Національний технічний університет України «КПІ ім. Ігоря Сікорського»

Факультет Інформатики та Обчислювальної Техніки

Кафедра Автоматизованих Систем Обробки Інформації та Управління

Лабораторна робота №3

з дисципліни «Комп'ютерна лінгвістика»

Тема:

# «**ОЦІНКА БЛИЗЬКОСТІ ДОКУМЕНТІВ МЕТОДОМ ШИНГЛІВ**»

Виконав:

студент гр. ІС-61

Іваницький Олег

Викладач: Фіногенов О.Д

Київ – 2019

## Мета роботи: вивчення методів визначення подібних (близьких за текстом) документів, зокрема для знаходження плагіату.

\*У ході лабораторної роботи було досліджено лише вхідні файли на збіжність між собою.У погодженні з Вами аналіз уривку з твору з іншими творами не проводився.

**Варіант завдання:**

## Для виконання лабораторної роботи були вибрані письменники: Микола Гоголь («Тарас Бульба», «Вій», «Ніч проти різдва») та Іван Нечуй-Левицький(«Гетьман Іван Виговський», «Кайдашева сім’я», «Два брати»).

|  |  |  |  |
| --- | --- | --- | --- |
| Автор  (Твір) | Обсяг фрагменту  (слів, укр.) | Обсяг фрагменту перекладу  (слів, рос.) | Обсяг  Автоматич-ного перекладу  (слів, рос.) |
| Гоголь «Тарас Бульба» | 3502 | 3364 | 3567 |
| Гоголь «Вій» | 3305 | 3273 | 3347 |
| Гоголь «Ніч проти різдва» | 2846 | 2752 | 2860 |
| Нечуй-Левицький «Гетьман Іван Виговський» | 3633 | 3636 | 3638 |
| Нечуй-Левицький «Кайдашева сім’я» | 2996 | 3115 | 3122 |
| Нечуй-Левицький «Два брати» | 1996 | 2044 | 2026 |

**Короткі теоретичні відомості:**

Мико́ла Васи́льович Го́голь (рос. Николай Васильевич Гоголь; прізвище при народженні Яно́вський, з 1821 року — Го́голь-Яно́вський; 20 березня [1 квітня] 1809 року, Сорочинці, нині Великі Сорочинці, Миргородський район, Полтавська область — 21 лютого [4 березня] 1852 року, Москва) — український письменник з роду Гоголів-Яновських, класик російської літератури, вважається найвизначнішим з представників її «української школи». Був одним із засновників реалізму та заклав соціально-критичний («гоголівський») напрямок в російськомовній літературі.

Іва́н Нечу́й-Леви́цький (справжнє прізвище — Леви́цький, 13 [25] листопада 1838, Стеблів, Богуславський повіт, Київська губернія, Російська імперія — 2 квітня 1918, Київ, Київська губернія, УНР) — український письменник, етнограф, фольклорист, педагог.

**Опис програми:**

Програма зчитує тексти творів з файлів у форматі .txt. Відбувається очищення (канонізація даних), що включає приведення до нижнього регістру, видалення стоп-слів\*, надлишкових пробілів та знаків пунктуації. Після цього з кожного документу формуються набори шинглів довжиною у три слова. Кожен шингл хешується 32-бітною хеш-функцією. Далі будується сигнатура документу: для 84 різних хеш-функцій знаходяться хеш-значення усіх шинглів, після чого серед них знаходиться мінімальне, яке і додається до сигнатури. В результаті кожен файл представлений списком з 84 значень хеш-функцій. Для порівняння файлів між собою порівнюються їхні сигнатури: кількість однакових хеш-значень ділиться на загальну їх кількість. В кінці виводяться результати схожості файлів між собою.

\*Список стоп-слів, що використовуються у програмі (отримані з модулю stopwords бібліотеки nltk): и,в,во,не,что,он,на,я,с,со,как,а,то,все, она,так,его,но,да,ты,к,у,же,вы,за,бы,по,только,ее,мне,было,вот, от, меня,еще,нет,о,из,ему,теперь,когда,даже,ну,вдруг,ли,если,уже, или,ни,быть,был,него,до,вас,нибудь,опять,уж,вам,ведь,там,потом, себя,ничего,ей,может,они,тут,где,есть,надо,ней,для,мы,тебя,их, чем,была,сам,чтоб,без,будто,чего,раз,тоже,себе,под,будет,ж, тогда,кто,этот,того,потому,этого,какой,совсем,ним,здесь,этом, один,почти,мой,тем,чтобы,нее,сейчас,были,куда,зачем,всех, никогда,можно,при,наконец,два,об,другой,хоть,после,над,больше, тот,через,эти,нас,про,всего,них,какая,много,разве,три,эту,моя, впрочем,хорошо,свою,этой,перед,иногда,лучше,чуть,том,нельзя, такой,им,более,всегда,конечно,всю,между,это,з,й,що,та,із,але, цей,коли,як,чого,хоча,нам,якось,чи,це,від,їх,інших,ти,він,вона, воно,ми,ви,вони,якого,яких,яким,є,чому,чим,де,десь,собі,свій.

**Результати оцінки схожості документів:**

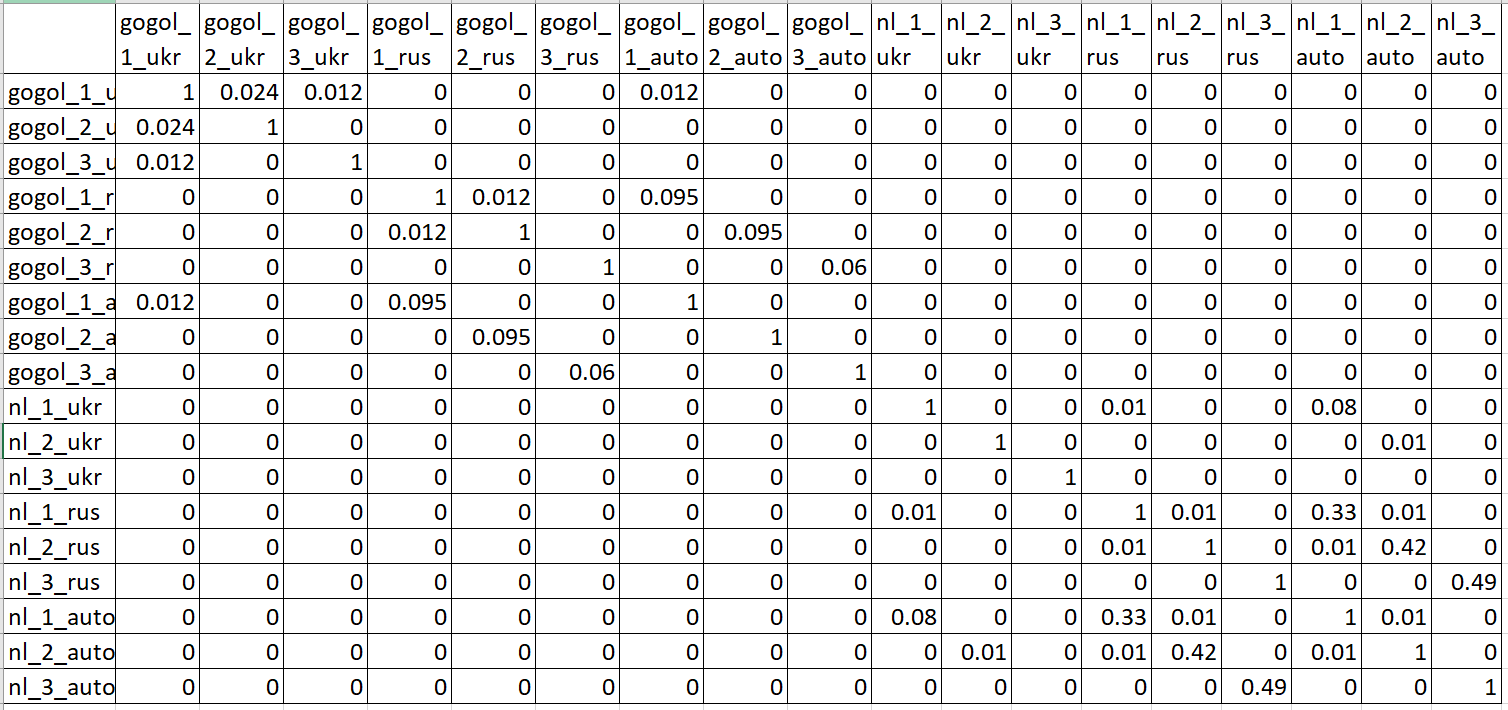


Рис.1 Результати аналізу текстів на схожість методом шинглів

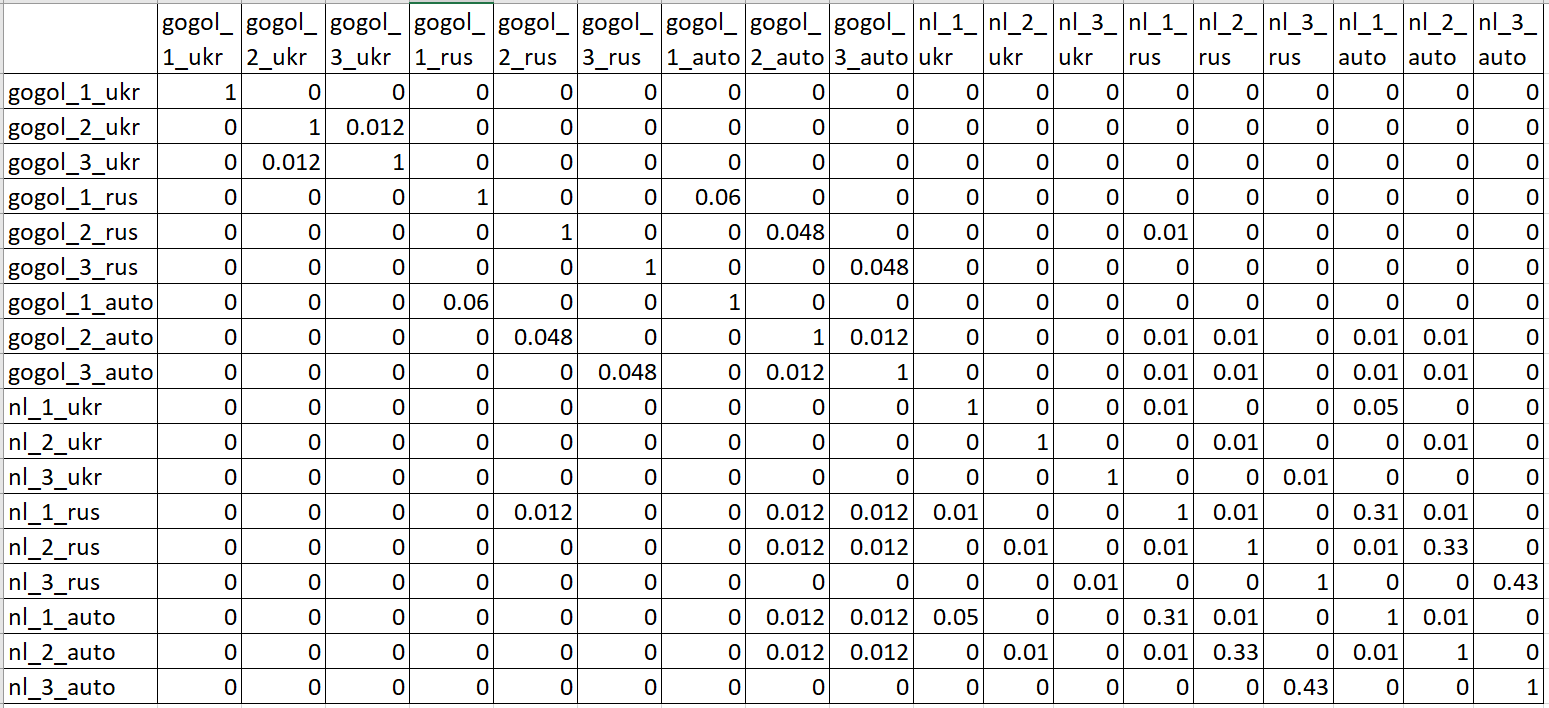


Рис.2 Результати аналізу текстів на схожість модифікованим методом шинглів (сортування слів у алфавітному порядку)

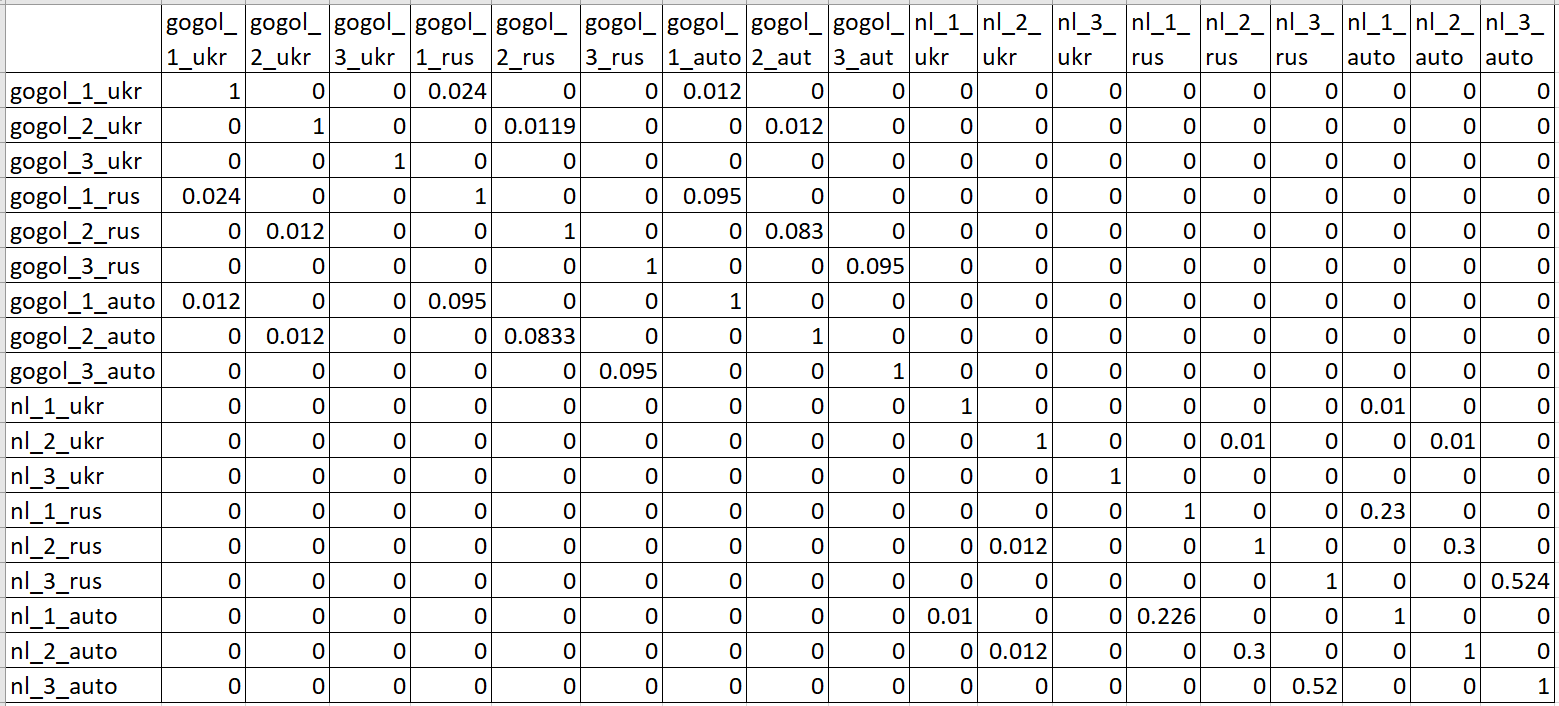


Рис.3 Результати аналізу текстів на схожість метдом шинглів після видалення стоп слів

**Висновки:** проаналізувавши усі 18 файлів, можна побачити досить невеликий відсоток схожості, що може свідчити, про, наприклад, не надто якісну роботу Google перекладача, що, насправді, досить логічно, адже переклад літературинх творів потребує значних зусиль та додаткових знань спеціаліста. Більші результати схожості можна спостерігати лише у деяких творів у випадку сортування слів у кожному шинглі за алфавітом, проте навіть цей відсоток не дозволяє зробити висновок про ідентичність документів.