Techniki analizy sieci społecznych
Projekt 2 – Grupowanie ukraińskich kanałów w serwisie Telegram

Jakub Ciemięga
Maciej Jabłoński
Natan Orzechowski

Semestr zimowy 2022

Spis treści

1 Wstęp

Temat projektu brzmiał: „Dokonaj grupowania ukraińskich kanałów serwisu Telegram, łącząc cytujące te same źródła zewnętrzne. Wykonaj badania dla różnych przekrojów profili (najpopularniejsze, oficjalne, tematyczne, regionalne itp.).

Kod źródłowy projektu w języku Python znajduje się w repozytorium GitHub.

2 Realizacja projektu

2.1 Wytypowanie źródeł danych i technologii ich pobrania i przechowania

Kanały w serwisie Telegram funkcjonują jako tablice jednostronnej komunikacji; dla użytkownika widoczne są jako regularne okno konwersacji, jednak jedynie właściciel (lub grupa administratorów) mogą wysyłać treści na danym kanale, który może być subskrybowany przez dowolną liczbę użytkowników.

W ramach projektu stworzona została lista 106 ukraińskich kanałów. Lista została stworzona ręcznie, przy użyciu następujących źródeł:

- https://detector.media/monitorynh-internetu/article/202968/2022-09-20-from-trukha-to-gordon-the-most-popular-channels-of-the-ukrainian-telegram/
- https://imi.org.ua/en/news/eight-of-ten-most-popular-telegram-channels-in-ukraine-are-anonymous-imi-research-i41304
- https://telegram.wroclaw.pl/lista-kanalow-i-grup-informujacych-o-wojnie-na-ukrainie/
- https://caluniv.in/ukraine-telegram-group-channel
- https://telegramguide.com/ukraine-telegram-group-link/
- wyszukiwanie nazw ukraińskich miast bezpośrednio w serwisie Telegram.

Do pobrania danych wykorzystano narzędzie telegram-api. Na podstawie stworzonej listy nazw kanałów w pliku tekstowym, umożliwiło ono pobranie w formacie JSON:

- metadanych kanału (nazwa, id, liczba subskrybentów a także szerego przydatnych danych statystycznych, np. czy konto jest zweryfikowane itp.),
- wszystkich postów z kanału (łącznie z datą, treścią i dodatkowymi metadanymi).

Następnie na podstawie pobranych w ten sposób plików stworzono strukturę słownikową, gdzie zbiorem kluczy był zbiór nazw kanałów, a każdemu kluczowi przypisana była odrębna struktura słownikowa, w której zapisano dane danego kanału (np. liczbę subskrybentów, liczbę postów itp.) W szczególności, każdemu kanałowi przypisano również listę cytowanych domen internetowych oraz listę liczności cytowań poszczególnych domen, aby umożliwić tworzenie połączeń między kanałami na tej podstawie.

Cytowane domeny zostały wyznaczone poprzez wyszukanie adresów URL we wszystkich wiadomościach danego kanału przy użyciu wyrażenia regularnego[1]. Następnie znalezione adresy zostały przetworzone tak, by zawierać jedynie nazwy domen (np. „youtube.com"). Wyjątkiem stanowiły tu adresy portalu Twitter, dla których zapisano również nazwy cytowanych użytkowników (sama informacja o cytowaniu postów z portalu Twitter nie byłaby zbyt użyteczna).

Uzyskana w ten sposób struktura słownikowa była przechowywana w pliku tekstowym w formacie JSON i stanowiła podstawę dla działań wykonanych na dalszych etapach projektu.

[1] https://stackoverflow.com/questions/839994/extracting-a-url-in-python

2.2 Tworzenie połączeń między danymi

Na podstawie uzyskanego wcześniej pliku *channels.json* zawierającego słownik z danymi poszczególnych kanałów dokonano łączenia par kanałów na różne sposoby. Oznaczone są one jako *Mode1*, *Mode2*, *Mode3*, *Mode4*. Ich opis przedstawiony zostanie poniżej.

Jako lista połączeń zwracana jest lista list w formacie $[channel1\_id, channel2\_id, weight]$, zawierająca parę kanałów oraz wagę połączenia pomiędzy nimi. Jeśli dwa kanały nie są połączone według metryki, to taka sublista nie jest dodawana do wynikowej listy połączeń. Jeśli waga połączenia wynosi 0, to znaczy, że kanały nie są połączone.

Dodatkowo zdecydowano się na pogrupowanie kanałów ze względu na ich atrybuty. Wybrane atrybuty to *verified* (przyjmuje wartości *True*, *False*; czy kanał został zweryfikowany przez Telegram) oraz *subscribers_count* (liczba subskrybentów - pozwala podzielić kanały na popularne i małopopularne). Niestety nie udało się ustalić dla kanałów np. ich lokalizacji lub innej podobnej metryki pozwalającej na ich podział.

2.2.1 Opis atrybutów wykorzystanych do tworzenia połączeń

Do tworzenia połączeń między kanałami użyto następujących atrybutów kanału:

- verified – czy zweryfikowany: *True*/*False*
- subscribers_count – liczba subskrybentów (na 19.12.2022)
- cited_sources – lista źródeł/domen w formie ["source1", "source2", ...]
- citation_count – liczba cytowań poszczególnych domen z listy cited_sources w formie [3, 15, 1, ...]

Dodatkowo dla każdego kanału wyliczone zostały takie metryki, jak:

- n_all_citations - liczba wszystkich cytowań
- n_all_sources - liczba wszystkich źródeł/domen

2.2.2 Mode1

Dwa kanały są połączone, jeśli cytują minimum **X** wspólnych źródeł (minimum **X** różnych domen).

Waga połączenia **= 1**.

2.2.3 Mode2

Dla każdego z kanałów ($i = 1, 2$) wyliczany jest współczynnik

$$\text{Waga}_i = \frac{\text{liczba cytowań ze wspólnych źródeł}_i}{\text{liczba wszystkich cytowań}_i}$$

mówiący jaki procent cytowań na kanale i pochodzi ze źródeł, które występują w obu kanałach jednocześnie.

Następnie wybierana jest niższa z dwóch wartości (procentowych):

$$\text{Waga} = min(\text{Waga}_1, \text{Waga}_2)$$

Wtedy dwa kanały są połączone, jeśli **Waga > X**. **Waga** jest wagą połączenia.

2.2.4 Mode3

Dwa kanały są połączone, jeśli cytują minimum jedno wspólne źródło.

Waga połączenia **= liczba wspólnych źródeł**.

2.2.5 Mode4

Dla każdego z kanałów ($i = 1,2$) wyliczany jest współczynnik

$$\text{Waga}_i = \frac{\text{liczba wspólnych źródeł}_i}{\text{liczba wszystkich źródeł}_i}$$

mówiący jaki procent źródeł kanału i jest wspólny dla obu kanałów.

Następnie wybierana jest niższa z dwóch wartości (procentowych):

$$\text{Waga} = min(\text{Waga}_1, \text{Waga}_2)$$

Wtedy dwa kanały są połączone, jeśli **Waga > X**. **Waga** jest wagą połączenia.

2.3 Analiza danych i Wnioski

Znajdują się w dalszej części, w PDFie z notatnika Jupyter.