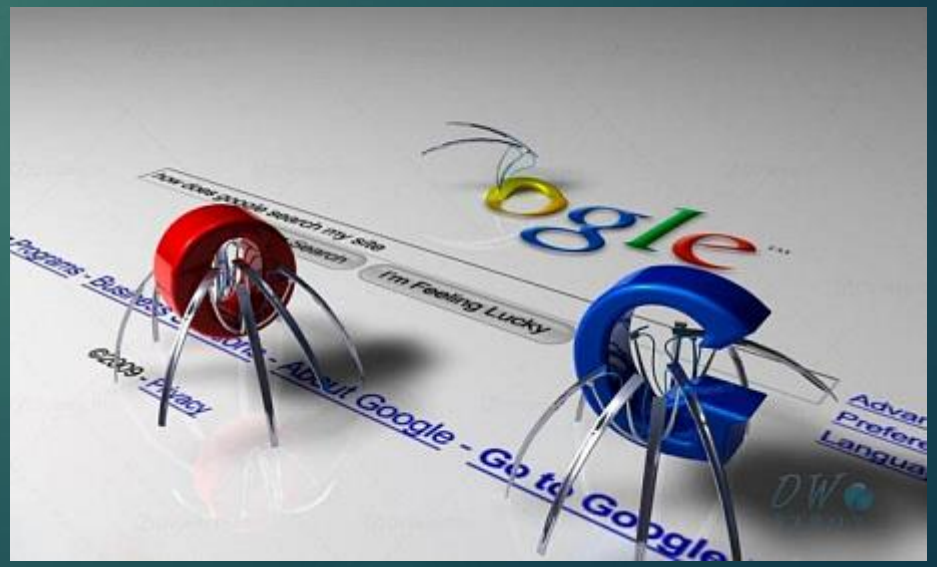
Виявлення веб-сканера(англ. crawler bot, spider)

Що таке веб сканер(веб робот)?

Це програма, яка використовується для автоматичного обходу веб-сторінок та отримання даних із них. Створювалася така технологія, як частина пошукової системи.

1993- перший веб-робот World Wide Web Wanderer.



Типи веб сканерів

- В основному веб сканери можна поділити за впливом на роботу сервера, де розміщена веб-сторінка:
- 1) Веб сканер, як частина пошукової системи, індексування сайтів
- 2) Веб сканер, який використовується для збирання інформації для подальшого використання, у своїх цілях.

Детальніше про веб сканери 2 типу

Приклади використання:

- порівняння цін та товарів, що продаються різними електронними комерційними сайтами
- Збирання поштових адрес
- Збирання статистик команд, гравців і т.п. із різних спортивних веб-сторінок
- І т.д.

Чому потрібно виявляти веб-сканери?

Веб сканери потрібно виявляти, щоб уникати такі ситуації:

- 1) Негативний вплив на ефективність роботи серверів
- 2) Ускладнення аналізу трафіку на веб-сторінці
- 3) Шахрайство з рекламою

Методи виявлення веб сканера

- **Метод опорних векторів**
- **Дерева рішень**
- **Нейронні мережі**

Популярні метрики, які використовуються для виявлення

- Відсоток відповідей 2xx
- Відсоток відповідей 3xx
- Відсоток запитів сторінок
- Відсоток нічних запитів
- Середній час між двома послідовними запитами
- І т.д.

Розробка системи. Датасет

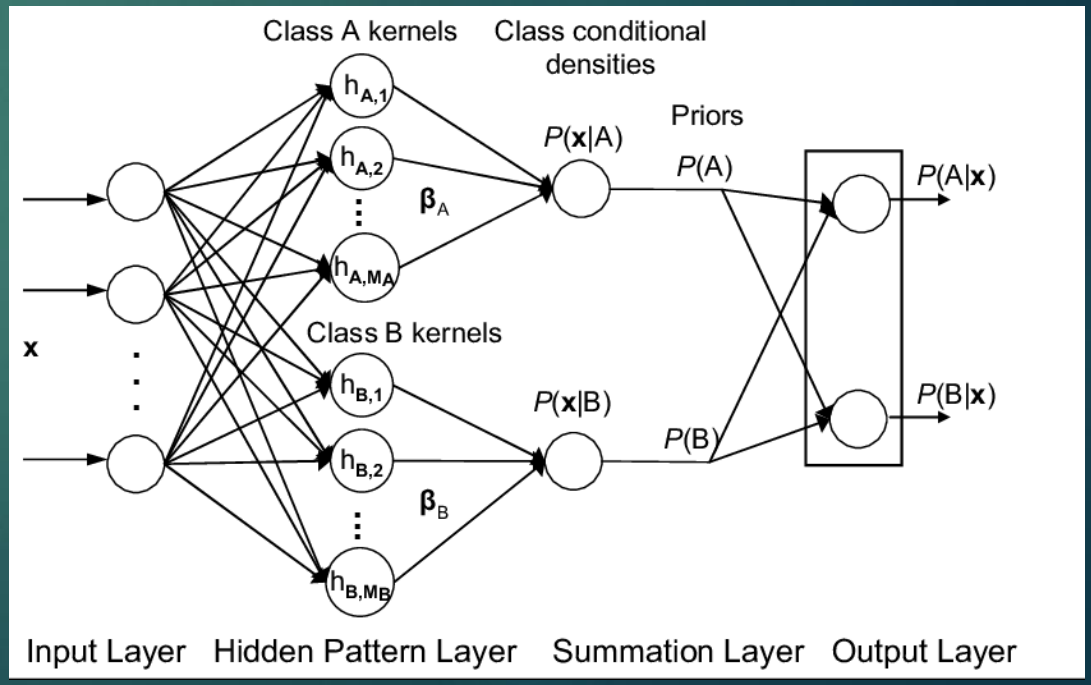
	ip	200_proc	300_proc	page_proc	400_proc	night_proc	number_requests	request_image	type_user
0	239.251.203.224	46.947	35.969	14.783	2.889	12.041	188	3.423	human
1	171.34.22.137	33.287	39.908	29.128	2.276	18.383	135	7.203	human
2	168.160.249.78	20.108	31.654	19.284	5.370	39.179	85	5.675	human
3	26.16.149.35	42.247	40.343	27.748	2.496	3.374	180	22.798	human
4	55.109.152.144	48.710	28.189	27.885	6.286	34.247	135	10.978	human

- Структура датасету:

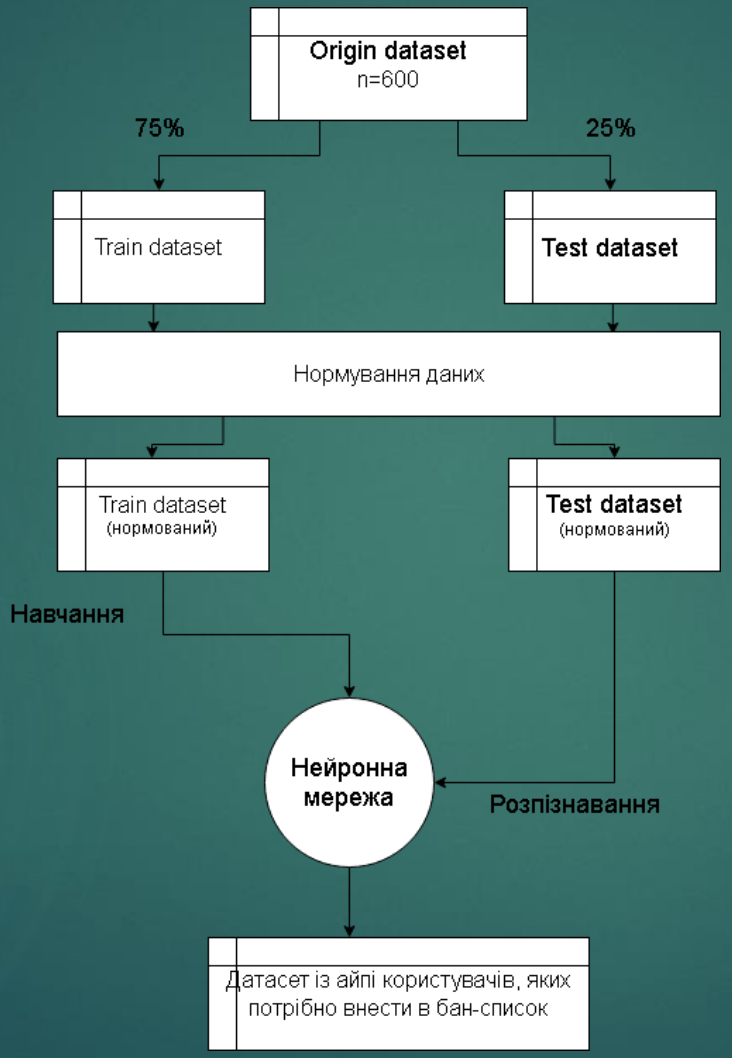
1. Ip користувача
2. Відсоток відповідей сервера 2xx кодами
3. Відсоток відповідей сервера 3xx кодами
4. Відсоток відповідей сервера 4xx кодами
5. Відсоток запитів html, php сторінок
6. Відсоток нічних запитів
7. Кількість запитів за одну сесію
8. Відсоток запитів на зображення
9. Тип користувача

Розробка системи. Метод

- Методом для розв'язання нашої задачі, було обрано, ймовірісні нейронні мережі. Завдяки тому, що у нашому випадку вибірка не велика, даний тип мереж буде достатньо швидким



Розробка системи. Експеримент



Розробка системи. Результат експерименту.

- У початковому датасеті знаходилося 600 прикладів. Після поділу на навчальні та тестові датасети, у першому 450, у другому 150. Після нормування даних було проведено навчання нейронної мережі. Потім було проведено розпізнавання тестової вибірки. Результати такі: із 150 тестових прикладів, правильно було класифіковано 142 випадки. Система вірно розпізнала всіх ботів, але 8 людей вона класифікувала як ботів. Тому можна говорити, що приблизна похибка нашої системи 5.333 відсотка.

Дякую за увагу

Література

- https://www.researchgate.net/publication/241184288_Real-time_web_crawler_detection