

**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE**

Wydział Informatyki, Elektroniki i Telekomunikacji
Katedra Informatyki



PROJEKT INŻYNIERSKI

**SYSTEM DO ANALIZY I MONITOROWANIA
PORTALI OGŁOSZENIOWYCH**

**PIOTR GÓRALCZYK, PAWEŁ MIKOŁAJCZYK
DOMINIK RUSIECKI**

OPIEKUN:
dr inż. Jacek Dajda

Kraków 2015

OŚWIADCZENIE AUTORA PRACY

OŚWIADCZAM, ŚWIADOMY(-A) ODPOWIEDZIALNOŚCI KARNEJ ZA PO-
ŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZY PROJEKT WYKONAŁEM(-AM)
OSOBIŚCIE I SAMODZIELNIE W ZAKRESIE OPISANYM W DALSZEJ CZĘŚCI
DOKUMENTU I ŻE NIE KORZYSTAŁEM(-AM) ZE ŹRÓDEŁ INNYCH NIŻ
WYMIENIONE W DALSZEJ CZĘŚCI DOKUMENTU.

.....

PODPIS

1. Cel prac i wizja produktu

Charakterystyka problemu, motywacja projektu (w tym przegląd istniejących rozwiązań prowadząca do uzasadnienia celu prac), ogólna wizja produktu, krótkie studium wykonalności i analiza zagrożeń.

Problemem, z którym stara się mierzyć system stworzony w ramach niniejszej pracy inżynierskiej, jest obserwacja wielu źródeł informacji w Sieci w poszukiwaniu interesujących użytkownika zagadnień z danej tematyki. Klient wyraził pragnienie posiadania narzędzia, umożliwiającego mu monitorowanie stron i portali internetowych pod kątem zdefiniowanego przez niego zbioru informacji, a ponadto udostępniającego metody analizy tych informacji.

Istnieją narzędzia odpowiadające na podobne potrzeby - Google Search pozwala na przeszukiwanie informacji w Internecie w odpowiedzi na podane hasło, a Google Analytics może służyć do analizowania trendów w Sieci. Nie dają one jednak użytkownikowi możliwości w bardziej specyficzny sposób definiowania sposobu ekstrakcji danych ze stron. Nie są też w stanie gromadzić zebranych informacji i w wygodny sposób przeprowadzać na nich analizy.

W ramach tworzonego projektu powstałby system, dający użytkownikowi swobodę definiowania rodzaju informacji, jakie będą przezeń poszukiwane, a także szczegółowego sposobu ich ekstrakcji ze źródła. Zakładana jest pewna znajomość przez użytkownika mechanizmów dopasowania wzorca, jak XPath lub wyrażenia regularne. Proces ekstrakcji oparty będzie o web-crawling przy użyciu jednego z rozwiązań open-source lub z wykorzystaniem naszego własnego crawlera. Wyekstrahowane informacje przechowywane będą w postaci dedykowanej bazy danych, z której będą korzystały narzędzia do analizy. Na wspomnianą analizę ma się składać np. wyszukiwanie w zgromadzonych danych słowa kluczowego lub pokazywanie trendów czasowych.

Wizja wypracowana w wyniku naszej współpracy z klientem wydawała się do spełnienia. Część, która mogła nam sprawić dużo problemów, to moduł analizy, wymagający zarówno złożonych algorytmów po stronie logiki systemu, jak i atrakcyjnego sposobu prezentacji w formie odpowiedniej technologii frontendowej. Nasze doświadczenie na tym polu było niewielkie. Być może dlatego ta część systemu została zaimplementowana tylko po części.

Ponadto, podczas specyfikacji wizji systemu zidentyfikowaliśmy następujące zagrożenia:

- Ograniczenia dotyczące crawlingu w regulaminach portali ogłoszeniowych.
- Brak doświadczenia zespołu z nierelacyjnym modelem bazy danych.
- Konieczność współpracy z różnymi formatami ogłoszeń. Konieczność dostosowywania się do zmian w tych formatach w czasie działania systemu.
- Uzależnienie od pracy portali może nastąpić brak dostępu do informacji, na który nie mamy wpływu.
- Stosunkowo szybkie tempo gromadzenia danych wykorzystywanych przez system, co może prowadzić do problemów wydajnościowych.
- Trudność w implementacji algorytmów wyszukiwania powiązań między ogłoszeniami odpowiadających oczekiwaniom klienta.

2. Zakres funkcjonalności

Kontekst użytkowania produktu (aktorzy, współpracujące systemy) oraz najważniejsze wymagania funkcjonalne i нефunkcjonalne.

3. Wybrane aspekty realizacji

Przyjęte założenia, struktura i zasada działania systemu, wykorzystane rozwiązania technologiczne wraz z krótkim uzasadnieniem ich wyboru.

4. Organizacja pracy

Struktura zespołu (role poszczególnych osób), krótki opis i uzasadnienie przyjętej metodyki i/lub kolejności prac, planowane i zrealizowane etapy prac ze wskazaniem udziału poszczególnych członków zespołu, wykorzystane praktyki i narzędzia w zarządzaniu projektem.

5. Wyniki projektu

Najważniejsze wyniki (co konkretnie udało się uzyskać: oprogramowanie, dokumentacja, raporty z testów/wdrożenia, itd.) i ocena ich użyteczności (jak zostało to zweryfikowane — np. wnioski klienta/użytkownika, zrealizowane testy wydajnościowe, itd.), istniejące ograniczenia i propozycje dalszych prac.

W wyniku prac udało się uzyskać część funkcjonalności wymaganej początkowo przez klienta. Proces monitorowania odbywa się zgo dnie z założeniami, natomiast, z powodu problemów technicznych, metody analizy zgromadzonych danych ograniczone są do wyszukiwania po słowach kluczowych. Stanowi to dobry wektor dla dalszego rozwoju systemu.

Ponadto dla systemu powstała szczegółowa dokumentacja techniczna. Zawiera ona opis technicznych aspektów systemu - architektura, wykorzystane technologie.

Oprócz tego, stworzyliśmy również dokumentację procesową, opisującą przebieg naszych prac podczas tworzenia systemu, jak i dokumentację użytkownika, mającą ułatwić mu zaznajomienie się z systemem i sprawne jego zainstalowanie oraz użytkownię.