Autorzy:
- Piotr Grabarski
- Szymon Szymborski

Temat: Wyszukaj w Wikipedii i źródłach pokrewnych wydarzenia, których opisy różnią się znacznie w różnych wersjach językowych. Można ograniczyć analizę do określonej klasy wydarzeń.

Koncepcja wykonania

1. Dane:

● Źródła danych

Podstawowym źródłem danych użytych w projekcie będzie wikipedia, czyli wielojęzyczna encyklopedia internetowa zawierająca otwarte treści. Wielojęzyczność wikipedii umożliwia otrzymywanie danych w różnych językach. Otwartość treści pozwala natomiast na ich użycie w szerokim zakresie. Połączenie tych dwóch cech pozwoli na porównania opisów wydarzeń w różnych wersjach językowych.
Źródłem pokrewnym do wikipedii jest natomiast portal wikicytaty, który zawiera cytaty dotyczące konkretnych kategorii wydarzeń. Również jest on wielojęzyczny oraz zawiera otwarte treści.
Poddane analizie będą wydarzenia historyczne w ich szerokim znaczeniu.

● Technologie pobrania danych

Wikipedia pozwala na pobranie danych za pomocą specjalnie przygotowanego API. Obecność takiego rozwiązania pozwala skupić się na danych i ich analizie, a nie na ich zdobywaniu. Pakiet wikipedia znajduje się w paczkach Pythona. W celu uzyskania danych na temat konkretnego hasła należy użyć polecenia *search*. Aby uzyskać skrót wiedzy o danym haśle należy użyć polecenia *summary*. W celu uzyskania całkowitych danych o żądanym haśle należy użyć polecenia *page*, które zwróci całość artykułu, który dostępny jest na internetowej stronie encyklopedii. Taka całość artykułu posiada takie pola jak choćby: *content*, *url*, *references*, *title*, *categories*, *links*. Wyjątkowo pomocna w realizacji zadania jest również obecność polecenia *set_lang*, które to pozwala na zmianę domyślnego języka otrzymywanych danych na dowolny. Po zmianie języka wszystkie opisane powyżej funkcje będą zwracały swoje wartości w nowym języku.
Analogiczne API udostępniane jest przez serwis wikicytaty (ang.: wikiquote). Dzięki modułowi języka Python o nazwie wikiquote możliwe jest wyszukiwanie i przeglądanie wszystkich cytatów dostępnych na internetowej wersji serwisu.

- Sposoby przechowywania danych

Dane uzyskane w powyższy sposób są możliwe do dalszego przetwarzania bez dodatkowej obróbki czy przetwarzania. Uzyskanie zawartości artykułu pozwoli na natychmiastowe przejrzenie jego zawartości. W ten sam sposób możliwe jest postąpienie z artykułami na ten sam temat występującymi w innych językach. Ta zaleta pozwoli na brak konieczności przechowywania poszczególnych danych i pozwoli skupić się na ich analizie.
Również przechowywanie danych dotyczących cytatów o wydarzeniach historycznych nie narzuca potrzeby ich przechowywania, gdyż możliwe jest ich uzyskanie w czasie rzeczywistym i natychmiastowa dalsza obróbka.

2. Metodyka uzyskania zadowalającego złączenia

Dane już podczas ich uzyskiwania są komplementarne do siebie. Jest to spowodowane faktem, iż będą służyć do porównania różnic w opisach danych względem poszczególnych wersji językowych wikipedii. Wydarzenie historyczne miało jednakowy przebieg niezależnie od języka w jakim zostało opisane. Jednakowe również były daty jego wystąpienia, czy strony w nim uczestniczące.
Łączenia cytatów z wydarzeniami będzie polegało na dopasowaniu autorów cytatów oraz nazwisk w nich wymienionych do ich wystąpień w artykułach o danych wydarzeniach na stronach wikipedii. Wielojęzyczne cytaty mogą dotyczyć innych postaci, które mogły, lecz nie musiały pojawiać się w opisie wydarzeń historycznych na stronach wolnej encyklopedii.

3. Metodyka analizy danych

Pierwszym ze sposobów na analizę danych, który wydaje się najprostszy jest badanie odnośników do innych artykułów oraz stron, czyli sekcji *links* w uzyskanym artykule. Zdarza się iż wpisy na tematy wydarzeń opisane na nieanglojęzycznych wersjach wikipedii odwołują się do angielskich źródeł. Taka analiza danych mógłby poprzeć powyższą tezę. Przykładowo zarówno polska jak i angielska wikipedia na stronie dotyczącej lądowania na Księżycu posiadają odwołania do angielskiego artykułu “One Small Step for Man”.
Kolejnym ze sposobów na analizę uzyskanych danych jest zliczanie wystąpień poszczególnych słów i badanie liczby ich wystąpień zależnie od wersji językowej strony. Przykładowo na polskiej wersji wikipedii w artykule opisującym Bitwę Warszawską słowo Wisła (i jego odmiany) jest wspomniane 22 razy. Artykuł o tym samym wydarzeniu w angielskiej wersji językowej wymienia słowo Vistula wyłącznie 15 razy.
Kolejną z potencjalnych różnic w opisach wydarzeń może być ich obszerność. Zależnie od wersji językowej wikipedii artykuły znacznie różnią się ilością informacji. Przykładowo opis osiemnastowiecznego wydarzenia jakim była Herbatka bostońska ma obszerny opis na angielskiej wersji strony podczas gdy na polskiej jest on zaledwie wspomniany.
Te przykładowe sposoby porównań pozwolą na uzyskanie różnic występujących pomiędzy poszczególnymi wersjami językowymi wikipedii.

Różnice w wersjach językowych artykułów w przypadku portalu na którym każdy może dodawać treści z pewnością okażą się znaczne. Jedną z ciekawych różnic między wersjami językowymi dotyczy artykułów spornych. Są to takie artykuły, które odnoszą się do wydarzeń będących konfliktami pomiędzy dwoma różnymi państwami. Artykuły w językach tych państw będą szczególnie podatne na różnice merytoryczne dotyczące przebiegu i moralności danego wydarzenia. Jako przykład tego typu wydarzenia można podać aneksję Krymu przez Rosję i jego opis w Rosyjskiej lub Ukraińskiej wersji wikipedii.
W celu wykorzystania wikicytatów do analizy danych potrzebna będzie metoda uzyskująca dane na temat obecności imion lub nazwisk wymienionych w cytatach bądź będących ich autorami osób. Na potrzeby projektu możliwe jest uproszczone założenie, że imię i nazwisko autora jest jego rzeczywistymi danymi, a występujące w cytacie słowa pisane wielką literą niebędące początkiem zdania są wspomnieniem postaci o której autor się wypowiada. Tak uzyskane dane o postaciach historycznych są możliwe do wyszukiwania w wikipedii. Ich obecność jest możliwa do sprawdzenia w artykułach za pomocą dostępnych danych. Pozwoli to na sprawdzenie czy rzeczywiście taka postać miała powiązanie z wydarzeniem.
Innym sposobem na analizę danych jest użycie *Free Google Translate Api* do tłumaczenia tekstu - czyli Pythonowej paczki *googletrans* i przetłumaczenie na starcie całego dokumentu do wybranego języka. Następnie za pomocą algorytmów typu *string similarities,* można określić stopień podobieństwa dwóch tekstów. Przypominałoby to narzędzie antyplagiatowe. Do określenia podobieństwa można użyć gotowych algorytmów znajdujących się np. w paczce do Pythona - strsimpy. Dla porównania można zaimplementować również własny prosty algorytm, sprawdzający liczbę tych samych słów oraz ciągów słów o różnej długości.

4. Definicja:

● Scenariusza użycia aplikacji

Użytkownik będzie mógł korzystać z aplikacji przy pomocy CLI (Command Line Interface).
CLI pozwoli na wybór źródła informacji a następnie na wpisanie dowolnego hasła obecnego w wikipedii lub w wikicytatach, oraz opcjonalnie na wybór języków (domyślnie będą to wszystkie języki, które posiadają definicję szukanego hasła). Po uruchomieniu programu użytkownik uzyska podsumowanie dotyczące różnic między poszczególnymi wersjami językowymi. Taka funkcjonalność oprócz analizy różnic pozwoli również ludziom będącym poliglotami na dobór wersji językowej według aktualnych preferencji na podstawie analizy różnic.

● Technologii wykonania aplikacji

Aplikacja, która powstanie w wyniku prac nad projektem będzie napisana w języku Python.
Obsługa CLI, będzie prosta, intuicyjna i opisana po przekazaniu argumentu *--help*. Do parsowania argumentów użyta zostanie biblioteka *argparse.*

Aplikacja wykorzystywała będzie pakiety:

- Wikipedii - wikipedia-1.4.0
- Wikicytatów - wikiquote-0.1.14
- Google tłumacza - googletrans-2.4.0

Do porównywania tekstów użyty zostanie pakiet *strsimpy*.
Wszystkie wymienione pakiety można pobrać za pomocą managera pakietów Pythonowych - *pip*.

● Ergonomii aplikacji

Aplikacja stworzona w wyniku podejmowanych działań będzie charakteryzować się znaczną ergonomią. Jej działanie zostanie opracowane w sposób, który pozwoli na korzystanie z niej po zapoznaniu się z podstawowym readme. Będzie ona posiadać prosty interfejs CLI, którego wszystkie parametry pojawią się po użyciu polecenia *help*. Wyniki, które zostaną uzyskane w efekcie przetwarzania uzyskanych danych zostaną przekazane użytkownikowi w sposób, który pozwoli na ich bezproblemowy odbiór. Aplikacja będzie sprawdzała, czy interfejs wiersza poleceń został użyty prawidłowo i w przypadku pomyłki będzie wyświetlała stosowny komunikat. Np.: jeśli szukane hasło nie ma podanej przez nas wersji językowej, użytkownik zostanie o tym poinformowany i wyświetlony zostanie zbiór języków, w których szukane wydarzenie jest opisane.

● Architektury aplikacji

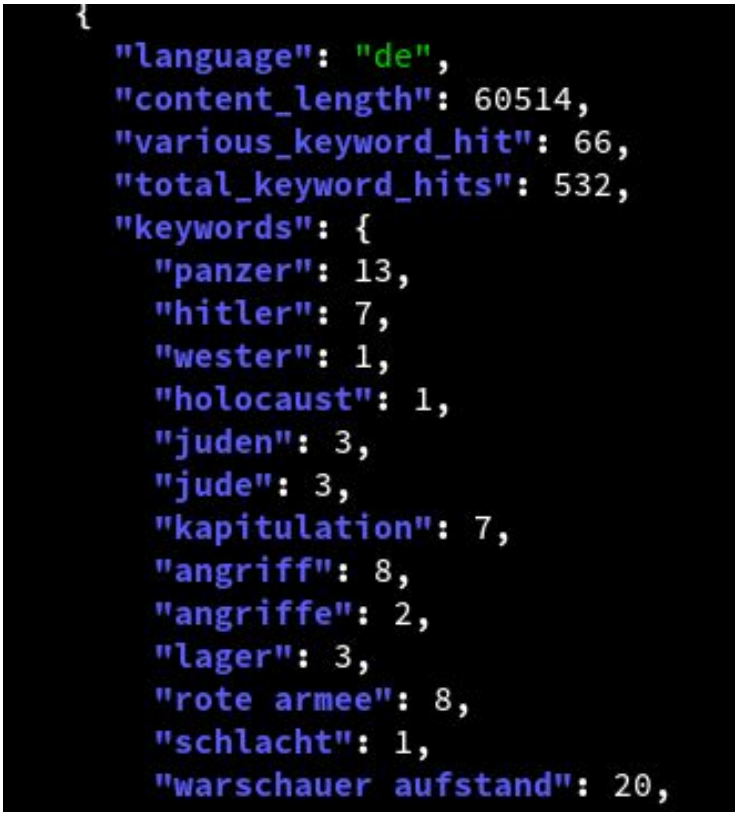
Aplikacja będzie korzystała z środowiska wirtualnego, w którym pobrać trzeba będzie paczki obecne w pliku requirements.txt. Napisana będzie w całości w języku Python.
Aplikacja nie będzie korzystała z bazy danych, gdyż będzie działała w schemacie - pobierz dane, przetwórz dane, zwróć rezultat.
Jeśli aplikacja miałaby się rozrosnąć, to warto rozważyć konteneryzację aplikacji (docker). W przypadku dalszego rozwoju aplikacji możliwe jest również przekształcenie jej do aplikacji internetowej dostępnej pod adresem jako witryna internetowa. Za pomocą prostego interfejsu użytkownika możliwe byłoby przekazanie żądania danych, a jako rezultat użytkownik otrzymałby oczekiwane dane przedstawione na stronie internetowej. Taki kierunek rozwoju pozwoliłby na spopularyzowanie aplikacji i możliwość użycia jej w dowolnym momencie.

Wykonanie

5. Odstępstwa od koncepcji:

Zgodnie z sugestiami otrzymanymi od prowadzącego postanowiliśmy zaniechać pomysłu automatycznego tłumaczenia i porównywania haseł. Wymyśleliśmy i zapisaliśmy listę około 160 słów kluczowych związanych z tematem Drugiej Wojny Światowej. Wykorzystaliśmy narzędzia do webscrappingu (requests i lxml) oraz do Stemmowania(nltk). Kolejnym odstępstwem od wstępnej koncepcji wykonania zadania była również rezygnacja z użycia wikicytatów. Podczas prac nad implementacją postanowiliśmy skupić się na wikipedii oraz danych z niej uzyskiwanych.

Dużą wartością dodaną względem początkowo zakładanej koncepcji jest użycie bazy danych NOSQL - ElasticSearch, uruchamianej za pomocą docker-compose. Pozwola ona na wygodne przechowywanie danych uzyskanych z aplikacji. Przykładowy zrzut ekranu danych przechowywanych w bazie przedstawia się następująco:



W finalnej wersji działanie programu polega na analizowaniu różnych wersji językowych wikipedii pod kątem obecności związanych z wydarzeniami historycznymi słów kluczowych odpowiednich dla danego języka i porównywania zależności między nimi. Aplikacja ma możliwość generowania plików *keywords.txt* dla różnych języków poprzez

użycie API google translate. Istnieje również możliwość dodawania własnych słów związanych z poruszanym tematem. Tak przygotowane słowa kluczowe są porównywane z wynikami uzyskanymi z wikipedii, a następnie przedstawiane są wychwycone różnice.

6. Instrukcja instalacji i obsługi aplikacji

Aplikacja posiada plik *README.md*, który zawiera dokładny opis instalacji oraz uruchomienia programu. Uruchomienie programu wymaga obecności dwóch parametrów - zbioru języków w jakich będzie przeprowadzona analiza oraz nazwy wydarzenia, które będzie analizowane. Parametry te są przekazywane poprzez argumenty wejściowe przy uruchamianiu programu.

Kody wykorzystywanych języków odpowiadają kodom zawartym w różnych wersjach językowych wikipedii. Pełna lista wersji językowych i ich kodów znajduje się pod adresem: https://pl.wikipedia.org/wiki/Wikipedia:Lista_wersji_językowych

W przypadku braku podania pliku *keywords.txt* dla wybranego języka program ma możliwość wygenerowania pliku językowego z poszukiwanymi słowami na podstawie tłumacza google. W ten sposób możliwa jest również analiza tych wersji językowych w których użytkownik nie posiada znajomości języka.

Program podczas działania wysyła zapytania do wikipedii poprzez api, a następnie przeszukuje uzyskaną odpowiedź wyszukując kluczowych słów i analizując częstość ich występowania w zależności od wersji językowej encyklopedii.

Rezultaty działania są prezentowane na standardowym wyjściu programu i zapisywane do bazy danych.

Dane, które dotychczas zebraliśmy w aplikacji znajdują się w folderze esdata, który jest mount'owany do ElasticSearch'a uruchomionego w dokerze.

7. Źródła:

- https://stackabuse.com/getting-started-with-pythons-wikipedia-api/
- https://pypi.org/project/wikipedia/
- https://pypi.org/project/googletrans/
- https://docs.python.org/3/
- https://en.wikipedia.org/
- https://www.mediawiki.org/wiki/API:Client_code
- https://github.com/luozhouyang/python-string-similarity#levenshtein
- https://en.wikipedia.org/wiki/Levenshtein_distance
- https://docs.python.org/3/howto/argparse.html