	0.0284826	0.0663677	0.134049	0.147205
R.M. Ballantyne	0.0256525	0.0259007	0.132092	0.146816
U.S. Copyright Office	0.0474249	0.0434929	0.230926	0.241641
Robert Louis Stevenson	0.0660438	0.0500744	0.13504	0.150064
Jules Verne	0.194515	0.232865	0.214297	0.223655
W.H.G. Kingston	0.0310675	0.0344322	0.140399	0.153805
George Sand	0.0345426	0.0521552	0.345482	0.354845
Anthony Trollope	0.0312901	0.0393127	0.144474	0.157774
Charles Dickens	0.137723	0.201273	0.147524	0.156707
G. A. Henty	0.0299935	0.0315441	0.141307	0.155087
Mór Jókai	0.23157	0.221696	0.266655	0.265208
Fergus Hume	0.0346095	0.0355302	0.132482	0.146382
Alexandre Dumas	0.10102	0.152488	0.283127	0.291188
E. Phillips Oppenheim	0.0273646	0.0319321	0.136483	0.150508
William Le Queux	0.0312424	0.0289607	0.131785	0.145958

Табл. 1: Порівняння відстаней до еталонів. Монограми.

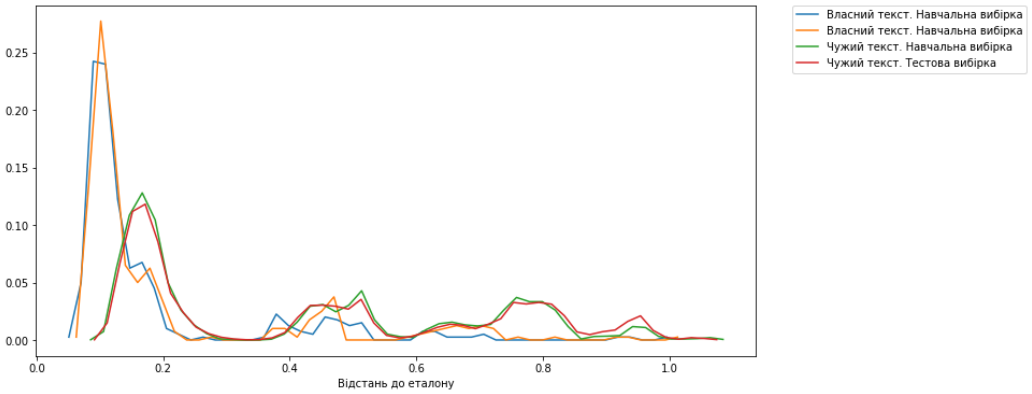

Рис. 2: Щільність розподілу відстаней. Монограми.

Автор	Середня відстань власних текстів до еталону		Середня відстань чужих текстів до еталону	
	Навчальна вибірка	Тестова вибірка	Навчальна вибірка	Тестова вибірка
George Manville Fenn	0.0994828	0.100713	0.337656	0.364495
Sir Walter Scott	0.0857056	0.160734	0.312519	0.337815
R.M. Ballantyne	0.0842319	0.0831463	0.303563	0.333081
U.S. Copyright Office	0.142319	0.131085	0.518372	0.536536
Robert Louis Stevenson	0.153465	0.126039	0.305945	0.336367
Jules Verne	0.434099	0.507919	0.467984	0.481571
W.H.G. Kingston	0.0986361	0.101119	0.319829	0.347821
George Sand	0.0953532	0.131873	0.720437	0.736496
Anthony Trollope	0.0978328	0.120733	0.335552	0.363841
Charles Dickens	0.289434	0.424478	0.322943	0.339882
G. A. Henty	0.0958608	0.0976739	0.32082	0.349905
Mór Jókai	0.473379	0.450451	0.564311	0.563759
Fergus Hume	0.103849	0.110262	0.308154	0.336897
Alexandre Dumas	0.228394	0.34131	0.612664	0.625097
E. Phillips Oppenheim	0.0855664	0.0948044	0.316374	0.344707
William Le Queux	0.0926607	0.0948494	0.306627	0.334875

Табл. 2: Порівняння відстаней до еталонів. Біграми.

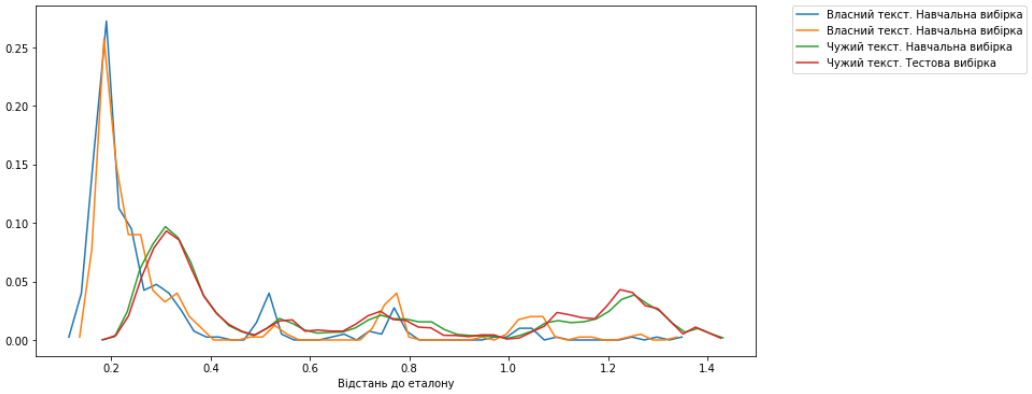

Рис. 3: Щільність розподілу відстаней. Триграми.

Автор	Середня відстань власних текстів до еталону		Середня відстань чужих текстів до еталону	
	Навчальна вибірка	Тестова вибірка	Навчальна вибірка	Тестова вибірка
George Manville Fenn	0.180418	0.18861	0.596302	0.623662
Sir Walter Scott	0.171463	0.283575	0.524401	0.549622
R.M. Ballantyne	0.172195	0.169814	0.51445	0.545968
U.S. Copyright Office	0.283822	0.276685	0.815953	0.833536
Robert Louis Stevenson	0.259684	0.228195	0.511306	0.545252
Jules Verne	0.655446	0.770669	0.789989	0.801207
W.H.G. Kingston	0.194357	0.204415	0.539513	0.568905
George Sand	0.162869	0.224804	1.15696	1.16934
Anthony Trollope	0.185929	0.224242	0.572796	0.603712
Charles Dickens	0.466389	0.668956	0.530902	0.546114
G. A. Henty	0.189261	0.19251	0.539592	0.571259
Mór Jókai	0.632626	0.749113	0.671995	0.671159
Fergus Hume	0.192996	0.206572	0.530133	0.561459
Alexandre Dumas	0.368591	0.561702	1.02425	1.02995
E. Phillips Oppenheim	0.170172	0.19042	0.546378	0.576854
William Le Queux	0.194836	0.198912	0.520049	0.550307

Табл. 3: Порівняння відстаней до еталонів. Біграми.

	Навчальна вибірка	Тестова вибірка
монограми	0.695	0.6875
біграми	0.7375	0.765
триграми	0.7725	0.785

Табл. 4: Точність для тексту в 200000 символів. Метод з використанням щільності функції розподілу

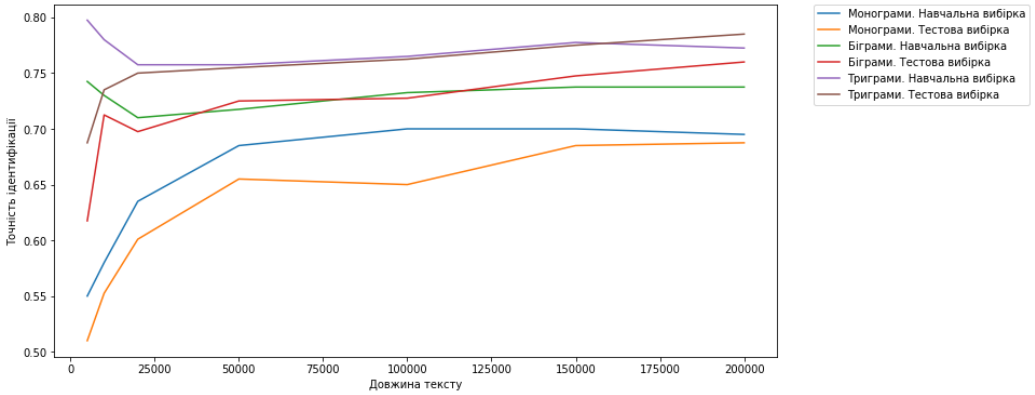

Рис. 4: Точність ідентифікації для різної довжини текстів.
На рисунку ?? зображена точність ідентифікації авторів для різної довжини n-грам.

На рисунку ?? зображена точність ідентифікації авторів для різної довжини текстів. Можна помітити, що для точність ідентифікації текстів за допомогою біграм та триграм на тестовій вибірці місцями перевищує точність ідентифікації на навчальній вибірці. Це пояснюється тим, що до тестової вибірки потрапили більш вузькі за своїм стилем тексти. 0.7725 0.785 0.7375 0.76 0.695 0.6875

Можна зробити такі висновки: - Починаючи з довжини тексту в 20000 сиволів зміна похибки точністі ідентифікації для біграм та трирам незначна(1-2%), а починаючи з 50000 сиволів зміна похибки для монограм також незначна.

5.4 Метод з використанням p-статистики

На рисунку ?? зображена залежність точності ідентифікації автора залежно від довжини тексту для K=20. Спостерігаємо, що загалом зі збіль-

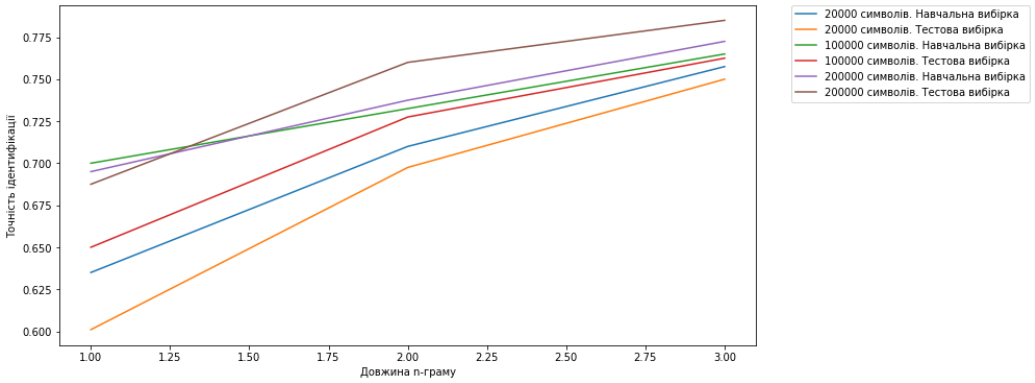

Рис. 5: Точність ідентифікації для різної довжини n-грам

	Навчальна вибірка	Тестова вибірка
монограми	0.7725	0.7075
біграми	0.83	0.8055
триграми	0.8325	0.815

Табл. 5: Точність для тексту в 200000 символів. P-статистика.

шенням довжини текстів, збільшується точність ідентифікації. Також між тестовими та навчальними вибірками є розрив в 5-10% точності.

5.5 Кластерізація

Так як автори можуть писати в різних стилях, кластерізація текстів автора з подальшим знаходженням еталону для кожного автора може бути досить ефективною. В якості методу кластерізації було обрано ієрархічний метод кластерізації, так як в метод можна передати максимальну відстань між кластерами. У якості параметру використовулася міра поділу авторів(Таблиці ??, ??, ?? для монограм, біграм та триграм відповідно). У таблицях ??,??, ?? знаходяться кількість кластерів по авторам для моногрма, біграм та триграм відповідно.

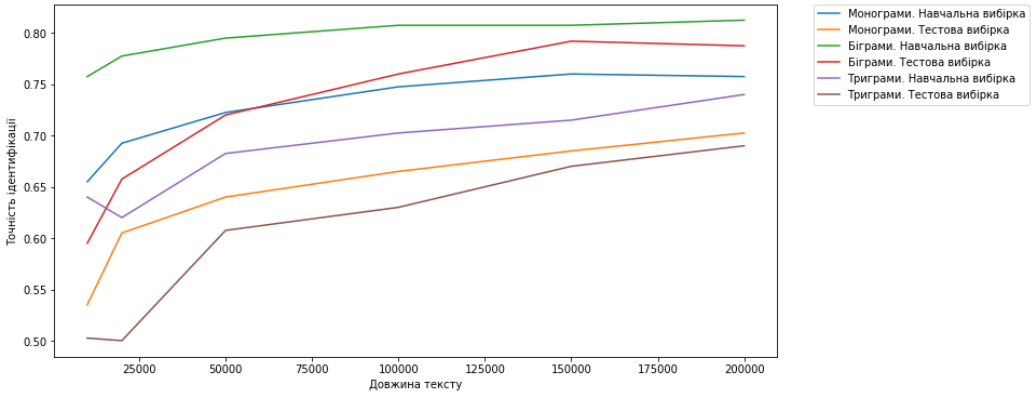

Рис. 6: Точність ідентифікації для різної довжини текстів. K=20

На рисунку ?? зображена залежність точності ідентифікації залежно від кількості частин на які розбивався текст. При збільшенні кількості частин тексту точність ідентифікації точність ідентифікації триграмами різко зменшується, що означає що, довжина тексту недостатня для отримання високох точності. При збільшенні кількості частин тексту точність ідентифікації для біграм та моногрaм зменшується незначно. Для Біграм найбільш ефективним є розбиття K=7, для моногрaм K=15, для триграм K=3. На рисунку ?? зображено залежність точності ідентицікації для відповідних розбиттів. Бачимо що для триграм навіть при розбитті на 3 частини 200000 символів в тексті недостатньо для досягнення околу статистичної границі.

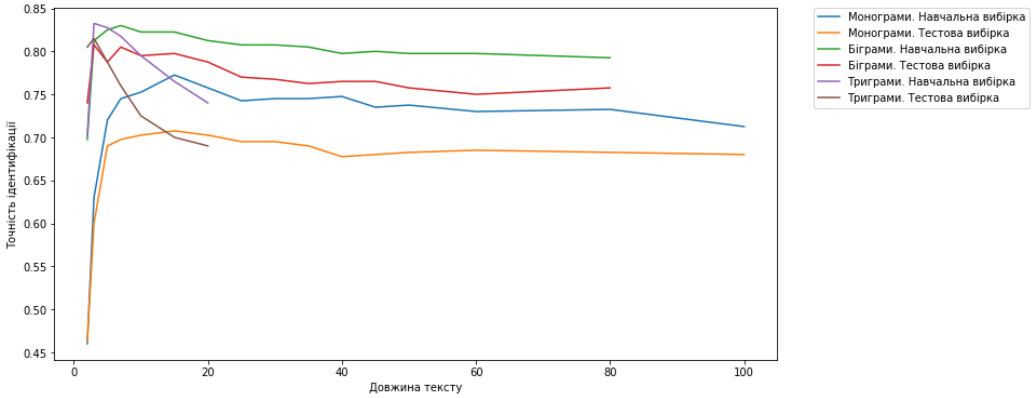

Рис. 7: Точність ідентифікації різних розбиттів. Добвжина текстів 200000

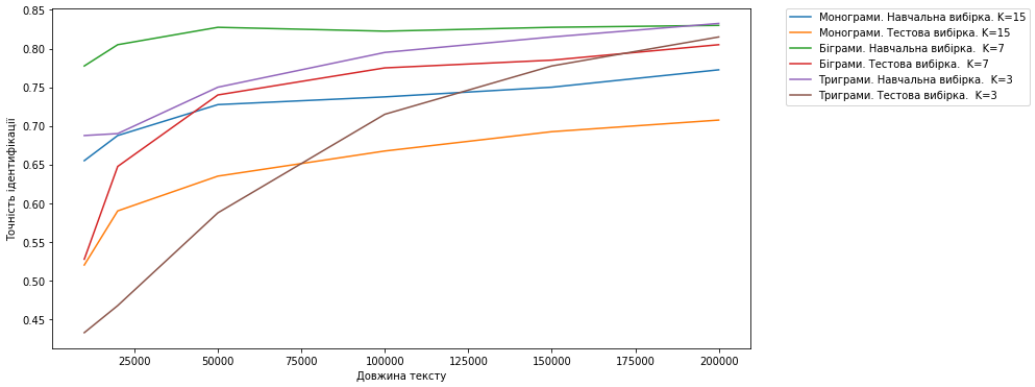

Рис. 8: Точність ідентифікації для різної довжини текстів. K=15 для монограм, K=7 для біграм, K=3 для триграм

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням r-статистики
George Manville Fenn	0.0463045	0.187912
Sir Walter Scott	0.0375501	0.158608
R.M. Ballantyne	0.0318236	0.143223
U.S. Copyright Office	0.0657424	0.176374
Robert Louis Stevenson	0.0543053	0.209341
Jules Verne	0.178568	0.365568
W.H.G. Kingston	0.0397744	0.157326
George Sand	0.0549845	0.249267
Anthony Trollope	0.0426615	0.172711
Charles Dickens	0.0823775	0.224908
G. A. Henty	0.04078	0.177473
Mór Jókai	0.247522	0.335897
Fergus Hume	0.0374213	0.170879
Alexandre Dumas	0.0929762	0.259524
E. Phillips Oppenheim	0.0411207	0.17619
William Le Queux	0.0387453	0.172344

Табл. 6: $\hat{\rho_\alpha}$. Монограми

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням p-статистики
George Manville Fenn	6	4
Sir Walter Scott	4	6
R.M. Ballantyne	5	3
U.S. Copyright Office	1	1
Robert Louis Stevenson	5	8
Jules Verne	4	4
W.H.G. Kingston	4	7
George Sand	1	1
Anthony Trollope	5	3
Charles Dickens	3	5
G. A. Henty	2	1
Mór Jókai	2	2
Fergus Hume	5	9
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	5	6

Табл. 7: Кількість кластерів. Триграми

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням p-статистики
George Manville Fenn	0.0463045	0.187912
Sir Walter Scott	0.0375501	0.158608
R.M. Ballantyne	0.0318236	0.143223
U.S. Copyright Office	0.0657424	0.176374
Robert Louis Stevenson	0.0543053	0.209341
Jules Verne	0.178568	0.365568
W.H.G. Kingston	0.0397744	0.157326
George Sand	0.0549845	0.249267
Anthony Trollope	0.0426615	0.172711
Charles Dickens	0.0823775	0.224908
G. A. Henty	0.04078	0.177473
Mór Jókai	0.247522	0.335897
Fergus Hume	0.0374213	0.170879
Alexandre Dumas	0.0929762	0.259524
E. Phillips Oppenheim	0.0411207	0.17619
William Le Queux	0.0387453	0.172344

Табл. 8: $\hat{\rho_\alpha}$. Біграми

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням p-статистики
George Manville Fenn	1	1
Sir Walter Scott	2	3
R.M. Ballantyne	3	3
U.S. Copyright Office	1	10
Robert Louis Stevenson	6	7
Jules Verne	4	2
W.H.G. Kingston	6	4
George Sand	1	1
Anthony Trollope	1	3
Charles Dickens	3	4
G. A. Henty	1	1
Mór Jókai	2	1
Fergus Hume	13	1
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	7	7

Табл. 9: Кількість кластерів. Триграми

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням p-статистики
George Manville Fenn	0.236771	0.269608
Sir Walter Scott	0.24953	0.266926
R.M. Ballantyne	0.200519	0.258232
U.S. Copyright Office	0.363961	0.275157
Robert Louis Stevenson	0.237309	0.264058
Jules Verne	0.724427	0.349427
W.H.G. Kingston	0.209656	0.268868
George Sand	0.221383	0.338605
Anthony Trollope	0.246643	0.27525
Charles Dickens	0.312767	0.285331
G. A. Henty	0.225755	0.274972
Mór Jókai	0.508067	0.321032
Fergus Hume	0.231127	0.264058
Alexandre Dumas	0.333833	0.348594
E. Phillips Oppenheim	0.218364	0.270533
William Le Queux	0.22903	0.26859

Табл. 10: $\hat{\rho_\alpha}$. Триграми

Автор	$\hat{\rho_\alpha}$ для метод з використанням щільності функції розподілу	$\hat{\rho_\alpha}$ для метод з використанням p-статистики
George Manville Fenn	1	1
Sir Walter Scott	2	21
R.M. Ballantyne	3	4
U.S. Copyright Office	1	9
Robert Louis Stevenson	6	7
Jules Verne	3	4
W.H.G. Kingston	6	5
George Sand	1	3
Anthony Trollope	1	1
Charles Dickens	3	5
G. A. Henty	6	1
Mór Jókai	2	7
Fergus Hume	2	12
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	8	5

Табл. 11: Кількість кластерів. Триграми

	Навчальна вибірка	Тестова вибірка
монограми	0.9125	0.75
біграми	0.98	0.8925
триграми	0.99	0.9175

Табл. 12: Точність методу з використанням щільності функції розподілу з кластерізацією

	Навчальна вибірка	Тестова вибірка
монограми	0.97	0.8125
біграми	0.91	0.8525
триграми	0.9875	0.805

Табл. 13: Метод з використанням p-статистики. K=15 для монограм, K=7 для біграм, K=3 для триграм

6 Аналіз результатів

6.1 Час виконання

Метод з використанням щільності функції розподілу виявився значино швидшим за метод з використанням p-статистики, за рахунок того, що для p-статистики треба розрахувати $\frac{K(K-1)}{2}$ довірчих інтервалів для кожного n-граму, у порівнянні з суммою різниць частот n-граму.

6.2 Точність ідентифікації

Для монограм, біграм та триграм метод з використанням p-статистики дає кращі результати в точності на $3-4\%$(для текстів довжиною більше 50000). Однак для невеликих за розміром текстів p-статистика дає гірші результати, ніж метод з використанням щільності функції розподілу.

З використанням кластерізації точність на тестовій виборці зросла приблизно на 5% для монограм, та на приблизно на 10% для біграм та триграм у випадку методу з використанням щільності функції розподілу, та найкращу точність в 91.75% було отримано для триграм. У випадку методу з використанням p-статистики точність для монограм та триграм зросла приблизно на 5%, а для триграм майже не змінилася, та найкращу точність в 85.25% було отримано для біграм.
З отриманих результатів можна зробити висновки: - у випадку кластерізації метод з використанням щільності функції розподілу дає кращі результати. Тому саме розподіл частот деякого n-граму не дає більше інформації о стилю, в якому пише автор, ніж середнє цього розподілу. - В данному контексті, виходячи з попереднього пункту, P-статистику, можна вважати деякою емпіричною версією аналогією довірчого інтервалу, так як без кластерізації p-статистика давала кращі результати, а з кластерізацією - гірші. - розподіл n-грамів дійсно змінюється зі зміною стилю, однак, для отримання більших точностей, необхідно шукати інші маркери

7 Висновки

У результаті дослідження було реалізовано 2 методи ідентифікації невідомого автора твору, який належить до бібліотеки відомих авторів, також реалізовано метод кластрізації текстів та проведено тестування методів з та без кластерізацією. Також було запропоновано метод критерій для відбору n-грам, які б найкраще слугували маркером для ідентифікації автора. Для тестування використовувалися 800 текстів 16 авторів. В результаті було виявлено, що методу, що використовує щільність функції розподілу підходить для ідентифікації авторів творів як великих текстів(50000+ симолів) так и малих(10000+ символів). А метод, що використовує p-статистику підходить тільки для використання на великих за обсягов творах. З кластерізацією текстів були отримані значно кращі результати на тестовій вибірці для обох методів.

8 Література

[1] Борисов Л.А. та Орлов Ю.Н. та Осминин К.П. Идентификація автора текста за розподілом частот буквосполучень. 2013.

[2] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. 2017.

[3] Marcia Fissette. Author identification in short texts. 2010.

[4] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. 2007.

[5] Giacomo Inches, Morgan Harvey, and Fabio Crestani. Finding participants in a chat: Authorship attribution for conversational documents. 2013.

[6] Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, and Walter Daelemans. Overview of the author identification task at pan-2018. 2018.

[7] D.A. Klyushin, S.I. Lyashko, and S.S. Zub. Author identification in short text.

[8] Granichin O., Kizhaeva N., Shalymov D., and Volkovich Z. Writing style determination using the knn text model. 2015.

[9] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 19, 1948.

[10] Stamatatos. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 2009.

[11] Liuyu Zhou. News authorship identification with deep learning. 2016.