Získávání a zpracování textových korpusů z internetu

Crawling, deduplikace, odstraňování boilerplate

Vít Suchomel

Centrum zpracování přirozeného jazyka
Fakulta informatiky, Masarykova univerzita

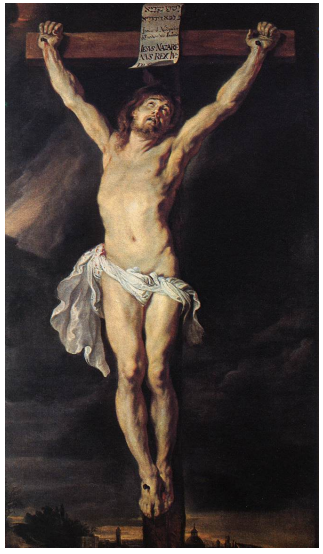
13. března 2017

Korpus

Co je to korpus?

Korpus

Co je to korpus?



Jazykový korpus

- Rozsáhlý soubor
- *autentických* textů (psaných nebo mluvených)
- převedený do *elektronické podoby*,
- v němž je možné jednoduše vyhledávat *jazykové jevy* (zejména slova a slovní spojení)
- a zobrazovat je v jejich *přirozeném kontextu*.[1]

[1] Zdroj: Český národní korpus, `http://wiki.korpus.cz/doku.php/pojmy:korpus`

Velikost jazykových korpusů

- Počet unikátních slov – velikost slovníku z korpusu, *types*
- Počet výskytů slov – velikost korpusu, *tokens*

Velikost jazykových korpusů

- Počet unikátních slov – velikost slovníku z korpusu, *types*
- Počet výskytů slov – velikost korpusu, *tokens*
- *Type-Token Ratio*: $\frac{types}{tokens}$

Velikost jazykových korpusů

- Počet unikátních slov – velikost slovníku z korpusu, *types*
- Počet výskytů slov – velikost korpusu, *tokens*
- *Type-Token Ratio*: $\frac{types}{tokens}$
- Je TTR větší u češtiny nebo angličtiny? Kolikrát?

Velikost jazykových korpusů

- Počet unikátních slov – velikost slovníku z korpusu, *types*
- Počet výskytů slov – velikost korpusu, *tokens*
- *Type-Token Ratio*: $\frac{types}{tokens}$
- Je TTR větší u češtiny nebo angličtiny? Kolikrát?
- TTR češtiny ku angličtině $\doteq 1,87$ (český web 2012, anglický web 2013)

Korpusová lingvistika

- Věda zkoumající *jazyk* (lingvistika)
- pomocí *jazykových korpusů*.

Query **korpus** **17,177** > GDEX **17,177** (3.39 per million)

Page 1 of 1,718 Go Next | Last

sportovniobchod.cz	Ke **korpusu** pouze připevníte 4 nohy a stolní fotbal je okamžitě provozu schopný.
kultura21.cz	Ti manuálně zruční mohou sestavit **korpus** stoličky z dvanácti stejných připravených dílů.
umimeudelatdomov.cz	Také je možná varianta smrkového **korpusu** s bílými předními dvířky.
ceskahospodynka.cz	Krémem bez ořechů potřeme vrchní **korpus** a boční stěny.
stylove-kuchyne.cz	Vyšší tloušťka používaných desek **Korpusy** skříněk jsou vyrobeny z konstrukčních desek o síle 18 mm.
jacques.cz	Všechny dřevěné díly postele jsou z bukového masivu, **korpus** má tloušťku 30 mm.
ok1blog.blog.cz	Tělo je vyrobeno z hliníkového **korpusu** , a části mechanismu jsou skutečně kovové.
apetitonline.cz	**Korpus** dejte znovu minimálně na hodinu chladit.
korpus.cz	Celková velikost **korpusu** InterCorp se tak již blíží k 50 milionům slov.
mix.cz	Tmavý **korpus** je silný a intenzivní a s jedním soustem roste chuť na další.

Ukázka dat v korpusu – XML vertikál

```
<dokument zanr="blog"
  nazev="Dovolená v Paříži" datum="2011-10-28"
  url="http://karel.bloguje.cz/dovolena-v-parizi">
<odstavec nadpis="1">
<veta>
Po
sedmi
letech
v
kouzelné
Paříži
!
</veta>
</odstavec>
...
</dokument>
```

Ukázka anotací v korpusu – XML vertikál s anotacemi

```
<dokument zanr="blog" nazev="Dovolená v Paříži">
<veta nadpis="1">
Po        po        k7c6           0  8
sedmi     sedm      k4c6           1  7
letech    léto      k1gNnPc6       2  7
v         v         k7c6           3 10
kouzelné  kouzelný  k2eAgFnSc6d1   4  9
<entita druh="město">
Paříži    Paříž     k1gFnSc6       5  9
</entita>
!         !         kx             6 11
<NP>                               7  8
<PP>                               8 11
<NP>                               9 10
<PP>                              10 11
<S>                               11  -
</veta>
</dokument>
```

Použití korpusů

- Obecně: data ke studiu přirozeného jazyka
- Lexikografové: slovníky
- Lingvisté: jazykové analýzy, změny jazyka
- Sociologové: jak a o čem píšeme, která témata jsou aktuální
- Marketingoví experti: hodnocení značek a výrobků v textech
- Statistické nástroje ZPJ: jazykové modely pro značkovače, analyzátory, překladové systémy, prediktivní psaní,...

Použití jazykového korpusu v lexikografii

korpus

czTenTen [2012] freq = 17,176 (3.38 per million)

a_modifier		
	7,932	**0.46**
dortový +	213	9.31
dortový korpus		
piškotový +	145	8.90
piškotový korpus		
vychladlý +	139	8.56
. Vychladlý korpus		
upečený +	193	8.46
. Upečený korpus		
závislostní	65	7.94
Pražský závislostní korpus		
anotovaný	40	7.27
anotovaný korpus		
předpečený	32	7.01
na předpečený korpus		
diachronní	33	6.96
Diachronní korpus		
synchronní	55	6.90

gen_1		
	1,923	**0.11**
skříňka +	203	6.44
Korpusy skříněk		
skříň +	150	4.71
korpus skříně		
buben	34	4.67
korpus bubnu		
komoda	9	4.67
Kristus	87	4.63
s korpusem Krista		
skříňka	6	4.48
dort	41	4.05
korpus dortu		
věnec	20	3.91
korpus věnce ,		
svítidlo	16	3.35
trezor	8	3.25
varhany	8	3.06

prec_verb		
	784	**0.05**
upéct	41	5.40
potřít	19	4.30
péct	24	3.29
peču korpus		
disponovat	16	0.66
disponuje korpusem typu		
tvořit	48	0.11
tvoří korpus		

prec_na		
	657	**0.04**
rozetřít	21	6.43
rozetřeme na korpus		
navršit	6	5.88
nalít	44	4.80
natřít	25	4.77
natřeme na korpus		

post_verb		
	1,413	**0.08**
prokrojit	14	8.20
korpus prokrojíme		
rozkrojit	32	7.85
korpus rozkrojíme		
rozříznout	46	7.64
korpus rozřízneme		
prořízнout	9	5.34
vyklopit	14	5.29
korpus vyklopíme		
potřít	38	5.27
korpus potřeme		
pomazat	8	5.16
rozetřít	9	5.07
pokapat	7	4.90
postlat	14	4.80
upéct	14	3.83
vychladnout	7	3.71

Srovnání korpusů podle původu textů – tradiční korpusy

- Příklady: British National Corpus, Corpus of Contemporary American English, Google Books, Český národní korpus
- Vznik: na objednávku, obsahová komise
- Zdroje: nejčastěji tištěná média, potom skenování knih, přepisy rozhovorů

Srovnání korpusů podle původu textů – tradiční korpusy

- Příklady: British National Corpus, Corpus of Contemporary American English, Google Books, Český národní korpus
- Vznik: na objednávku, obsahová komise
- Zdroje: nejčastěji tištěná média, potom skenování knih, přepisy rozhovorů
- Výhody: úplná kontrola nad zdroji (kvalitní a bohaté informace o datech: autor, název, rok vydání, žánr, styl, oblast), známe rozložení typů textů v korpusu, zaručená kvalita textů (možnost opravy chyb)

Srovnání korpusů podle původu textů – tradiční korpusy

- Příklady: British National Corpus, Corpus of Contemporary American English, Google Books, Český národní korpus
- Vznik: na objednávku, obsahová komise
- Zdroje: nejčastěji tištěná média, potom skenování knih, přepisy rozhovorů
- Výhody: úplná kontrola nad zdroji (kvalitní a bohaté informace o datech: autor, název, rok vydání, žánr, styl, oblast), známe rozložení typů textů v korpusu, zaručená kvalita textů (možnost opravy chyb)
- Nevýhody: nákladnost (jednání s vlastníky dat), omezená velikost (nedostatečná pro některé účely)

Srovnání korpusů podle původu textů – tradiční korpusy

- Příklady: British National Corpus, Corpus of Contemporary American English, Google Books, Český národní korpus
- Vznik: na objednávku, obsahová komise
- Zdroje: nejčastěji tištěná média, potom skenování knih, přepisy rozhovorů
- Výhody: úplná kontrola nad zdroji (kvalitní a bohaté informace o datech: autor, název, rok vydání, žánr, styl, oblast), známe rozložení typů textů v korpusu, zaručená kvalita textů (možnost opravy chyb)
- Nevýhody: nákladnost (jednání s vlastníky dat), omezená velikost (nedostatečná pro některé účely)
- Shrnutí: reprezentativní vyvážený korpus daného jazyka

Srovnání korpusů podle původu textů – internetové korpusy

- Příklady: Web as Corpus, ClueWeb, TenTen corpora, Corpora from the Web
- Vznik: opakované stahování internetu
- Zdroje: texty na internetu

Srovnání korpusů podle původu textů – internetové korpusy

- Příklady: Web as Corpus, ClueWeb, TenTen corpora, Corpora from the Web
- Vznik: opakované stahování internetu
- Zdroje: texty na internetu
- Výhody: aktuální psaná podoba jazyka, velikost – pokrývá lépe/více jazykových jevů (ClueWeb 09: 70 mld. anglických slov), větší rozmanitost textů

Srovnání korpusů podle původu textů – internetové korpusy

- Příklady: Web as Corpus, ClueWeb, TenTen corpora, Corpora from the Web
- Vznik: opakované stahování internetu
- Zdroje: texty na internetu
- Výhody: aktuální psaná podoba jazyka, velikost – pokrývá lépe/více jazykových jevů (ClueWeb 09: 70 mld. anglických slov), větší rozmanitost textů
- Nevýhody: malá kontrola nad zdroji (neuspořádanost, nevíme, co stahujeme), nezaručená kvalita textů (ale množství správných tvarů převáží chyby), nežádoucí obsah, duplicity, spam

Srovnání korpusů podle původu textů – internetové korpusy

- Příklady: Web as Corpus, ClueWeb, TenTen corpora, Corpora from the Web
- Vznik: opakované stahování internetu
- Zdroje: texty na internetu
- Výhody: aktuální psaná podoba jazyka, velikost – pokrývá lépe/více jazykových jevů (ClueWeb 09: 70 mld. anglických slov), větší rozmanitost textů
- Nevýhody: malá kontrola nad zdroji (neuspořádanost, nevíme, co stahujeme), nezaručená kvalita textů (ale množství správných tvarů převáží chyby), nežádoucí obsah, duplicity, spam
- Shrnutí: velký korpus daného jazyka

Proč je velikost korpusů důležitá

- Většina jazykových jevů podléhá Zipfově rozložení

Proč je velikost korpusů důležitá

- Většina jazykových jevů podléhá Zipfově rozložení
- „There is no data like more data“ (Mercer, 1985)
- „More data usually beats better algorithms“

Proč je velikost korpusů důležitá

- Většina jazykových jevů podléhá Zipfově rozložení
- „There is no data like more data“ (Mercer, 1985)
- „More data usually beats better algorithms“
- Větší seznamy slov (více unikátních slov)
  $\Rightarrow$ lepší pokrytí slov jazyka
- Více vět s výskytem daného slova
  $\Rightarrow$ lepší příklady použití slov v kontextu
- Lepší pokrytí řídkých jazykových jevů
  $\Rightarrow$ více podkladů pro studium jazyka
- Více dat pro jazykové modely
  $\Rightarrow$ přesnější (?) jazykové modely s větším pokrytím

Ukázka: Slova rozvíjející frázi „deliver speech“

- BNC (96 M words): major (8), keynote (6).
- ukWaC (1,32 G words): keynote (125), opening (12), budget (8), wedding (7).
- enTenTen12 (11,2 G words): keynote (813), acceptance (129), major (127), wedding (118), short (101), opening (97), famous (80).
- enClueWeb09 (70,5 G words): keynote (3802), acceptance (1035), opening (589), famous (555), commencement (356), impassioned (335), inaugural (333).

Web crawler

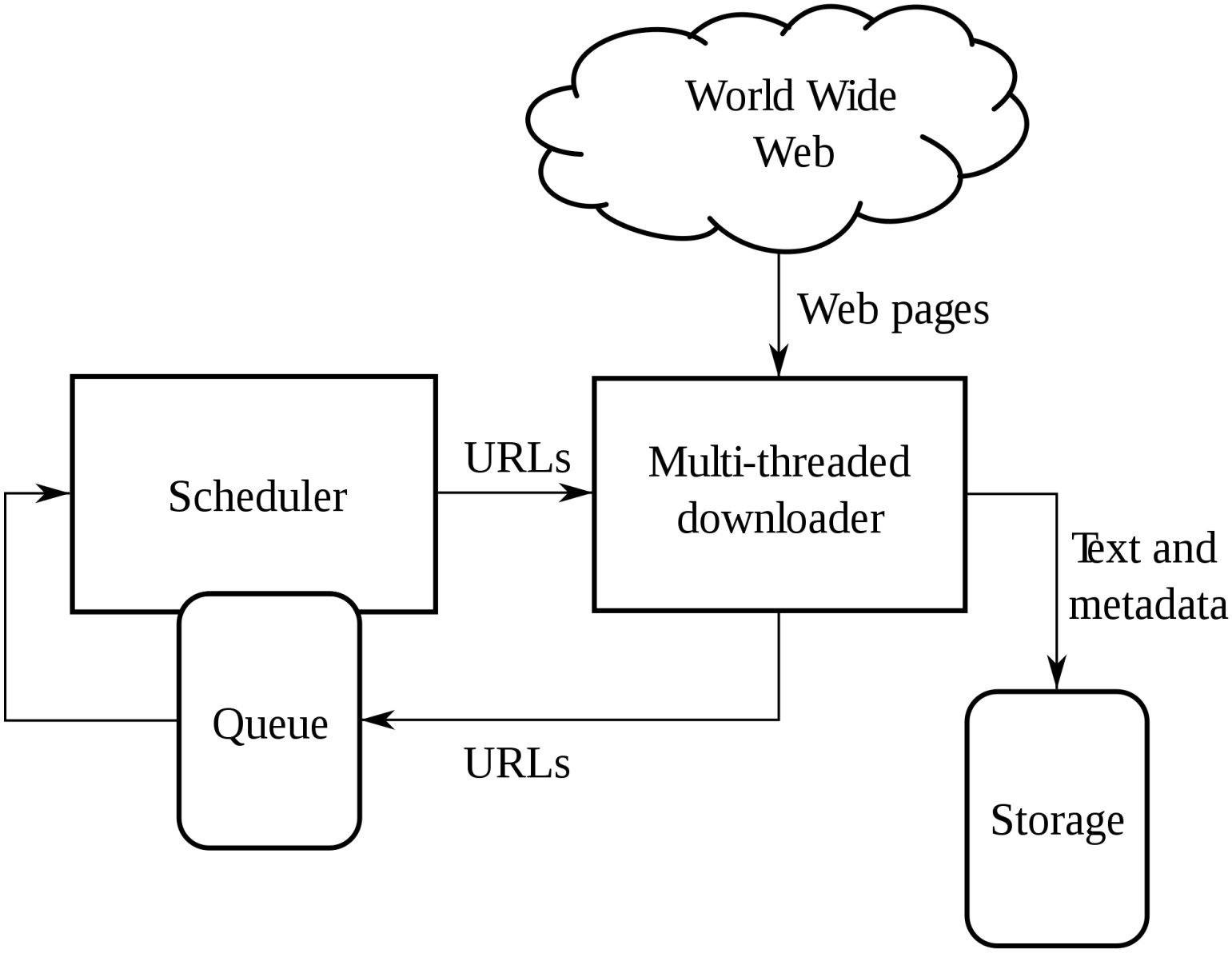
- Traverses the internet (graph of pages and links).
- Downloads documents (content & meta information).
- Stores documents (or their parts) in various formats for further use.
- Crawlers for various purposes:
  - GoogleBot – web indexing,
  - Linkcrawler – links, broken link checking,
  - Heritrix – general crawler, (Java, multiple treads),
  - SpiderLing – text corpora, (Python, multiple sockets).

Features a crawler should provide

- **Distributed:** Executable in a distributed fashion across multiple machines.
- **Scalable:** Scaling up the crawl rate by adding extra machines and bandwidth.
- **Performance and efficiency:** Efficient use of system resources (processor, memory, storage and network bandwidth).
- **Quality:** Biased towards fetching "useful" pages first.
- **Freshness:** Operate in continuous mode: obtain fresh copies of previously fetched pages, i.e. with a frequency that approximates the rate of change of that page. *Search engine crawler* $\rightarrow$ *the index contains a fairly current representation of each indexed web page.*
- **Extensible:** Cope with new data formats, new fetch protocols, various data processing needs. Modular architecture.

Source: Manning, Raghavan, Schutze: Introduction to Information Retrieval, Chapter 20

Basic crawler design

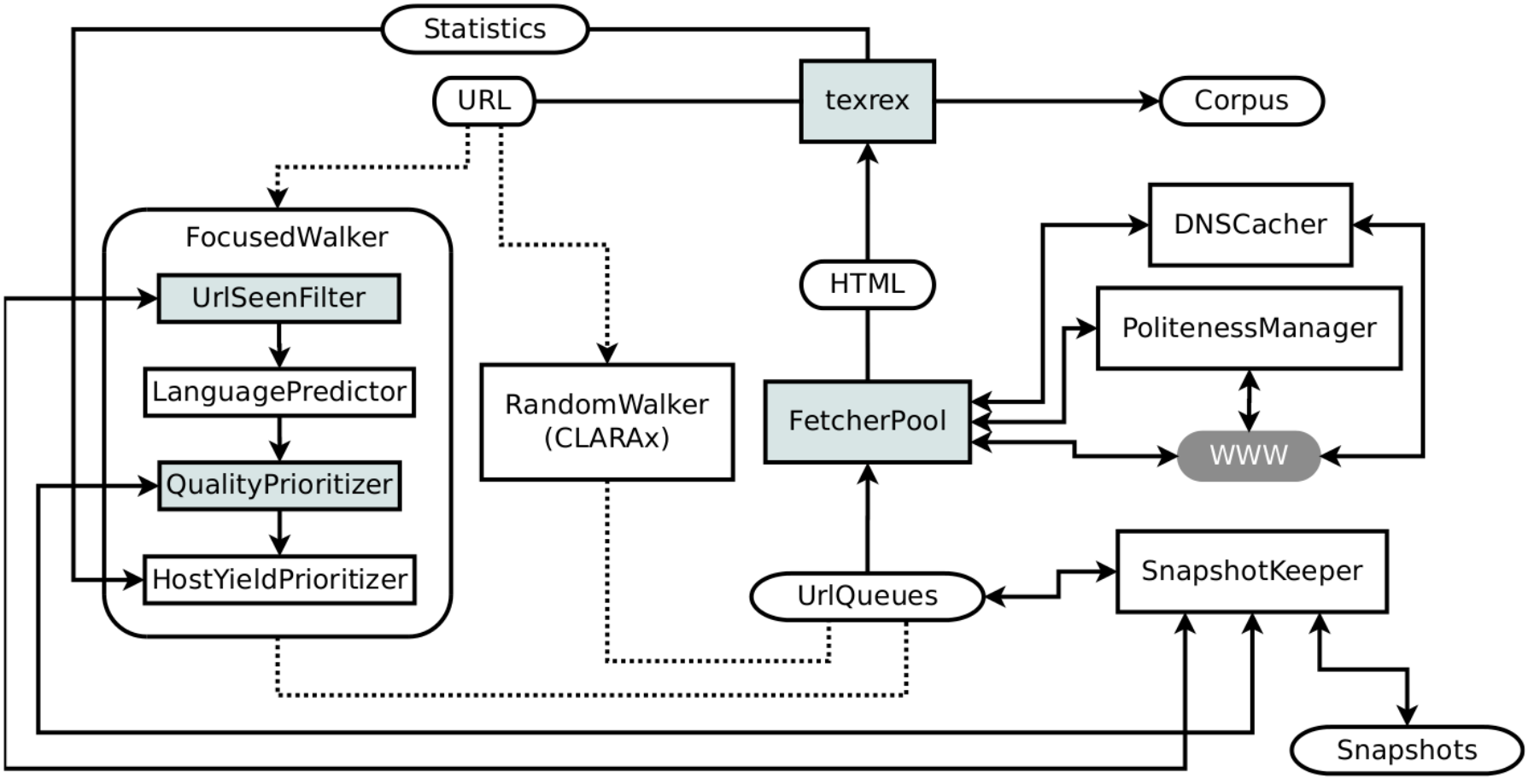


Source: http://en.wikipedia.org/wiki/Web_crawler

Advanced crawler implementation details

- Distributed vs. extensible.
- Multi-threaded synchronous vs. multi-socketed asynchronous.
- Web traversal policy:
  - depth vs. breadth,
  - domain selection,
  - domain distance,
  - focused crawling (topic oriented) vs. general crawling,
  - yield ratio.

Focused crawler design

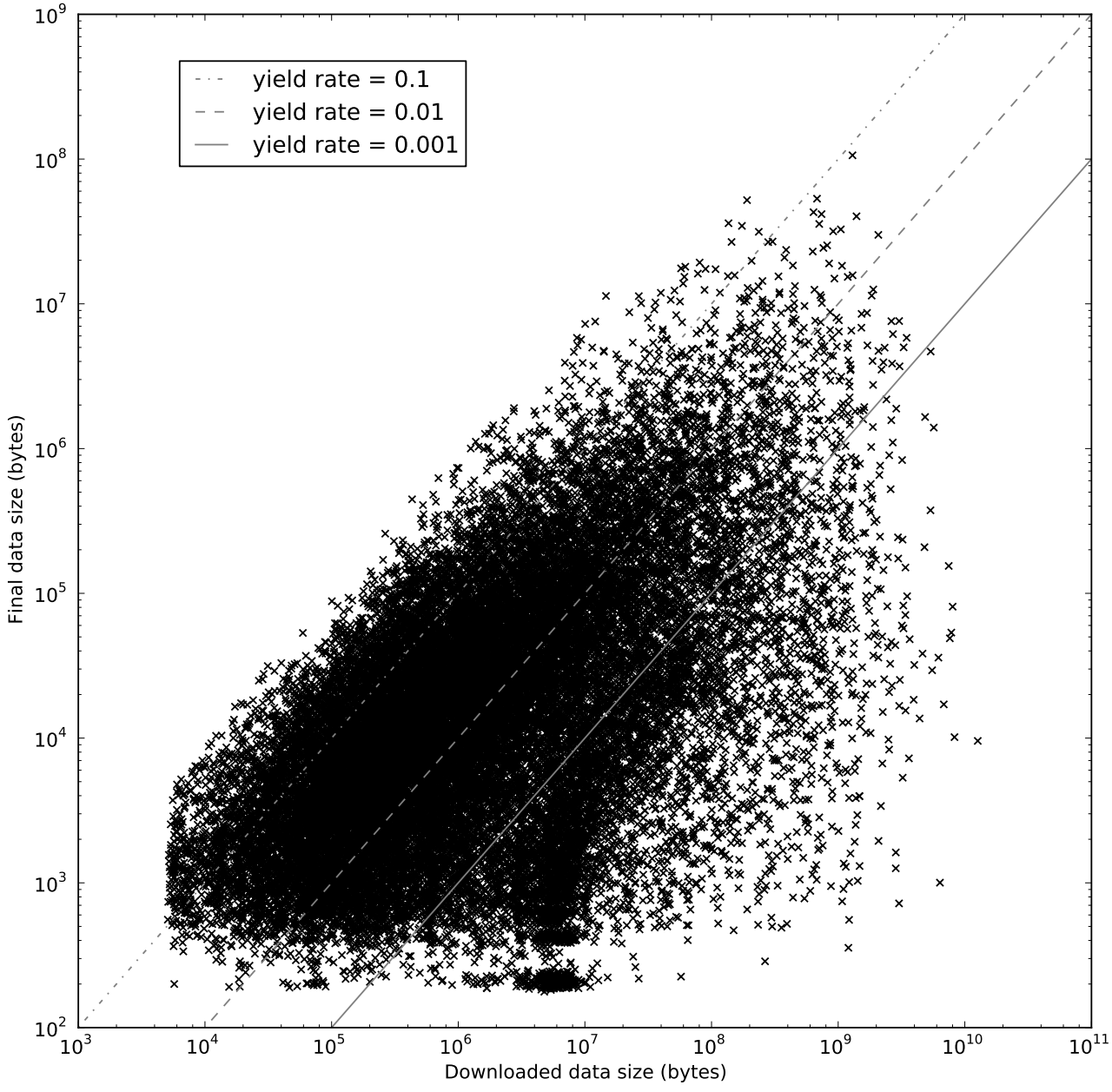


Source: Roland Schafer, Adrien Barbaresi, Felix Bildhauer. Focused Web Corpus Crawling. 9th Web as Corpus Workshop, 2014.

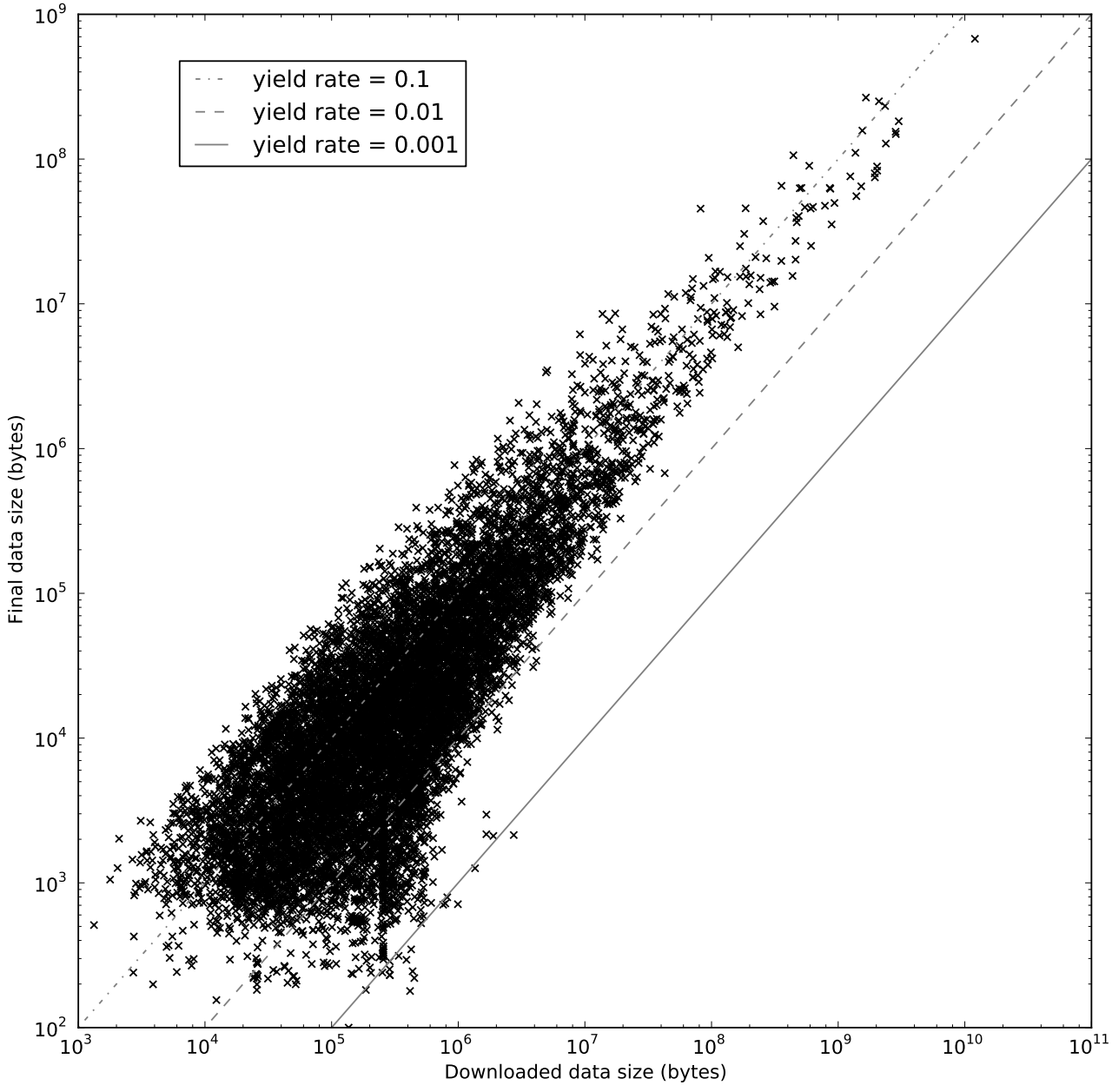
SpiderLing – crawler pro textové korpusy

- důraz kladen na efektivitu stahování
- *míra výtěžnosti* $= \frac{\text{velikost výsledných dat}}{\text{velikost stažených dat}}$
- crawler průběžně vyhodnocuje výtěžnost webových domén, zaměřuje se na „textově bohaté“ a odkládá stahování (nebo vůbec nestahuje) z neperspektivních webů
- cílem je sestavit korpusy velikosti $\geq 10^{10}$ slov pro všechny významné jazyky

General unfocused crawling efficiency (Heritrix)



Domain yield ratio optimised efficiency (SpiderLing)



Konkordance – autentické použití slova v textu

questions our archives. Explore the links we provide in each category. Use a	search	engine like Google. Questions should be specific . Questions like
Hello the best wordpress :p I have a request but it's not about WG =S I	search	Davichi - love & war mv gif but i don't find =/ I ask you because maybe
shook his head, "No." "Jump up and down." "What?" "If I	search	you, all I find I keep?" "You ain't searching me, man." Marco
gotten from searching sources to answer my curiosity. Using Google Scholar	search	engine, with key words: winter health tips, I compiled and present to
helps us to understand and trade the market swing. For example, we may	search	the chart for a reversal or breakout pattern that spells opportunity,
eventually lose interest in this type of price action and jump ship in	search	of a more exciting trading vehicle. The market loses broad sponsorship
numbers, astrological dates and prayer wheels have all been enlisted in the	search	for that elusive trading edge. Most traders believe Fibonacci fits
geographic area covered by each publication. Google News Archive	Search	-- Google's "News archive search provides an easy way to search and explore
publication. Google News Archive Search -- Google's "News archive	search	provides an easy way to search and explore historical archives. Users
Archive Search -- Google's "News archive search provides an easy way to	search	and explore historical archives. Users can search for events, people,
provides an easy way to search and explore historical archives. Users can	search	for events, people, ideas and see how they have been described over time
overview of the results by browsing an automatically created timeline.	Search	results include both content that is accessible to all users and content
Google has developed to put related stories together in the same news	search	result. Excellent place to search for news articles by keywords.
related stories together in the same news search result. Excellent place to	search	for news articles by keywords. Technorati --Search Technorati's
's database of over 1.9 weblogs and get up-to-date information on your	search	terms. British Library Newspaper Library -- The only large, integrated
monitor breaking news from over 3,000 sources, 24 hours a day. You can also	search	for news by zipcode and receive a grouping of news from numerous news
from numerous news sources in your area. NewspaperArchive.com --	Search	over 12.3 million newspaper pages. Not a complete resource to search
-- Search over 12.3 million newspaper pages. Not a complete resource to	search	but a good place to start research. Today's Front Pages -- "Today
morale and possibly paralyze you from taking necessary actions in your job	search	. If that isn't bad enough, it can also stop you from being hired! Here
save you a lot of hassle later on and you should do it early in your job	search	. In fact, this is something we all should do at least once a year: Find
on their own in the shortest time possible. Discover more insider job	search	secrets by visiting http://www.jobchangesecrets.com Art Canvas
quickest way to locate the information you want is to use the Label and	Search	Features. The Label pulldown is by far the easiest and quickest, just
the item of information that you require. You can use the very powerful	search	engine. Remember that each section on shows 'x' number of posts
section that you are currently viewing. Searching You can very easily	search	the section that you are currently in. To use, you simply enter
section that you are currently in. To use, you simply enter in your	search	item and hit Enter or click the Search Button . Your results are then

Konkordance – co je špatně?

Homepage What is this? | Create Your Own Homepage Change My **Search** To: Google Search MSN Search Yahoo! Search Ask

this? | Create Your Own Homepage Change My Search To: Google **Search** MSN Search Yahoo! Search Ask.com Search Wikipedia

Own Homepage Change My Search To: Google Search MSN **Search** Yahoo! Search Ask.com Search Wikipedia English

Change My Search To: Google Search MSN Search Yahoo! **Search** Ask.com Search Wikipedia English Yahoo! Answers

 Google Search MSN Search Yahoo! Search Ask.com **Search** Wikipedia English Yahoo! Answers Answers.com

a z . zaz DIRECT Everything You're Looking For Browse, **Search** , Find... Get Found (List Yourself) * Bookmark Your Favorites

france24.com Google News Google News Aggregated headlines and a **search** engine of many of the world's news sources. news.google.com

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier Founded in 1997, Jiangsu Xinye Chemical

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier O-Fluorobenzoyl Chloride Enquiry

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier Flutriafol Enquiry ...

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier O-Fluorobenzoyl Chloride Enquiry

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier O-Fluorobenzoyl Chloride Enquiry

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. **Search** products of this supplier Flutriafol Enquiry ...

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

 Contact Infomation Online Inquiry P-Chlorophenol **Search** products of this supplier Product Type: Agrochemicals

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

 Contact Infomation Online Inquiry O-Chlorophenol **Search** products of this supplier Product Type: Agrochemicals

 Gold Suppliers All Products All Suppliers Advanced **Search** Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates

Contact Infomation Online Inquiry 2,4-Dichlorophenol **Search** products of this supplier Product Type: Agrochemicals

Odstraňování nežádoucího obsahu

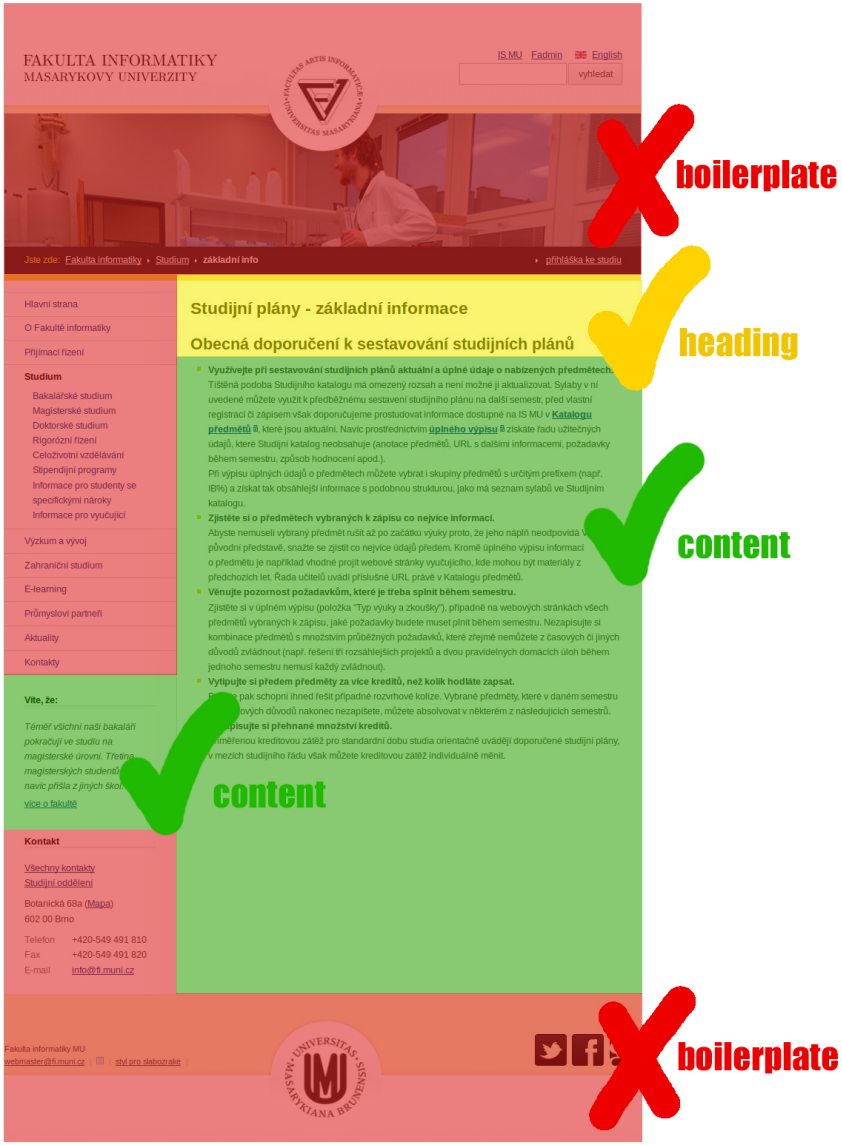
Nežádoucí obsah

- html značky, styly, poznámky
- negramatické věty: navigace, reklamy, tabulky, příliš krátké úseky,...

Používáme nástroj jusText
(`http://nlp.fi.muni.cz/projects/justext`)

- rozdělení na odstavce
- slovník častých slov v daném jazyce
- klasifikace odstavce podle délky, hustoty slov ze slovníku, hustoty odkazů, třídy okolních odstavců
- demo na stránce nástroje

Ukázka boilerplate



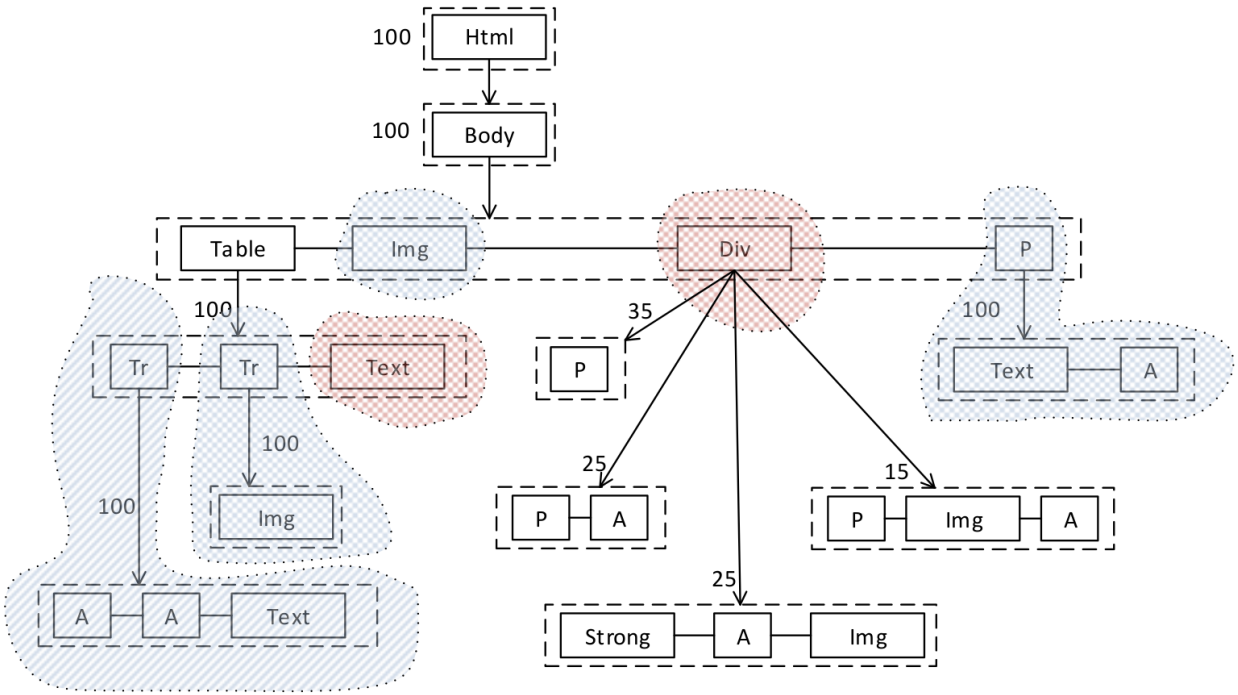
Zdroj: http://corpus.tools/attachment/wiki/Justext/Algorithm/cs_classification_example.png

Boilerplate removal approaches

- Machine learning (SVM, CRF, neural networks, n-gram models):
  - Annotated web pages required for training.
  - Victor (CRF),
  - Ncleaner (n-grams).
- Heuristics:
  - Rules for including/excluding sections of text.
  - BTE (tag density),
  - Boilerpipe (link/text ratio),
  - jusText (link/text ratio, frequent words, context sensitive – smoothing).

Site Style Tree (Yi, Lan, Bing Liu, Xiaoli Li, 2003)

- Represents both layout and content of a web page.
- Node importnace = node entropy over the whole Site Style Tree.



Zdroj: Ján Švec: Inteligentní detekování struktury webu, p. 32. Online: http://is.muni.cz/th/420072/fi_m/.

jusText

Context sensitive paragraph classification:

Demo: `http://nlp.fi.muni.cz/projects/justext/`

Konkordance – co je ještě špatně?

Homepage What is this? | Create Your Own Homepage Change My Search To: Google Search MSN Search Yahoo! Search Ask
this? | Create Your Own Homepage Change My Search To: Google Search MSN Search Yahoo! Search Ask.com Search Wikipedia
Own Homepage Change My Search To: Google Search MSN Search Yahoo! Search Ask.com Search Wikipedia English
Change My Search To: Google Search MSN Search Yahoo! Search Ask.com Search Wikipedia English Yahoo! Answers
 Google Search MSN Search Yahoo! Search Ask.com Search Wikipedia English Yahoo! Answers Answers.com
a z . zaz DIRECT Everything You're Looking For Browse, Search , Find... Get Found (List Yourself) * Bookmark Your Favorites
france24.com Google News Google News Aggregated headlines and a search engine of many of the world's news sources. news.google.com
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier Founded in 1997, Jiangsu Xinye Chemical
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier O-Fluorobenzoyl Chloride Enquiry
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier Flutriafol Enquiry ...
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier O-Fluorobenzoyl Chloride Enquiry
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier O-Fluorobenzoyl Chloride Enquiry
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Infomation Online Inquiry Jiangsu Xinye Chemical Co., Ltd. Search products of this supplier Flutriafol Enquiry ...
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
 Contact Infomation Online Inquiry P-Chlorophenol Search products of this supplier Product Type: Agrochemicals
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
 Contact Infomation Online Inquiry O-Chlorophenol Search products of this supplier Product Type: Agrochemicals
 Gold Suppliers All Products All Suppliers Advanced Search Browse Categories Popular Searches : Urea Fertilizer Dyes Intermediates
Contact Infomation Online Inquiry 2,4-Dichlorophenol Search products of this supplier Product Type: Agrochemicals

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí
  - Licenční ujednání

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí
  - Licenční ujednání
- Jak podobné texty jsou příliš podobné?

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí
  - Licenční ujednání
- Jak podobné texty jsou příliš podobné?
  - Členění textu: dokumenty, odstavce, věty, slova

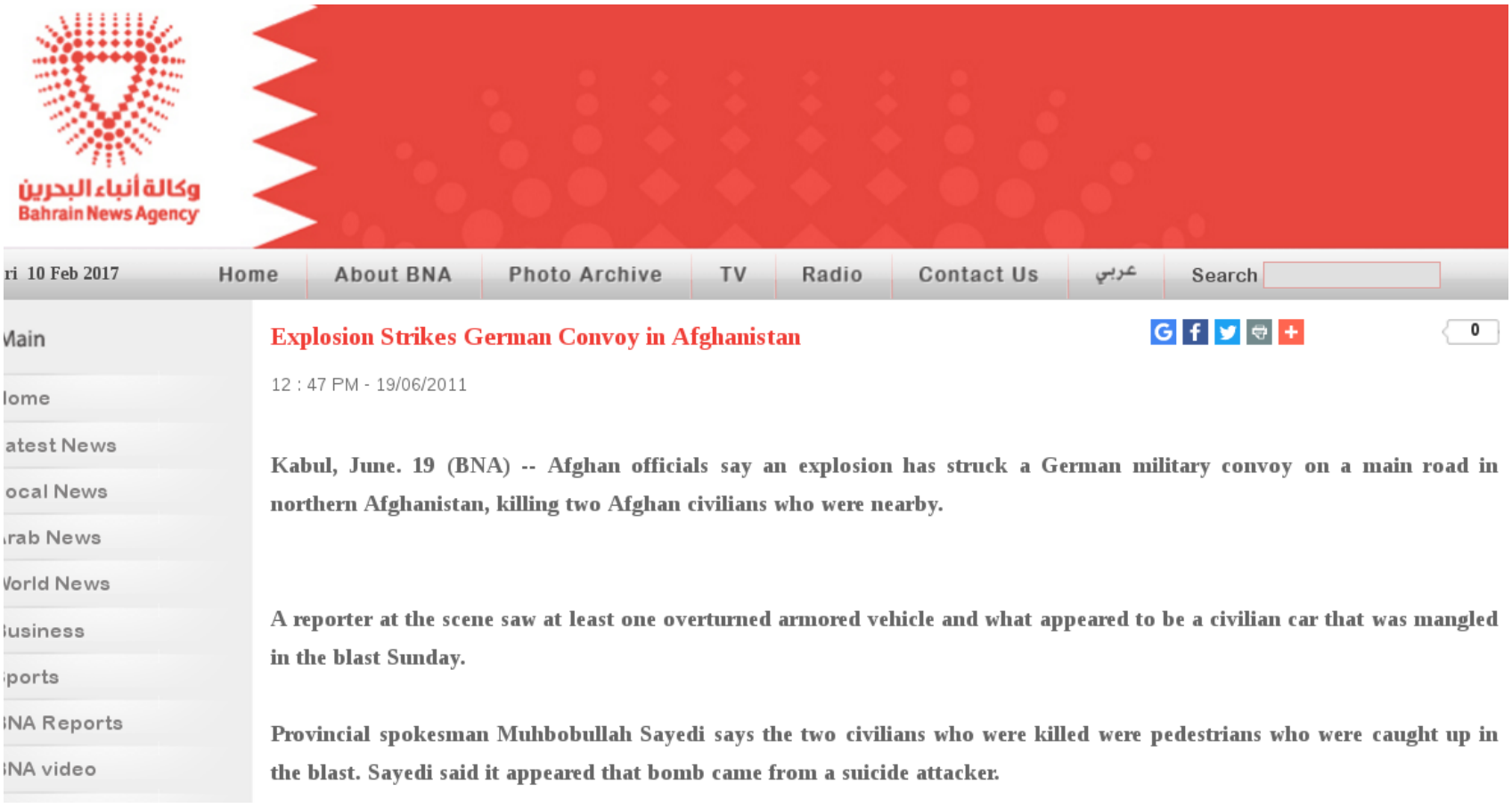
Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí
  - Licenční ujednání
- Jak podobné texty jsou příliš podobné?
  - Členění textu: dokumenty, odstavce, věty, slova
  - Jak můžeme měřit podobnost textů?

Duplicitní texty na internetu

- Jak duplicitní texty vznikají?
  - Hlavičky, patičky, navigace, copyright
  - Kopírování – informace na více místech
  - Levné získání obsahu vykradením textu z původních zdrojů
  - Média přejímají agenturní zprávy
  - Minimální změny: sportovní výsledky, předpověď počasí
  - Licenční ujednání
- Jak podobné texty jsou příliš podobné?
  - Členění textu: dokumenty, odstavce, věty, slova
  - Jak můžeme měřit podobnost textů?
  - Jak nastavit práh podobnosti?

Ukázka agenturní zprávy – Bahrain News Agency



وكالة أنباء البحرين
Bahrain News Agency

ri 10 Feb 2017 | Home | About BNA | Photo Archive | TV | Radio | Contact Us | عربي | Search

Main
lome
atest News
ocal News
rab News
Vorld News
usiness
ports
NA Reports
NA video

Explosion Strikes German Convoy in Afghanistan

0

12 : 47 PM - 19/06/2011

Kabul, June. 19 (BNA) -- Afghan officials say an explosion has struck a German military convoy on a main road in northern Afghanistan, killing two Afghan civilians who were nearby.

A reporter at the scene saw at least one overturned armored vehicle and what appeared to be a civilian car that was mangled in the blast Sunday.

Provincial spokesman Muhbobullah Sayedi says the two civilians who were killed were pedestrians who were caught up in the blast. Sayedi said it appeared that bomb came from a suicide attacker.

Ukázka agenturní zprávy – Dawn

DAWN

Explosion strikes German convoy in Afghanistan

AP — PUBLISHED Jun 19, 2011 08:35am

KUNDUZ, Afghanistan: A suicide car bomber struck a German military convoy in northern Afghanistan on Sunday, detonating explosives that killed three Afghan civilians and overturned at least one armored vehicle, according to officials and witnesses.

Taliban spokesman Zabiullah Mujahid claimed responsibility for the attack.

The bomber blew his vehicle up shortly before 10 a.m. local time on a busy road on the edge of Kunduz city, near the airport.

Ukázka agenturní zprávy – San Diego Union Tribune

Explosion strikes German convoy in Afghanistan

By KATHY GANNON, Associated Press

JUNE 18, 2011, 11:37 PM | KUNDUZ, AFGHANISTAN

A suicide attacker blew up his explosives-laden car next to a German military convoy in northern Afghanistan on Sunday, killing three Afghan civilians, officials and witnesses said.

Taliban spokesman Zabiullah Mujahid claimed responsibility for the attack.

The bomber detonated his vehicle shortly before 10 a.m. local time on a busy road on the edge of Kunduz city, near the airport, the Afghan Interior Ministry said in a statement. An Associated Press reporter at the scene saw at least one overturned armored vehicle and what appeared to be a civilian car that was mangled in the blast.

Ukázka agenturní zprávy – Frum Forum

FrumForum.tm

search FF

Search FF GO

HOME BLOGGER INDEX BOOKCLUB ABOUT US CONTACT US BEST OF FF RSS

Taliban Targets German Troops

June 19th, 2011 at 12:18 pm FRUMFORUM NEWS | No Comments | Share | Print

The Associated Press **reports:**

A suicide attacker blew up his explosives-laden car next to a German military convoy in northern Afghanistan on Sunday, killing three Afghan civilians, officials and witnesses said.

Taliban spokesman Zabiullah Mujahid claimed responsibility for the attack.

The bomber detonated his vehicle shortly before 10 a.m. local time on a busy road on the edge of Kunduz city, near the airport, the Afghan Interior Ministry said in a statement. An Associated Press reporter at the scene saw at least one overturned armored vehicle and what appeared to be a civilian car that was mangled in the blast.

Three civilians were killed and 11 were wounded in the attack, the ministry said.

Germany's military said two German soldiers were lightly wounded and treated at a nearby base. Two vehicles were damaged, according to a German military spokesman, who declined to be named in line with department policy. The spokesman said the military could not immediately confirm whether the attack was a suicide assault or a roadside bomb.

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?
  - Lidé často zkopírují jen části textu

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?
  - Lidé často zkopírují jen části textu
  - Nebo dokument zkopírují a změní

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?
  - Lidé často zkopírují jen části textu
  - Nebo dokument zkopírují a změní
  - Nebo zkopírují a rozšíří vlastním textem

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?
  - Lidé často zkopírují jen části textu
  - Nebo dokument zkopírují a změní
  - Nebo zkopírují a rozšíří vlastním textem
  - Problém je třeba řešit na nižší úrovni, než celý dokument

Deduplikace

Odstranění opakujících se úseků textu

- Zcela identické texty – snadné?
  - Text převedeme na číslo pomocí *hašovací funkce*
  - Srovnáváme haš nového textu se všemi předchozími
- A co podobné texty – obtížné?
  - Lidé často zkopírují jen části textu
  - Nebo dokument zkopírují a změní
  - Nebo zkopírují a rozšíří vlastním textem
  - Problém je třeba řešit na nižší úrovni, než celý dokument
  - Nápady?

Dokument jako vektor četností slov

- Hlavní myšlenka: počet výskytů slova je důležitý

Dokument jako vektor četností slov

- Hlavní myšlenka: počet výskytů slova je důležitý
- Spočítáme *relativní četnost* slov

Dokument jako vektor četností slov

- Hlavní myšlenka: počet výskytů slova je důležitý
- Spočítáme *relativní četnost* slov
- Dokument reprezentujeme *vektorem četností slov*

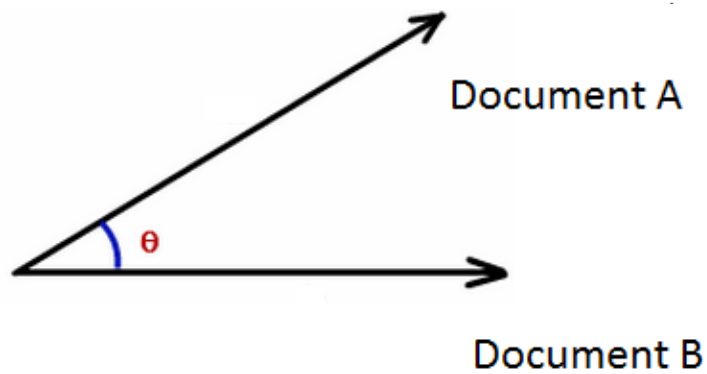
Dokument jako vektor četností slov

- Hlavní myšlenka: počet výskytů slova je důležitý
- Spočítáme *relativní četnost* slov
- Dokument reprezentujeme *vektorem četností slov*
- Jak zjistíme podobnost vektorů v n-rozměrném prostoru?

Dokument jako vektor četností slov

- Hlavní myšlenka: počet výskytů slova je důležitý
- Spočítáme *relativní četnost* slov
- Dokument reprezentujeme *vektorem četností slov*
- Jak zjistíme podobnost vektorů v n-rozměrném prostoru?
- Mírou podobnosti dvou vektorů může být *kosinová podobnost*

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$



Dokument jako překrývající se n-tice slov

- Hlavní myšlenka: okolí slova je důležité

Dokument jako překrývající se n-tice slov

- Hlavní myšlenka: okolí slova je důležité
- Rozdělíme dokument na překrývající se *n-tice slov*
  Hospodin je můj, je můj pastýř, můj pastýř nebudu, pastýř nebudu mít, nebudu mít nedostatek

Dokument jako překrývající se n-tice slov

- Hlavní myšlenka: okolí slova je důležité
- Rozdělíme dokument na překrývající se *n-tice slov*
  Hospodin je můj, je můj pastýř, můj pastýř nebudu, pastýř nebudu mít, nebudu mít nedostatek
- Dokument reprezentujeme množinou hašů těchto n-tic

Dokument jako překrývající se n-tice slov

- Hlavní myšlenka: okolí slova je důležité
- Rozdělíme dokument na překrývající se *n-tice slov*
  Hospodin je můj, je můj pastýř, můj pastýř nebudu, pastýř nebudu mít, nebudu mít nedostatek
- Dokument reprezentujeme množinou hašů těchto n-tic
- Jak zjistíme podobnost množin?

Dokument jako překrývající se n-tice slov

- Hlavní myšlenka: okolí slova je důležité
- Rozdělíme dokument na překrývající se *n-tice slov*
  Hospodin je můj, je můj pastýř, můj pastýř nebudu, pastýř nebudu mít, nebudu mít nedostatek
- Dokument reprezentujeme množinou hašů těchto n-tic
- Jak zjistíme podobnost množin?
- Mírou podobnosti dvou množin může být *Jaccardova podobnost*

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

Dokument jako překrývající se n-tice slov

V CZPJ používáme následující parametry deduplikace pomocí překrývajících se n-tic slov:

- Úroveň jemnosti: odstavce,
- velikost úseků: sedmice slov
- hranice úseků: začátek a konec věty,
- práh podobnosti sedmic: 50 %
- vyhlazování.

Dokument jako překrývající se n-tice slov

V CZPJ používáme následující parametry deduplikace pomocí překrývajících se n-tic slov:

- Úroveň jemnosti: odstavce,
- velikost úseků: sedmice slov
- hranice úseků: začátek a konec věty,
- práh podobnosti sedmic: 50 %
- vyhlazování.

Odstavec je ponechán,

- pokud alespoň 50 % sedmic slov v daném odstavci nebylo zaznamenáno dříve,
- nebo leží bezprostředně mezi dvěma ponechanými odstavci.

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov
  - Stačí změnit pořadí slov ve větách

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov
  - Stačí změnit pořadí slov ve větách
- Obelstěte obě metody deduplikace zároveň

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov
  - Stačí změnit pořadí slov ve větách
- Obelstěte obě metody deduplikace zároveň
  - Stačí zaměnit některá slova za synonyma,
  - případně celý text převyprávět

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov
  - Stačí změnit pořadí slov ve větách
- Obelstěte obě metody deduplikace zároveň
  - Stačí zaměnit některá slova za synonyma,
  - případně celý text převyprávět
- Ukažte, že existuje text s unikátními větami, který s naším nastavením metody překrývajících se n-tic slov zahodíme

Srovnání a slabá místa metod deduplikace

Zkonstruujte text, kterým uvedené metody obelstíte

- Metoda založená na relativní četnosti slov
  - Stačí zkopírovat dva různé dokumenty za sebe
- Metoda založená na překrývajících se n-ticích slov
  - Stačí změnit pořadí slov ve větách
- Obelstěte obě metody deduplikace zároveň
  - Stačí zaměnit některá slova za synonyma,
  - případně celý text převyprávět
- Ukažte, že existuje text s unikátními větami, který s naším nastavením metody překrývajících se n-tic slov zahodíme
  - Tvoří-li unikátní věty méně než 50 % sedmic slov, celý odstavec zahodíme

Využití k odhalování plagiátů – podezřelá diplomová práce

LIDOVKY.cz

DOMOV | SVĚT | BYZNYS | SPORT | KULTURA | CESTOVÁNÍ | RELAX | DESIGN | DOBRÁ CHUŤ | NÁZORY | LIDÉ | VIDEO | SPECIÁLY

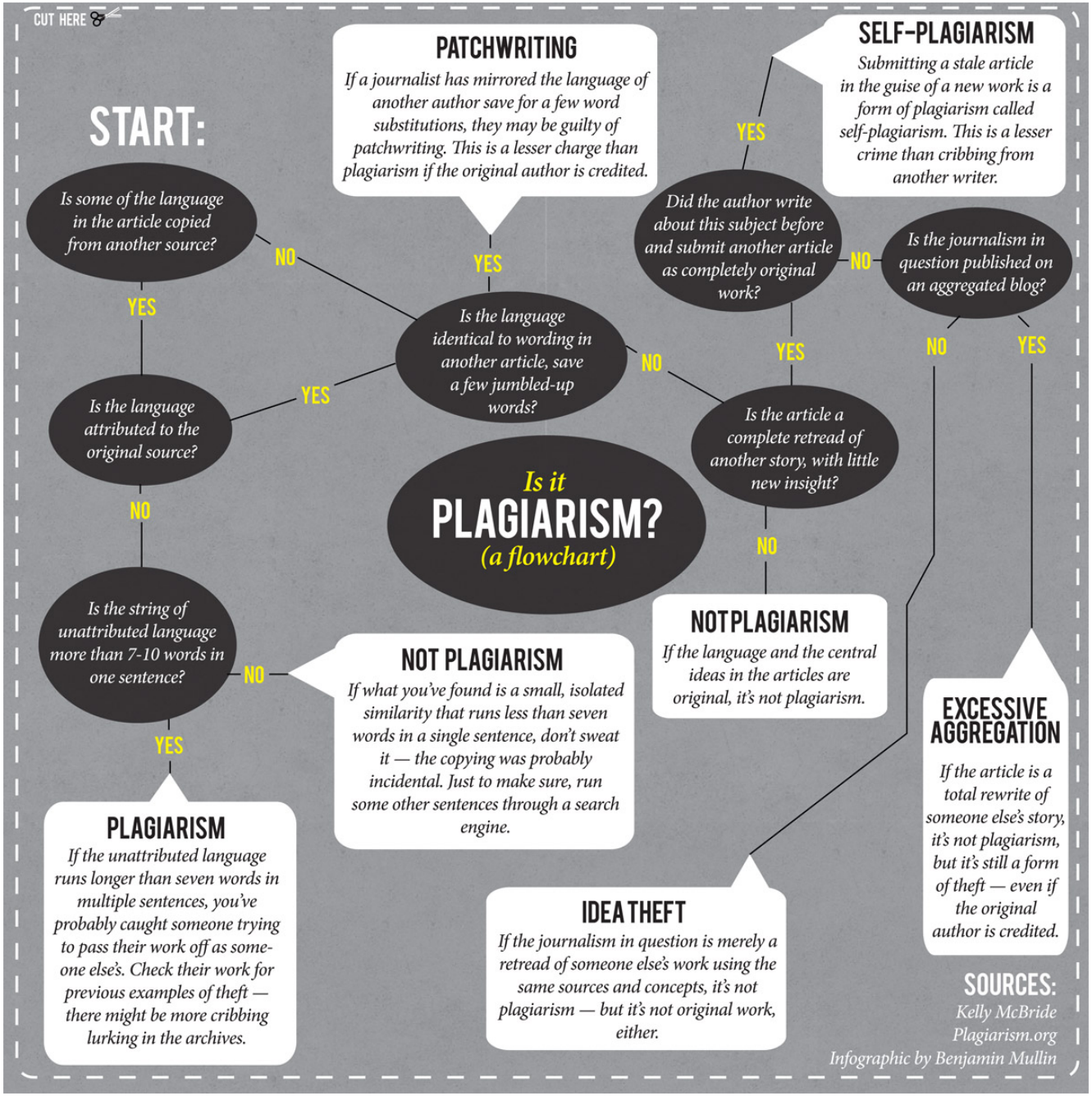
OSUDY 20. STOLETÍ | VOLEBNÍ SPECIÁLY | DOKUMENTY | ARCHIV ZPRÁV | SAMIZDATOVÉ LN | 25 LET SVOBODY

4. ŘÍJNA 2016 8:52 | LIDOVKY.CZ > ZPRÁVY > DOMOV

Piráti: Chovanec opsal úvod bakalářské práce. Snaha o předvolební skandálek, brání se ministr

PRAHA Ministr vnitra Milan Chovanec (ČSSD) okopíroval značnou část své bakalářské práce, tvrdí Piráti a dokazují to pasážemi bez řádného citování. Práce pochází z roku 2009, kdy Chovanec studoval na Právnické fakultě Západočeské univerzity v Plzni. Ministr se brání, že se Piráti snaží jen o „předvolební skandálek“ a za svým studiem si stojí.

Problém plagiátorství je však složitější



Issues of Building Language Resources from the Web

Particular tasks:

- Language identification,
- Character encoding detection,
- Efficient web crawling,
- Boilerplate removal,
- De-duplication (removal of identical or nearly identical texts),
- Fighting web spam,
- Authorship recognition & plagiarism detection,
- Storing & indexing of large text collections.

NLPC & Lexical Computing corpus tools:
`http://corpus.tools/`

Postup získávání webových korpusů v CZPJ

- příprava jayzkově závislých modelů používaných v dalších krocích — učení na dokumentech z Wikipedie
- spuštění crawleru (`SpiderLing`)
- zpracování a vyhodnocování během běhu crawleru
  - detekce znakové sady dokumentu (`Chared`)
  - filtrování jazyka (vektor trigramů znaků)
  - odstraňování nežádoucího obsahu (`Justext`)
  - kontrola duplicitních dokumentů
  - vyhodnocování průběžné výtěžnosti webových domén
- zpracování získaných dat
  - odstranění podobných odstavců (`Onion`)
  - tokenizace (`Unitok` nebo jiný nástroj)
  - značkování morfologické a syntaktické — externími nástroji, jsou-li dostupné
  - zakódování a nahrání do korpusového manažeru (`Manatee`/`Sketch Engine`)