Building Large Text Corpora from the Web

Vít Suchomel

Lexical Computing Ltd.
vit.suchomel@sketchengine.co.uk

SKEW 3
March 21, 2012

Introduction

- A need for large text corpora
  - lexicographers, linguists
  - better coverage of rare language phenomena
- The web as a corpus
  - general web crawlers (Heritrix)
  - SpiderLing – a new web crawler for text corpora

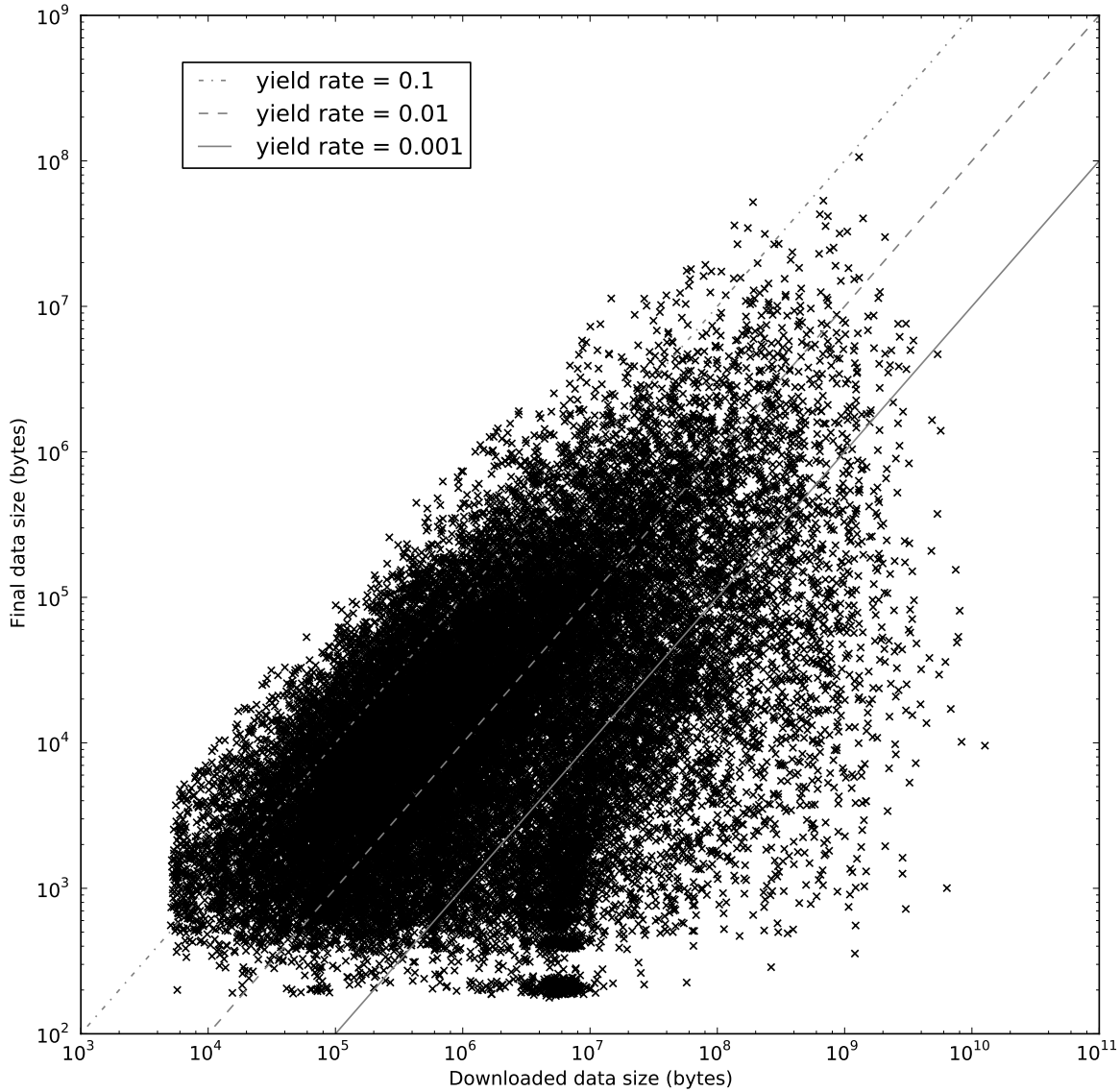
Selected corpora sizes in SketchEngine in mid 2011

language	millions of tokens
Arabic (Arabic web)	174
Chinese (zhTenTen)	2,107
Czech (czes)	465
English (enTenTen)	3,269
French (frWaC)	1,629
German (deTenTen)	2,845
Italian (itTenTen)	3,077
Japanese (JpWaC)	409
Portuguese (ptTenTen)	948
Russian (Russian web)	188
Spanish (esTenTen)	2,459
Turkish (TurkishWaC)	42

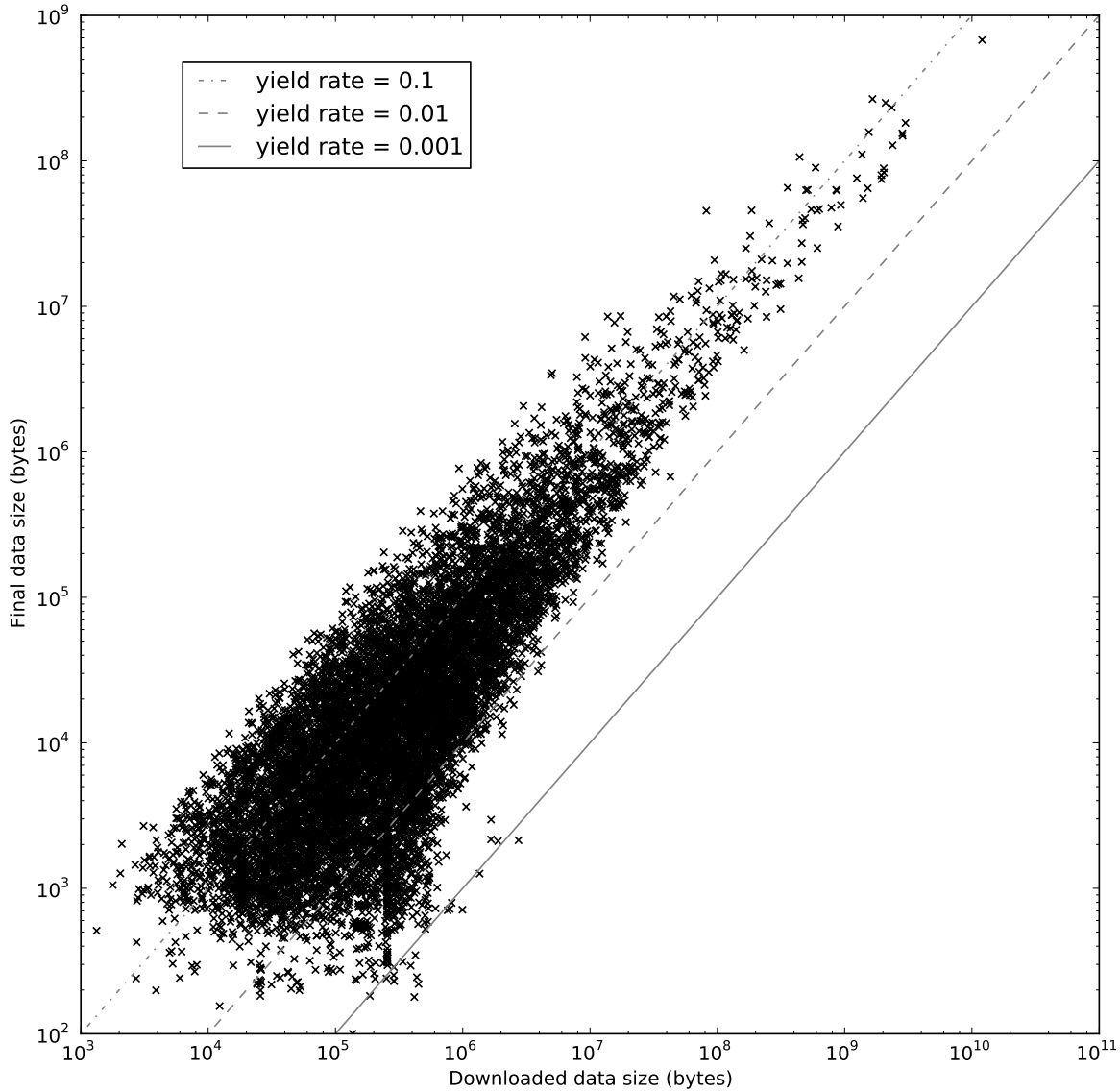
SpiderLing – a new web crawler for text corpora

- aimed to improve the crawling efficiency
- $\textit{yield rate} = \frac{\textit{final data}}{\textit{downloaded data}}$
- focus on the text–rich web domains
- **Target:** $\geq 10^9$ words for main languages

Yield rate by web domains (Heritrix, Portuguese web)



Yield rate by web domains (SpiderLing, Czech web)



Websites' yield rate threshold function in SpiderLing

The yield rate threshold for a domain is computed using the following function:

$$t(n) = 0.01 \cdot (\log_{10}(n) - 1)$$

where n is the number of documents downloaded from the domain.

# of documents	yr threshold
10	0.00
100	0.01
1000	0.02
10000	0.03

Corpus building pipeline

- preparation of language specific models based on Wikipedia articles: byte trigrams, character trigrams, wordlist
- web crawling using SpiderLing
- processing during crawling
  - character encoding detection (tool Chared, byte trigrams based)
  - language filtering (character trigrams based)
  - boilerplate removal (tool Justext)
  - duplicate documents removal
- postprocessing
  - similar paragraphs removal (tool Onion, 7-tuples of words, 50 % similarity threshold)
  - tokenization (tool Unitok)
  - part of speech tagging (3rd party taggers where available)
  - compilation in the Sketch Engine

Textual data downloaded by SpiderLing so far

language	raw data [GB]	corpus size [GB]	corpus size [10^9 tokens]	crawling time [days]
Am. Spanish	1874	44.14	8.7	14
Arabic	2015	58.04	6.6	28
Czech	ca. 4,000		5.8	ca. 40
French	3273	being	processed	15
Japanese	2806	61.36	11.1	28
Russian	4142	197.5	20.2	14
Turkish	2700	26.21	4.1	14

Future work

- analyzing the topics and genres of the downloaded texts (eventually ballancing the downloaded content in this respect)
- topic sensitive crawling