05 – Building Language Resources from the Web

IA161 Natural Language Processing in Practice

Vít Suchomel

NLP Centre, FI MU, Brno

October 17, 2023

Outline

Lots of Text Data Can Be Downloaded From the Web. . .

Source: blesk.cz, 2021-10-26

. . . And Every Web Page. . .

BLESK.CZ

≡ Menu | Celebrity | Zprávy | Politika | Volby | Krimi | Sport | Fotbal | PREMIUM

Pokus o vraždu matky a syna v Dolních Chabrech: Stíhání odloženo! Útočník není trestně odpovědný

26. října 2021 • 09:38

Sdílej:

Scéna jako z hororu se odehrála koncem dubna v Dolních Chabrech. Došlo tam k pobodání 14letého chlapce a jeho maminky, útočit měl podle informací Blesku synův stejně starý kamarád. Policisté věc 21. října odložili - trestní stíhání totiž nebylo možné zahájit kvůli nízkému věku pachatele, potvrdil Blesku mluvčí městského státního zastupitelství Aleš Cimbala.

Vyhledat kraj, okres, město, ... VYHLEDAT

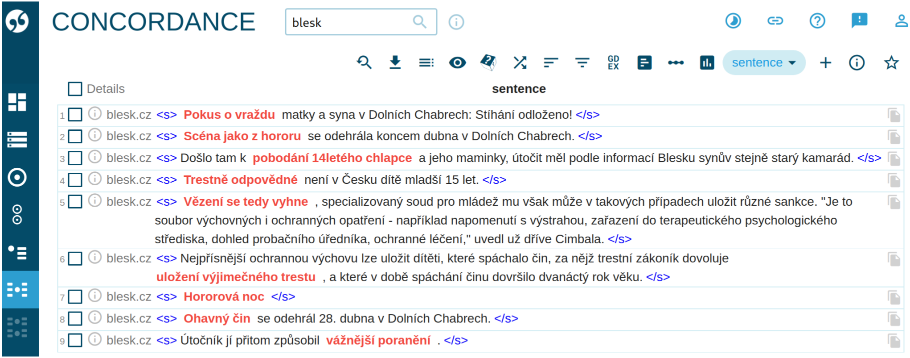
Hvězda StaraDance Mirai Navrátil: Dohra tokání s Šoposkou na večírku

JDI NA HLAVNÍ STRÁNKU BLESK.CZ

- Stav Zemana má posoudit nové lékařské konzilium. ...
- Britská královna Alžběta II. ruší program: Na ...
- Šok pro záchranáře: Při překotném porodu byl ...
- Kostková má covid: ČT řekla, co bude se StarDance!

. . . Requires Various Kinds of Cleaning. . .

. . . To Get Natural Fluent Sentences for a Text Corpus



Text Corpus

A corpus is a set of texts in a natural language.

Text Corpus

A corpus is a set of texts in a natural language.

Statistical NLP:

a large amount of language use data

situated within its textual context

Corpus Use

- generally: data for studying natural language
- linguists: analyses of language phenomena, language changes over time
- lexicographers, teachers: dictionaries, word meanings, examples of a typical use
- sociologists: style and theme, hot topics
- marketing experts: brands/product evaluation, sentimet analysis
- statistical NLP: language models for taggers, analysers, translation systems, predictive writing,. . .

Text Sources

- printed media: books, newspapers, magazines, poetry collections
- internet: articles, presentations, blogs, discussions, socnet messages (tweets, fb)
- speech: transcription of speech recordings, movie subtitles
- other: personal correspondence, school essays

Corpus Size Matters ...

Most language phenomena follow the Zipfian distribution.

$\Rightarrow$ The more data the better.

Corpus Size Matters ...

Most language phenomena follow the Zipfian distribution.
⇒ The more data the better.

Example: Modifiers of phrase deliver speech (frequency):

- 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).
- enTenTen15 (15,7 G words): keynote (3673), opening (684), welcome (413), key (257), major (255), acceptance (233), powerful (229), commencement (226), inspiring (210), inaugural (146).
- enClueWeb09 (70,5 G words): keynote (3802), acceptance (1035), opening (589), famous (555), commencement (356), impassioned (335), inaugural (333).

...But the Size Is Not Everything

A significant fraction of all web pages are of poor utility. [1]

[1] [Manning et al., 2008, Chapter 20]

. . . But the Size Is Not Everything

A significant fraction of all web pages are of poor utility. [1]

Why are qualitative aspects so important?

- web is the most used data source to obtain enough source texts – Web as Corpus
- web is garbage (by definition) – garbage as corpus?
- building language resources from the web requires extensive post-processing

[1] [Manning et al., 2008, Chapter 20]

Selected Issues of Building Web Corpora

- language identification
- character encoding detection
- efficient web crawling
- boilerplate removal
- de-duplication (removal of identical or nearly identical texts)
- fighting web spam
- text classification (topic, genre, language variety)
- authorship recognition & plagiarism detection
- storing & indexing of large text collections

Brno Processing Pipeline

1. web crawler SpiderLing – Suchomel, Pomikálek (2012)
2. character encoding detection (byte trigram model) – Pomikálek, Suchomel (2012)
3. language identification (character trigram model)
4. boilerplate removal – Pomikálek (2011)
5. text tokenisation – Michelfeit, Suchomel (2014)
6. near-duplicate paragraph removal – Pomikálek (2011)
7. discerning (similar) languages – Suchomel (2019)
8. all data is stored and indexed by corpus manager Sketch Engine – Kilgarriff, Rychlý, Smrž, Tugwell (2004)

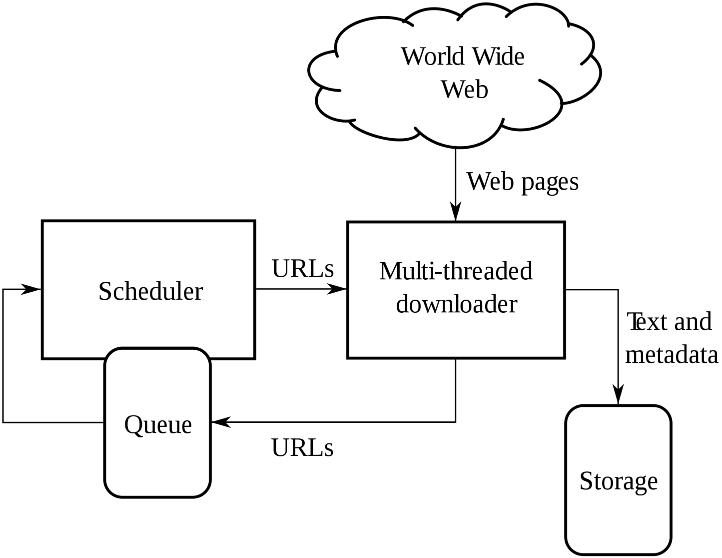
NLPC & Lexical Computing corpus tools: `http://corpus.tools/`

Outline

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 links checking,
  - Heritrix – general crawler, (Java, multiple treads),
  - SpiderLing – text corpora, (Python, multiple sockets).

Basic crawler design

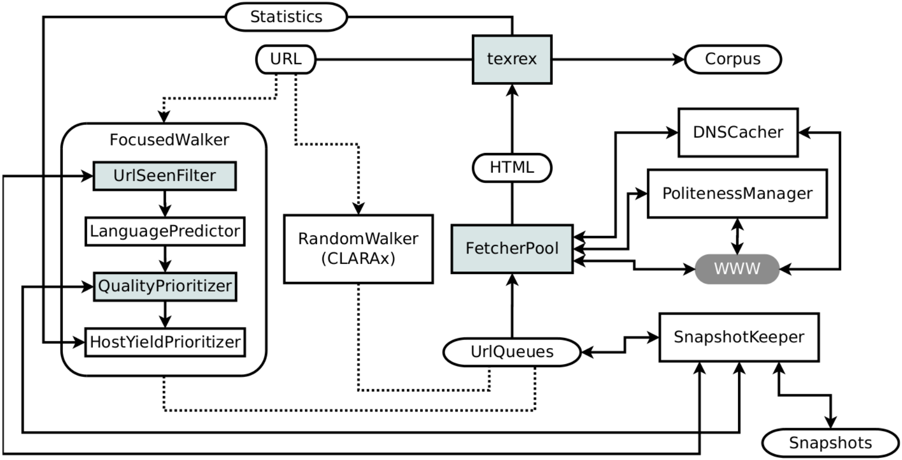


Source: http://en.wikipedia.org/wiki/Web_crawler

Advanced crawler implementation details

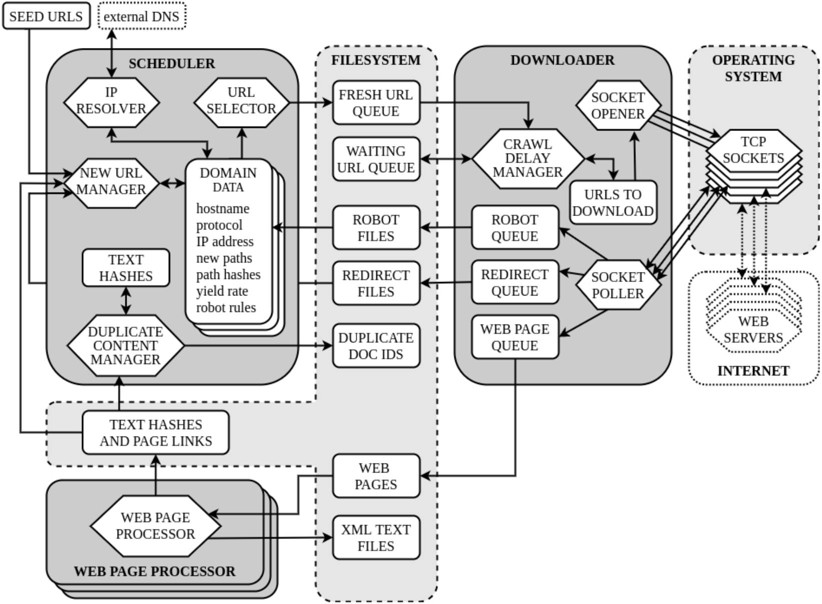
- Distributed vs. extensible.
- Multi-threaded vs. multi-socketed.
- Web traversal policy:
  - depth vs. breadth,
  - domain selection,
  - domain distance,
  - focused crawling (topic oriented) vs. general crawling,
  - yield ratio.

Focused crawler design



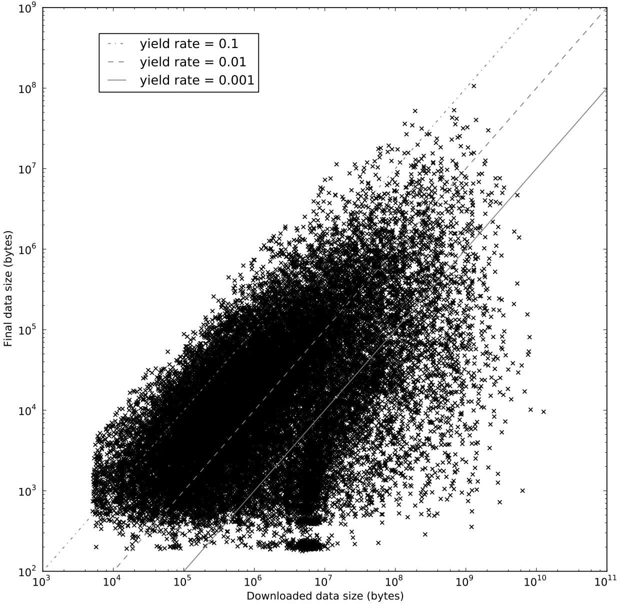
Source: Roland Schafer, Adrien Barbaresi, Felix Bildhauer. Focused Web Corpus Crawling. 9th Web as Corpus Workshop, 2014.

Asynchronous focused crawler design

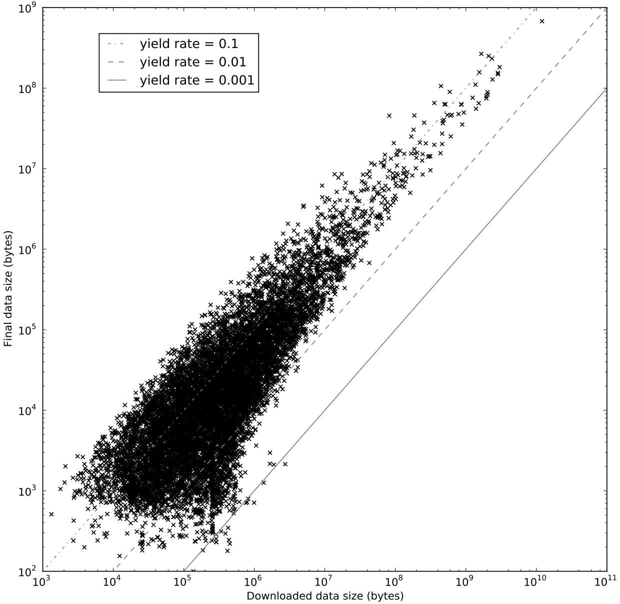


Source: Suchomel: Better Web Corpora For Corpus Linguistics And NLP, doctoral thesis, Masaryk university, Brno, 2020

General unfocused crawling efficiency (Heritrix)



Domain yield ratio optimised efficiency (SpiderLing)



Outline

Issues of Language Identification of Text from the Web

- multiple languages in a single web page, e.g. Maori/English
- similar languages, e.g. Danish vs. Norwegian
- language varieties, e.g. European vs. Brasilian Portuguese

Solution

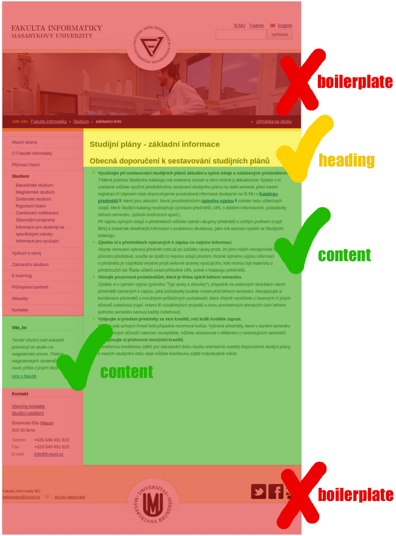
- langid.py [Lui and Baldwin, 2012], 97 languages
  - naive Bayes classifier over byte n-grams ($1 \leq n \leq 4$)
- Google Compact Language Detector v. 2 (CLD2) – 2013, $\sim 80 \rightarrow \sim 250$ languages
  - character 4&8-grams
  - easy to use and fast
- Google Compact Language Detector v. 3 (CLD3) – 2021
  - character 1-3-grams, neural network model
- FastSpell – 2021
  - FastText (character 3&6-grams and word embeddings)
  - Hunspell (spelling differences of similar languages)
- HeLI-OTS [Jauhiainen et al., 2022]
  - log word freq and character 1-6-grams
  - high recall for less common languages

Outline

What is boilerplate

- Repeated parts of a web page (not containing a new text) – header, footer, navigation.
- Uninteresting text (too short or not continuous) – advertisement, lists of items, article previews.
- Discussions should be separated from the main article text.

What is boilerplate



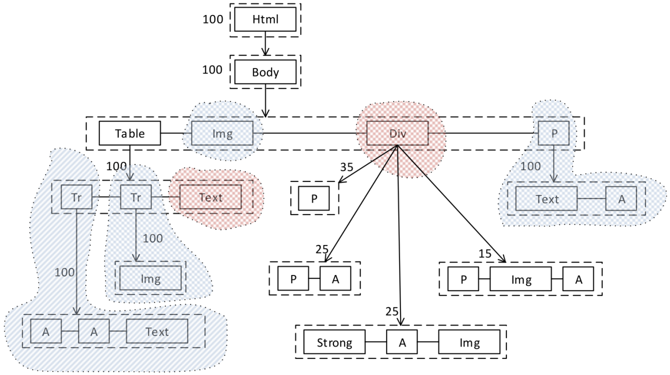
Source: http://corpus.tools/attachment/wiki/Justext/Algorithm/cs_classification_example.png

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 et al., 2003]

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



Source: 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/`

Outline

What Is Wrong with this Text?

Now on the web stores are very aggressive price smart so there genuinely isn't any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).

What Is Wrong with this Text?

Now on the web stores are very aggressive price smart so there genuinely isn't any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).

Web spam, computer generated text – Not a good evidence of natural language phenomena

Web Spam Definition – Text Corpus Point of View

Good content: fluent, natural, consistent text (regardless its purpose)
Bad content – computer generated text

- machine translation
- keyword stuffing
- phrase stitching
- synonym replacement
- automated summaries
- any incoherent text

Varieties of spam removable by existing tools dealt with by other means

- duplicate content
- link farms
- redirection

Approaches to Web Spam Removal

1. trustworthy websites only
2. website rules in the crawler: distance from the seeds, hostname
3. supervised classification
4. semi-manual filtering of websites

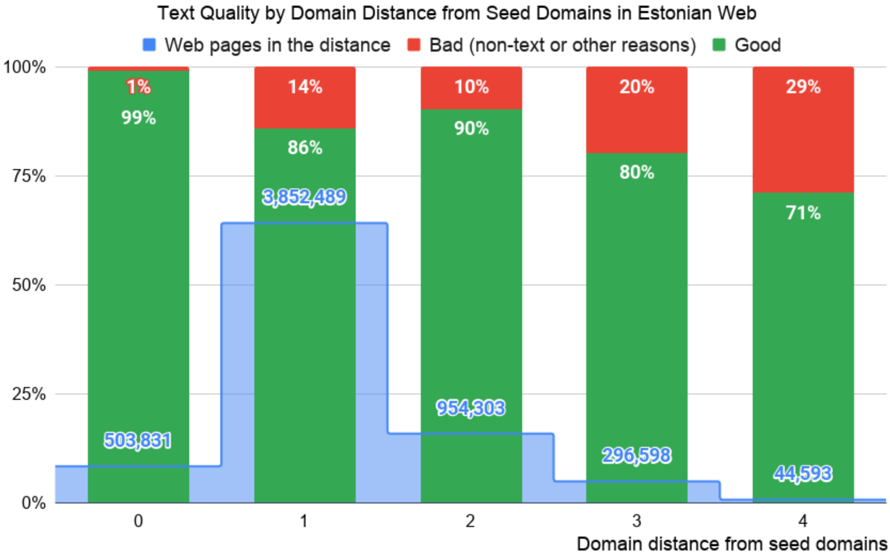
Suchomel: Better Web Corpora For Corpus Linguistics And NLP, doctoral thesis, Masaryk university, Brno, 2020

[Suchomel, 2020]

Trustworthy Websites Only

- works well but not perfect
- limited amount/size of trustworthy sources $\Rightarrow$ unsuitable for small languages

Website Distance from the Seed (Trustworthy) Websites



Supervised Classification – Data & Method

- 146 spam pages of 1630 manually classified web pages
- various web sources, 2006 to 2015
  - phrase and sentence level incoherency
  - frequent spam topics: medication, financial services, essay writing
  - other non-text, various techniques

- FastText supervised classifier (Mikolov, 2016)
- applied to a large English web corpus from 2015
- 35 % most ‘spam-like’ documents removed
- recall: 70.5 %
- precision: 71.5 %

Note: As of 2023, Large pretrained LMs fine-tuned to non-text removal would perform better!

Supervised Classification – Evaluation – Wordlist

	Original corpus	Clean corpus	Kept
Document count	58,438,034	37,810,139	64.7 %
Token count	33,144,241,513	18,371,812,861	55.4 %

Phrase	**Original hits/M**	**Clean hits/M**	**Kept**
viagra	229.71	3.42	0.8 %
cialis 20 mg	2.74	0.02	0.4 %
aspirin	5.63	1.52	14.8 %
oral administration	0.26	0.23	48.8 %
loan	166.32	48.34	16.1 %
payday loan	24.19	1.09	2.5 %
cheap	295.31	64.30	12.1 %
interest rate	14.73	9.80	36.7 %
essay	348.89	33.95	5.4 %
essay writing	7.72	0.32	2.3 %
pass the exam	0.34	0.36	59.4 %
slot machine	3.50	0.99	15.8 %
playing cards	1.01	0.67	36.8 %
play games	3.55	3.68	53.9 %

Supervised Classification – Evaluation – Collocates/Lexicography

Top collocate objects of verb 'buy' before and after spam removal

Original corpus				**Cleaned corpus**		
lemma	**frequency**	**score**		**lemma**	**frequency**	**score**
viagra	569,944	10.68		ticket	52,529	9.80
ciali	242,476	9.56		house	28,313	8.59
essay	212,077	9.17		product	37,126	8.49
paper	180,180	8.93		food	24,940	8.22
levitra	98,830	8.33		car	20,053	8.18
uk	93,491	8.22		book	27,088	8.09
ticket	85,994	8.08		property	17,210	7.88
product	105,263	8.00		land	15,857	7.83
cialis	71,359	7.85		share	12,083	7.67
car	75,496	7.75		home	22,599	7.63
house	70,204	7.61		item	12,647	7.40
propecia	55,883	7.53		good	9,480	7.37

Semi-manual Website Filtering

Data:

- 1,000 Estonian 2019 web sites, manually checked by Kristina Koppel (Tartu University)
- 16 % marked as computer generated non-text, mostly machine translated, 6 % marked as poor quality

Method:

- FastText supervised classifier
- probability threshold set to aim for a high recall

Evaluation:

- 100 positive & 100 negative random pages for manual evaluation
- recall: 97.1 %, precision: 66.7 %
- quite efficient method – just several man-days of manual work

Outline

De-duplication

- Quite straightforward for full duplicates.
- What about similar documents?
- People copy just parts of the document: original vs. copy
- Or copy and modify: original vs. modified
- Or copy and extend: original vs. extended

N-gram shingling algorithm [Manning et al., 2008, Chapter 19]

- Shingles of length of n words.
- N-grams represented by hashes.

Image source: Wikipedia user ANKAWÜ, cropped

onion – One INstance ONly[2]

Algorithm inspired by Broder's shingling algorithm:

- Make n-grams of words for every structure,
- every n-gram is represented by its hash,
- the current structure is a duplicate $\Leftrightarrow$ at least p % of n-gram hashes is duplicate (has been observed before).
- Default options: structure = paragraph, $n = 7$, $p = 50$, smoothing.

[2] Pomikálek, Jan. Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk university, 2011.

Outline

Main and related tasks in plagiarism detection

- Plagiarism detection: Given a document, identify all plagiarized sources and boundaries of re-used passages.
- Author identification: Given a document, identify its author.
- Author profiling: Given a document, extract information about the author (e.g. gender, age).

Stamatatos et al. Overview of the pan/clef 2015 evaluation lab. In Experimental IR Meets Multilinguality, Multimodality, and Interaction, pages 518–538. Springer. 2015.

External vs. Intrinsic plagiarism detection

According to PAN 2021 [Bevendorff et al., 2021]

External plagiarism detection

Given a set of suspicious documents and a set of source documents the task is to find all text passages in the suspicious documents which have been plagiarized and the corresponding text passages in the source documents.

Intrinsic plagiarism detection

Given a set of suspicious documents the task is to identify all plagiarized text passages, e.g., by detecting writing style breaches. The comparison of a suspicious document with other documents is not allowed in this task.

Plagiarism techniques [Potthast et al., 2015]

- Manual paraphrasing = human retelling. Similar to copywriting.
- Random text operations. Random shuffling, insertion, replacement, or removal of characters, phrases or sentences. Replacement of characters with look-alike UTF characters.
- Semantic word variation. Random replacement words with synonyms, antonyms, hyponyms, or hypernyms.
- Part-of-speech-preserving word shuffling. Shuffling of phrases while maintaining the original POS sequence.
- Improvement of previous synthetic techniques: Insertions, replacements and variations may be obtained from context documents.
- Machine translation, cyclic translation. Automatic translation of a text passage from one language via a sequence of other languages to the original language.
- Summarization. Summaries of long text passages.
- Improvement of machine translation and summarization techniques: Manually corrected output.

Basic techniques for revealing similar documents[3]

Bag of words

Full fingerprint methods

Overlapping substrings of length k in words from the beginning of the document.

Selective Fingerprint methods

Non-overlapping substrings of length k in words from the beginning of the document.

Rarest-in-document

All substrings are sorted according to their document frequency, then the rarest are selected as representatives of the document.

Selected Anchors

The document is reduced to pre-selected short chunks of characters.

Symmetric Similarity measure

$SS(X, Y) = \frac{|d(X) \cap d(Y)|}{|d(X) \cup d(Y)|}$ where $d(X)$ is a set of fingerprints of X.

[3] According to HaCohen-Kerner at al. Detection of simple plagiarism in computer science papers. In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 421-429. Association for Computational Linguistics, 2010.

Outline

Task: Plagiarists vs. plagiarism detectors

Either:

Create 5 documents (with a similar topic) and 5 plagiarisms of these documents, 10 documents total.[4]

- 100 words $\leq$ document lenght $\leq$ 500 words
- 20 % $\leq$ plagiarism content $\leq$ 90 %
- POS tagged text:
  - Czech: `asteria04:/opt/majka_pipe/majka-czech_v2.sh | cut -f1-3`.
  - English: `asteria04:/opt/treetagger_pipe/tt-english_v2.1.sh`.
- For each plagiarism:
  1. describe plagiarism technique(s) used
  2. which detection methods might be able to reveal it – give reasons
  3. which detection methods might not be able to reveal it – give reasons

The minimal homework.

[4] For the sake of simplicity: A plagiarism cannot have more sources here.

Task: Plagiarists vs. plagiarism detectors

Or:

Select a detection algorithm and implement it in Python.

- Input format: A POS tagged vertical consisting of structures `doc` with attributes `author`, `id`, `class`, `source`. Pair author, id is unique. Class is "original" or "plagiarism". Source is the id of the source (in case of plagiarism) or own id (in case of original).[5]
- Output format: One plagiarism per line: id `TAB` detected source id `TAB` real source id. Evaluation line: precision, recall F1 measure.
- `./plagiarism_simple.py < training_data.vert`
- Your script will be evaluated using data made by others.
- Describe which plagiarism detection technique(s) were implemented.

The right homework if you want to learn something.

[5] For the sake of simplicity: A plagiarism cannot have more sources here.

Task: Input data example

```
<doc author="Já První" id="1" class="original" source="1">
<s>
Dnes	dnes	k6eAd1
je	být	k5eAaImIp3nS
pěkný	pěkný	k2eAgInSc4d1	pěkný
den	den	k1gInSc4	den
<g/>
!	!	k?
</s>
</doc>
<doc author="Já První" id="2" class="plagiarism" source="1">
<s>
Dnes	dnes	k6eAd1
je	být	k5eAaImIp3nS
ale	ale	k9
pěkný	pěkný	k2eAgInSc4d1	pěkný
den	den	k1gInSc4	den
<g/>
!	!	k?
</s>
</doc>
```

Task: Output example

```
Doc set by John Smith
506     501     501
507     501     501
508     502     502
509     502     503
510     502     504
Set precision: 0.60, recall: 1.00, F1: 0.75

...

Overall precision: 0.82, recall: 1.00, F1: 0.90
```

References I

Bevendorff, J. et al. (2021).
Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection.
In D. Hiemstra, MF. Moens, J. M. R. P. M. P. F. S., editor, *Advances in Information Retrieval (ECIR 2021)*. Springer.

Jauhiainen, T., Jauhiainen, H., and Lindén, K. (2022).
HeLI-OTS, off-the-shelf language identifier for text.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Lui, M. and Baldwin, T. (2012).
langid. py: An off-the-shelf language identification tool.
In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

References II

- Manning, C. D., Raghavan, P., Schütze, H., et al. (2008).
  Introduction to information retrieval, volume 1.
  Cambridge university press.
- Potthast, M., Hagen, M., Göring, S., Rosso, P., and Stein, B. (2015).
  Towards data submissions for shared tasks: first experiences for the task of text alignment.
  Working Notes Papers of the CLEF, pages 1613–0073.
- Suchomel, V. (2020).
  Better Web Corpora for Corpus Linguistics and NLP.
  Doctoral thesis, Masaryk university, Brno.
- Yi, L., Liu, B., and Li, X. (2003).
  Eliminating noisy information in web pages for data mining.
  In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 296–305. ACM.