Dokumentacja wstępna projektu
Projekt łączony z przedmiotów *Systemy Agentowe* oraz *Wstęp do eksploracji danych tekstowych w sieci WWW*

Andrzej Dawidziuk, Tomasz Kogowski, Rafał Okuniewski

10 stycznia 2018

1 Zadanie projektowe

Zadanie projektowe polega na wyszukiwaniu tematycznych wiadomości w Internecie oraz ich analizie semantycznej pod kątem podobieństwa i tematyki. Na potrzeby projektu analizowane będą portale informacyjne udostępniające treści w języku angielskim.

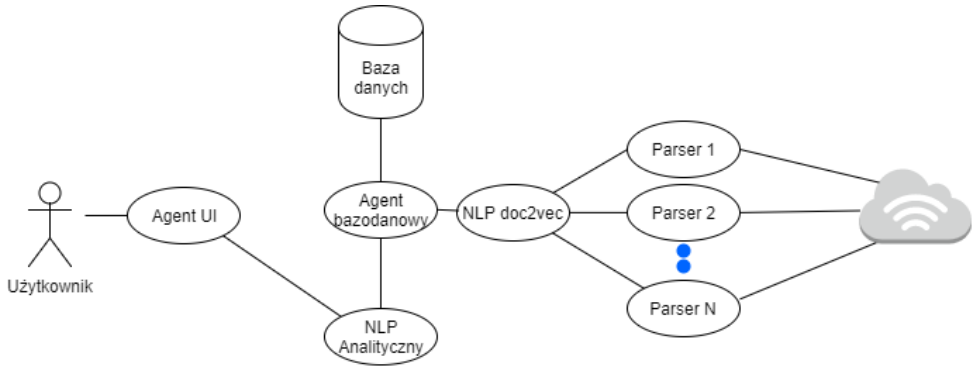
2 Opis projektu

Projekt ma na celu znajdowanie na podstawie analizy semantycznej podobnych artykułów na różnych portalach informacyjnych. Artykuły te pochodzić będą z kilku internetowych portali informacyjnych udostępniających kanały RSS, takich jak CNN, BBC czy The Economist. Dokładna lista analizowanych stron zostanie wykreowana w trakcie implementacji projektu.

Kanały RSS posłużą do znalezienia wiadomości zebranych pod jednym tematem. Będą one analizowane co pewien czas, a na podstawie linków zgromadzonych na kanałach pobierane będą artykuły. Następnie treści artykułów poddane zostaną analizie z wykorzystaniem algorytmu Doc2Vec, zaś wyniki tej analizy (wektor liczb rzeczywistych), wraz z tagiem opisującym tematykę wiadomości oraz czas jej opublikowania będą systematycznie składowane w bazie danych.

Na żądanie użytkownika wiadomości z podanej przez niego tematyki oraz okresu czasowego będą analizowane pod kontem podobieństwa. Jeśli system znajdzie jeden lub więcej podobnych do siebie artykułów, zwróci mu pogrupowane wyniki przeszukiwania wraz z miarą podobieństwa.

3 System agentowy



Rysunek 1: System agentowy

Tabela 1 zawiera słownik pojęć używanych w dalszych częściach dokumentu.

Agenty	Opis
UI	Odpowiedzialny za odebranie wejścia od użytkownika oraz prezentację wyników.
NLP analityczny	Odpowiedzialny za obliczenia skali podobieństw.
bazodanowy	Odpowiedzialny za zapisywanie oraz odczytywanie danych z bazy danych.
NLP doc2vec	Odpowiedzialny za przekształcenie danych otrzymanych od parserów do wektorów numerycznych.
parser	Odpowiedzialny za pobranie artykułów o zadanej tematyce z zadanego serwisu.

Tabela 1: Opis agentów.

4 Wykorzystane technologie

4.1 Framework Akka

W celu implementacji systemu wieloagentowego zostanie wykorzystany framework Akka oraz język Scala. Framework ten jest jednym z najpopularniejszych środowisk do tworzenia systemów agentowych. Zaimplementowane z jego wykorzystaniem zostaną moduły odpowiadające za parsowanie stron i kanałów RSS a także agent bazodanowy.

4.2 Doc2vec

Do ustalania semantycznego podobieństwa dokumentów zostanie użyta technika Doc2Vec. Wektory generowane będą z wykorzystaniem następujących frameworków języka Python:

- NLTK
- Gensim

Gensim zapewnia określanie podobieństwa wektorów, ale ze względu na ograniczenia tej technologii (m. in. konieczność przechowywania wszystkich porównywanych wektorów w pamięci), do tego zadania może zostać użyta inna biblioteka, np. NumPy.