Syndykacja danych

Krótki opis usługi

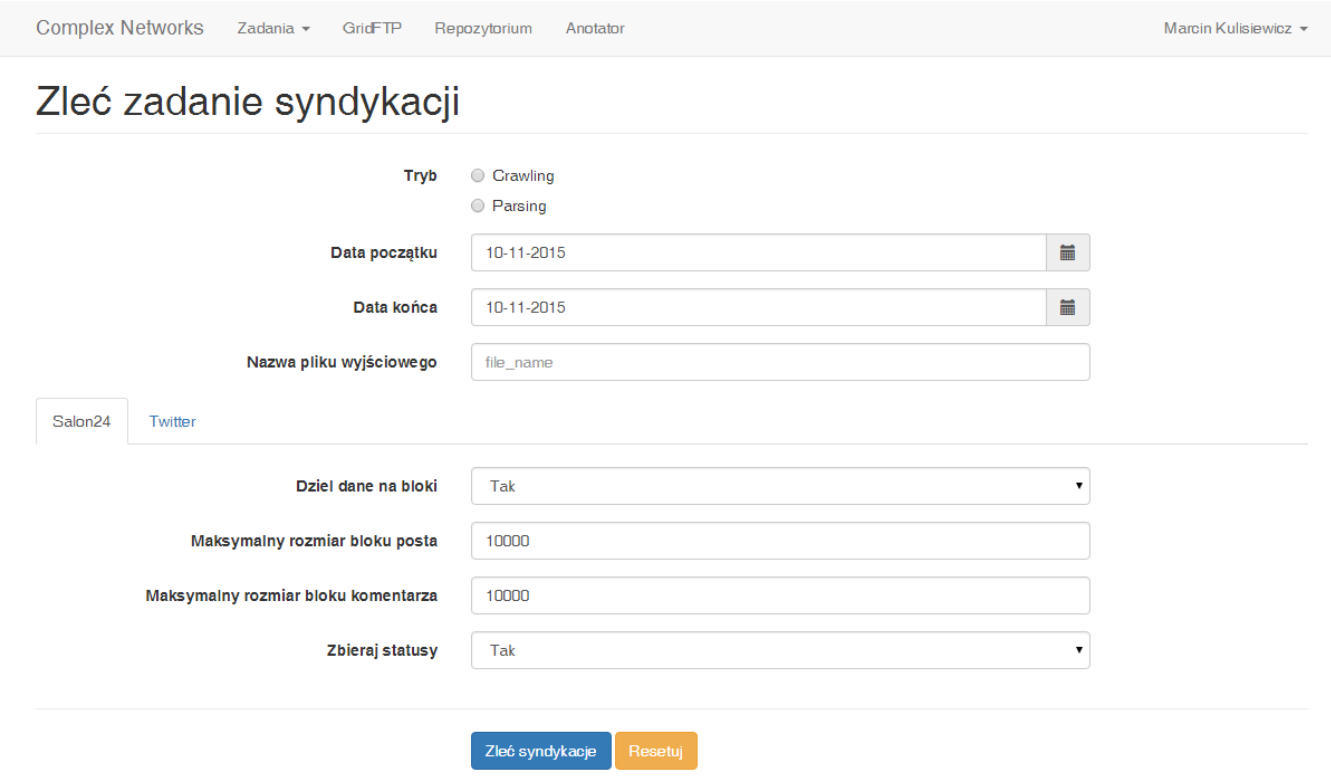
Usługa syndykacji danych jest przeznaczona dla wszystkich badaczy i naukowców którzy są zainteresowani pozyskiwaniem dużych zbiorów danych z mediów społecznościowych. Usługa oferuje zbieranie danych z serwisu społecznościowego Twitter oraz portalu publicystycznego Salon24. Oprócz tego usługa oferuje funkcję AnnotationHelper, która pozwala na opisanie zbiorów danych dodatkowymi atrybutami - poprzez ręcznie adnotowane klasy. Całość usługi pozwala na uzyskanie wartościowych zbiorów danych użytecznych do badań nad sieciami społecznościowymi. Zbiory danych są przechowywane w jednolitej strukturze danych, która zapewnia kompatybilność z innymi usługami platformy Complex Networks.

Aktywowanie usługi

Aby korzystać z usługi *Syndykacja danych* należy posiadać konto w infrastrukturze PLGrid, a następnie złożyć wniosek o dostęp do usługi w portalu https://portal.plgrid.pl/.

Pierwsze kroki

Po aktywacji usługi w portalu PLGrid, należy wejść na stronę platformy Complex Networks https://cn.plgrid.pl/. Przed skorzystaniem z usługi użytkownik może zostać poproszony o zalogowanie się z użyciem loginu i hasła do infrastruktury PLGrid. Następnie, w celu przeprowadzenia syndykacji danych z wybranego serwisu społecznościowego, należy wybrać z menu (u góry strony) [Zadania] -> [Zleć syndykację].



Usługa syndykacji działa w dwóch trybach: *Crawling* oraz *Parsing*.

Crawling

Ten tryb służy do pobierania surowych danych z wybranego źródła. Każde źródło zostanie zapisane w postaci plików HTML w repozytorium danych (więcej na stronie o repozytorium).

Podstawowymi parametrami usługi jest data początkowa oraz data końcowa okresu, z jakiego mają pochodzić dane. Usługa sprawdza datę publikacji i decyduje czy należy ją pobrać do zbioru użytkownika. Użytkownik ma także możliwość sparametryzowania nazwy pliku wyjściowego.

Pozostałe parametry są zmienne ze względu na heterogeniczność źródeł danych. Poniżej przedstawione są parametry poszczególnych źródeł [**parametr_zalecany**/parametr]:

Twitter

Początkowy użytkownik: nazwa użytkownika (login) którego posty mają zostać zebrane.

Parsing

Ten tryb służy do przetworzenia plików zebranych w trybie *Crawling* do jednolitej dla wszystkich źródeł struktury danych. Parsowane dane również są umieszczane w repozytorium w postaci pliku tekstowego w formacie JSON. Schemat struktury danych dostępny jest tutaj.

Dodatkowe parametry:

Salon24

Dziel dane na bloki - [**Tak**/Nie] parametr określający czy parsowane dane mają być dzielone na bloki. To źródło danych jest bardzo duże i przy parsowaniu dużej jego części może dojść do sytuacji gdy zabraknie pamięci operacyjnej, aby zapisać dane do repozytorium.

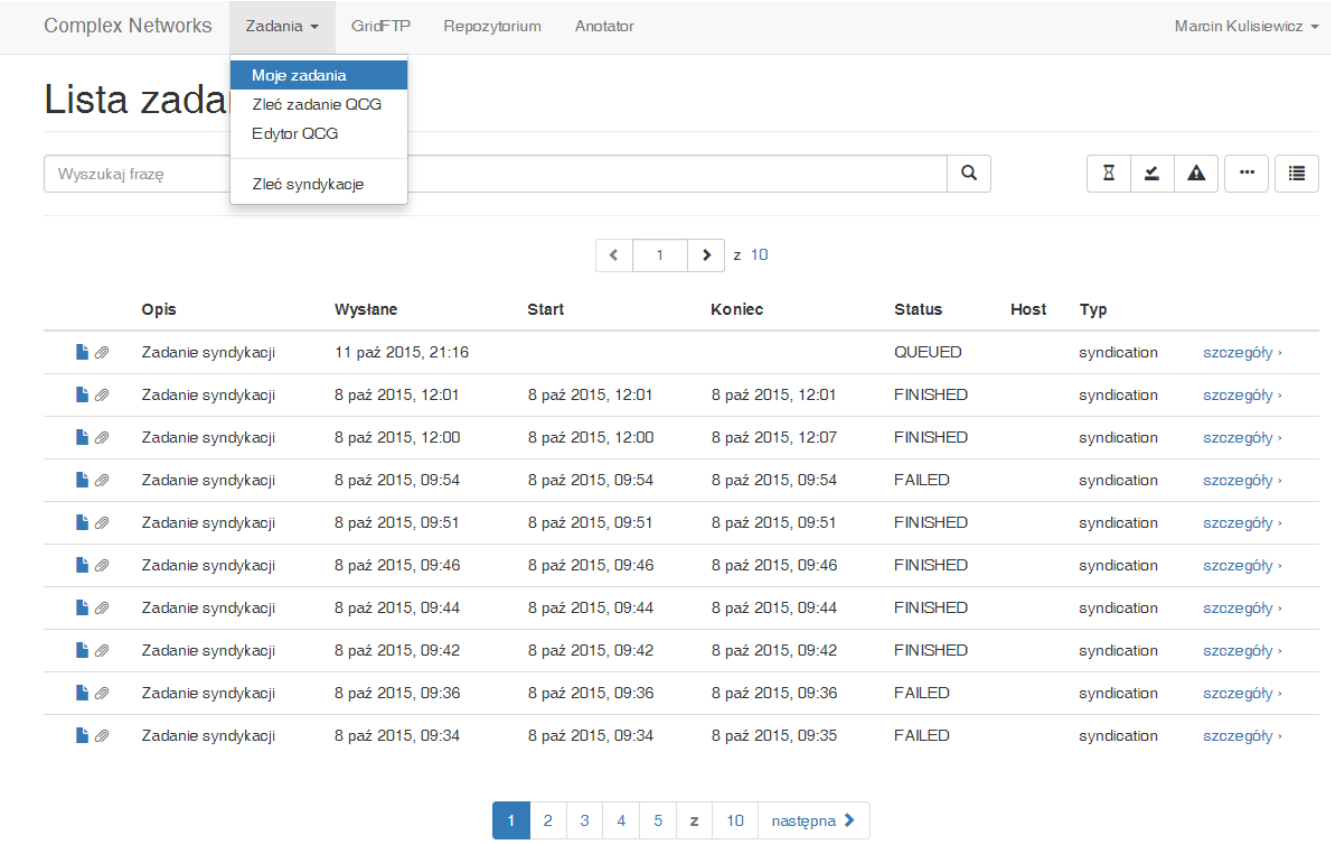
Maksymalny rozmiar bloku posta - maksymalna ilość postów w jednym bloku. Parametr ma zastosowanie jeśli i tylko jeśli parametr *Dziel dane na bloki* ma wartość TAK.

Maksymalny rozmiar bloku komentarza - maksymalna ilość komentarzy w jednym bloku. Parametr ma zastosowanie jeśli i tylko jeśli parametr *Dziel dane na bloki* ma wartość TAK.

Zbieraj statusy - [TAK/**NIE**] parametr określa czy parsowane posty mają mieć zbierane również dane o statusie społecznym (statusy w mediach społecznościowych Facebook, Twitter, Google+). Wymaga to jednak dodatkowego łączenia się z tymi serwisami, co znacznie spowalnia pracę usługi.

Podgląd uruchomionych zadań

Po uruchomieniu zadania, możemy sprawdzić jego status przechodząc do listy zadań w zakładce menu górnego [Zadania] -> [Moje zadania]



Bezpośrednio po uruchomieniu zadanie będzie posiadało status **QUEUED**.

Statusy zadań:

- **QUEUED** - zadanie czeka w kolejce na uruchomienie,
- **RUNNING** - zadanie jest w trakcie wykonywania,
- **FAILED** - zadanie zakończyło się błędem w trakcie wykonywania,
- **FINISHED** - zadanie zakończyło się poprawnie.

Jeżeli zadanie ma status **FINISHED**, możliwe jest pobranie wyników zadania. Aby pobrać wyniki należy wybrać w górnym menu pozycję Repozytorium i przejść do Community *Complex Networks* i kolekcji *Syndykacja*. Więcej informacji w rozdziale podręcznika opisującym Repozytorium.

Zaawansowane użycie

Aby skorzystać z adnotowania danych patrz podręcznik użytkownika AnnotationHelper.