Urząd Statystyczny
w Kielcach

Warszawa, 23-24 marca 2023 r.
Konferencja
Statystyka w służbie polityki
społeczno-gospodarczej:
nowe badania, metody i narzędzia

Web scraping jako dodatkowe źródło danych w badaniach statystyki publicznej

Jak szybko zebrać trudnodostępne dane o rynku prostytucji?

Urząd Statystyczny w Kielcach
Małgorzata Sobieraj

PLAN PREZENTACJI

- O Ośrodku Badań Gospodarki Nieobserwowanej
  obowiązek szacowania działalności nielegalnej
- O źródłach danych
- O narzędziach zbierania danych
  dlaczego web scraping?
  etapy wdrożenia web scrapingu
- Co dalej?

OGN

Poznajmy się …

OŚRODEK BADAŃ GOSPODARKI NIEOBSERWOWANEJ

Specjalizacje Urzędu Statystycznego w Kielcach

- Statystyka handlu i usług
- Gospodarka nieobserwowana
  - szara gospodarka
  - działalność nielegalna
    - prostytucja
    - przemyt papierosów i alkoholu
    - narkobiznes

PODSTAWY PRAWNE

- Rozporządzenie Parlamentu Europejskiego i Rady (UE) nr 549/2013 z dnia 21 maja r. w sprawie europejskiego systemu rachunków narodowych i regionalnych w Unii Europejskiej
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 20198/516/2013 z dnia 19 marca 2019 r. w sprawie harmonizacji dochodu narodowego brutto w cenach rynkowych oraz uchylające dyrektywę Rady 89/130/EWG, Euratom i rozporządzenie Rady (WE, Euratom)

DEFINICJA DZIAŁALNOŚCI NIELEGALNEJ

DZIAŁALNOŚĆ NIELEGALNA *illegal activity*

Działalność produkcyjna polegająca na wytwarzaniu towarów i/lub usług, zabroniona przez prawo.

Obejmuje ona:

- produkcję wyrobów i usług, których sprzedaż, rozprowadzanie lub posiadanie są zabronione przez prawo
- działalność produkcyjną, która jest zwykle legalna, lecz staje się nielegalna, gdy jest wykonywana przez producentów nie mających do tego prawa (na przykład praktyka medyczna bez licencji)
- zgodnie z zaleceniami Eurostatu w pierwszej kolejności kraje UE powinny uwzględniać w rachunkach narodowych: prostytucję, narkomanię i przemyt

PROSTYTUCJA – ASPEKTY PRAWNE

Prostytucja w Polsce jest legalna

Przepisy Kodeksu karnego: czerpanie korzyści z prostytucji jest karalne

- Art. 203 nakłanianie do uprawiania prostytucji
- Art.204 §1 czerpanie korzyści majątkowych z uprawiania prostytucji przez inną osobę (sutenerstwo)
- Art.204 §2 wykorzystywanie małoletnich
- Art.204 §4 uprowadzenia za granicę

OGN

Analiza źródeł danych

PRZEGLĄD ŹRÓDEŁ DANYCH

Badania własne

prowadzone na reprezentatywnej próbie populacji generalnej – najlepsze źródło danych

- badania podaży usług
- badania popytu na usługi

Informacje z obcych źródeł

- jednorazowe publikacje
- raporty policji
- prasa, internet, telewizja
- raporty z badań ośrodków naukowych

OCENA PRZYDATNOŚCI DANYCH

Trudności

- dane wyrywkowe
- dane dotyczące tylko jednego aspektu badanego zjawiska
- badania prowadzone na niewielkiej grupie, której liczebność trudno ustalić
- dane pochodzące z różnych lat, które trudno jest porównywać ze sobą

DANE O CENACH USŁUG

Pierwsze próby...

Lp.	Rok	Ceny	Opis	Źródło	Wiarygodność
1	1998	100/godz.	Zarobki prostytutki w porządnej agencji. Ona bierze z tego 50 zł	Ośka- Internet	wysoka
2	1998	4 000 zł	Zarobki miesięczne prostytutki	Ośka- Internet	wysoka
3	1998	150/godz.	Rosjanka w agencji. Ona bierze z tego 30 zł.	Ośka- Internet	wysoka
4	2000	40-50 zł	cena usług tirówek	polityka.onet.pl - Internet	średnia
5	2000	800-1000 zł	dochód sutenera od jednej tirówki	polityka.onet.pl - Internet	średnia
6	2000	20-25	ilość klientów tirówki dziennie	polityka.onet.pl - Internet	średnia
7	2000	4,5 mld zł rocznie	Szacunki La Strady - wartość rynku prostytucji w Polsce	polityka.onet.pl - Internet	średnia
8	2000	800 zł	Szacunki La Strady; średni dzienny dochód sutenera od jednej prostytutki	polityka.onet.pl - Internet	średnia
9	2001	500-1000 zł dziennie	zarobki TIR-ówek	Internet	średnia
...					

OGN

Narzędzia do zbierania danych

NOWOCZESNE TECHNIKI ZBIERANIA DANYCH

BIG DATA

Duże, złożone zbiory danych, pochodzących zwłaszcza z nowych źródeł. Zbiory te są tak obszerne, że tradycyjne oprogramowanie do przetwarzania danych po prostu nie jest w stanie nimi zarządzać. W szczególności - analiza i ocena danych w celu uzyskania użytecznych informacji na potrzeby kolejnych analiz.

DATA MINNING

Eksploracja danych, pozyskiwanie danych, drążenie danych, wydobywanie danych. Idea eksploracji danych polega na wykorzystaniu szybkości komputera do znajdowania ukrytych dla człowieka prawidłowości w danych zgromadzonych w hurtowniach danych.

MACHINE LEARNING

Uczenie maszynowe analizuje wzorce i korelacje. Uczy się na ich podstawie i optymalizuje się w miarę upływu czasu. Pozwala: tworzyć modele szybciej, precyzyjniej, w pełni automatycznie, analizować większą ilość danych, o bardziej złożonych strukturach, otrzymywać dokładniejsze, o wiele bardziej użyteczne wyniki.

WEB SCRAPING

Z języka angielskiego - Web scraping, web harvesting lub ekstrakcja danych internetowych to wydobywanie danych ze stron internetowych. Oprogramowanie Web scraping może uzyskać bezpośredni dostęp do sieci WWW za pomocą protokołu Hypertext Transfer Protocol lub przeglądarki internetowej.

WIĘCEJ O WEB SCRAPINGU

Zalety

wypracowanie przewagi konkurencyjnej

- szybkość pozyskiwania danych
- elastyczność - dostosowanie scraper'a do swoich potrzeb
- wykorzystanie do uczenia swoich sieci neuronowych
- skalowalność - pozyskiwanie danych na dużą skalę
- niski koszt uzyskania informacji - może wystarczyć nawet najmniejszy serwer VPS za kilka złotych
- integralność danych - eliminacja błędów ludzkich podczas przenoszenia danych
- ustrukturyzowane dane - dzięki nałożonej strukturze dane mogą być poddawane obróbce przez inne oprogramowanie np. służące do analizy danych

WIĘCEJ O WEB SCRAPINGU

Wady

- znalezienie gotowego rozwiązania lub stworzenie go w krótkim czasie przy bardzo specyficznych danych może przekroczyć nasze chęci lub możliwości
- konieczność monitorowania i kontrolowania pod kątem poprawności zbieranych danych (szczególnie na początkowym etapie pozyskiwania informacji)
- niektóre strony internetowe „nie lubią” być scrapowane i starają się blokować podejrzane działania (omijanie zabezpieczeń – powszechnie korzysta się z puli serwerów proxy)

OGN

Wdrożenie web scrapingu

PO PIERWSZE: POMYSŁ

Burza mózgów

- Problemy z regularnym dostępem do danych
- Chęć uniezależnienia się od zewnętrznych źródeł danych dotyczących cen za usługi seksualne (dane firmy Sedlak &Sedlak)
- Dysponowanie „własnym” źródłem danych

PO DRUGIE: DYSKUSJA

Pytania, kwestie sporne...

- Jak znaleźć słowa kluczowe?
- Jakich stron internetowych szukać?
- Jak „ominąć” etycznie nieakceptowalne treści/obrazy?
- Jak zapisywać potencjalne pozyskane dane?

„Zacznij od robienia tego, co konieczne,
potem zrób to, co możliwe,
a nagle stanie się to, co niemożliwe.”

Franciszek z Asyżu

PO TRZECIE: REALIZACJA

Etapy prac – przygotowanie do pozyskania danych

- „zgoda” w sprawie słów kluczowych (tagów) i stron internetowych
- analiza zdefiniowanych źródeł danych (budowy stron)
- wytypowanie strony o hierarchicznej, dobrze zorganizowanej budowie zawierającej ogłoszenia prostytutek
- opracowanie algorytmu komputerowego pozwalającego na automatyczne pozyskanie interesujących nas informacji
- napisanie skryptów w języku PYTHON pozwalających na maszynowe gromadzenie danych

PO TRZECIE: REALIZACJA

Etapy prac – dane pozyskane

- proces ekstrakcji danych
- zapis wyodrębnionych danych do tymczasowej bazy danych
- transformacja pozyskanych danych do postaci umożliwiającej załadowanie do określonej struktury danych
- załadowanie przekształconych danych do bazy danych (grupowanie danych)
- walidacja pozyskanych danych

REALIZACJA

Dane pozyskane – pierwsze efekty...

```
nr_pokaz;telefon;wiek;ceny;waga;wzrost;miasto;biust;język;1 godzina;15 min;30 min;cała noc453108;573 
332617;517 985 912;21;'120 zł';54;170;Warszawa;4;-;120 zł;;;
392495;793 100 734;24;'500 zł';50;180;Olesno;4;angielski,;500 zł;;;
398686;507 584 190;21;'200 zł', '130 zł', '160 zł', '1600 zł';52;170;Bydgoszcz;3;angielski,
299540;576 960 347;36;'150 zł', '40zł';56;153;Wrocław;4;-;150 zł;;;
268066;576 043 509;38;'200 zł', '150 zł';45;172;Katowice;2;angielski,       niemiecki,
393978;739 694 749;23;'200 zł', '1600 zł';54;167;Warszawa;3;angielski,;200 zł;;;1600 zł
456032;733 412 977;21;'150 zł';54;169;Kraków;3;-;150 zł;;;
452533;518 804 176;49;'140 zł', '70 zł', '100 zł';80;166;Bytom;5;-;140 zł;70 zł;100 zł;
450841;577 3[illegible]23;'200 zł', '2000 zł';50;167;Poznań;3;angielski,;200 zł;;;2000 zł
382031;537 [illegible]'100 zł', '100 zł', '100 zł', '1500 zł';60;178;Szczecin;3;-;100 zł;100 zł;100 z
141414;798 [illegible]0 zł', '100 zł', '100 zł';60;168;Katowice;4;angielski,;150 zł;100 zł;100 zł;
269125;57[illegible] 466;[illegible]0 zł', '2000 zł';50;172;Kraków;3;angielski,;200 zł;;;2000 zł
421[illegible] 794;41;'150 zł', '120 zł';75;170;Kraków;5;-;150 zł;;120 zł;
[illegible]1081;664 [illegible];35;'200 zł', '130 zł';60;170;Nowy Targ;3;-;200 zł;;130 zł;
441841;798 [illegible]2 [illegible]28;'150 zł', '100 zł', '120 zł', '1200 zł';57;172;Łódź;5;niemiecki,;150 zł;100 zł;
273146;66[illegible]0 0 [illegible]22;'150 zł', '100 zł', '120 zł', '1200 zł';50;168;Leszno;3;-;150 zł;100 zł;120 zł;
```

REALIZACJA

Dane pozyskane – porządkowanie

nr_pokaz	telefon	wiek	ceny	miasto	biust	język	1 godzina	15 min	30 min	cała noc
219038	797 752 655	26	0 , 0 , 150 , 80 , 100 , 80 , 0 , 0 , 0 , 0	Kraków	4	angielski, francuski,	150	80	100	80
95896	723 798 447	39	100 , 80 , 100	Toruń	5	angielski, niemiecki,	100		80	100
260000	576 708 463	19	200 , 200 , 200 , 200	Katowice	5	angielski,	200	200	200	200
359992	576 708 463	20	200 , 200 , 200 , 200	Katowice	5	angielski,	200	200	200	200
450481	884 630 996	27	150 , 100 , 150 , 200	Piła	2	-	150	100	150	200
459422	504 488 660	20	300 , 300 , 2300 , 200, 50	Lublin	2	angielski, niemiecki, hiszpański,	300			300
417657	533 784 723	22	250 , 300	Bielsko-Biała	2	angiel	2			300
317341	739 059 758	22	100 , 80 , 300 , 50	Rybnik	3	-			80	300
443721	884 045 071	28	400, 400, 200 , 150 , 400 , 400, 400, 400, 400, 400, 400, 30	Oleśnica	4	angielski, niemiecki,			150	400

PO CZWARTE: MAMY TO!

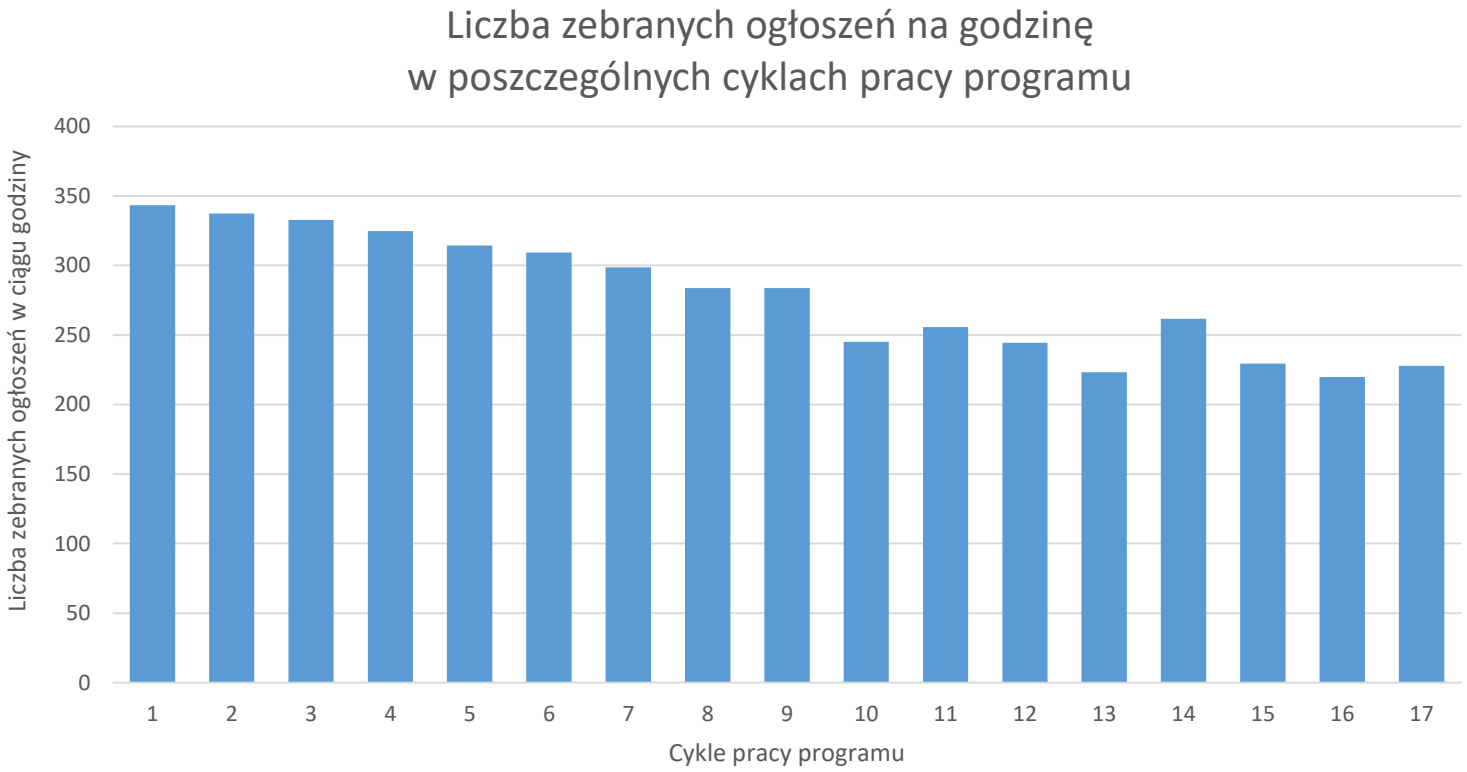
Zebrano 12733 (10380) rekordów danych

- stawki: 15min, 30min, 1 godzina, cała noc
- kategorie: wiek, wzrost, waga, znajomość języków obcych
- podział terytorialny: miejscowość, województwo

Wykorzystano ceny usług do szacunku przychodów

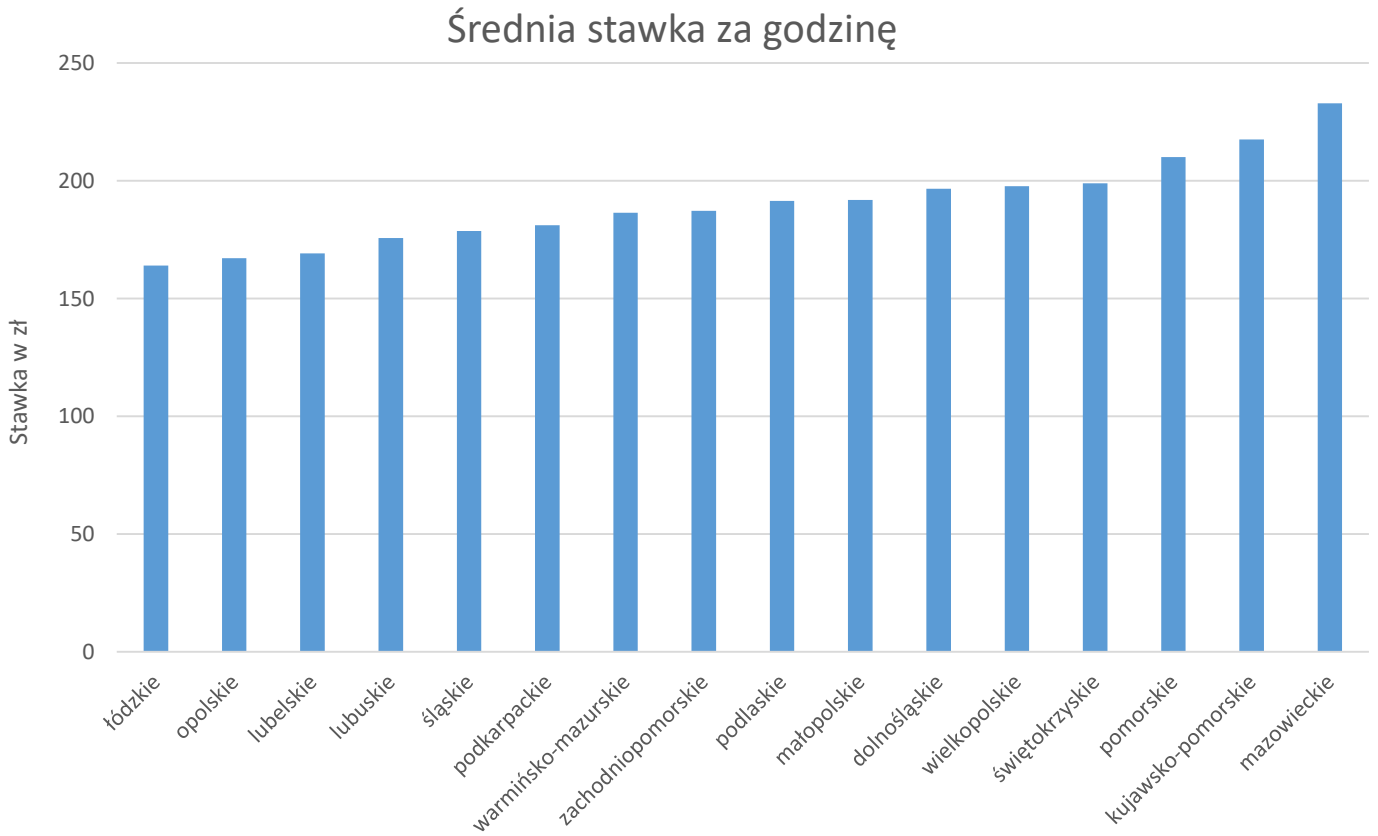
PREZENTACJA WYNIKÓW

Efektywność działania programu



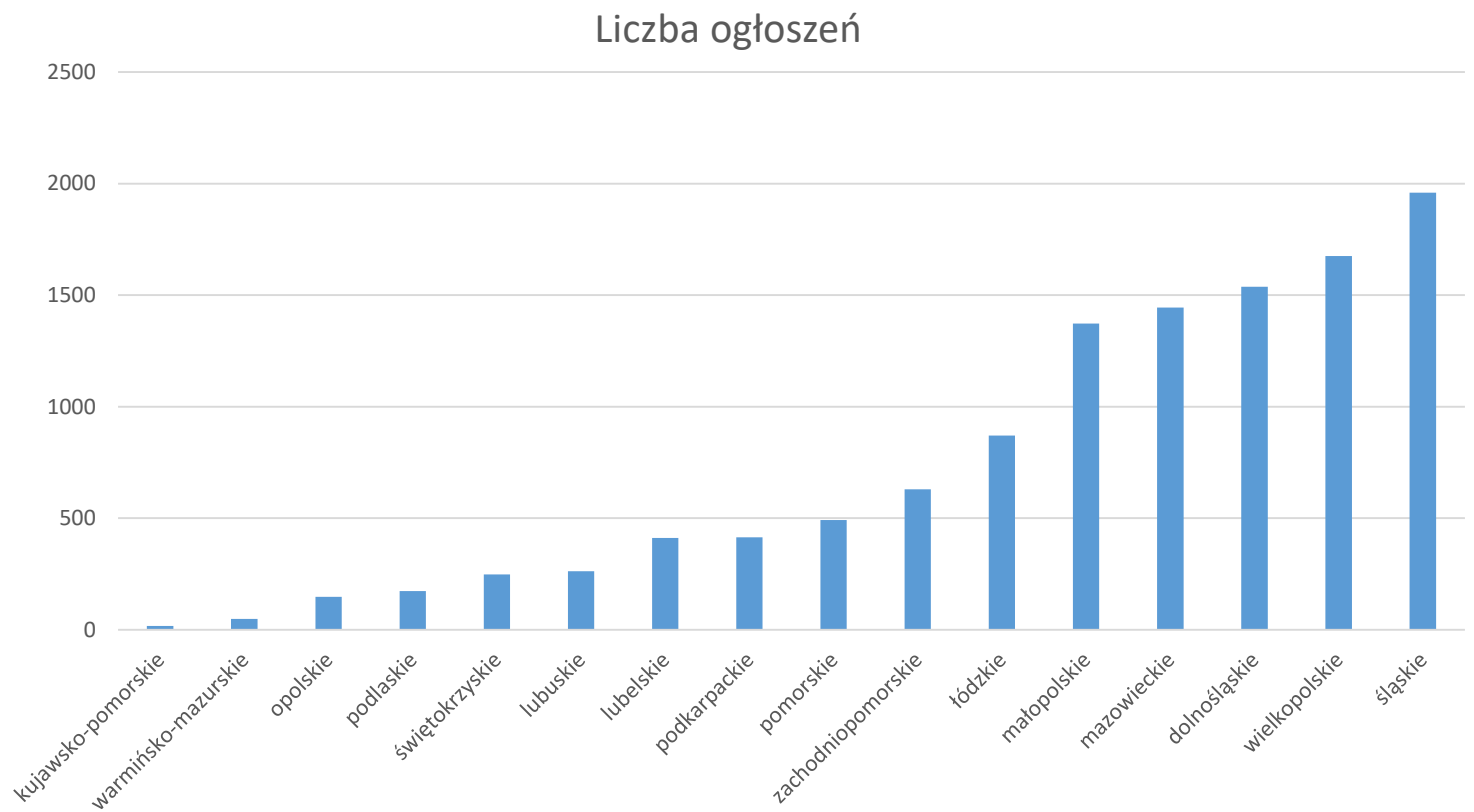
PREZENTACJA WYNIKÓW

Stawki za godzinę usług wg województw



PREZENTACJA WYNIKÓW

Liczba zebranych ogłoszeń na godzinę w poszczególnych cyklach pracy programu



SZACUNKI DZIAŁALNOŚCI NIELEGALNEJ W RACHUNKACH NARODOWYCH

Prostytucja

Wyszczególnienie	2017	2018	2019	2020
PRODUKT KRAJOWY BRUTTO (łącznie z gospodarką nieobserwowaną)	100,0	100,0	100,0	100,0
Ogółem gospodarka nieobserwowana	12,5	12,1	11,1	10,4
działalność nielegalna	0,4	0,4	0,4	0,4
prostytucja	0,04	0,04	0,04	0,04

Źródło: *„Rachunki narodowe wg sektorów i podsektorów instytucjonalnych w latach 2017-2020”, Warszawa, lipiec 2022*

POWRACAJĄCY PROBLEM

Czy musimy zaczynać od nowa...?

Brak uniwersalnej metodologii

- niestabilność źródeł danych
- triangulacja źródeł danych
- brak realnych możliwości potwierdzenia jakości uzyskanych danych

Wstrzymanie web scrapingu

- nieodwołalne zamknięcie portalu z ogłoszeniami erotycznymi Roksa.pl
- zatrzymanie 11 osób związanych z Roksa.pl, w tym właściciela firmy i jego brata
- zarzuty cyber-sutenerstwa (Roksa pobierała opłatę 50 zł miesięcznie za każde ogłoszenie)
- zarzuty kuplerstwa

PODSUMOWANIE

„Wszystko jest trudne zanim stanie się łatwe”
J.W. Goethe

Przydatność web scrapingu do pozyskiwania danych na potrzeby szacunków prostytucji

- określenie przeciętnych stawek za usługi
- zbadanie struktury zjawiska prostytucji w Polsce
- dostarczenie informacji na temat podaży usług prostytucji w Polsce
- pozyskanie informacji na temat popytu na usługi prostytucji w Polsce
- dostęp do darknetu jako alternatywa...?

„Zacznij od robienia tego, co konieczne....
Franciszek z Asyżu

Dziękuję za uwagę

m.sobieraj@stat.gov.pl