Wydział Elektroniki
i Technik Informacyjnych

POLITECHNIKA WARSZAWSKA

Systemy agentowe
Wstęp do eksploracji danych tekstowych w sieci WWW

System przeszukiwania projektów open source na podstawie danych z repozytoriów kodu

Sprawozdanie z projektu

Paweł Karwacki *259820*
Maciej Krasowski *259831*
Łukasz Kilaszewski *259822*

28 maja 2018

Wstęp

Podstawowym założeniem projektu było przygotowanie systemu agentowego, który będzie w stanie przeszukiwać projekty open source. Po podaniu wyszukiwanej frazy, system miał zwracać listę najbardziej dopasowanych projektów wraz z odnośnikami prowadzącymi do repozytoriów.

Kod źródłowy przygotowanego rozwiązania dostępny jest pod adresem `https://github.com/luktor99/SAG-WEDT`.

Dane do przeszukiwania

Jako źródło danych do przeszukiwania wybraliśmy platformę GitHub, głównie ze względu na ilość zgromadzonych tam publicznych repozytoriów oraz dostępność darmowego interfejsu API. Niestety interfejs ten posiada ograniczenie w postaci limitu 5000 zapytań na godzinę. Ponadto, pobieranie repozytoriów w czasie rzeczywistym wiąże się z dużym spowolnieniem działania systemu ze względu na opóźnienia generowane przez liczne zapytania HTTP.

Biorąc pod uwagę powyższe niedogodności, zdecydowaliśmy się na wcześniejsze pobranie pewnej puli repozytoriów. Dzięki temu właściwa aplikacja może uzyskiwać dostęp do wielu repozytoriów z minimalnymi opóźnieniami. Aby wybrać najbardziej interesujące repozytoria, pobieranie obejmuje te o największej ilości "gwiazdek", czyli najbardziej wartościowe z punktu widzenia społeczności serwisu GitHub. Dane składowane są w folderze projektu w folderze `gh-database`. Za pobieranie oraz dostęp do zapisanych repozytoriów odpowiada klasa `GHInterface`, której implementacja znajduje się w pliku `src/ghinterface.py`.

Ilość pobranych repozytoriów może być ustawiana za pomocą parametru `top_repos_to_download` w pliku `src/config.py`. Domyślnie wynosi ona 10000.