МЕТОДИ ПОПЕРЕДНЬОГО АНАЛІЗУ ТЕКСТОВИХ ДАНИХ НА МОВІ PYTHON

Студент групи ДА-81мп
Будьонний Данило Юрійович
Керівник: проф., д.т.н. Рогоза В.С.
Консультант: асистент Яременко В.С.

ОБ'ЄКТ ДОСЛІДЖЕНЬ

- Методи попереднього аналізу текстових даних

ПРЕДМЕТ ДОСЛІДЖЕНЬ

- Методи для вирішення задач багатокласової класифікація текстових даних, що надходять у режимі реального часу

МЕТА РОБОТИ

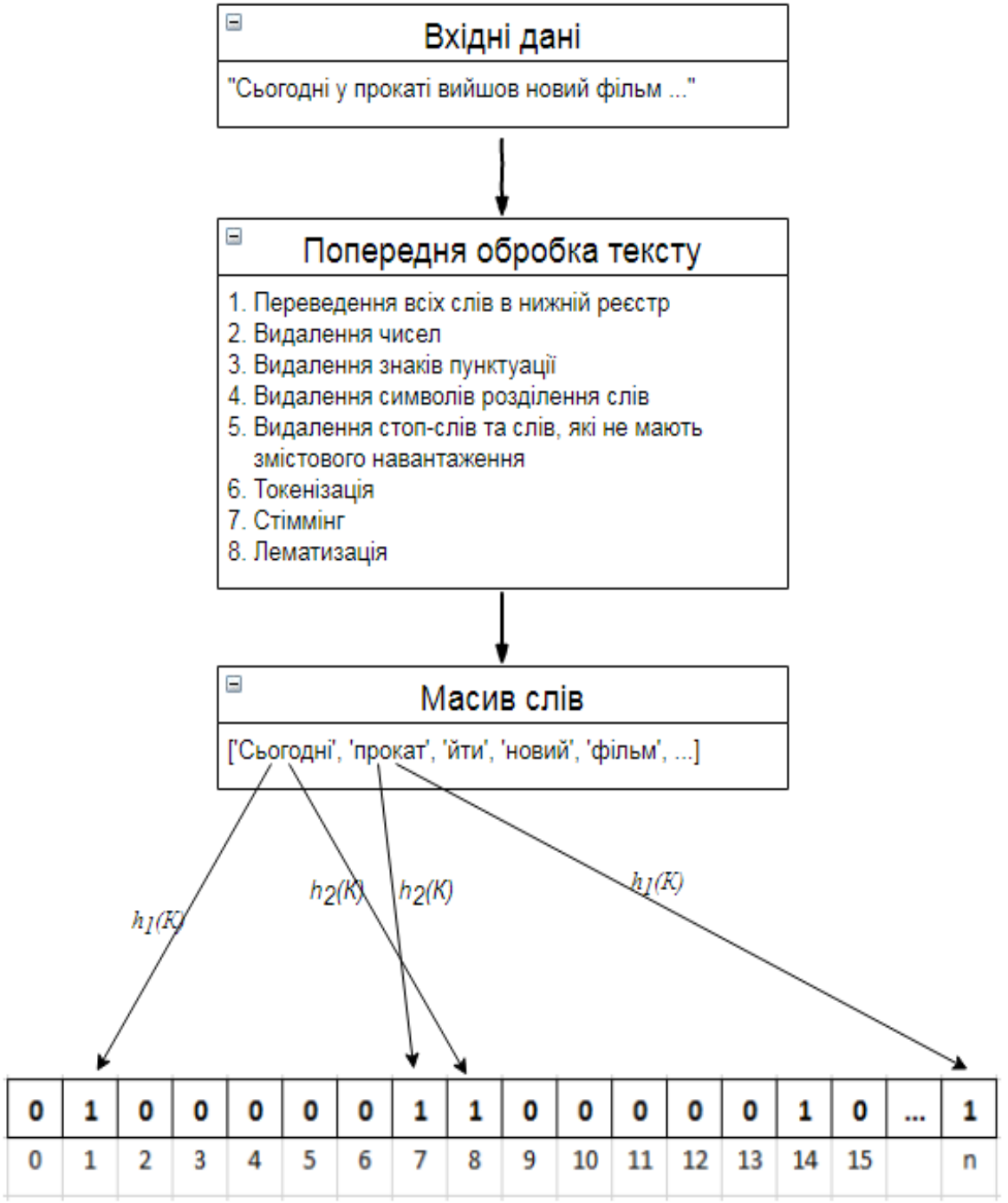
- Розробка моделі класифікації текстових даних у реальному часі на основі фільра Блума
- Реалізація моделі
- Аналіз результатів роботи системи
- Порівняння з існуючими рішеннями

СКЛАДОВІ ЧАСТИНИ ФІЛЬТРА БЛУМА

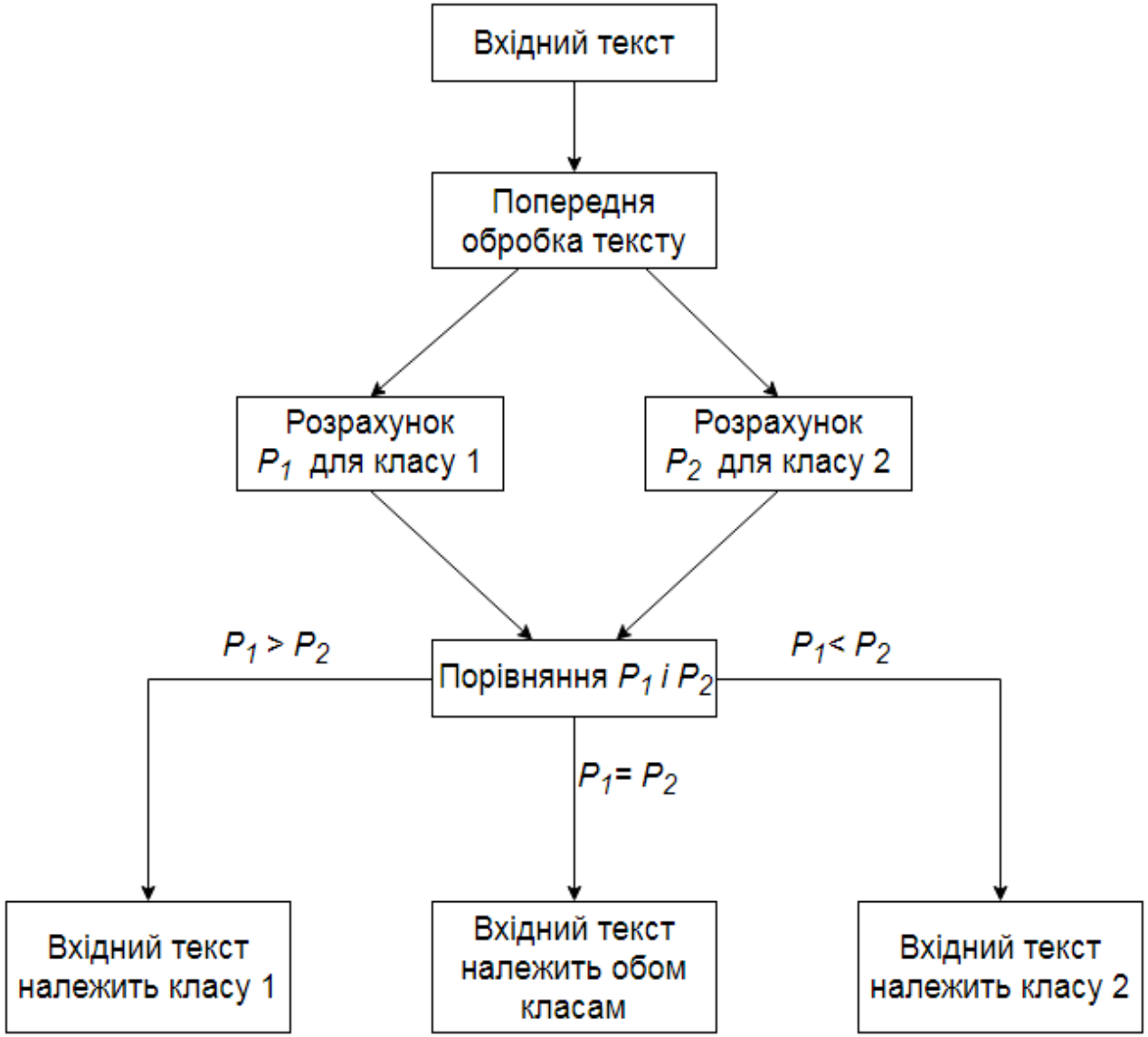
- Масив *n* біт, спочатку рівних *0*.
- Набір хеш-функцій h_1, h_2, …, h_k, кожна з яких відображає значення «ключа» в n комірок, відовідно до *n* бітів масиву.
- Множина *S*, що містить *m* ключів.

Значення	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Номер біта	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

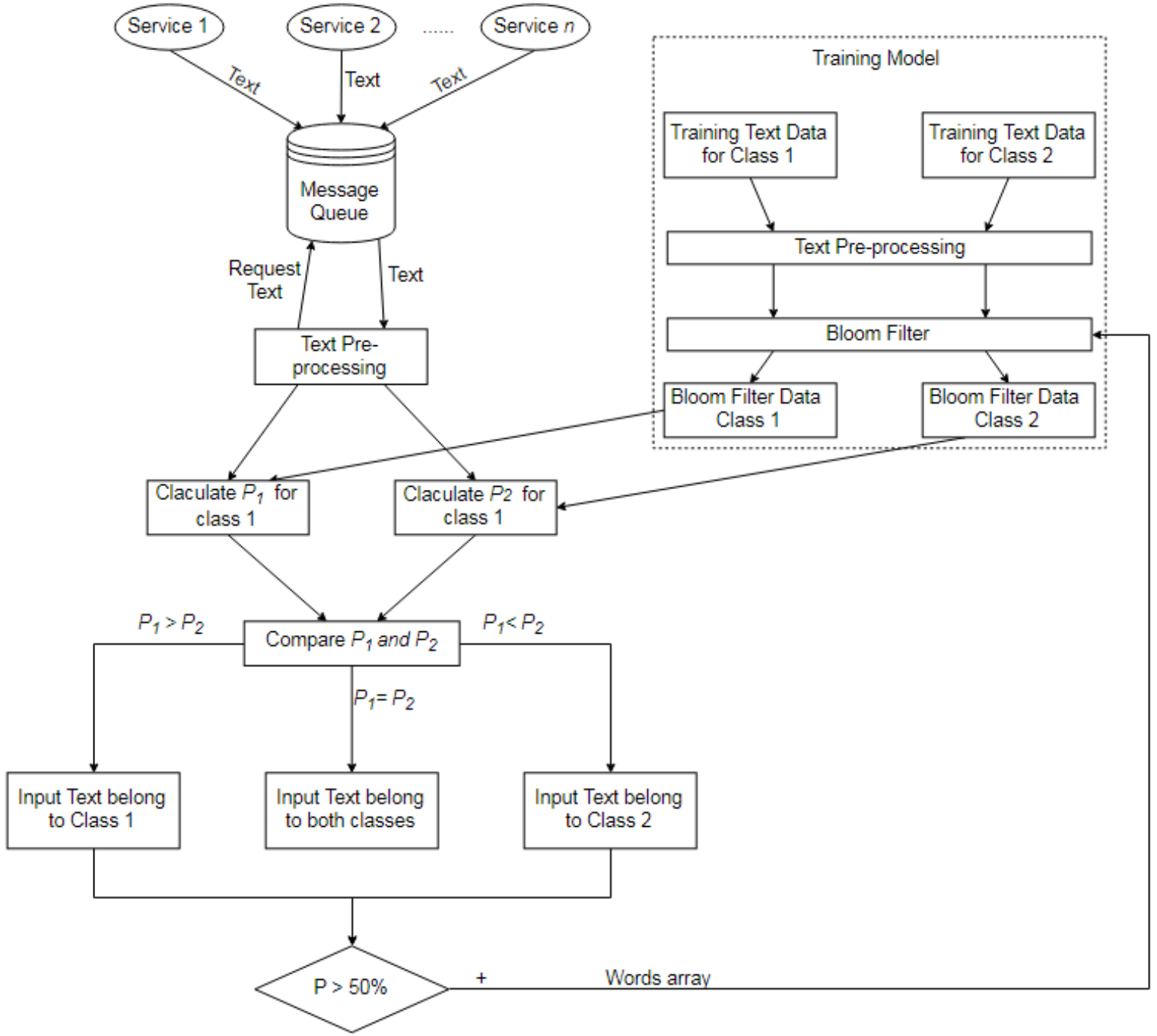
НАВЧАННЯ МОДЕЛІ



РОБОТА МОДЕЛІ КЛАСИФІКАЦІЇ



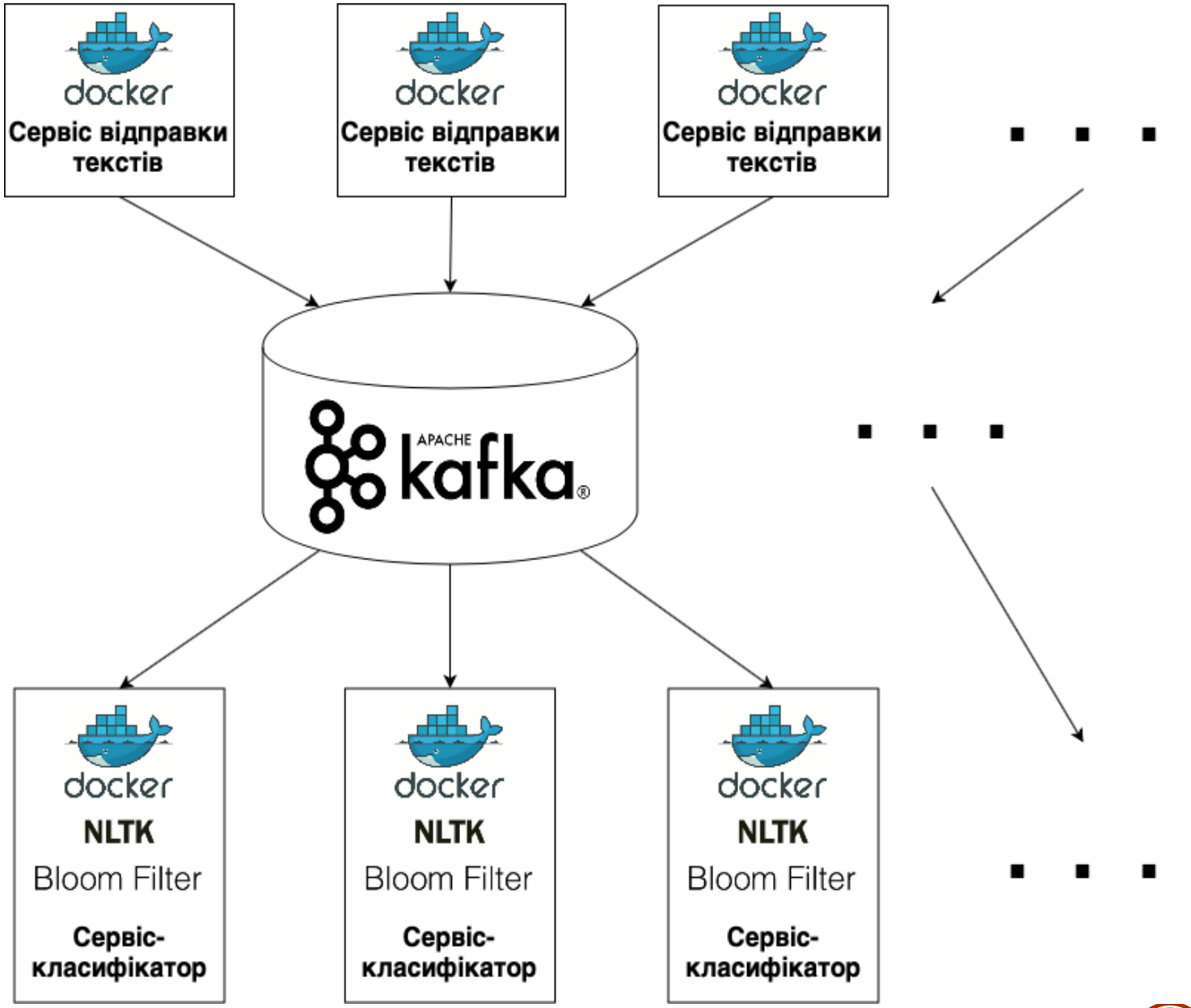
РОБОТА СИСТЕМИ



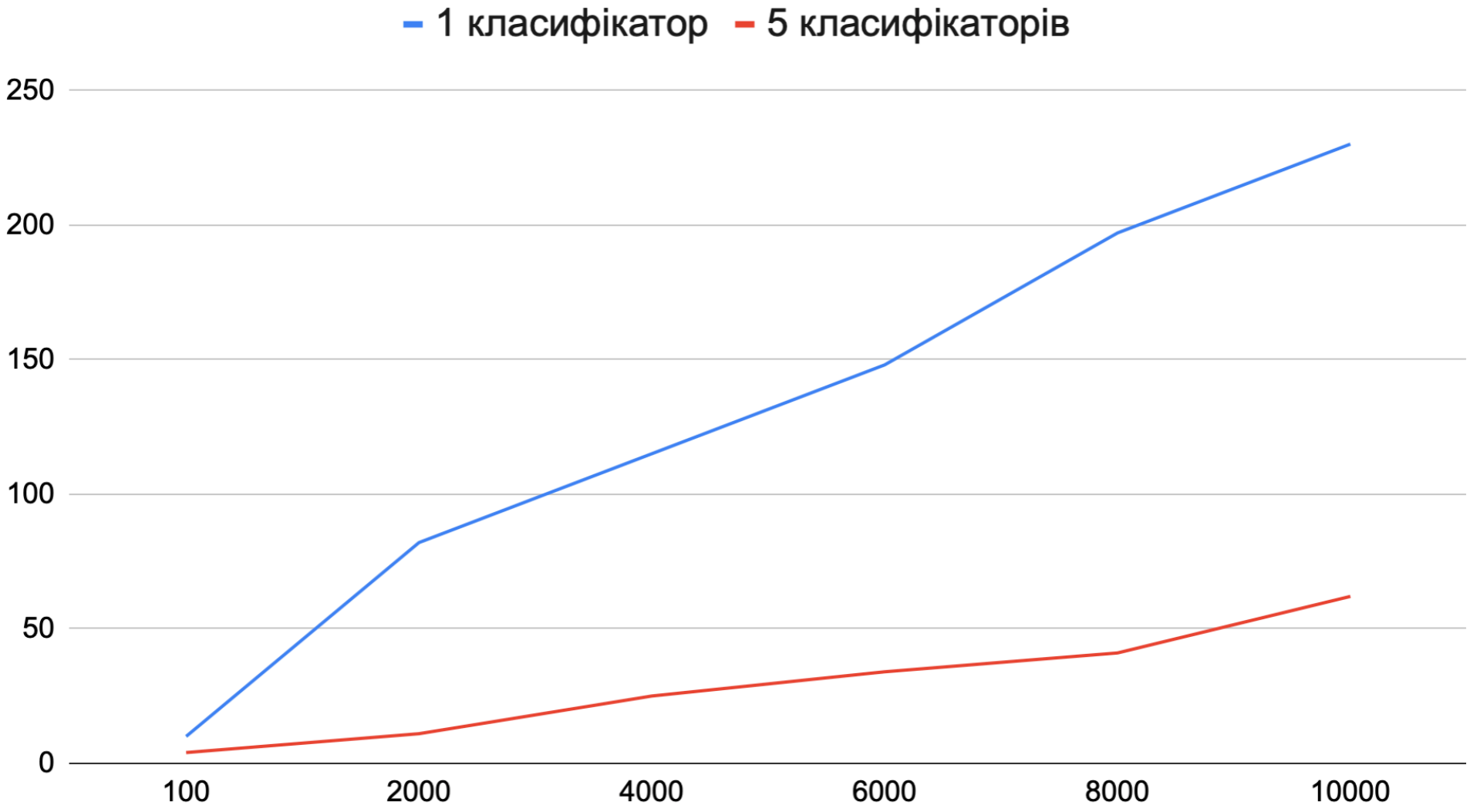
ВИМОГИ ДО АРХІТЕКТУРИ СИСТЕМИ

- обробка текстів у режимі реального часу
- жодне повідомлення не повинне бути втрачене
- відмовостійкість
- масштабування

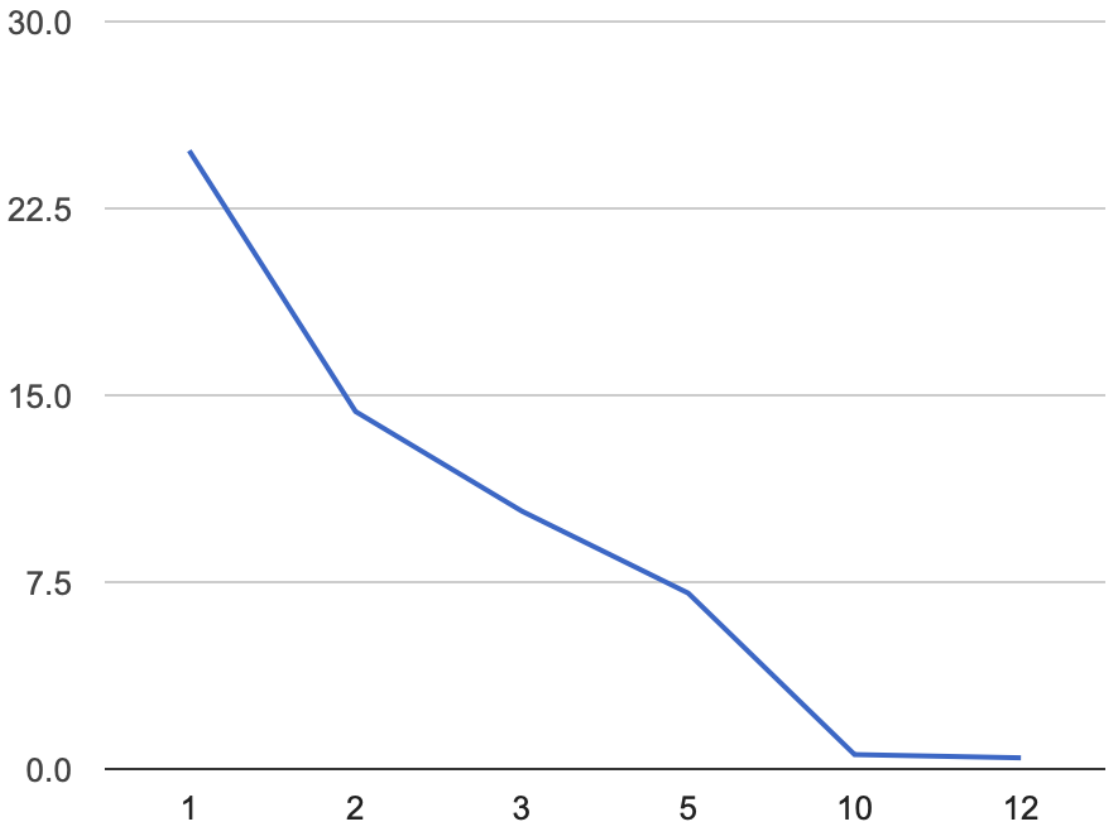
АРХІТЕКТУРА СИСТЕМИ



ГРАФІК ЗАЛЕЖНОСТІ ЗАТРИМКИ (У СЕКУНДАХ) ВІД КІЛЬКОСТІ ПОВІДОМЛЕНЬ



ЗАЛЕЖНІСТЬ КІЛЬКОСТІ КЛАСИФІКАТОРІВ ВІД ЗРОСТАННЯ ЧЕРГИ ДЛЯ ОБРОБКИ 1000 ПОВІДОМЛЕНЬ



ПОРІВНЯННЯ З КЛАСИФІКАТОРОМ MONKEYLEARN

	Час на обробку повідомлення	Похибка
Класифікатор з фільтром Блума	0,14 сек	0%
Класифікатор MonkeyLearn	1,4 сек	6.6%

ПЕРЕВАГИ

- Швидкість навчання моделі (~1с)
- Підхід з використанням фільтра Блума використовує невелику кількість пам'яті (до 0.1 Кб для кожного класу класифікації) в залежності від розміру хеш-таблиці
- Кросплатформенність та простота маштабування завдяки контейнеризації
- У 10 разів швидше від класифікатора MonkeyLearn для заданого набору текстових даних

НЕДОЛІКИ

- Перенавчання моделі
- Класифікації можуть бути хибними для класів, які мають багато спільних слів

МАЙБУТНІ НАПРЯМИ РОБОТИ ТА ДОСЛІДЖЕНЬ

- Створити реалізацію збереження навченої моделі
- Впровадження можливості спільного доступу всіх сервісів класифікації до однієї моделі
- Реалізація можливості автоматичного створення та видалення компонентів системи при різних навантаженнях
- Вирішення проблеми динамічного розширення виділеної пам’яті під масив бітів

ВИСНОВКИ

- Реалізована модель для класифікації на основі фільра Блума
- Запропонована система для класифікації повідомлень у реальному часі
- Проведено аналіз системи
- Порівняно модель класифікації з класифікатором від MonkeyLearn

ПУБЛІКАЦІЇ

Автори: Будьонний Д. Ю., Яременко В. С.

Назва: ПІДХІД ДО ВИКОРИСТАННЯ ФІЛЬТРА БЛУМА ДЛЯ БАГАТОКЛАСОВОЇ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДАНИХ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ

Науковий журнал: Комп'ютерно-інтегровані технології: освіта, наука, виробництво. 2019. №36

Режим доступу до ресурсу: http://ki.lutsk-ntu.com.ua/node/146/section/24

Дякую за увагу