Raport początkowy z prac nad projektem “Wyszukiwarka dziwnych wydarzeń”

1. Wstępne ustalenia
Zespół składa się z następujących osób:

- Piotr Ożga (lider)
- Grzegorz Krukar
- Adam Janda
- Leszek Koziatek

Zespół zobowiązuje się do wykonania projektu “Wyszukiwarka dziwnych wydarzeń” w ramach hackatonu na przedmiocie Inteligentne Systemy Informatyczne. Termin oddania projektu to 23 kwietnia 2014 roku.

2. Wymagania
Projekt zakłada spełnienie następujących wymagań:

- Stworzenie bazy danych o zjawiskach paranormalnych na podstawie źródeł znalezionych w internecie (blogi, strony tematyczne, wikipedia)
- Stworzenie wyszukiwarki, która na podstawie zebranych materiałów będzie wyszukiwała wydarzenia lub treści interesujące użytkownika
- Stworzenie podstawowego layoutu, który ułatwi obsługę narzędzia przez użytkownika końcowego

3. Ograniczenia
Projekt będzie miał następujące ograniczenia:

- Baza danych będzie się składała tylko i wyłącznie z treści w języku polskim
- Projekt będzie dostępny jedynie na lokalnych maszynach

4.Plan działań i technologie

Crawlowanie wikipedii
Crawlowanie treści na wikipedii będzie odbywało się z pomocą oficjalnego api silnika mediawiki. Kategorią startu będzie kategoria “Zjawiska paranormalnie”. Zostaną wyciągnięte treści artukułów pokrewnych (na podstawie dolinkowanych i kategorii), oraz linki zewnętrzne, do serwisów, które mogą oferować ciekawe z punktu widzenia wyszukiwarki treści. Owe linki zostaną przecrawlowane z pomocą tradycyjnego crawlera. Szacowana liczba wyciągniętych haseł z wikipedii: 200. Do crawlowania zostanie użyta biblioteka nodemw odpalana w środowisku Node.JS.

Wyszukiwarka

Będziemy używać wyszukiwarki Sphinx albo Solr. Obie są wydajnymi, otwarto źródłowymi aplikacjami serwerowymi. Sphinx korzysta z licencji GPLv2, natomiast Solr z licencji Apache2. Oznacza to, że oba systemy mogą być używane za darmo, Solr może być dowolnie modyfikowany natomiast modyfikacja Sphinxa wymaga wykupienia licencji.
W internecie panują różne opinie co do tego, która wyszukiwarka jest lepsza, jednak w naszym projekcie nie ma to, aż tak dużego znaczenia ponieważ obie wyszukiwarki spełniają nasze wymagania.

Solr

Indeksowanie

- Indeksowanie dokumentów XML, CSV, Word, PDF
- Pełne i częściowe indeksowanie bezpośrednio z baz danych SQL
- Deduplikacja danych
- Synonimy
- Tokenizacja (usuwanie znaków specjalnych)
- Sprowadzanie do formy podstawowej

Wyszukiwanie

- Filtrowanie
- Sortowanie
- Podświatlanie trafień
- Liczniki
- Grupowanie

Pełna obsługa, czyli wyszukiwanie, indeksowanie i modyfikacja danych odbywa się przy pomocy interfejsu HTTP/XML. Biblioteki dostępne są dla wielu popularnych języków programowania, a dane wyjściowe mogą być zwracane w kilku formatach: XML/XSLT, JSON, Python, Ruby lub PHP.

Porównianie: http://db-engines.com/en/system/Solr%3BSphinx

5.Zadania wykonane przez ostatni tydzień

-Przegląd wykorzystywanych technologii

-Implementacja crawlera i clawling witryny niewiarygodne.pl

-Początek prac nad crawlerem dla wikipedii

-Utworzenie repozytorium kodu na serwisie github