Webowy system do automatycznego gromadzenia korpusów tekstu z Internetu i ich ręcznej anotacji meta danymi

Tytuł w języku angielskim: Web-based System for Gathering Text Corpora from Internet and Facilitating Manual Annotation of the Corpora

Spis treści

[Dział I 4](#_Toc452237677)

[Streszczenie 4](#_Toc452237678)

[Synopsis 4](#_Toc452237679)

[Wstęp: Cel, motywacje, ograniczenia 5](#_Toc452237680)

[Motywacje 5](#_Toc452237681)

[Cel pracy 5](#_Toc452237682)

[Opis problemu 6](#_Toc452237683)

[Ograniczenia pracy 6](#_Toc452237684)

[Kluczowe pojęcia i terminy używane w tej pracy 7](#_Toc452237685)

[Korpus językowy 7](#_Toc452237686)

[Anotacja 7](#_Toc452237687)

[Crawler 7](#_Toc452237688)

[Drzewo DOM 7](#_Toc452237689)

[Selektor 7](#_Toc452237690)

[Algorytm pozycyjny 7](#_Toc452237691)

[CCL 7](#_Toc452237692)

[Premorph 7](#_Toc452237693)

[Inforex 8](#_Toc452237694)

[Corpo-grabber 8](#_Toc452237695)

[WCRFT 9](#_Toc452237696)

[Blog Reader 9](#_Toc452237697)

[Przegląd literatury 10](#_Toc452237698)

[Internet Documents: A Rich Source for Spoken Language Modeling 10](#_Toc452237699)

[Efﬁcient Web Crawling for Large Text Corpora 10](#_Toc452237700)

[A DOM Tree Alignment Model for Mining Parallel Data from the Web 11](#_Toc452237701)

[Dział II 12](#_Toc452237702)

[Corpo-grabber 12](#_Toc452237703)

[Dlaczego skrypty PHP? 12](#_Toc452237704)

[Elementy 12](#_Toc452237705)

[Interfejs 13](#_Toc452237706)

[Działanie 14](#_Toc452237707)

[Badanie 17](#_Toc452237708)

[Teoria 17](#_Toc452237709)

[Praktyka 17](#_Toc452237710)

[Wyniki badań 17](#_Toc452237711)

[Dział III 18](#_Toc452237712)

[Wnioski z badań 18](#_Toc452237713)

[Podsumowanie 18](#_Toc452237714)

[Źródła 18](#_Toc452237715)

# Dział I

## Streszczenie

Celem tej pracy jest opisanie rozwiązania problemu, jakim jest skonstruowanie narzędzia do ekstrakcji korpusów lingwistycznych z zasobów internetowych, jak na przykład blogi, strony informacyjne, itp. Narzędzie to ma być przystępne dla użytkownika nieobytego w technologiach informatycznych. Przedstawię jakie rozwiązania do tej pory się pojawiły, moją propozycja rozwiązania problemu oraz analizę, czy moje rozwiązanie faktycznie może być aplikowane do tego problemu.

W ramach rozwiązania problemu postanowiłem podejść do niego z innej strony niż zwykle. Mając doświadczenie w tworzeniu stron internetowych wiem, że jeżeli rozpatrujemy strukturę drzewiatą wielu stron internetowych z tej samej domeny, to konkretne elementy, które nas interesują, są zazwyczaj tymi samymi gałęziami w drzewie strony. W związku z tym chciałem sprawdzić teorię, czy można dokonywać ektrakcji tekstu ze stron internetowych oznaczając konkretne fragmenty drzewa strony (DOM). Badanie potwierdzające moją teorię przeprowadziłem porównując moją metodę ekstrakcji tekstu z metodą konkurencyjną, która polega na ekstrakcji elementów ze strony internetowej przy pomocy selektorów CSS.

### Synopsis

In this paper I shall present a solution to the problem of constructing tool for extracting linguistic corpora from internet resources, like blogs, informational sites, etc. The tool must be handy for a user that isn’t common with informational technologies. I shall present what kind of solutions have been created, my solution to the problem, and analysis if my solution can actually be used to resolve this problem.

Regarding my problem, I decided to resolve it from different approach, compared to what was used before. Having experience in web design, I know that when we interpret many web pages from the same domain as tree structure, specific elements that we could be interested in are usually the same branches in these trees. Because of that I decided to test my theory, if it will be possible to extract text from web pages marking certain tree site parts (DOM). The research that confirmed my theory were done by comparison to another method, that involves in extracting elements of web page using CSS selectors.

## Wstęp: Cel, motywacje, ograniczenia

### Motywacje

Analiza syntaktyczna języka jest jedną z ważnych dziedzin informatyki. Przede wszystkim może nam służyć do analizy części mowy, zapis (może być różny, w zależności od akcentu), regularności, kolejności słów w zdaniu, znaczenia i problemów związanych z językiem. Tymi i innymi problemami zajmuje się dział nauki zwany lingwistyką, ale również lingwistyka informatyczna i inżynieria języka naturalnego.

Lingwistycy do analizy języka pisanego - czy mówionego - potrzebują mieć materiały do analizy. Często są to nagrania audio, ich transkrypcje, książki, czasopisma, nagrania video i inne. Do celów analizy przy pomocy narzędzia informatycznego potrzebna jest ich wcześniejsza cyfryzacja. Nagrania audio muszą być transkrybowane na medium cyfrowe, transkrypcje w formie analogowej muszą zostać przeanalizowane przy pomocy programu do optycznej interpretacji tekstu (Optical Character Recognition), tak samo czasopisma czy książki. Jednakże, cyfryzacja to nie wszystko: często typowe formaty transkrypcji danych są niewystarczające. Do tego celu teksty w formie cyfrowej muszą zostać przekształcone do formatu pliku zwanego korpusem.

Korpus językowy to skompilowany zbiór danych tekstowych dostępnych w formie elektronicznej, stanowiący materiał do badań. Jest zawsze w formie cyfrowej, czasem ten plik posiada również wstępną analizę syntaktyczną tekstu w nim zawartego. Politechnika Wrocławska opracowała własny format pliku obsługującego korpusy. Nazywa się on CCL i oprócz przechowywania korpusów, przechowują również anotacje istotne w procesie analizy lingwistycznej.

Lingwistyka informatyczna boryka się z wieloma problemami, jest to np. tworzenie modelu języka w oparciu o ciągle zmieniające się zasady. Jednym z nich jest pozyskiwanie materiałów do analizy, a przede wszystkim ich trudnodostępność. Jednakże jest miejsce, w którym są duże zasoby tekstu do analizy lingwistycznej, z której polscy naukowcy do tej pory nie czerpali bardzo. Jest nią internet.

Nie jest problemem pozyskiwanie korpusów z internetu. Każdy może wejść na dowolną stronę internetową i pobrać jej zawartość, wyszczególnić tekst i przekonwertować na format korpusu. Problemem jest natomiast pozyskanie dużych ilości tekstu w opraciu o pojedynczy, ogólny adres internetowy. Praca ta ma na celu przybliżenie algorytmów, które rozwiązują ten problem.

Praca ta również ma na celu przyjrzenie się konkurencyjnym algorytmom, które służą do gromadzenia korpusów językowych ze stron internetowych. Mam nadzieję, że dzięki tej pracy przyszłe pokolenia naukowców z dziedziny lingwistyki będą mogli korzystać z mojego narzędzia bez problemów, oraz że będą rozumieli jego działanie.

### Cel pracy

Celem pracy jest opracowanie systemu – narzędzia badawczego – do gromadzenia korpusów tekstowych z internetu, gdzie korpus tekstowy to duży zbiór tekstów gromadzonych w celach badawczych w ramach różnych dziedzin nauki. System powinien pracować automatycznie, ale pod kontrolą naukowca. System powinien umożliwić zdefiniowanie zakresu pobieranych danych, wydobycie tekstu wraz z jego semantyczną strukturą z dokumentów (np. z podziałem na akapity, nagłówki itd.), zapis do stosowanych formatów korpusowych, w tym formatu CCL opracowanego w Grupie Naukowej G4.19 PWr. oraz późniejszy opis tekstów meta-danymi na różnych poziomach, np. dokumentów, zdań, słów itd.

### Opis problemu

W celu pracy opisuje się następujące elementy, jak automatyczna praca, ale pod kontrolą naukowca. To oznacza zdefiniowanie, które elementy będą automatyczne, a które pod kontrolą naukowca. Przyjąłem, że automatyczne w systemie będzie pobieranie zasobów internetowych, a pod kontrolą naukowca będzie wyznaczenie wzorca ekstrakcji.

Dodatkowo ważnym problemem w pracy jest zdefiniowanie takich reguł, które dla każdego naukowca będą jasne i klarowne. To wywiera ograniczenia na narzędziu, jak:

1. Jasne i klarowne reguły
2. Prosty interfejs

To wyklucza wykorzystanie metod ekstrakcji tekstu, jak definiowanie reguł przy pomocy selektorów.

### Ograniczenia pracy

Ograniczyłem się jedynie do zdefiniowania nowej metody analizowania strony internetowej (pozycyjnej). Porównuję to podejście z konkurencyjnym podejściem (według selektorów). W ramach badania analizuję oba podejścia na przykładzie wybranych blogów internetowych, oraz dwa podejścia mieszane. Badania ograniczyłem do ekstrakcji następujących elementów:

1. Notki – element opisujący wszystkie niżej wymienione elementy
2. Autor, tytuł, data wpisu i tekst wpisu

Do ekstrakcji tekstu uznałem, że element notki jest konieczny do wyznaczenia w celu wykrywania nieprawidłowości w strukturze.

Do badania wybrałem około 40 blogów i stron internetowych, które najpierw pobierałem, a potem ręcznie oznaczałem. Blogi wybierałem samodzielnie z listy stron, które były problematyczne przy algorytmie konkurencyjnym.

Narzędzie opracowywałem na platformie Windows, w związku z tym, część narzędzi skryptowych, jak skrypty linuxowe opracowane przez pracowników Politechniki Wrocławskiej, nie mogą zostać przeze mnie użyte i muszę wówczas opracować alternatywny algorytm.

## Kluczowe pojęcia i terminy używane w tej pracy

### Korpus językowy

Potrzebne fajne źródło słownikowe

### Anotacja

Potrzebne fajne źródło słownikowe

### Crawler

Potrzebne fajne źródło słownikowe

### Drzewo DOM

Potrzebne fajne źródło słownikowe

### Selektor

Potrzebne fajne źródło słownikowe

### Algorytm pozycyjny

Jest to opracowane przeze mnie podejście do analizy strony internetowej. Zamiast przedstawiać każdy element na stronie jako selektor o znanym typie, klasie bądź identyfikatorze, oznaczam każdy element jako ścieżka, jaką trzeba przebyć od korzenia do tego elementu. Pozwala to podejście definiowanie łatwych i jednoznacznych reguł w przypadku:

1. Nieistniejących nazw klas bądź identyfikatorów dla konkretnych elementów
2. Gdy element jest tej samej klasy lub (co jest błędne według W3C) ma ten sam identyfikator w drzewie
3. Gdy tekst jest oddzielony selektorami nowej linii („br”) a interesująca nas dana jest w jednym akapicie lub drugim

Ale o tych zaletach więcej opiszę we wnioskach z badań.

### CCL

CCL jest to format korpusowy opracowany na Politechnice Wrocławskiej. CCL jako format powstał w opraciu o format XCES (który w dużej mierze służy również do gromadzenia i anotacji korpusów lingwistycznych). Jest to plik tekstowy w formacie XML. Poza przechowywaniem tekstu źródłowego, format ten może przechowywać interesujące lingwistyków własności tekstu, takie jak:

1. Podziały na akapity i zdania
2. Podziały na tokeny i informacje bez-spacjowe
3. Anotacje morfosyntaktyczne
4. Anotacje do styli fragmentów (chunk-style) z możliwymi dyskontynuacjami
5. Nagłowki syntaktyczne anotacji
6. Właściwości tokenów – a co za tym idzie – anotacji

<http://nlp.pwr.wroc.pl/redmine/projects/corpus2/wiki/CCL_format>

### Premorph

Istnieje wiele form pośrednich pomiędzy dowolnym dokumentem zawierającym tekst a formatem CCL. Można zwykły dokument tekstowy podzielić na akapity i zapisać w formacie tekstowym. Można dokument podzielony na akapity zapisać w formacie html (z podziałem na akapity przy pomocy selektora „p”). Istnieje jeszcze format premorph, z którego korzystałem.

Nazwa premorph pochodzi z właściwości, jaką ma ten format: jest to plik XML sformatowany w podobny sposób, jak pliki CCL z jedną różnicą: nie ma podziału na słowa i nie ma anotacji morfo-syntaktycznej. Plik premorph różni się od pliku CCL następującymi elementami

1. Podział na akapity
2. Brak podziału na zdania, tokeny.
3. Nie zawiera jakichkolwiek anotacji morfo syntaktycznych.

A co je łączy?

1. Zapis w XML-u
2. Podział na listy chunków (akapity)
3. Nazwa akapitu (chunku)

### Inforex

Inforex jest narzędziem utworzonym na Politechnice Wrocławskiej. Służy ono do analizy morfo syntaktycznej korpusów w formacie CCL, oraz do ręcznej anotacji elementów mowy jak nazwy jednostkowe, anafory, nadawanie sensu słowom niejednoznacznym i wszelkiego rodzaju relacje pomiędzy jednostkami. System również wspiera ręczne czyszczenie tekstu oraz automatyczną analizę, w której się zawiera segmentacja tekstu, analiza morfo syntaktyczna, selekcja słów do późniejszego nadania znaczenia w przypadku słów niejednoznacznych.

Marcińczuk, Michał; Kocoń, Jan; Marcin, Ptak and Kaczmarek, Adam, 2010, *Inforex*, CLARIN-PL digital repository, <http://hdl.handle.net/11321/13>.

Nie wykluczone, że będę wspominał niejednokrotnie o tym systemie na przestrzeni tego dokumentu, ze względu na to, że główną inspiracją do napisania tej pracy był właśnie system Inforex. Z tego też powodu moje narzędzie będzie konwertowało dokumenty do formatu CCL.

### Corpo-grabber

Jest to narzędzie, podobnie jak moje, służące do automatycznego gromadzenia korpusów internetowych z Internetu. Korzysta ono z takich narzędzi jak:

1. Httrack – do pozyskiwania stron internetowych
2. Fdupes – Do deduplikacji tekstów identycznych
3. Onion – Do deduplikacji tekstów podobnych.
4. WCRFT – Do konwersji tekstu podzielonego na akapity do formatu CCL. Więcej o nim w podrozdziale.

Moje narzędzie opierałem na corpo-grabberze, nawet je podobnie nazywałem. W początkowej fazie projektu korzystałem zarówno z httracka i narzędzi deduplikacyjnych takich samych jak w tym projekcie, jednakże z czasem i doświadczeniem zacząłem korzystać z innych.

### WCRFT

Wrocław CRF Tagger – jest to narzędzie stworzone na Politechnice Wrocławskiej, tak jak powyższe narzędzia, do analizy morfo syntaktycznej tekstu. Korzysta z algorytmu CRF: Conditional Random Fields, który służy do rozpoznawania wzorców i w nauczaniu maszynowym. Ma szereg różnych zastosowań, jednym z nich jest łatwa konwersja tekstów w języku polskim na korpusy lingwistyczne, z której korzysta moje narzędzie w formie finalnej.

<https://en.wikipedia.org/wiki/Conditional_random_field>

<http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki>

### Blog Reader

Jest to kolejne narzędzie opracowane na Politechnice Wrocławskiej, służące do ekstrakcji tekstu ze stron internetowych już pobranych. Korzysta z podejścia selektorowego do ekstrakcji tekstu. Podejście jest proste: definiujemy elementy, jakie chcemy pobrać ze strony i nadajemy im selektor CSS, według którego należy wyszukiwać tych elementów, które nas interesują. Informatycy są w stanie bez problemu wyznaczyć wystarczające reguły do opracowania selektora, jednakże osoby nieobyte w technologiach IT nie były w stanie ich zdefiniować, co było wielką wadą tego programu.

Moje narzędzie opierałem na Blog Readerze do celów badawczych. Była to najlepsza znana mi metoda ekstrakcji tekstu, gdzie chcemy pozyskać konkretną treść ze stron internetowych.

## Przegląd literatury

Literatura poruszająca problem pozyskiwania korpusów tekstu z internetu jest uboga, jednakże jest wiele prac naukowych dotykających problematyki zautomatyzowanego pobierania zasobów internetowych, oraz dotyczących wykorzystania korpusów tekstu w różnych dziedzinach wiedzy. W tym rozdziale postaram się przedstawić najistotniejsze źródła, z których korzystałem przy pracy, opiszę co zawierają oraz opiszę stosunek tych prac z moją pracą.

### Internet Documents: A Rich Source for Spoken Language Modeling

W tej pracy zaproponowano metodologię pozyskiwania dokumentów internetowych w celu pozyskania tekstów do modelowania języka mówionego. Aktualne modele językowe często są oparte na tekstach pisanych i/lub nadużywanym Czarnoksiężniku z krainy Oz, albo na eksperymentalnych dialogach. Przy pomocy tekstów pozyskanych z internetu, potem przeczytanych przez eksperymentatorów, a następnie przeanalizowanych przez algorytmy rozpoznawania mowy otrzymano 15% wzrost skuteczności w rozpoznawaniu słów.

Moja praca będzie w dużej mierze będzie opierała się na pozyskiwaniu tekstów z internetu. Moim zadaniem jednak będzie przekształcanie ich na korpusy tekstów w formacie CCL, gdzie praca ta opierała się głównie na pozyskaniu tekstów do celów późniejszego ich odczytania. Analiza morfo-syntaktyczna byłaby im niepotrzebna.

### Efﬁcient Web Crawling for Large Text Corpora

Praca ta skupia się na efektywnym pobieraniu zasobów internetowych. Autorzy proponują różne rozwiązania, które usprawniają proces pobierania, ale również umożliwiają przy pomocy różnych metod umożliwić łatwą ekstrakcję tekstu. Autorzy mieli dwie główne motywacje: pobieranie danych ze źródeł zawierających najwięcej treści oraz możliwe unikanie duplikatów danych.

Pierwszy problem, czyli uzyskiwanie lepszych wyników pobierań, rozwiązali poprzez napisanie własnego crawlera, którego nazwali pajączek (spiderling). Zdefiniowali zależność pomiędzy ilością danych pobranych, ilością danych istotnych (czyli „gołego" tekstu) oraz ilością dokumentów pobranych. Wyznaczenie tej zależności umożliwiło im zredukowanie czasu pobierania danych oraz zwiększyło stosunek danych istotnych do danych pobranych (yield rate) poprzez odrzucanie domen, z których nie dało się wyekstraktować wystarczającej ilości danych.

Drugi problem, jakim jest unikanie duplikatów danych, uzyskali przez napisany przez nich program – jusText. Dokonywali deduplikacji w dwóch momentach: w momencie pobierania danych i po ekstrakcji tekstu. Nadawali dokumentom sumy kontrolne, które porównywali z pozostałymi dokumentami. W przypadku duplikatu, jeden z dokumentów usuwali a sumę zapisywali. Korzystali również z programu onion do deduplikacji tekstów zbliżonych.

W moim rozwiązaniu korzystam z gotowych rozwiązań do crawlowania. Nie potrzebuję wykrywać, czy strona zawiera treść. Zakładam, że naukowiec korzystający z mojego narzędzia wie jakiej treści oczekuje ze strony pobieranej, i że pobiera ją nie bez celu. Dlatego zależności wyznaczone przez autorów nie będą mnie interesowały.

Drugie rozwiązanie z ich pracy jest dla mnie przydatne. Duplikaty danych występują niemalże na każdej stronie internetowej. Choćby przykładem niech będą notki na blogach, które pojawiają się zarówno jako osobne strony, na stronach z komentarzami, oraz na stronach archiwum. Proces deduplikacji zatem jest potrzebny dla mnie po pobieraniu danych, ale również przed, jak i po analizie danych.

### A DOM Tree Alignment Model for Mining Parallel Data from the Web

Pracę tę znalazłem po sformułowaniu mojego problem. Opisuje ona inny problem niż mój na podstawie podobnych do moich założeń.

W tym dokumencie jest postawiona teza, że strony internetowe dwujęzykowe często posiadają taką samą strukturę strony i linków, niezależnie od języka, który się wybierze. Autorzy zbadali i zauważyli, że 10% spośród 150 tysięcy stron internetowych z domeny .de jest dwujęzykowa. W związku z tymi obserwacjami, powstało wiele crawlerów służących do pozyskiwania danych dwujęzycznych. Między innymi twórcy tej pracy, którzy zastosowali podobieństwo stron internetowych do ekstrakcji zdań w obu językach.

Ja również korzystam z tego podejścia. Jest ono bardzo wygodne i zastanawiające jest, że nikt nie zrobił tak prostego badania, jakim jest sprawdzenie jednolitości struktury stron internetowych.

# Dział II

## Corpo-grabber

Corpo-grabber jest moim podejściem do tematu, jakim jest rozwinięcie istniejącej już architektury na Politechnice Wrocławskiej. Moje narzędzie jest webowym skryptem w PHP. Ma ono za zadanie przy dyrektywach użytkownika pobrać wybraną stronę internetową, następnie po zakończonym procesie pobierania ma dokonać wstępnej deduplikacji stron internetowych, po czym wyświetla pobrany projekt. Następnie zadaniem użytkownika jest dla przykładowej pobranej strony internetowej zdefiniowanie interesujących go elementów strony. Po tym, mój program po głuższej analizie prezentuje wyniki pobierania.

### Dlaczego skrypty PHP?

Odpowiedź jest bardzo prosta. Podczas analizy strukturalnej strony internetowej potrzebna jest idealna wizualizacja strony internetowej. W związku z tym potrzebne jest narzędzie, które umie idealnie prezentować dane. Jest nią przeglądarka internetowa. W związku z tym warstwa prezentacyjna strony internetowej powinna być realizowana przy pomocy skryptów PHP.

Jednakże fragment części logicznej strony jest również zaprogramowana przy pomocy skryptów PHP. Wynika to z bardzo prostego faktu – do języka PHP napisanych zostało wiele interesujących skryptów, które są w stanie analizować drzewa DOM. Korzystam z nich i są bardzo pomocne. Jednakże do celów optymalizacyjnych powinno się te narzędzia napisać w zewnętrznym skrypcie, jak w języku Java (dla pracy na różnych systemach) czy Python (by konturować trendy na Politechnice Wrocławskiej).

### Elementy

Program dzieli się na parę elementów, które są wyznaczone przez interfejs:

1. Część dotycząca pobierania
2. Część dotycząca oznaczania stron

Z czego część druga ma jeszcze następujące podczęści:

1. Część dotyczącą wyboru podstrony
2. Podgląd podstrony
3. Widok drzewa podstrony
4. Analiza drzewa podstrony

Program korzysta z takich zewnętrznych programów jak:

1. WGET – program służący do pobierania stron internetowych
2. PHPHtmlParser – wtyczka PHP służąca do analizy drzewa DOM
3. Simple MVC - Prosta biblioteka MVC do PHP.

WGET jest wykorzystywany tylko w części pierwszej. Natomiast PHPHtmlParser jest wykorzystywany w części drugiej w podczęści 3 i 4. Sens takiego wykorzystania opiszę w podrozdziale „Działanie”.

### Interfejs

W ciągu całego procesu tworzenia oprogramowania korzystałem z wielu różnych podejść do tworzenia strony internetowej. Chciałem z początku podejść do tematu przy pomocy bardzo ubogiego interfejsu. Chciałem po prostu umożliwić wybieranie strony internetowej do pobrania i wyświetlić wynik. Jednakże wkrótce po tym okazało się, że nie da się spełnić wszystkich oczekiwań na prostym interfejsie. W związku z tym skorzystałem z biblioteki MVC jaką jest Simple MVC. Umożliwiła mi ona proste tworzenie interfejsu i prostej logiki.

W związku z tym, postanowiłem nagłówek przeznaczyć na łącza do podstawowych części strony:

1. Strona główna
2. Pobieranie pojedynczej strony
3. Pobieranie wielu stron z jednej domeny (Pobieranie zaawansowane)
4. Anotacja strony pobranej (Ekstrakcja wzorca)

#### Strona główna

Zawiera krótką adnotację dotyczącą przeznaczenia strony internetowej. Opisuje to, co możemy znaleźć na konkretnych podstronach.

#### Pobieranie pojedynczej strony

Z początku myślałem, że pobieranie pojedynczej strony internetowej będzie czymś istotnym na stronie internetowej. W związku z tym, po wprowadzeniu adresu strony algorytm automatycznie pobierał wybraną stronę internetową a następnie wyświetlał efekty ekstrakcji przy predefiniowanych parametrach. Jednakże nie były to parametry, które interesowały językowców, dlatego musiałem zrezygnować z tego pomysłu. Nadal jednak istnieje w projekcie.

#### Pobieranie zaawansowane

Jest to efekt pracy związany z adaptacją corpo-grabbera do projektu webowego. Interfejs zawiera podstawowe elementy, jak:

1. Nazwa projektu
2. Adres URL

W wersjach wcześniejszych występował jeszcze element trzeci: czas pobierania. Wynikało to ze specyfiki narzędzia, które wykorzystywałem. Z początku korzystałem z narzędzia zwanego Httrack (tak samo jako pierwotne narzędzie Corpo-grabber). Jednakże w pewnym momencie przestało spełniać moje oczekiwania, dlatego zacząłem korzystać z innego narzędzia, jakim jest WGET.

Element „nazwa projektu” pozwalał nam zdefiniować nazwę projektu do pobierania. Nie jest to element obowiązkowy, jako że w przypadku nie zdefiniowania nazwy projektu, moje narzędzie definiuje domyślną nazwę jako (kolejno): datę pobierania i stronę internetową (pozbawioną prefiksu http:// i www., oraz wszystkich nielegalnych znaków pod systemem Windows).

Element „adres URL” jest obowiązkowy. Służy ono do zdefiniowania adresu pobierania treści. Od tego adresu rozpoczynane jest pobieranie a domeną treści jest uznawane wszystko przed pierwszym wystąpieniem w adresie znaku ukośnik „/”

#### Ekstrakcja wzorca

Jest to efekt mojej samodzielnej pracy. Element ten zawiera 4 główne części:

1. Menu wybierania projektu i podstrony
2. Widok podglądu strony
3. Widok drzewa strony
4. Menu boczne

Ogólnie, ta podstrona pozwala nam na wyznaczenie wzorca, według którego możemy dokonać ekstrakcji danych ze strony. Wszystkie te elementy służą temu celu.

Menu wybrania projektu i podstrony pozwala nam na zdefiniowanie projektu, który nas interesuje (można wybrać każdy przeszły, jeżeli nasze reguły kiedyś nam się nie spodobały). W tym menu również możemy zdefiniować, którą stronę chcemy wybrać, aby było widać w widoku podglądu strony i drzewie strony.

Menu boczne jest zawiera informacje i metody pozwalające oznaczać interesujące nas elementy. W formie badawczej występują tam dwa elementy: przyciski do zaznaczania elementów na drzewie strony oraz pola tekstowe do (odpowiadających tym przyciskom) wyznaczania reguł ekstrakcji elementów z drzewa. Służyło to głównie porównaniu metod ekstrakcji: mojej i konkurencyjnej.

W formie ostatecznej, która zostanie przekazana Politechnice Wrocławskiej do celów dalszego rozwoju, nie będzie już w menu bocznym pól tekstowych. Służyły one jedynie badaniu i ich istnienie nie ma sensu w praktycznym użyciu.

### Działanie

Tak jak wspomniałem w rozdziale o interfejsie i elementach, korzystałem z wtyczki Simple MVC do administrowania strukturą model – widok – kontroler. Rozdzieliła wówczas elementy strony na podkatalogi (w katalogu app):

1. Controllers – zawieta kontrolery
2. Core – zawiera dane konfiguracyjne, w tym definicję możliwych adresów na stronie
3. Templace i Views – zawierające widok strony z różnych perspektyw
4. PHPHtmlParser – zawiera wtyczkę do obsługi drzewa DOM
5. Pozostałe katalogi – zawierają dane potrzebne przez skrypty Simple MVC

Dodatkowo moje narzędzie wyznaczyło dodatkowy katalog, jakim jest katalog tmp. Ten katalog zawiera wszystkie pobierania, jakie wykonali użytkownicy. W każdym katalogu znajduje się katalog Web. Jest to efekt wstecznej kompatybilności z programem httrack, który podczas pobierania stron, pliki html umieszczał w katalogu web a pozostałe dane, jak obrazki umieszczał w innych katalogach. W katalogu web po dokonaniu analizy drzewa moim algorytmem jest tworzony katalog premorph, w którym umieszczone są wszystkie utworzone wówczas pliki xml a następnie pakowane do pliku zip do downloadu przez użytkownika. Dodatkowo, jeżeli zostały zdefiniowane parametry do badania, w katalogu dla każdego badania pojawiał się osobny plik zip oraz plik report.html, który zawierał dane szczegółowe dotyczące badania.

#### Algorytm pobierania: pojedyncza strona

Algorytm ten polega na skorzystaniu z wbudowanej biblioteki PHP zwanej CURL. Przy odpowiednich parametrach, biblioteka ta po chwili zwraca wynik pobierania dla zadanej strony. Następnie przy pomocy algorytmu ekstrakcji tekstu wyciągam wszystkie elementy tekstowe ze strony z „podziałem na akapity”. Podział na akapity jest tu w cudzysłowie, gdyż algorytm dzielenia na akapity w przypadku analizowania całej strony wyciąga wówczas elementy niepożądane, na przykład wszystkie linki z paneli bocznych są traktowane jako osobny akapit.

#### Algorytm pobierania: wiele stron

Algorytm ten polega na wywołaniu zewnętrznego programu zwanego WGET. WGET ma wiele opcji do pobierania, jedną z nich jest między innymi pobieranie rekursywne, gdzie parametrem ograniczającym zagłębianie się w linki jest głębokość rekursji. Jednakże w ramach działania programu nie zmieniałem go i ma on wartość 5.

Oto polecenie, przy pomocy którego wywołuję program wget w programie:

*"wget --recursive -nd --page-requisites --html-extension --convert-links restrict-file-names=windows -P \"$path\" -A html,htm -w 1 --random-wait domains $URL\_short –no-parent $URL -o \"$path\_log\" -nv"*

Gdzie:

1. $path – to ścieżka absoultna katalogu, do którego jest pobierana strona
2. $URL – adres strony taki, jaki został wprowadzony przez użytkownika
3. $URL\_short – adres strony sformatowany przeze mnie do celów uzyskania domeny przeszukiwań. Usuwam prefiks „http://”, „www.”, oraz sufiks „/(…)”.
4. $path\_log – adres pliku, do którego należy zapisywać logi programu.

Dodatkowo, przy pomocy oprogramowania jQuery wykonuję cykliczne zapytanie do serwera, aby pokazał plik logów. Serwer w tym czasie dokonuje deduplikacji danych i przedstawia logi programu w odwróconej kolejności linii (dzięki czemu ostatni linie z logów pojawiają się jako pierwsze na stronie).

W przypadku wyłączenia strony w przeglądarce katalog nie jest analizowany w kierunku deduplikacji, co może poważnie wydłużyć czas otwierania projektu przy późniejszej próbie deduplikacji strony.

#### Algorytm deduplikacji

Jest on wywoływany w kilku momentach: w trakcie pobierania oraz w trakcie otwierania projektu. W programie został zastosowany podstawowy mechanizm deduplikacji. W przypadku wystąpienia dwóch plików w tym samym projekcie o tej samej wartości hashu MD5, jeden z nich jest usuwany. W przypadku identycznego rozmiaru plików, jeden z nich jest usuwany.

Dodatkowo, żeby jeszcze zmniejszyć rozmiar plików i ułatwić ich podgląd usuwane z plików html są wszystkie skrypty. Zostawia to wówczas drobne artefakty na stronie, ale nie są one zbyt wielką przeszkodą.

#### Algorytm tworzący drzewo

Jest bardzo prosty. Tworzę drzewo html poprzez rekursję: aktualny element otrzymuje od elementu nadrzędnego ścieżkę, w której ten element się znajduje. Wówczas on zapisuje sobie, że jest elementem o identyfikatorze [ta ścieżka].

Pojedynczy element w ścieżce jest zapisywany poprzez zapisanie typu elementu, jego wszystkich klas (oddzielonych spację pomiędzy sobą i kropką od typu) i ilości elementów danego typu na ścieżce napotkanych spośród dzieci tego samego rodzica (oddzielonych dwukropkiem). Przykładowo, selektor „div” o klasie „class” będący drugim elementem typu „div” u danego rodzica będzie miał w ścieżce adnotację o następującym wyglądzie:

*div.class:2*

W przypadku, gdy dziecko jest elementem tekstowym, zamiast typu jest ten element oznaczany jako typ „text”.

Każdy z elementów w ścieżce rozdzielam myślnikiem. To wymusza na mnie, że wszystkie elementy w ciągu ścieżki nie mogą zawierać ani kropki, dwukropka ani myślnika. W przypadku ich wystąpienia, w kodzie ścieżki są one zastąpione podkreślenie: „\_”. Przykładowy element, będący już elementem drzewa html:

*root-html.v2:1-body.loading variant\_birds:1-div.navbar section:1-div.widget Navbar:1-div:1*

Jeżeli w tej gałęzi znajdują się podgałęzi (dzieci), to każdy element otrzymuje identyfikator z dopisanym elementem opisującym tę podgałąź. Dla przykładu, jeżeli dziećmi powyższej gałęzi są dwie tablice o klasach „table-inner” i „table-inner2”, to dzieci te otrzymają identyfikatory:

*root-html.v2:1-body.loading variant\_birds:1-div.navbar section:1-div.widget Navbar:1-div:1-table.table\_inner table\_inner2:1*

*root-html.v2:1-body.loading variant\_birds:1-div.navbar section:1-div.widget Navbar:1-div:1-table.table\_inner table\_inner2:2*

I zostaną umieszczone w drzewie strony kolejno z adnotacja tekstową zawierającą typ i klasę. Finalnie, przykładowy element wygląda następująco a dzieci pojawiłyby się pod selektorem „br”:

*<div id="root-html.v2:1-body.loading variant\_birds:1-div.navbar section:1-div.widget Navbar:1-div:1"><b>div</b><br></div>*

Ten algorytm umożliwia mi utworzenie strony internetowej, w której elementy tekstowe są zagnieżdżone w strukturze, która jest zbliżona do tej na stronie. Co więcej, każdy element mojego drzewa zachowuje informację o tym, jakie elementy są po drodze. Dzięki temu oznaczając konkretny element jako zawierający dane, nie muszę przechodzić przez całe drzewo ponownie w celu zidentyfikowania jakie elementy są po drodze.

#### Oznaczanie elementów drzewa

#### Algorytm przechodzenia drzewa

#### Algorytm ekstrakcji tekstu

## Badanie

### Teoria

### Praktyka

### Wyniki badań

# Dział III

## Wnioski z badań

## Podsumowanie

## Źródła

NIEPOSORTOWANE

<http://dl.acm.org/citation.cfm?id=775182>

Możliwe źródła:

<http://www.teachingenglish.org.uk/article/analysing-language> - elementy języka

<http://korpusy.net/> - definicja korpusu

<http://nlp.pwr.wroc.pl/redmine/projects/corpus2/wiki/CCL_format> - o formacie CCL

<http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf> - definicje dot. korpusów

<https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-000D-F696-9> - jusText, Pomikalek Jan

Alternatywy do corpograbbera i inforex:

<http://www.textise.net/>

<https://www.diffbot.com/>

<https://github.com/kohlschutter/boilerpipe> // fulltext extraction

<https://code.google.com/p/boilerpipe/wiki/Benchmarks>

<http://www.crummy.com/software/BeautifulSoup/>

<http://rodp.me/2015/how-to-extract-data-from-the-web.html>

<http://www3.nd.edu/~tweninge/pubs/WH_TIR08.pdf>

<https://hal.archives-ouvertes.fr/inria-00326150/document>

<https://books.google.pl/books?id=vFHOx8wfSU0C&printsec=frontcover&dq=text+mining&hl=en&sa=X&ved=0ahUKEwj3usnaldjLAhVI7xQKHRmpDqwQ6AEIODAD#v=onepage&q=text%20mining&f=false>