# Plan

1. Wstęp

O tym, co praca będzie zawierała: programy innych twórców, technologie użyte przeze mnie

O formacie CCL, O serwisie Inforex, O programie corpo grabber

1. Dział 1: technologie używane przez innych w podobnych problemach
2. Dział 2: Opis algorytmu do wyłapywania korpusów
3. Dział 3: Opis algorytmu do tagowania korpusów
4. Badania skuteczności
5. Wnioski

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

# 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.

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.

# 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 metody pozyskiwania korpusów lingwistycznych na podstawie stron internetowych. Narzędzie to ma być przystosowane do współpracy z użytkownikiem, który nie ma wiele wspólnego ze współczesną technologią. Nakłada to następujące ograniczenia:

- interfejs musi być przejrzysty

- użytkownik musi wiedzieć, co robi

- użytkownik nie musi wiedzieć z czego składa się strona internetowa, aby program działał

# 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.

# 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.

**Ilość czy liczba? \/**

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 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 zarówno przed, jak i po analizie danych.

# Kluczowe pojęcia i terminy używane w tej pracy

Korpus językowy

Anotacja

CCL

Inforex

Corpo Grabber

## CCL

CCL jako format powstał w opraciu o format XCES (który w dużej mierze służy również do gromadzenia i anotacji korpusów lingwistycznych). 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