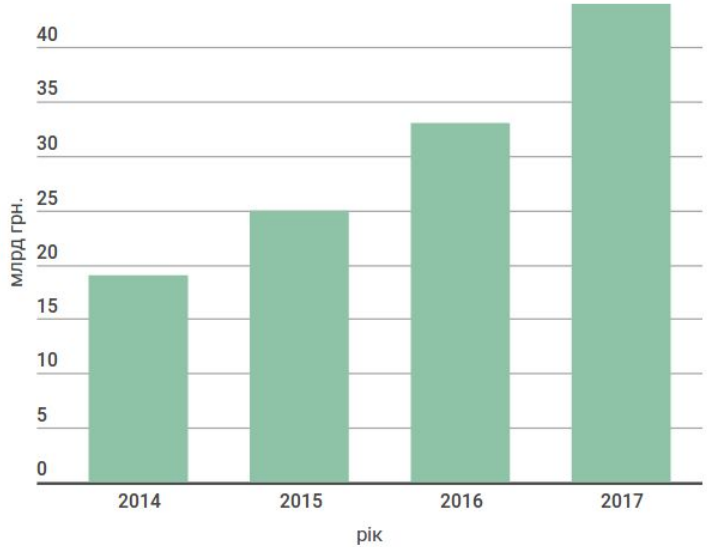
Міністерство освіти і науки України
Національний тейхнічний університет України «КПІ ім. Сікорського»
Інститут прикладного та системного аналізу
Кафедра математичних методів системного аналізу

Генерація пошукових запитів на основі ключових слів для автоматизації роботи фахівців з оптимізації сайтів

Виконала студентка групи КА-33
Німець А.В.
Науковий керівник:
К.т.н, доцент Тимошенко Ю. О.

Актуальність роботи

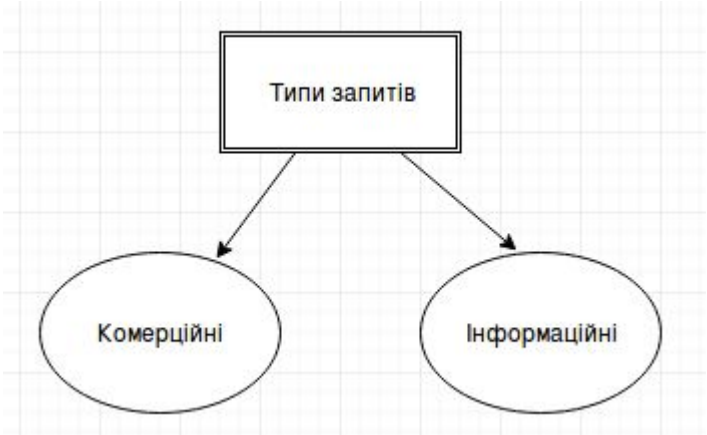
- Інтернет - є ефективним маркетинговим інструментом та місцем продажу.
- 95% користувачів звертає увагу на перші 10 сайтів у пошуковій видачі.
- SEO-спеціалісти займаються оптимізацією сайтів з метою просування сторінок у пошуковій видачі.



Об'єм українського ринку інтернет-торгівлі

Предметна область

- SEO - спеціалісти часто займаються формуванням **семантичного ядра**.
- Запорука ефективного ядра: просування сторінки **або** за **інформаційними, або** за **комерційними** запитами.
- Доступні інструменти не передбачають класифікації запитів.
- На формування семантичного ядра **SEO-спеціаліст витрачає від 1 дня до тижня**.



Об’єкт, предмет і мета роботи

- Об’єкт дослідження: сайти та методи просування в пошукових системах.
- Предмет дослідження: методи класифікації текстових документів для фільтрації пошукових запитів на основі ключових слів за типом.
- Мета роботи: *створити програмний модуль для генерації пошукових запитів з можливістю їх фільтрації за типом (інформаційні та комерційні) на основі ключових слів.*

Основні задачі

- **Парсинг запитів.** Виконується з пошукової системи “Google” з використанням API компанії DataRoot.

- **Побудова класифікатора запитів.** Використано методи з бібліотеки scikit-learn для мови програмування Python.

Вибірка

- 224817 елементів:
  - 183123 - комерційні запити;
  - 41694 - інформаційні запити.
- Після балансування вибірки її розмір - 83388 елементів.
- Розбито на тренувальну і тестову у співвідношенні 70% / 30%

	x	y
0	термобелье tramp купить	1
1	electrolux eww 51685 swd отзіві	0
2	духовые шкафы samsung купить	1
3	матрац ортопедический киев	1

Фрагмент вибірки

Побудова класифікатора. Попередня підготовка даних

Концепція “Bag-of-words”: порядок слів та граматичний зв’язок не мають значення.

Крок 1. Стемінг
Скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс. Використано стемер Портера.

Крок 2. Фільтрація від стоп-слів
Стоп-слова - слова, що не несуть інформаційного навантаження. Використано слованик бібліотеки nltk.

Крок 3. Формування матриці TF-IDF
В результаті кожен запит представляється вектором розмірністю 10407.

TF-IDF

Для слова ***i*** в документі ***j***:

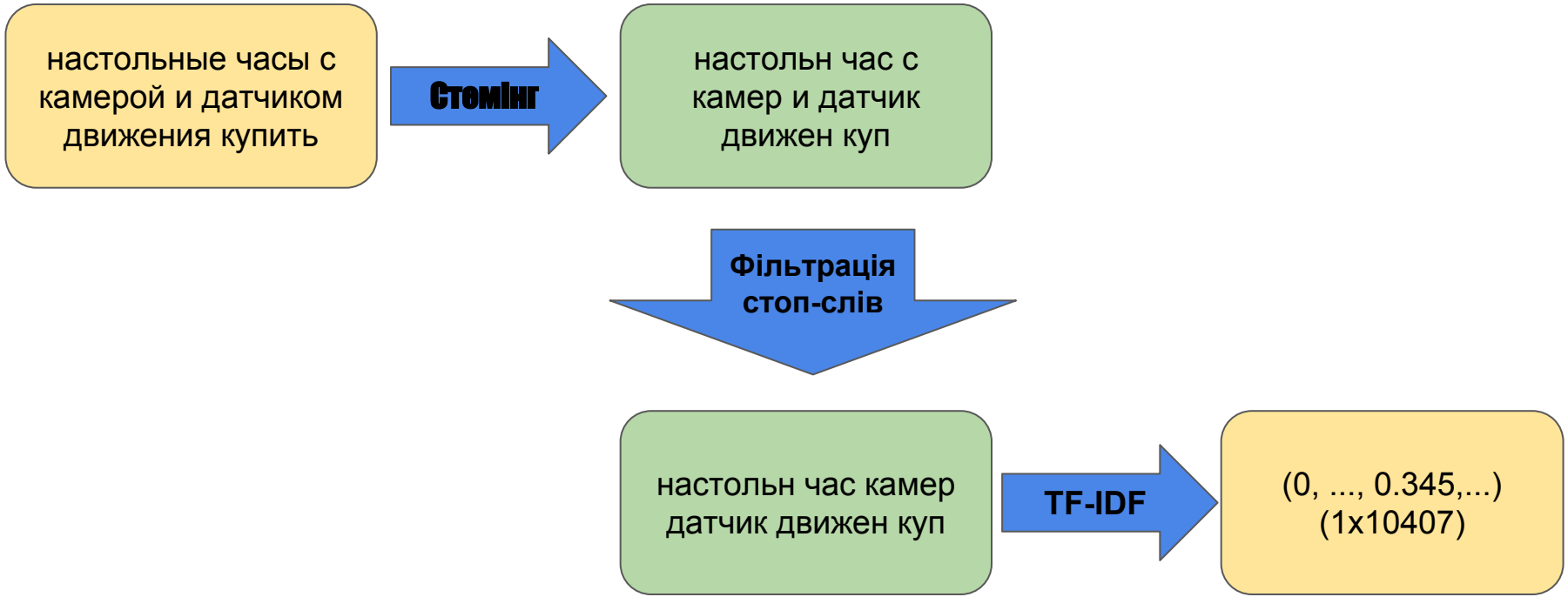
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

де tf_{ij} - кількість появ ***i*** в ***j,***
df_i - кількість документів,
у яких зустрічається ***j,***
N - загальна кількість документів.

Використано додаткові параметри:

- **верхня границя**- термін зустрічається не більше 80% документів;
- **нижня границя** - термін зустрічається більше ніж в 2х документах.

Приклад перетворення запиту



Побудова класифікатора

Використано наступні алгоритми:

- **пасивно-агресивний класифікатор;**
- **AdaBoost;**
- **логістична регресія.**

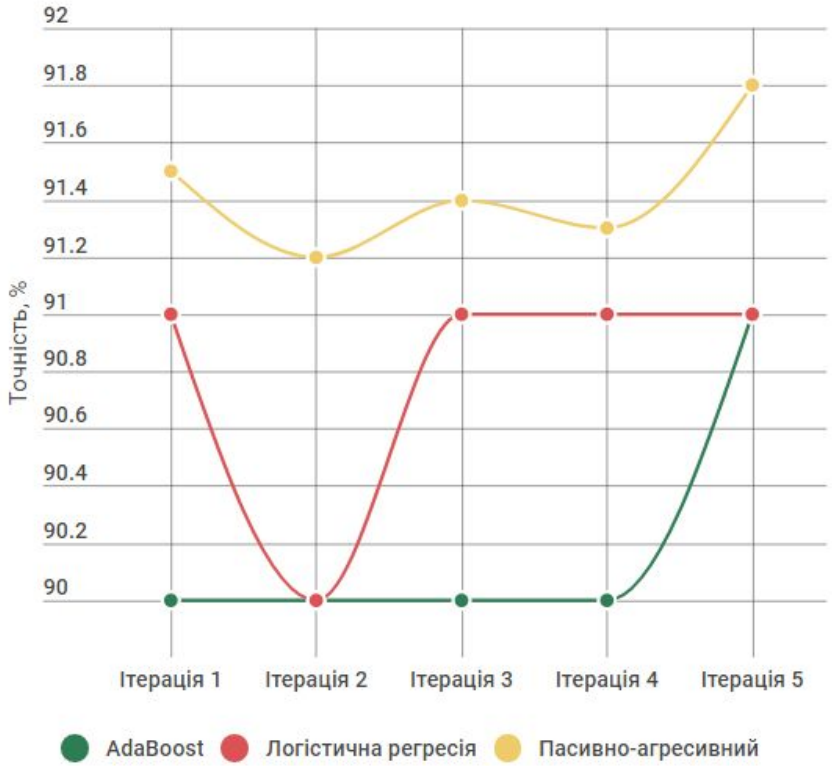
$$\tau_t = \min\left\{ C\,,\ \frac{\ell_t}{\|\mathbf{x}_t\|^2} \right\}$$

Функція оновлення пасивно-агресивного класифікатора

$$Q(w) = \sum_{i=1}^{m} \ln\Big(1+\exp(-y_i\langle x_i, w\rangle)\Big) \to \min_w.$$

Задача мінімізації в логістичній регресії

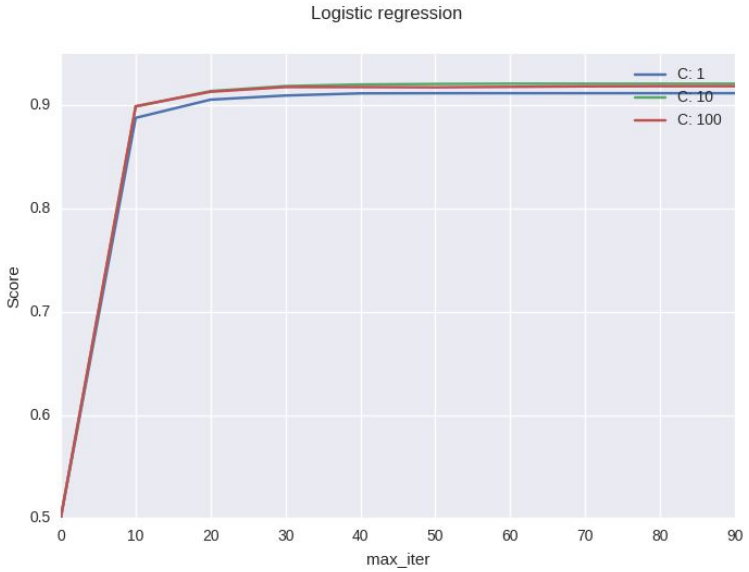
Побудова класифікатора. Крос-валідація



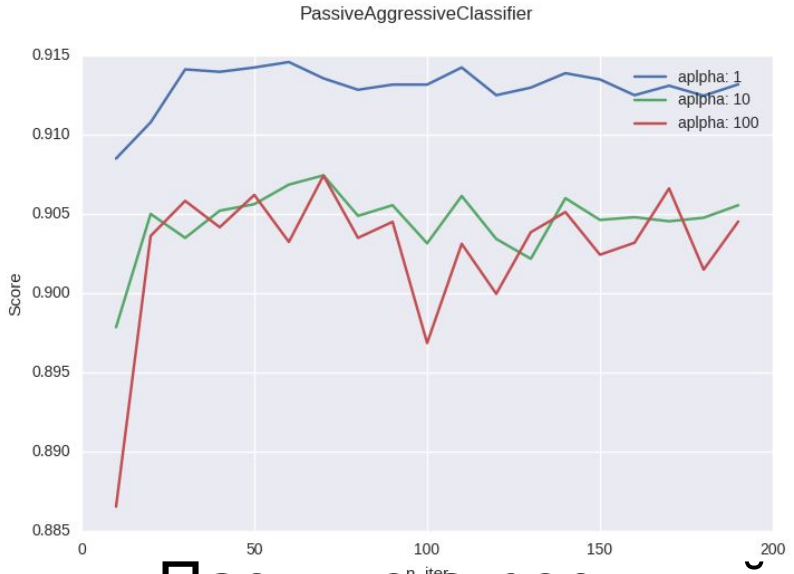
Висновки:

- моделі на даних будуються адекватно;
- найстабільніша модель - пасивно-агресивний класифікатор.

Підбір оптимальних параметрів

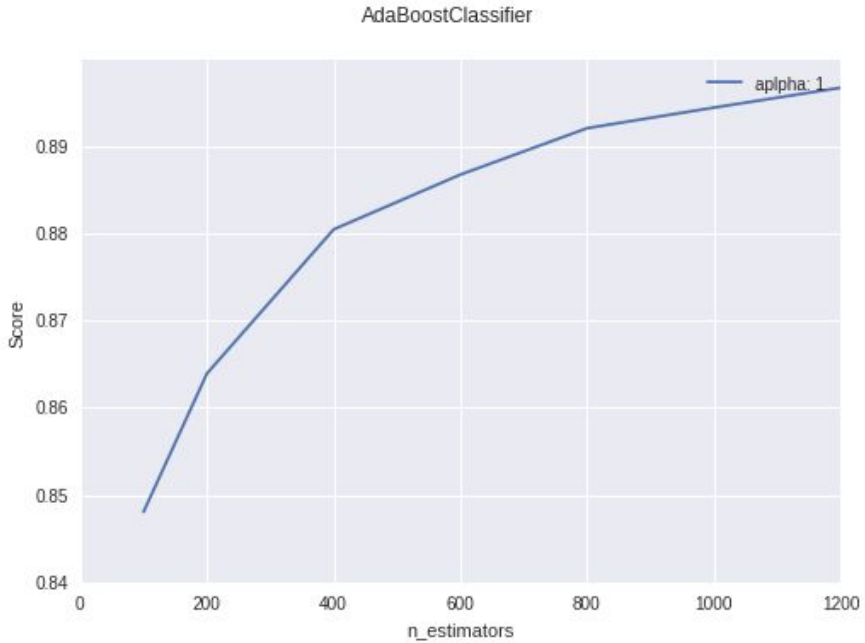


Лінійна регресія



Пасивно-агресивний класифікатор

Підбір оптимальних параметрів



AdaBoost

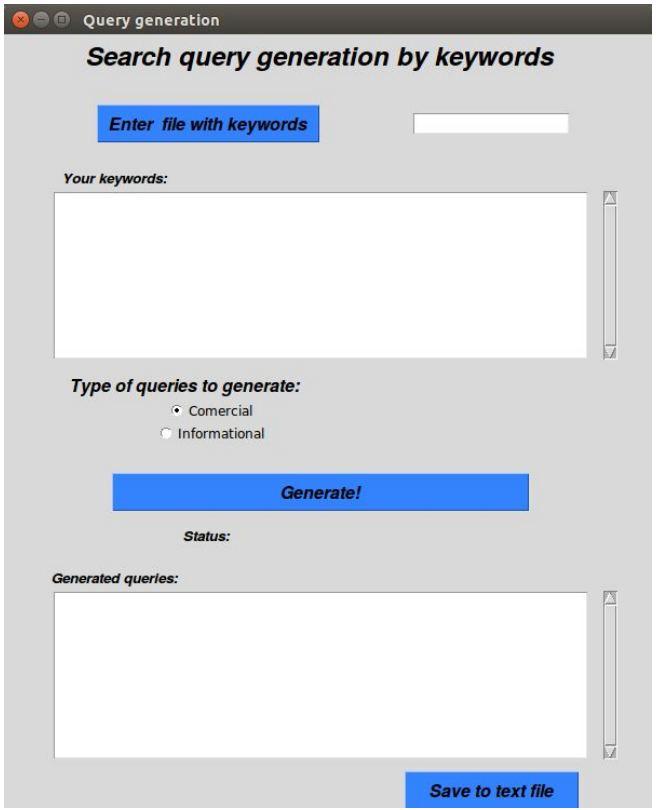
Оптимальні параметри

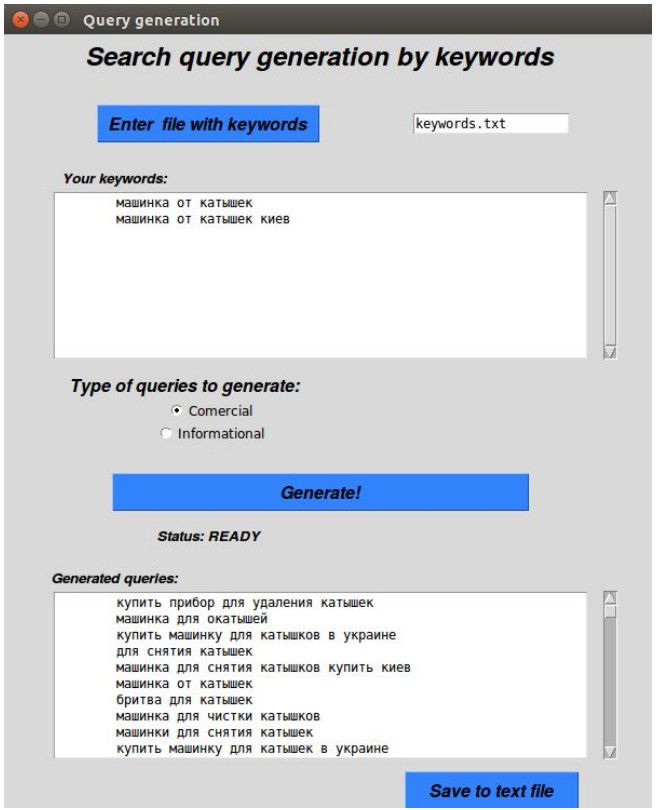
- AdaBoost:
  - оптимальна кількість класифікаторів - 2400.
- Лінійна регресія:
  - оптимальна кількість ітерацій - 32,
  - параметр регуляризації - 10.
- Пасивно-агресивний класифікатор:
  - оптимальна кількість ітерацій - 70,
  - параметр регуляризації - 1.

Побудова класифікатора. Метрики

- **Пасивно-агресивний класифікатор**
  Точність: 92 %
  Матриця неточностей: 11062 1185
  864 11759
- **Перцептрон**
  Точність: 91 %
  Матриця неточностей: 11270 1284
  959 11504
- **Логістична регресія**
  Точність: 91,6 %
  Матриця неточностей: 11161 1393
  707 11756

Інтерфейс користувача





Приклад роботи

За ключовими словами **“машинка от катышек”, “машинка от катышек киев”** та **комерційним типом** було згенеровано наступні пошукові запити (наведено фрагмент):

- купить прибор для удаления катышек
- машинка для окатышей
- купить машинку для катышков в украине
- для снятия катышек
- машинка для снятия катышков купить киев
- машинка от катышек
- бритва для катышек
- машинка для чистки катышков
- машинки для снятия катышек
- купить машинку для катышек в украине
- машинки для снятия катышков
- прибор для удаления катышек с одежды купить

Висновки

Виконано наступне:

- проведено дослідження існуючих підходів до вирішення задачі генерації пошукових запитів за ключовими словами;

- реалізовано модуль для парсингу пошукових запитів за допомогою API компанії DataRoot;

- розглянуто декілька алгоритмів для класифікації запитів, спосіб попереднього опрацювання даних;

- досліджено процедуру навчання класифікаторів та отримано найбільш оптимальні параметри для кожного з них, обрано модель з найбільшою точністю для класифікації запитів на комерційні та інформаційні;

- розроблено модуль для класифікації пошукових запитів;

- реалізовано графічний інтерфейс користувача.

Висновки

Даний програмний продукт вже використовується SEO-спеціалістами, адже:

- Має достатньо високу точність класифікації запитів (91%)
- Має зрозумілий інтерфейс
- Дозволяє виконувати задачу з підбору запитів швидше, ніж доступні аналоги завдяки автоматичній фільтрації.

Шляхи покращення

- При попередній обробці даних виконувати тегування типів товару, брендів та географічних назв.

- Детекція мови запитів та побудова окремих класифікаторі для української та англійської мов.

- Додати до функціоналу можливість підбору семантичного ядра для сторінки інтернет-магазину за вказаним посиланням.

Дякую за увагу!