Politechnika Wrocławska
Wydział Inormatyki i Telekomunikacji
Katedra Sztucznej Inteligencji

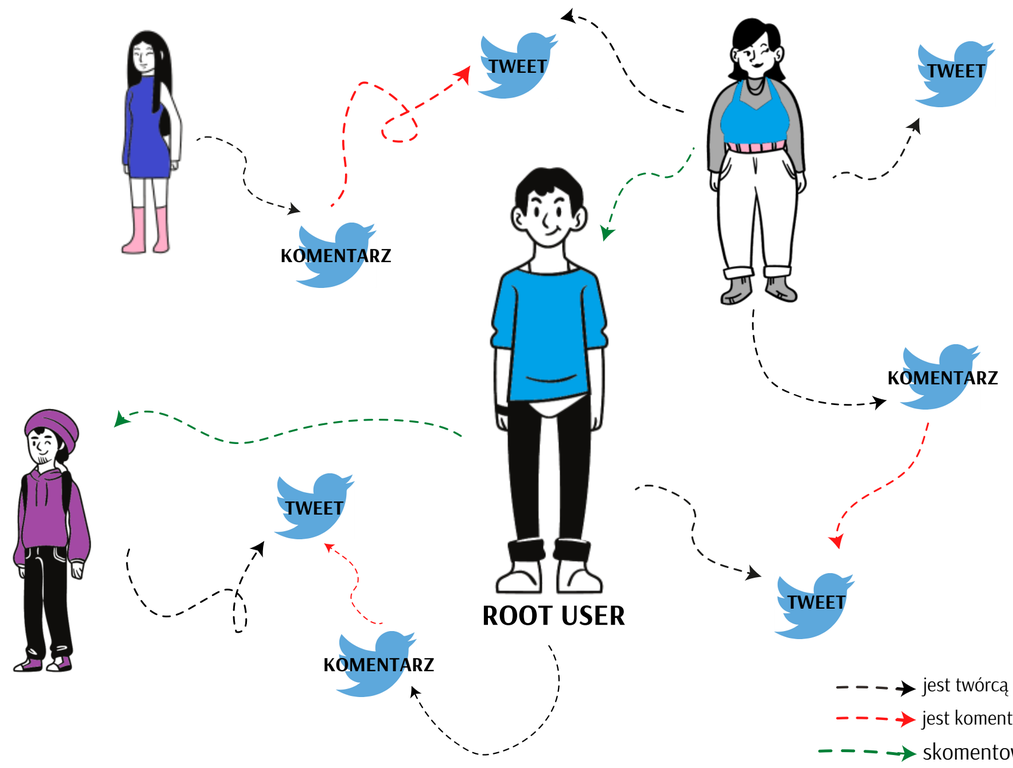
O czym ćwierka graf?

Joanna Waczyńska, Patryk Rygiel, Sztuczna Inteligencja, Wydział Informatyki i Telekomunikacji, Politechnika Wrocławska
Opiekunowie projektu: prof. Tomasz Kajdanowicz, w ramach kursu Analiza Mediów Społecznościowych.

IDEA

Personalizacja rekomendacji jest jednym z najbardziej kluczowych elementów w analizie mediów społecznościowych. Analizy, opiera się na informacji o użytkownikach oraz udostępnianych przez nich treści. Warto zauważyć, że na statystyki dotyczące postów (takie jak liczba lajków) wpływa nie tylko co piszemy, ale także w jakim "otoczeniu" się znajdujemy.

W tej pracy badamy wpływ wykorzystania informacji o zależnościach w sieci społecznościowej Twitter'a na budowanie lepszej reprezentacji użytkowników i postów w celu predykcji liczby lajków.

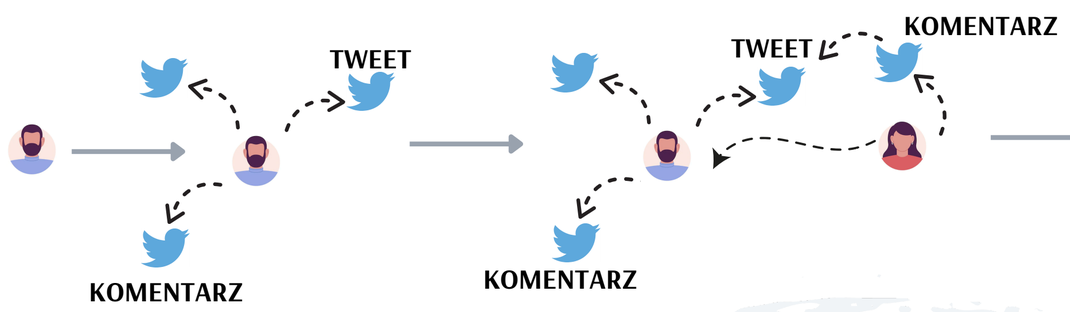


PROCES ZBIERANIA DANYCH

Sieci społecznościowe tworzą **heterogeniczną sieć informacyjną**, w której węzły mogą reprezentować użytkowników, udostępnianą przez nich treść, a krawędzie reprezentują interakcji między nimi [1]. My rozpatrujemy 3 typy relacji: "jest twórcą", "jest komentarzem", "skomentował".

Graf przedstawiający sieć informacyjną Twitter'a był budowany iteracyjnie w następujących krokach:

1. Wybór użytkownika bazowego - *root user*
2. Pobranie jego aktywności przy użyciu API Twittera [2]
3. Budowa sieci społecznej według poniższego schematu:



Schemat został powtórzony wielokrotnie. Tym sposobem pozyskaliśmy sieć z około **8 tysięcy użytkownikami oraz z ok. 220 tysiącami tweetami**.

Warto zwrócić uwagę na limitację jaką jest możliwość zbierania danych tylko z **ostatnich 7 dni**.

EKSPERYMENTY & WYNIKI

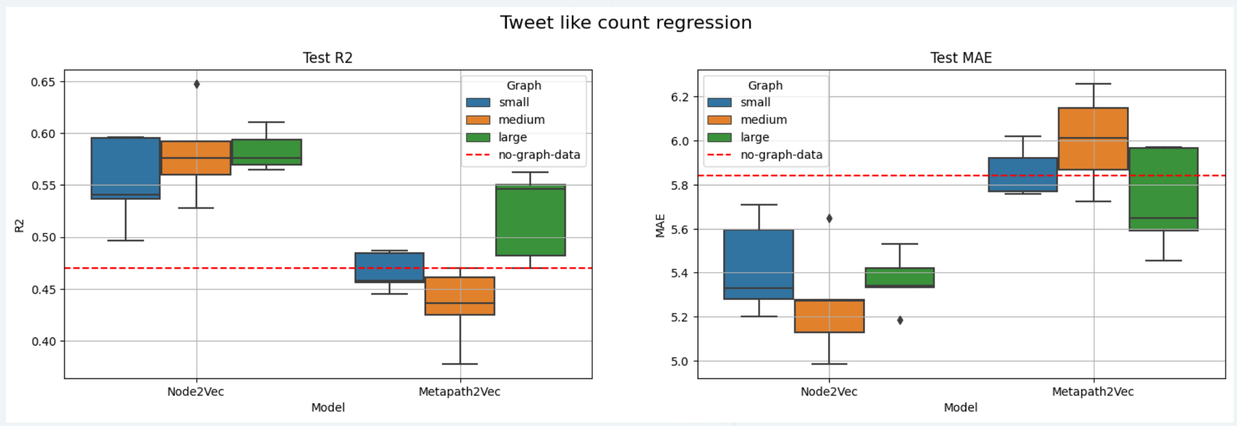
Zadanie:
Przewidywanie liczby lajków tweeta minimum po dwóch dniach od publikacji postu.

Ze względu na ograniczenia, podczas zbierania danych zdecydowaliśmy się na predykcję mało popularnych wpisów tj takich, które nie osiągają więcej niż 300 lajków.

Metodologia:

- Wybór i ekstrakcja cech Tweet'ów np. język, treść (reprezentacja tworzona za pomocą *all-MiniLM-L6-v2*); oraz cech autora np. ilość obserwujących, ilość postów.
- Budowa macierzy osadzeń węzłów w grafie przy użyciu **Node2Vec** (graf homogeniczny) oraz **Metapath2Vec** (graf heterogeniczny).
- Nauka regresora (*ExtraTreeRegressor)* na danych z osadzeniami grafowymi i bez.

Ponieważ zbieranie danych odbywało się iteracyjnie, a użytkownicy *"root users"* byli wybierani tak, aby zagęścić graf. Zgodnie z oczekiwaniami im większy graf jest rozważany tym przy zastosowaniu modelu **Node2Vec** dostajemy lepszą reprezentację wierzchołka - i bardziej stabilne wyniki regresji. Nie jest to jednak oczywiste w przypadku zastosowania **Metapath2Vec**. Warto zauważyć, że błąd MAE to tylko różnica 5 lajków. Jest to bardzo dobry i rozsądny rezultat.



WNIOSKI

Sieci społeczne są ogromną skarbnicą wiedzy. Warto mieć na uwadze, że na statystyki udostępnianej przez nas postów ma wpływ wiele więcej czynników niż tylko treść. Zagregowana informacja, pozyskana poprzez analizę naszego sąsiedztwa w sieci może znacząco poprawić jakości predykcji.

Limitacje:

- Ponieważ nie zajmujemy się grafem dynamicznym przyjmujemy, że aktywność wpisu ustabilizowuje się minimum 2 dni.
- Możliwe też, że użytkownik wystawił komentarz do tweeta, który powstał dawniej niż 7 dni temu. Istnieje więc duże ryzyko, że nie będziemy mieć dostępu do informacji reprezentującej wierzchołek.
- Wiele wpisów posiada informację wielomodalną (np. obrazek, filmik), której nie przetwarzamy - tracąc informację o kontekście.

[1] TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation, Ahmed El-Kishky et all
[2] https://developer.twitter.com/en/docs/twitter-api

SPRAWDŹ MNIE