Міністерство освіти і науки України
Департамент освіти і науки виконавчого органу Київської міської ради
(Київської міської державної адміністрації)

Комунальний позашкільний навчальний заклад
«Київська Мала академія наук учнівської молоді»

Відділення комп'ютерних наук.
Секція: інформаційні системи, бази даних та системи штучного інтелекту.

АВТОМАТИЗАЦІЯ ПРОЦЕСУ ЗБОРУ ТА СИСТЕМАТИЗАЦІЇ ПУБЛІЧНО ДОСТУПНОЇ ІНФОРМАЦІЇ З СОЦІАЛЬНИХ МЕРЕЖ

Роботу виконав:
дійсний член МАН
Полухін Андрій Вячеславович
Дата народження: 22 листопада 2003 року,
учень 11 класу ліцею № 142 м. Києва

Науковий керівник:
Троценко Юрій Володимирович,
провідний науковий співробітник
Інституту математики НАН України,
доктор фізико-математичних наук

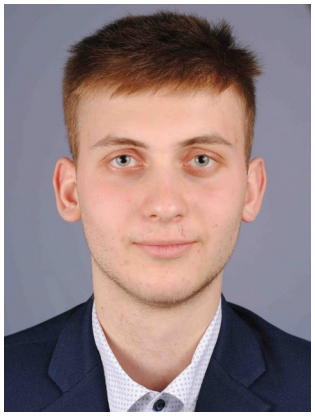
Педагогічний керівник:
Овчинникова Тетяна Анатоліївна,
учитель математики
ліцею № 142 міста Києва,
учитель-методист

КИЇВ – 2021

Комунальний позашкільний навчальний заклад
«Київська Мала академія наук учнівської молоді»

Анотація

Автоматизація процесу збору та систематизації публічно доступної інформації з соціальних мереж



Полухін Андрій Вячеславович,
учень 11 класу ліцею № 142 м. Києва

Науковий керівник: Троценко Юрій Володимирович, провідний науковий співробітник Інституту математики НАН України, доктор фізико-математичних наук.

Педагогічний керівник: Овчинникова Тетяна Анатоліївна, учитель математики вищої категорії, учитель-методист.

Дослідницьку роботу присвячено аналізу способів програмної взаємодії з популярними соціальними мережами з метою збору, аналізу та візуалізації даних про їх користувачів з метою створення автоматизованого програмного продукту.

Проаналізовано документацію API-взаємодії із соціальними мережами: Facebook, Twitter, LinkedIn, Instagram. З метою взаємодії з сервісами Google в ролі розробника створено акаунт Google Developers. Висвітлено специфіку роботи з технічними інструментами Google, а саме: Google Custom Search Engine.

Імплементовано методи сортування та систематизації даних. Відокремлено публічно доступні дані соціальних мереж від технічних. Сортування даних здійснювалося за допомогою використання технологій Rabin fingerprint та відсіювання хешів. Отриманий масив даних підготовлено до візуалізації.

Проаналізовано різні підходи відображення відсортованих даних з урахуванням необхідності відображати як текстову, так і графічну інформацію. Імплементовано відповідний алгоритм зі створенням PDF-звіту з використанням бібліотеки *fpdf2*. Створено графічний інтерфейс з допомогою бібліотеки PyQt5.

Ключові слова: соціальні мережі, дані, отримання даних, сортування даних, візуалізація, PDF-звіт.

ЗМІСТ

Вступ

Актуальність теми. Способи комунікації між людьми змінилися. Більшість публічно доступних даних можна знайти в популярних соціальних мережах: Facebook, Twitter, LinkedIn та Instagram, а також використовуючи ресурси пошукового двигуна Google Search.

Проте в мережі Інтернет знаходиться значний обсяг неструктурованої інформації, тому збір, фільтрація та аналіз отриманих даних є складним та часозатратним завданням. Отже, створення автоматизованої системи, функціональні можливості якої забезпечать збір, систематизацію та візуалізацію даних з соціальних мереж, є актуальною проблемою в наш час. Робота з такою системою дозволить споживачу отримати доступну інформацію про суб'єкт запиту з мінімальними витратами часу та енергії, що може використовуватися в професійних та соціальних сферах життя суспільства.

Мета дослідження полягає у створенні програмного продукту, який автоматизує збір, аналіз та візуалізацію публічно доступних даних з таких соціальних мереж: Facebook, Twitter, LinkedIn та Instagram, а також з пошукового двигуна Google Search.

Гіпотеза дослідження полягає в тому, що із сукупності публічно доступних даних можна автоматично вибрати їхню достатню кількість для отримання висновків щодо соціальних контактів, громадянської позиції, професійної діяльності людини й дати їй об'єктивну оцінку.

Завдання дослідження:

1. Проаналізувати документацією API Facebook, Twitter, LinkedIn та Instagram.
2. Визначити специфіку роботи сервісу Google Custom Search Engine та платформи Google Developers.
3. Автоматизувати агрегацію інформації з Facebook API Graph, Twitter API, LinkedIn API, Instagram API.
4. Розробити алгоритми відсіювання отриманих даних до публічно доступних.

5. Розробити алгоритми аналізу даних до інформації, уведеної користувачем, зокрема методами цифрового відбитку підрядків.
6. Відобразити систематизований масив даних у вигляді PDF-звіту.
7. Створити графічний інтерфейс для кінцевого користувача.
8. Провести верифікацію функціоналу програмного продукту.

Об’єкт дослідження — автоматизована система пошуку, систематизації та візуалізації даних, отриманих із соціальних мереж та ресурсами Google Search.

Предмет дослідження — функціонал опрацювання даних, отриманих через API соціальних мереж з використанням Google Custom Search Engine мовою Python та засобами Qt.

Методи дослідження – аналіз, синтез, узагальнення, систематизація.

Наукова новизна полягає в інтеграції методів збору, систематизації та візуалізації публічно доступних даних із соціальних мереж та Google Search з використанням бібліотек Python.

Особистий внесок: автором розроблені програмна логіка, інтерфейси взаємодії з різними соціальними мережами, а також графічний інтерфейс програмного продукту, при цьому використані сучасні методи розробки автоматизованих систем. На високому рівні реалізовано логічні моделі. Графічний інтерфейс відділений від логіки продукту.

Практичне значення дослідження полягає в автоматизованому знаходженні інформації про користувача, що дозволяє прослідкувати за темпами його інтернет-активності.

Даний програмний продукт був апробований Національною поліцією України («Додаток В») з метою здійснення інформаційної підтримки її діяльності із забезпечення публічної безпеки й порядку, охорони прав і свобод людини, а також інтересів суспільства й держави та протидії злочинності.

РОЗДІЛ 1

ВИБІР КОМПОНЕНТІВ АРХІТЕКТУРИ ПРОЄКТУ

1.1. Вибір платформи розробки проєкту

Першим етапом обрання платформи реалізації проєкту є вибір мови програмування. На сьогодні більшість мов програмування підтримують взаємодію з API соціальних мереж та сервісами компанії Google, проте саме Python має достатньо сторонніх бібліотек для автоматизації збору, систематизації та візуалізації отриманих даних. Графічна оболонка може бути реалізована з допомогою додаткової бібліотеки PyQt5, яка дозволяє модульно взаємодіяти з UI файлами. Модульна архітектура дозволяє відділити логіку проєкту від дизайну.

Вибір Python також зумовлений лаконічністю мови програмування, тому написаний код легше підлягає технічній підтримці та модернізації як і автором, так і сторонніми розробниками.

1.2. Взаємодія із соціальними мережами й Google Search

Оскільки ми ставимо за мету створити сервіс, який працюватиме тривалий час, будучи непомітним для широкої громади, то доцільно використовувати API соціальних мереж: Facebook, Twitter, LinkedIn та Instagram. Зазвичай взаємодія з API соціальної мережі зводиться до отримання відповідних прав розробника, які потім можна використовувати, маючи API ключ та секрет.

На етапі підготовки до роботи з сервісом Facebook API Graph [7] з’ясувалося, що для отримання необхідних маркерів доступу до профілів інших користувачів необхідно підтвердити компанію, від імені якої розробляється продукт. Оскільки наша робота має науково-дослідницьку мету, то на API запит на профіль іншого користувача в Facebook ми отримували серверну відповідь (рис. 1.1).

```
{
  "error": {
    "message": "Unsupported get request. Object with ID '<object_id>' does not exist, cannot be loaded due to missing permissions, or does not support this operation. Please read the Graph API documentation at https://developers.facebook.com/docs/graph-api",
    "type": "GraphMethodException",
    "code": 100,
    "error_subcode": 33,
    "fbtrace_id": "<fbrace_id>"
  }
}
```


Рис. 1.1. Відповідь Facebook API на запит на профіль іншого користувача

Підготовча робота з іншими соціальними мережами пройшла успішно й були використані такі сервіси: Twitter API v.1.1 [5], LinkedIn API [6] та Instagram API [8]. Отримані дозволи дали нам можливість використовувати доступ до загальнодоступних ресурсів.

Для програмного користування сервісом Google Search треба зробити два кроки: отримати API ключ та секрет для програмної взаємодії із сервісами Google на платформі Google Developers [9] та зареєструвати свій пошуковий сервіс, який шукатиме інформацію з Інтернету ресурсами компанії Google [10].

1.3. Систематизація зібраних даних

Зауважимо, що отримані за допомогою API та публічно доступні дані відрізняються. Метою систематизації є відсіювання технічних даних про користувача соціальної мережі до публічно доступних. Підготовка отриманої інформації до завершального етапу, візуалізації, із соціальної мережі LinkedIn відбувалася за певними правилами (рис. 1.2).


```
{"certifications": ["authority", "name", "timePeriod", "url"],
  "education": ["degreeName", "fieldOfStudy", "schoolName", "timePeriod"],
  "experience": ["company", "companyName", "locationName", "timePeriod", "title"],
  "skills": ["name"]
}
```


Рис. 1.2. Правила для вибору інформації з LinkedIn

Однак відповідь від API може не відповідати критеріям пошуку інформації, адже, наприклад, є користувачі з однаковим прізвищем. Саме тому був введений ще один етап відсіювання потенційних користувачів: за додатковою інформацією, яку

вводить кінцевий користувач. У цьому випадку найефективнішим став метод Rabin fingerprint та співставлення отриманих хешів за результатами реалізації алгоритму з роботи [1, частини 1 і 2].

Розглянемо приклад знаходження міри схожості двох текстів (рис. 1.3).

```
>>> import find_similarity_ratio
>>> first_string = "Найкращий математик України, переможець міжнародних олімпіад"
>>> second_string = "Найкращий український математик, переміг декілька міжнародних олімпіад"
>>> find_similarity_ratio(first_string, second_string)
0.7083333333333334
```

Рис. 1.3. Реалізація алгоритму співставлення хешів

1.4. Візуалізація систематизованих даних

Працювати з масивами даних є обтяжливим завданням для звичайного користувача, отже, доцільним є створення сервісу, який візуалізує систематизовану інформацію. З-поміж усіх способів відображення даних було обрано відображення інформації у вигляді PDF-звіту: такий файл не залежить від операційної системи користувача та може містити в собі як графічну, так і текстову інформацію.

З цією метою була обрана стороння бібліотека *fpdf2* ([11]), яка дозволяє створювати PDF-звіти методами мови програмування Python. Основною її перевагою є швидкість створення кінцевого файлу, підтримка необхідних графічних форматів (JPEG, PNG), автоматичний перехід на наступну сторінку та відступи.

Щодо графічного інтерфейсу, то у «Додатку А» представлений конструктор класу, що реалізує графічну оболонку проєкту.

Підсумовуючи етап вибору компонентів архітектури проєкту, для кожного з його логічних етапів, а саме: агрегації, систематизації, візуалізації даних, а також графічної оболонки, було обрано:

1. Відповідні API сервіси для взаємодії із соціальними мережами для агрегації даних.
2. Методи, за допомогою яких дані будуть систематизуватися з метою відсіяти технічні дані та обрати лише користувачів, які задовольняють умови пошуку.
3. Сторонню бібліотеку для відображення текстової та графічної інформації.
4. Графічну оболонку, дизайн якої може бути змінений за бажанням споживача.

РОЗДІЛ 2
СТВОРЕННЯ АРХІТЕКТУРИ ПРОЄКТУ

2.1. Агрегація сервісів збору, систематизації та візуалізації інформації

Робота кожного з етапів проєкту (збір, систематизація та візуалізація даних) вимагає часових затрат, а отже у науково-дослідницькій роботі доцільним є використання методів асинхронного програмування для збереження ресурсів споживача й рівномірного розподілення навантаження на ПК.

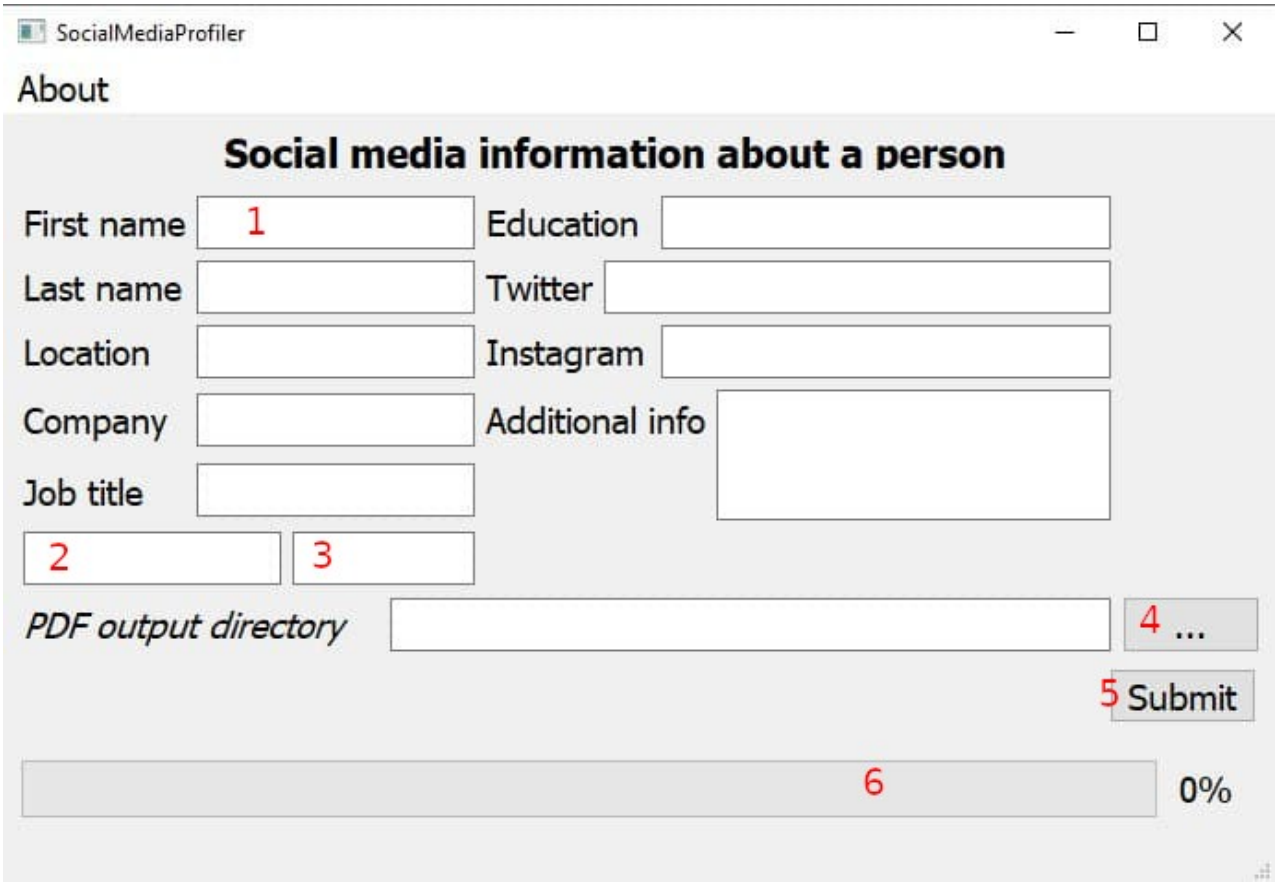
Наприклад, процес отримання даних з кожної із соціальних мереж: Twitter, LinkedIn, Instagram та сервіс Google Search не залежить один від одного, а тому наведені дії можна виконувати паралельно. Це ж стосується систематизації даних: Twitter, LinkedIn та Instagram використовують різні методи систематизації, а тому дані процеси також доцільно виконувати одночасно. Беручи до уваги, що відображення графічної та текстової інформації відбувається в одному PDF-файлі, то візуалізація відбувається синхронно. Розглянемо повну схему роботи сервісів (рис. 2.1).

Рис. 2.1. Повна схема роботи логічних сервісів програмного продукту

Як бачимо, необхідний масив даних для початку роботи алгоритму формується за результатами вводу інформації споживачем на графічному інтерфейсі. Для адекватної роботи наданої схеми треба налаштувати програму, інструкцію до налаштування можна знайти на README GitHub репозиторію проєкту.

2.2. Створення графічного інтерфейсу

У якості графічного інтерфейсу була використана стороння бібліотека PyQt5. Саме ця графічна оболонка є лаконічною та ефективною, що дозволяє масштабувати та змінювати дизайн за вподобанням споживача. Розглянемо графічний інтерфейс програмного продукту (рис. 2.2).



1 — звичайне поле вводу інформації, 2 — селектор для пошуку додаткової інформації, 3 — інформація для пошуку до додаткового селектору, 4 — кнопка вибору папки для PDF, 5 — кнопка “Підтвердити”, 6 — шкала виконання логічних сервісів проєкту.

Рис. 2.2. Графічний інтерфейс програмного продукту

Поля типу, як поле 1, вказують на введення інформації, тип якої описано на етикетці зліва. Наприклад, власне поле 1 створено для вводу імені (first name) людини. Поля 2 і 3 створені для пошуку додаткової інформації про користувача з допомогою сервісу Google Search, причому поле 2 призначено для селектора пошуку, а поле 3 створено для власне інформації пошуку. Наприклад, якщо необхідно знайти інформацію про користувача *pandrey2003* на сайті GitHub, то в поле 2 треба вписати селектор: “GitHub”, а в поле 3 ввести власне інформацію пошуку: “pandrey2003”.

Кнопка під номером 4 відкриває додатковий графічний діалог, який дозволяє обрати папку на ПК кінцевого користувача для збереження PDF-звіту з візуалізованою інформацією.

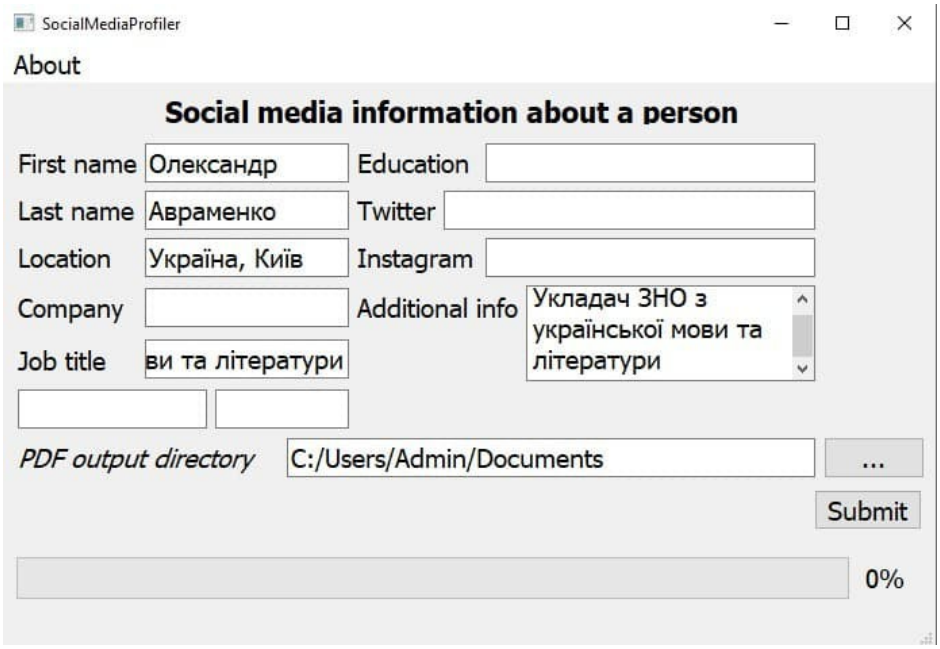
Натискання кнопки 5 забезпечує збирання всієї введеної користувачем GUI інформації та починає логічну частину проєкту: збір, систематизацію та візуалізацію даних. На неї доцільно натискати тоді, коли вся інформація про користувача, який розшукується, уже введена.

Шкала наповнення 6 відображає прогрес виконання отримання, аналізу та відображення інформації. Логічно, що 100% вказує на готовий PDF звіт у директорії, яку визначив кінцевий користувач.

2.3. Верифікація функціоналу програми

З метою перевірки працездатності нашої програми були розроблені критерії пошуку (рис. 2.3) та здійснена верифікація на даних таких людей:

- Олександр Авраменко;
- Поль Манандіз;
- Kevin Goldsmith;
- Дмитро Гордон.



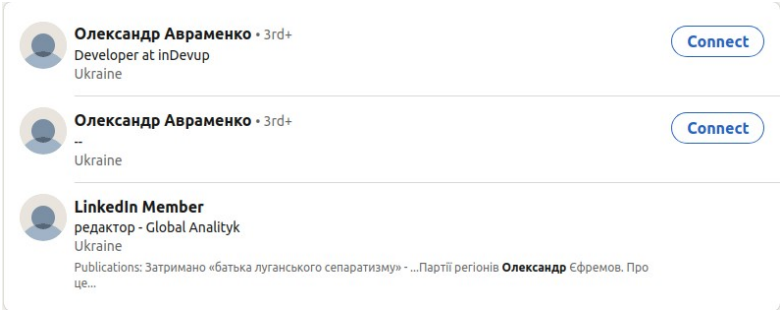
Про роботу введено “Вчитель української мови та літератури”

Рис. 2.3. Критерії пошуку інформації про О. Авраменка

У результаті була знайдена наступна інформація (рис. 2.4).

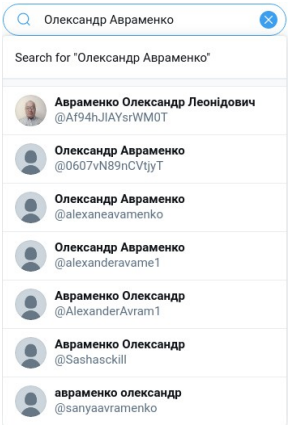
Рис. 2.4. Знайдена інформація про Олександра Авраменка

Наводячи курсор мишки на нікнейм Інстаграму та переходячи на відповідне посилання з PDF-звіту, пересвідчуємось, що цей профіль належить пану Олександру. Також ми перевірили, чи нема інформації про Олександра Авраменка в інших соціальних мережах (рис. 2.5 та рис. 2.6).



Профілі LinkedIn не задовольняють умовам запиту

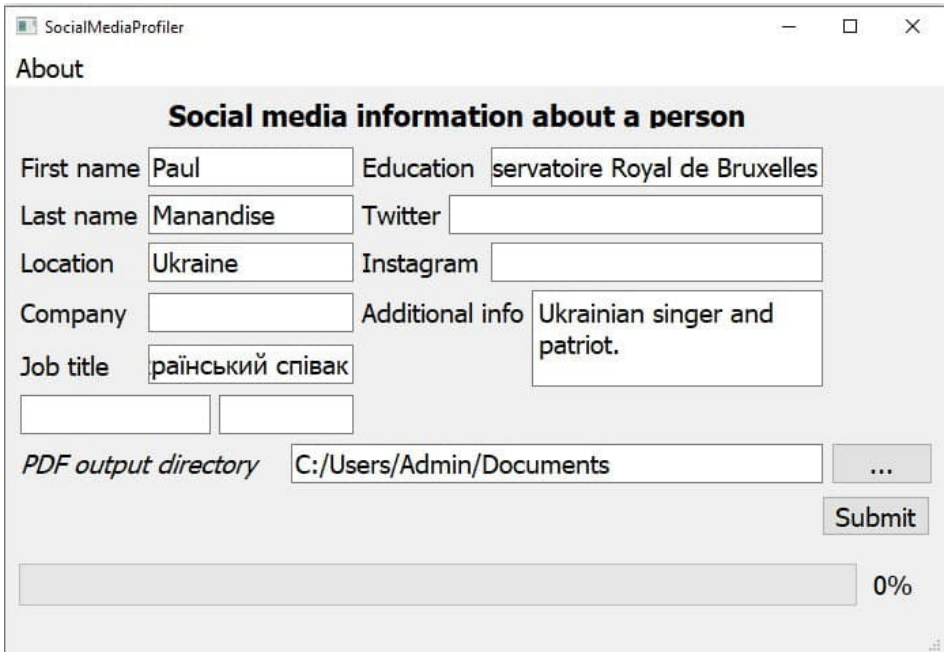
Рис. 2.5. Перевірка наявності профіля О. Авраменка в LinkedIn



Профілі Twitter не задовольняють умовам запиту

Рис. 2.6. Перевірка наявності профіля О. Авраменка у Twitter

Також нами був здійснений пошук даних про П. Манандіза (рис. 2.7).



Про роботу вказано “український співак”

Про освіту вказано “Conservatoire Royal de Bruxelles”

Рис. 2.7. Критерії пошуку інформації про П. Манандіза

Поль Манандіз народився в Бельгії й велику частину свого життя прожив у Франції, а тому його альма-матер вказана бельгійською, а ім’я, локація та додаткова інформація — англійською (міжнародна мова спілкування).

У процесі пошуку ми знайшли фейкові акаунти Поля Манандіза (рис. 2.8).

Рис. 2.8. Фейкові акаунти Instagram, які отримуємо на API запит про Манандіза

Натискаючи на вбудовані в PDF посилання, знаходимо тільки потрібний нам Instagram профіль (рис. 2.9).

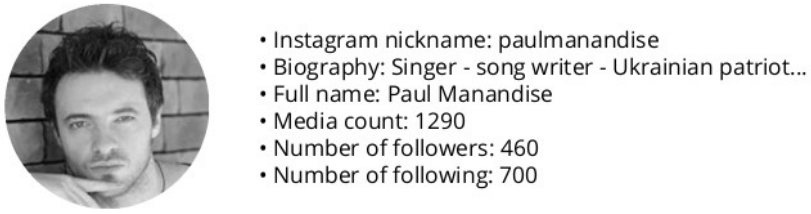


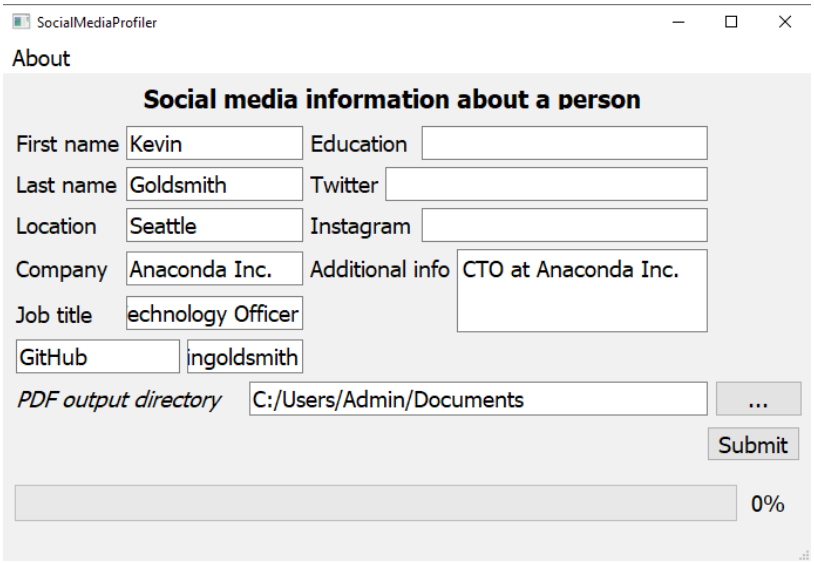
Рис. 2.9. Справжній Instagram профіль П. Манандіза

Аналогічно до Instagram, пошук у Twitter дав декілька результатів, зокрема і справжній профіль (рис. 2.10).

Рис. 2.10. Справжній Twitter профіль П. Манандіза

Пошук П. Манандіза в LinkedIn не дав результатів як вручну, так і програмно.

Програмно проведемо збір, систематизацію та візуалізацію даних про Kevin’a Goldsmith’a. Беручи до уваги, що відомий його GitHub профіль, знайдемо про нього інформацію з Google Search, використавши відповідні налаштування (рис. 2.11).



Про роботу написано “Chief Technology Officer”; GitHub: “kevingoldsmith”

Рис. 2.11. Критерії пошуку інформації про Kevin’a Goldsmith’a

Аналогічно до отриманої інформації про Поля, у соціальній мережі Instagram знайдено багато профілів на запит про Kevin’a Goldsmith’a, проте мануальна перевірка показала, що жоден із них не є справжнім. Водночас знайдено справжній профіль у Twitter (рис. 2.12).

Twitter

• Screen name: KevinGoldsmith
• Description: CTO @ Anaconda & Speaker (ex: Onfido, Avvo,...
• Full name:
• Location: Seattle
• Number of followers: 3595
• Number of following: 1456
• External URL: https://t.co/KYDeZCA0Q6
• Two last posts:
- RT @jodyrodgers: I've replaced the election Twitter news...
- RT @anacondainc: We look forward to leveraging @eodaGmbH's...

Рис. 2.12. Справжній Twitter профіль Kevin’a Goldsmith’a

До речі, дані були отримані незважаючи на те, що користувач використував емодзі Юнікод символи у своєму імені, а саме: “🅺evin 🅶oldsmith”.

Була також знайдена інформація в LinkedIn (рис. 2.13).

LinkedIn
Potential user(s)

Kevin Goldsmith
Headline: Chief Technology Officer (CTO) at....
Industry: Computer Software.
Location: .
Work experience:
• *Anaconda, Inc.*
Job title: Chief Technology Officer.
Company location: Austin, Texas, United States.
Time period: 10/2020 - Now.
• *Nimble Autonomy, LLC*
Job title: Founder / Principal.
Company location: Greater Seattle Area.
Time period: 07/2020 - Now.
• *Unit Circle Media*
Job title: Owner.
Company location: Greater Seattle Area.
Time period: 01/1991 - Now.
• *Onfido*
Job title: Chief Technology Officer.
Company location: London, United Kingdom.
Time period: 04/2019 - 06/2020.
• *cavnessHR*
Job title: Member, Board of Advisors.
Company location: Greater Seattle Area.
Time period: 06/2018 - 01/2020.

Education:
• *BS*
School name: Carnegie Mellon University.
Time period: 1988-1992.

Skills: Software Development, Mobile Applications, Agile Methodologies, Software Engineering, Scrum, C++, Software Design, Distributed Systems, Cloud Computing, Algorithms, Scalability, Mobile Devices, Architecture, Web Services, Windows, C, Objective-C, Digital Imaging, MySQL, GPGPU, Technical Leadership, Object Oriented Design, Digital Media, High Performance Computing, 3D, Multithreading, OpenGL, iOS development, Architectures, Digital Asset Management, Digital Image Processing, Lean Management, Kanban, Lean Thinking, Application Architecture, Human-computer Interaction, Parallel Algorithms, Language Design, Media Technology, Grid Computing, Parallel Computing, Extreme Programming, Lean Software Development, Leadership, Cross-functional Team Leadership, Organizational Leadership, Strategic Leadership, Thought Leadership, Charcuterie, Muay Thai.

Рис. 2.13. Дані про Kevin’a Goldsmith’a з соціальної мережі LinkedIn

Також ми отримали результати пошуку на Google Search (рис. 2.14).

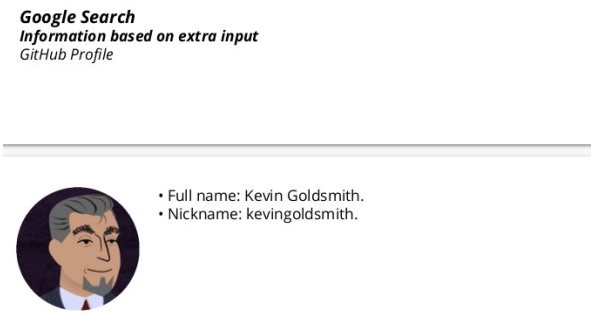
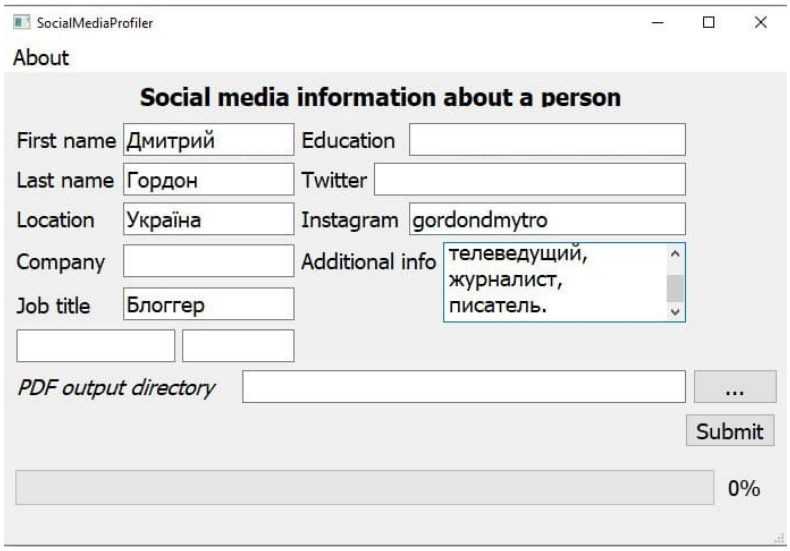


Рис. 2.14. Результати пошуку про Kevin’a Goldsmith’a з Google Search

Для наочності отриманий PDF-файл повністю представлений у «Додатку Б».

Візьмемо ще одну постать. Дмитро Гордон російськомовний, тому дані для пошуку виглядають дещо інакше (рис. 2.15).



Як додаткову інформацію записано “украинский телеведущий, журналист, писатель”

Рис. 2.15. Критерії пошуку інформації про Д. Гордона

Був знайдений його справжній профіль в Instagram (рис. 2.16).

Рис. 2.16. Справжній Instagram профіль Д. Гордона

Окрім цього, наша автоматизована система знайшла профіль Дмитра Гордона в Twitter (рис. 2.17).

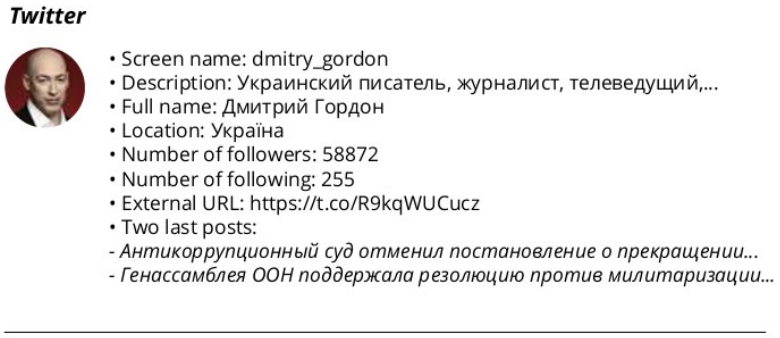


Рис. 2.17. Справжній Twitter профіль Д. Гордона

Перевірка вручну показує, що всі дані знайдено коректно. LinkedIn профіля ні нашим програмним продуктом, ні вручну знайти не вдалося.

Отже, алгоритми, описані в нашій науково-дослідницькій роботі автоматизують збір, систематизацію та візуалізацію даних. У ході верифікації можна зробити висновки, що якість отриманої інформації не поступається пошуку вручну, а програмний продукт мінімізує часові витрати кінцевого користувача.

2.4. Порівняння нашого програмного продукту з аналогами

Провівши аналіз ринку, ми виявили, що всі доволі популярні засоби для пошуку інформації про людину онлайн не орієнтовані на пошук інформації про українців, а мають практичну значущість переважно на американському й західноєвропейському ринках. Причому були виявлені недоліки кожного з програмних продуктів з точки зору або характеру знайдених даних, або ціни продукту, або якості отриманих даних.

Американський інструмент “Pipl” використовується для пошуку інформації про людину загалом, а не як інтернет-користувача. Засіб “Yoname” спеціалізується в соціальних мережах, проте шукає дані лише за ім’ям та прізвищем, тому систематизація інформації відбувається неефективно, а візуалізація відсутня. Сервіси “Jobster” та “Zoom info” шукають інформацію про людину як про робітника, але вони платні. Засоби “Intellius” та “Zaba Search” не знаходять інформацію про людину в соцмережах; вони є умовно безкоштовними, проте за повну інформацію доведеться заплатити. Розроблений же нами програмний продукт має MIT ліцензію.

РОЗДІЛ 3

АПРОБАЦІЯ ПРОГРАМНОГО ПРОДУКТУ

3.1. План проведення апробації

Беручи до уваги постійно щораз більшу кількість потенційних загроз, що використовуються зловмисниками зокрема з використанням мережі Інтернет, Департамент інформаційно-аналітичної підтримки Національної поліції України (далі — Департамент) вирішив апробувати наш програмний продукт, як додатковий альтернативний модуль до наявних відповідних технологій, що використовуються силовими структурами для попередження та виявлення можливих правопорушень, забезпечення підтримки правопорядку та громадської безпеки. Наш програмний продукт був використаний для порівняльного аналізу активностей користувачів соціальних мереж, які можуть негативно впливати на громадську свідомість своїми діями в наступних соціальних мережах, а саме: Instagram, Twitter, LinkedIn, та окремо — за допомогою пошукового двигуна Google Search.

Для виявлення потенційних правопорушень було вирішено виявити зв’язок щодо характеру інформації, яку людина поширює в одній соціальній мережі та, таким чином, швидко й ефективно знайти подібну інформацію в інших соціальних мережах. Використання вказаного підходу дозволило швидко виявити злочинні наміри людини в просторі мережі Інтернет, а саме: заклики до вчинення погромів, підпалів, знищення майна, насильства над людьми тощо.

Використовуючи програмний продукт, фахівці Департаменту одразу визначили характер інформації, що була представлена в PDF-файлі, а тому мали змогу ефективно звузити кількість громадян до тих, чиї заклики до розпалювання ненависті поширюються, зокрема, в мережі Інтернет.

Зауважимо, що пошук конкретної інформації про людину, яка має кримінальне минуле та негативно впливає на громадянську свідомість, полегшується тим фактором, що базова інформація про таку людину, а саме: прізвище, ім’я, по-батькові, дата народження, місце роботи та інше — уже є в спеціальних інформаційно-телекомунікаційних системах або можуть бути отримані в межах

співпраці Національної поліції з іншими правоохоронними та державними структурами.

Відповідно до поставленої мети був сформований такий план використання нашого програмного продукту:

1. Виділити групу потенційних зловмисників, чия активність розповсюджується переважно в мережі Інтернет.
2. Засобами вже зібраної інформації знайти якнайбільше інформації про людину як про потенційного кримінального злочинця.
3. Отримати ключі доступу до API соціальних мереж, сервісу Google Developers та двигуна Google Custom Search Engine задля коректного налаштування роботи програми.
4. Порівняти кількість та якість зібраної вручну інформації в соціальних мережах і пошукового двигуна Google Search та отриманої за допомогою нашого програмного забезпечення.

3.2. Процедура апробації

Першим та ключовим етапом процедури проведення апробації було виокремити групу людей, на якій буде випробувана наша автоматизована система. Виявилося, що, враховуючи популярність соціальних мереж у XXI сторіччі, 59 людей із вибірки в 64 особи мають інтернет-активність, що може бути проаналізована нашим програмним продуктом.

Другим кроком було отримати інформацію, яка б розширила межі пошуку та допомогла б одразу на програмному рівні відсіяти людей, які не підпадають під критерії пошуку відповідних державних структур. Як і очікувалося, була доведена необхідність отримати найбільшу кількість вихідної інформації про шуканого суб’єкта (детальніше в п. 3.3).

Процес отримання ключів доступу пройшов ефективно: залучені працівники отримали свої унікальні ключі доступу, що дозволило розширити кількість надісланих API запитів до соціальних мереж та Google Custom Search Engine. Фахівці Департаменту підкреслили зручність використання окремого .env-файлу

замість вписування всіх ключів доступу в код, що дозволило збільшити конфіденційність нашого програмного продукту.

Останній етап, порівняння якості та повноти інформації, отриманої нашою автоматизованою системою в порівнянні з інформацією, яка знаходилась вручну, був виконаний із залученням окремого працівника Департаменту. Його задача зводилася до порівняння вже зібраної Національною поліцією інформації та детального візуального знаходження потенційно упущеної нашим програмним продуктом інформації на деякій вибірці людей з метою отримання висновку стосовно ефективності/неефективності результатів нашої науково-дослідницької роботи в якості альтернативного допоміжного елементу для попередження можливих правопорушень та забезпечення правопорядку. Результати цього кроку плану наведені в п. 3.3.

Ще на етапі процедури апробації працівники Департаменту відзначили зручність та лаконічність нашого програмного продукту, що досягалося використанням графічної оболонки PyQt5, а також присутністю індикатора виконання програми, що дозволяє відстежувати логічний етап виконання пошуку.

3.3. Результати апробації

По-перше, була визнана висока кількість “цільової аудиторії”, тобто зловмисників, на яких орієнтується наш програмний продукт. Як видно на Рис. 3.1, з обраної вибірки результати нашої науково-дослідницької роботи можуть бути використані для більше, ніж 92% осіб. Такі результати підтверджують високу актуальність дослідження в сучасному інформаційному світі.

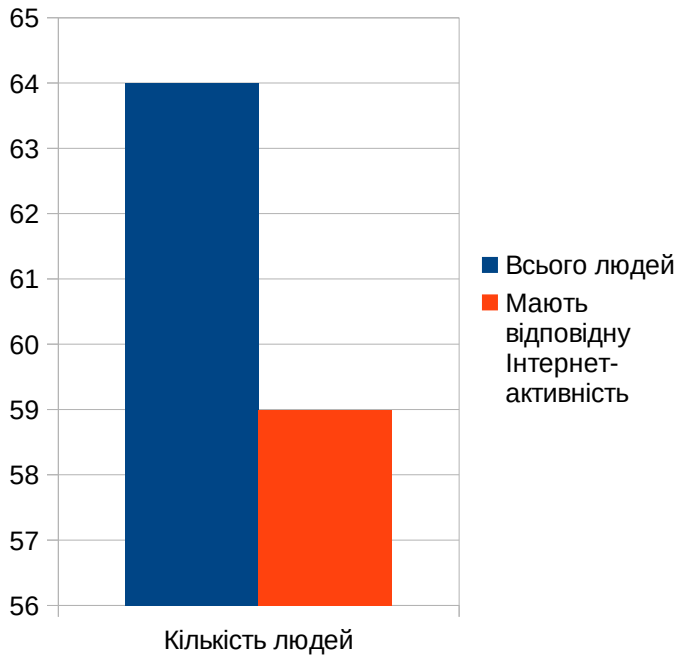


Рис. 3.1. Порівняння осіб з Інтернет-активністю до загального числа усієї вибірки

Зауважимо, що згідно з вимогами законодавства у сфері захисту персональних даних, а також захисту інформації, результати апробації не містять ніякої персональної інформації як про розшукуваних людей, так і про працівників, які здійснювали аналіз ефективності нашого програмного продукту. Уся наведена інформація використовується винятково в контексті математичної статистики.

При роботі з автоматизованими системами першорядна роль приділяється вхідній інформації, отож був проведений аналіз візуалізованої інформації в залежності від повноти заданої інформації. Пропонуємо ознайомитися з ними на Рис. 3.2.

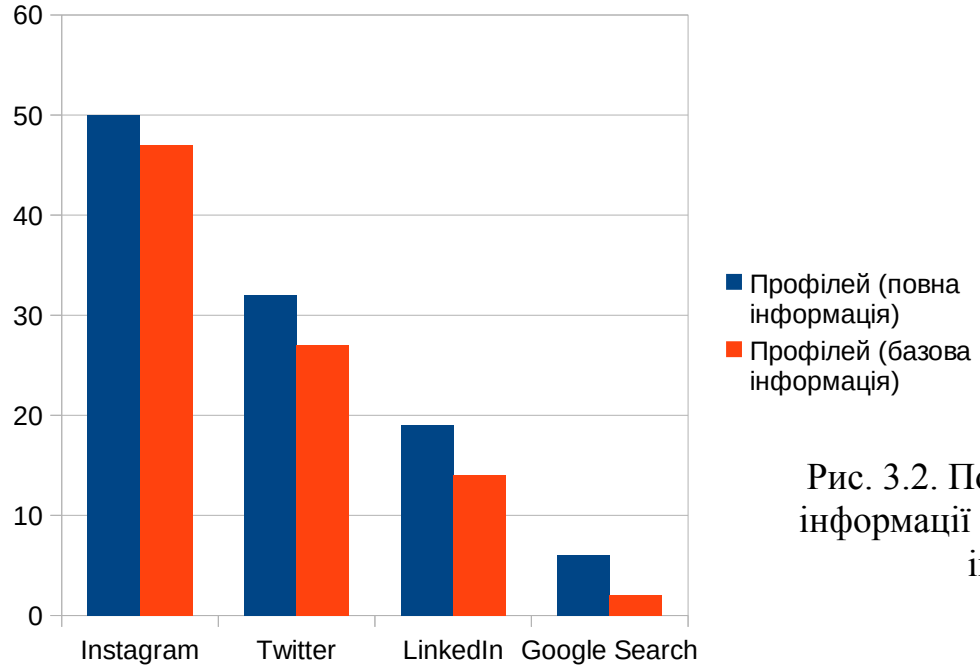


Рис. 3.2. Порівняння знайденої інформації залежно від повноти інформації

Як бачимо, наш програмний продукт показує набагато більше точної інформації при більшій кількості вхідних даних. Цікавим є аналіз наведеної гістограми з точки зору математичної статистики. Наприклад, соціальна мережа Instagram не вимагає великої кількості точної інформації, що видно зі збільшенням кількості знайдених осіб всього на 6,38%. Це мінімальне збільшення відображає культуру вказаної соціальної мережі: мають значення нікнейм та опис “про себе”, а справжні ім’я, прізвище та місцезнаходження найчастіше використовуються тільки маркетологами.

У соціальній мережі Twitter користувач, окрім подібної інформації в Instagram, зазначає (за бажанням) своє місцезнаходження. Наявність інформації про місце проживання людини дозволяє точніше знайти тих користувачів соціальних мереж, які вказали своє місце проживання. Математично це відобразилося в збільшенні кількості знайдених користувачів на 18,5%.

В Україні соціальна мережа LinkedIn поступається в популярності Instagram та Twitter, тож маємо, що дисперсія отриманих даних є суттєво більшою за вже проаналізовані соціальні мережі. LinkedIn — це соціальна мережа для роботи, отож вже на етапі реєстрації ця соціальна мережа вимагає від свого користувача заповнити конфіденційну інформацію: справжні ім'я, прізвище, місцезнаходження, освіту, досвід роботи. Отже, додаткова інформація призвела до збільшення точного знайдення профілів аж на 35,7%.

Зазначимо, що Google Search не соціальна мережа, тобто в ній не можна "зареєструватися", у ній можна знайти людину за умови її популярності та доволі точної інформації про неї. З усієї вибірки тільки 6 осіб виявилися настільки активними Інтернет-користувачами, щоб з'явитися в PDF файлі після вводу додаткової інформації. До цього система могла знайти тільки 2 особи, бо одна з умов, щоб бути знайденим у пошуковику (ідеться про точну інформацію) не виконувалась.

Останнім етапом апробації було винесення вердикту стосовно надійності нашої інформаційної системи проти мануального пошуку та відсіювання інформації. На Рис. 3.3 можна бачити гістограму порівняння кількості якісно знайдених осіб у

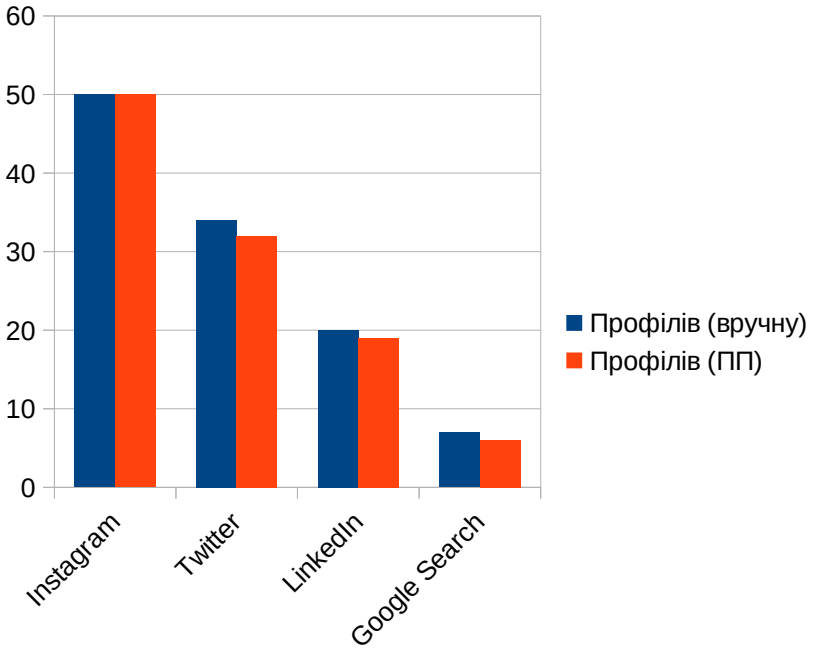


Рис. 3.3. Порівняння знайдених профілів вручну та з допомогою нашого ПП

соціальних мережах, зроблену вручну та з допомогою результатів нашої науково-дослідницької роботи.

Механізми отримання інформації із соціальної мережі Instagram та загальна інформація про можливі інтереси людей виявилися достатніми, щоб знайти усіх осіб у мережі Instagram.

Робота з Twitter виявилася не настільки ідеальною, оскільки 2 особи зазначили своє місце реєстрації в полі іншим, ніж було місце прописки. Тож, ті 2 людини відсіялися нашим алгоритмом як такі, що не задовольняють умови пошуку.

Щодо LinkedIn ми отримали такий фактичний результат: наш програмний продукт знайшов 19 осіб з 20 за час, що значно швидший за знаходження необхідної інформації “вручну”. Незнайденою людиною виявилася особа, яка вирішила змінити свою професію, а тому всю інформацію про університет, який вона закінчила, вона видалила.

Google Search виявився дуже чутливим до того, який запит він отримує. За допомогою ручного пошуку вдалося знайти 7-у людину, яку не знайшла наша автоматизована система. Це був розробник, чий GitHub-профіль не мав достатньої популярності, щоб бути поміченим нашою автоматизованою системою.

ВИСНОВКИ

Основним результатом наукового дослідження є створення автоматизованої системи для отримання, аналізу та візуалізації даних про користувачів соціальних мереж.

У роботі було проаналізовано літературні джерела та доступні сучасні системи зі збору даних про користувачів соціальних мереж, сформульовано мету й завдання дослідження. Під час проведення дослідження проаналізовано документацію API Facebook, Twitter, LinkedIn та Instagram. Створено акаунт Google Developers та Google Custom Search Engine. При цьому робота з Facebook API Graph унеможливилася через брак дозволів. Було автоматизовано агрегацію даних з Twitter API, з LinkedIn API, з Instagram API.

Окрім цього, у дослідженні проаналізовано алгоритми аналізу даних методами Rabin fingerprint та реалізацією алгоритму з роботи “Winnowing: Local Algorithms for Document Fingerprinting” Стенфордського університету, а також розроблено та описано алгоритми відсіювання отриманих даних до публічно доступних та алгоритми візуалізації в PDF-звіті. Для програмної візуалізації отриманого масиву даних у PDF-звіті використано бібліотеку *fpdf2*. Для графічного інтерфейсу була обрана бібліотека PyQt5. З метою ознайомлення з базовим функціоналом PyQt5 було опрацьовано офіційну документацію цієї бібліотеки.

Після створення програмного продукту автором було проведено його верифікацію на прикладі таких користувачів соціальних мереж: Олександр Авраменко, Поль Манандіз, Kevin Goldsmith та Дмитро Гордон, а також порівняно ефективність знаходження інформації програмними засобами та вручну.

Ефективність запропонованої методики була апробована працівниками Департаменту інформаційно-аналітичної підтримки Національної поліції України та

було відзначено, що розроблена автором програма має практичну значущість, оскільки дозволяє здійснювати автоматизовану систематизацію публічно доступних даних з соціальних мереж і може бути використана в подальшій практичній діяльності цього Департаменту, з метою прийняття превентивних заходів із запобігання вчинення правопорушень, розкриття та розслідування злочинів.

Отже, у науково-дослідницькій роботі було підтверджено гіпотезу, що із сукупності публічно доступних даних можна автоматично вибрати їхню достатню кількість для отримання висновків щодо соціальних контактів, громадської позиції, професійної діяльності людини й дати їй об’єктивну оцінку. Отримані результати можуть бути використані різними категоріями юридичних та фізичних осіб у професійній діяльності, у тому числі спецслужбами, з метою охорони прав і свобод людини, а також інтересів суспільства й держави.

Усі завдання дослідження виконано. Код проєкту розміщено на GitHub-репозиторії [12].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Saul Schleimer, Daniel S. Wilkerson, Alex Aiken. Winnowing: Local Algorithms for Document Fingerprinting. 2003, 10 с. URL: https://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf.
2. Ryan Mitchell. Web Scraping With Python: Collecting More Data From The Modern Web. O'Reilly, 2018, 308 с.
3. David Beazley, Brian K. Jones. Python Cookbook, Third Edition. O'Reilly, 2013, 706 с.

ІНТЕРНЕТ-РЕСУРСИ

4. Офіційна документація Python // Сервер Python Software Foundation. URL: https://docs.python.org/3/ (дата звернення: 17.06.2020).
5. Twitter API v.1.1 // Сервер компанії Twitter. URL: https://developer.twitter.com/en/docs/twitter-api/v1 (дата звернення: 25.06.2020).
6. LinkedIn API // Сервер компанії LinkedIn. URL: https://legal.linkedin.com/api-terms-of-use (дата звернення: 02.07.2020).
7. Facebook API Graph // Сервер компанії Facebook. URL: https://developers.facebook.com/docs/graph-api/ (дата звернення: 19.06.2020).
8. Instagram API // Сервер компанії Instagram. URL: https://developers.facebook.com/docs/instagram-basic-display-api (дата звернення: 20.07.2020).
9. Google Developers // Сервер компанії Google. URL: https://developers.google.com/ (дата звернення: 30.07.2020).
10. Programmable Google Search Engine // Сервер компанії Google. URL: https://programmablesearchengine.google.com/about/ (дата звернення: 02.08.2020).
11. fpdf2 // GitHub Pages. URL: https://alexanderankin.github.io/pyfpdf/ (дата звернення: 28.08.2020).
12. Код проєкту на GitHub // Сервер GitHub. URL: https://github.com/pandrey2003/social-media-profiler (дата звернення: 17.06.2020).

«Додаток А». Реалізація графічного інтерфейсу проєкту

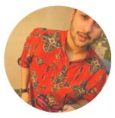
```
# Імпорт необхідних модулів
from PyQt5 import QtWidgets, uic
from PyQt5.QtWidgets import QMainWindow, QMessageBox
from app.backend.backend import main_backend


class Window(QMainWindow):
    def __init__(self):
        super().__init__()
        # Завантаження UI файлу
        uic.loadUi("app/design/ui/main.ui", self)
        self.setWindowTitle("SocialMediaProfiler")
        # Знаходження полей вводу
        self.first_name_input = self.findChild(QtWidgets.QLineEdit, "first_name_field")
        self.last_name_input = self.findChild(QtWidgets.QLineEdit, "last_name_field")
        self.location_input = self.findChild(QtWidgets.QLineEdit, "location_field")
        self.company_input = self.findChild(QtWidgets.QLineEdit, "company_field")
        self.job_title_input = self.findChild(QtWidgets.QLineEdit, "job_title_field")
        self.school_input = self.findChild(QtWidgets.QLineEdit, "school_field")
        self.twitter_input = self.findChild(QtWidgets.QLineEdit, "twitter_field")
        self.instagram_input = self.findChild(QtWidgets.QLineEdit, "instagram_field")
        self.additional_text = self.findChild(QtWidgets.QTextEdit, "additional_info_field")
        self.extra_field_name = self.findChild(QtWidgets.QLineEdit, "extra_field_name")
        self.extra_field_input = self.findChild(QtWidgets.QLineEdit, "extra_field_input")
        self.directory_text = self.findChild(QtWidgets.QLineEdit, "directory_text")
        # Знаходження кнопок
        self.select_button = self.findChild(QtWidgets.QPushButton, "select_directory")
        self.submit_button = self.findChild(QtWidgets.QPushButton, "submit_button")
        self.credits_from_menu = self.findChild(QtWidgets.QAction, "actionCredits")
        # Знаходження Qt ProgressBar
        self.progress_bar = self.findChild(QtWidgets.QProgressBar, "app_progress_bar")
        # Під’єднання кнопок до методів
        self.select_button.clicked.connect(self.__choose_directory_as_dialog)
        self.submit_button.clicked.connect(self.__send_input_to_backend)
        self.credits_from_menu.triggered.connect(self.__open_credits_message)
        # Власне методи див. в коді проекту
```

«Додаток Б». Повний PDF-звіт Kevin'a Goldsmith'a

Kevin Goldsmith
Seattle

Instagram
Potential users

• Instagram nickname: _kevin_goldsmith_
• Biography: 23 Creators and artists add colour to...
• Full name: Kevin Goldsmith
• Media count: 26
• Number of followers: 204
• Number of following: 297



• Instagram nickname: kkevin_goldsmith
• Biography: It's your PEG make it large Sarcastic?...
• Full name: Kevin Goldsmith
• Media count: 0
• Number of followers: 348
• Number of following: 8

• Instagram nickname: kevingoldsmith
• Biography:
• Full name: Kevin Goldsmith
• Media count: 376
• Number of followers: 372
• Number of following: 482

• Instagram nickname: kevin.goldsmith.146
• Biography:
• Full name: Kevin Goldsmith
• Media count: 0
• Number of followers: 2
• Number of following: 2

• Instagram nickname: goldsmith.kevin
• Biography:
• Full name: Kevin Goldsmith
• Media count: 0
• Number of followers: 0
• Number of following: 0

• Instagram nickname: kevingoldsmith68
• Biography:
• Full name: Kevin Goldsmith
• Media count: 0
• Number of followers: 0
• Number of following: 0

Twitter

• Screen name: KevinGoldsmith
• Description: CTO @ Anaconda & Speaker (ex: Onfido, Avvo,...
• Full name:
• Location: Seattle
• Number of followers: 3595
• Number of following: 1456
• External URL: https://t.co/KYDeZCA0Q6
• Two last posts:
- RT @jodyrodgers: I've replaced the election Twitter news...
- RT @anacondainc: We look forward to leveraging @eodaGmbH's...

LinkedIn
Potential user(s)

Kevin Goldsmith
Headline: Chief Technology Officer (CTO) at....
Industry: Computer Software.
Location: .
Work experience:
• *Anaconda, Inc.*
Job title: Chief Technology Officer.
Company location: Austin, Texas, United States.
Time period: 10/2020 - Now.
• *Nimble Autonomy, LLC*
Job title: Founder / Principal.
Company location: Greater Seattle Area.
Time period: 07/2020 - Now.
• *Unit Circle Media*
Job title: Owner.
Company location: Greater Seattle Area.
Time period: 01/1991 - Now.
• *Onfido*
Job title: Chief Technology Officer.
Company location: London, United Kingdom.
Time period: 04/2019 - 06/2020.
• *cavnessHR*
Job title: Member, Board of Advisors.
Company location: Greater Seattle Area.
Time period: 06/2018 - 01/2020.

Education:
• *BS*
School name: Carnegie Mellon University.
Time period: 1988-1992.

Skills: Software Development, Mobile Applications, Agile Methodologies, Software Engineering, Scrum, C++, Software Design, Distributed Systems, Cloud Computing, Algorithms, Scalability, Mobile Devices, Architecture, Web Services, Windows, C, Objective-C, Digital Imaging, MySQL, GPGPU, Technical Leadership, Object Oriented Design, Digital Media, High Performance Computing, 3D, Multithreading, OpenGL, iOS development, Architectures, Digital Asset Management, Digital Image Processing, Lean Management, Kanban, Lean Thinking, Application Architecture, Human-computer Interaction, Parallel Algorithms, Language Design, Media Technology, Grid Computing, Parallel Computing, Extreme Programming, Lean Software Development, Leadership, Cross-functional Team Leadership, Organizational Leadership, Strategic Leadership, Thought Leadership, Charcuterie, Muay Thai.

Google Search
Information based on extra input
GitHub Profile

• Full name: Kevin Goldsmith.
• Nickname: kevingoldsmith.

«Додаток В». Довідка про впровадження від Національної поліції України

НАЦІОНАЛЬНА ПОЛІЦІЯ УКРАЇНИ
ДЕПАРТАМЕНТ ІНФОРМАЦІЙНО-АНАЛІТИЧНОЇ ПІДТРИМКИ

вул. Богомольця, 10, м. Київ, 01601,
тел. 254-14-40, duty.it@police.gov.ua

01 грудня 2020 року № 8166/27/02/1-2020

На № _____ від _____

Про впровадження результатів наукового дослідження

ДОВІДКА

про впровадження результатів наукового дослідження в практичну діяльність Департаменту інформаційно-аналітичної підтримки Національної поліції України

У роботі учня 11-А класу ліцею № 142 м. Києва Полухіна Андрія Вячеславовича на тему: «Автоматизація процесу збору та систематизації публічно доступної інформації з соціальних мереж» запропоновано ряд підходів щодо збору та узагальнення інформації з метою підвищення ефективності її використання, у тому числі для попередження та боротьби зі злочинністю.

Запропонована методика має практичну цінність, оскільки дозволяє здійснювати автоматизовану систематизацію публічно доступної інформації з соціальних мереж для проведення подальшого її аналізу з метою прийняття превентивних заходів із запобігання вчинення правопорушень, розкриття та розслідування злочинів.

Проведений аналіз та розроблені на його базі пропозиції щодо підвищення ефективності використання зазначеної інформації можуть бути використані в подальшій практичній діяльності Департаменту інформаційно-аналітичної підтримки Національної поліції України для здійснення інформаційної підтримки діяльності поліції із забезпечення публічної безпеки і порядку, охорони прав і свобод людини, а також інтересів суспільства і держави та протидії злочинності.

Перший заступник начальника Департаменту **Вадим НАУМОВ**