TASS – Techniki Analizy Sieci Społecznych

Koncepcja

Analiza porównawcza sieci współpublikowania z siecią autorów patentów

Aleksander Łosiewicz

Daniel Petrykowski

15.12.2019

1. Wstęp

W ramach projektu zostanie zbadane podobieństwo sieci badaczy wygenerowanej na zasadzie wspólnego patentowania z siecią badaczy wygenerowaną na zasadzie współpublikowania. W raporcie końcowym skupimy się na przedstawieniu zarówno różnic jak i podobieństw pomiędzy obiema sieciami.

2. Dane

2.1. Sieć badaczy wygenerowana na zasadzie wspólnego patentowania

Przez prowadzących zostały dostarczone dane dla pewnej grupy patentów w postaci plików xml. Możemy z nich wyczytać między innymi informacje o numerze patentu, dacie publikacji, czy jego opis. Daie nie zawierają informacji o autorach patentów, które są potrzebne do stworzenia sieci. Informacja o autorach zostanie pozyskana ze strony Google Patents (https://patents.google.com/). W języku Python zostaną przygotowane skrypty łączące się ze stroną, wyszukujące dany patent i wyciągające dane jego autorów.

2.2. Sieć badaczy wygenerowana na zasadzie współpublikowania

Dla porównania sieci wynikającej z wspólnego patentowania i publikowania również potrzebna jest informacja o autorach publikacji. Zostanie ona uzyskana ze strony Google Scholar (https://scholar.google.com/). Również w języku Python zostaną przygotowane skrypty łączące się ze stroną, wyszukujące daną publikację i zwracające dane jej autorów.

3. Klucz łączący dane oraz metodyka uzyskania zadowalającego złączenia

Wspólnym kluczem łączącym wszystkie dane będą imiona i nazwiska autorów. Sieć wygenerowana na zasadzie wspólnego patentowania jest zdefiniowana przez prowadzących i na tej podstawie zostanie wygenerowana sieć wspólnego publikowania – będzie ona zawierała tylko autorów występujących w sieci wspólnego patentowania.

4. Metodyka analizy danych

Do stworzenia oraz analizy sieci zostanie użyty pakiet NetworkX. Każdy wierzchołek będzie przechowywał imię i nazwisko autora, a krawędzie będą obrazowały wspólne patenty/publikacje pomiędzy danymi autorami. Podczas tworzenia sieci zostaną z nich usunięte duplikaty krawędzi (powstające, kiedy dani autorzy stworzyli razem więcej niż jedną publikację/patent). Obie sieci zostaną porównane pod względem rzędu i rozmiaru, a także rozkładu stopni wierzchołków.

5. Repozytorium, postęp prac

Repozytorium zawierające aktualny postęp prac znajduje się pod linkiem https://github.com/danielpetrykowski/TASS. Na obecnym etapie prac powstały skrypty mające sprawdzić możliwości uzyskiwania danych o patentach i publikacjach ze stron Google Patents i Google Scholar. Są to jedne z największych publicznie dostępnych baz danych patentów o publikacji. Liczymy, że zawierają one informacje o autorach wszystkich polskich patentów i publikacji. W przypadku braku informacji o danym patencie informacja zostanie pozyskana ze strony Urzędu Patentowego Rzeczypospolitej Polskiej (https://grab.uprp.pl/SitePages/Start.aspx).

Skrypt pozyskujący informacje o patentach używa biblioteki google_patent_scraper (https://github.com/ryanlstevens/google_patent_scraper), a skrypt pozyskujący informacje o publikacjach biblioteki scholar (https://github.com/peterzjx/scholar.py). Ostatni skrypt został poprawiony tak, aby prawidłowo uzyskiwał informacje o publikacji w formacie bibtex, który zawiera

imiona i nazwiska wszystkich autorów publikacji. Oba skrypty korzystają z biblioteki Beautiful Soup (https://www.crummy.com/software/BeautifulSoup/bs4/doc/), która pozwala wyciągać i grupować informacje z plików HTML i XML. Do pobrania danych strony został użyty pakiet urllib (https://docs.python.org/3/library/urllib.html).