WebBootCaT

building instant corpora

Jan Pomikálek

NLPlab, FI MU, Brno

16 October 2006

Outline

- Text Corpora
- Corpus Managers
- Motivation
- Web as Corpus
- Keywords Extraction
- What the Future Holds
- Conclusion

Text Corpora

- large and structured sets of texts
- usually annotated (POS-tags, lemmas)
- monolingual vs. multilingual
- parallel corpora

Usage of Text Corpora

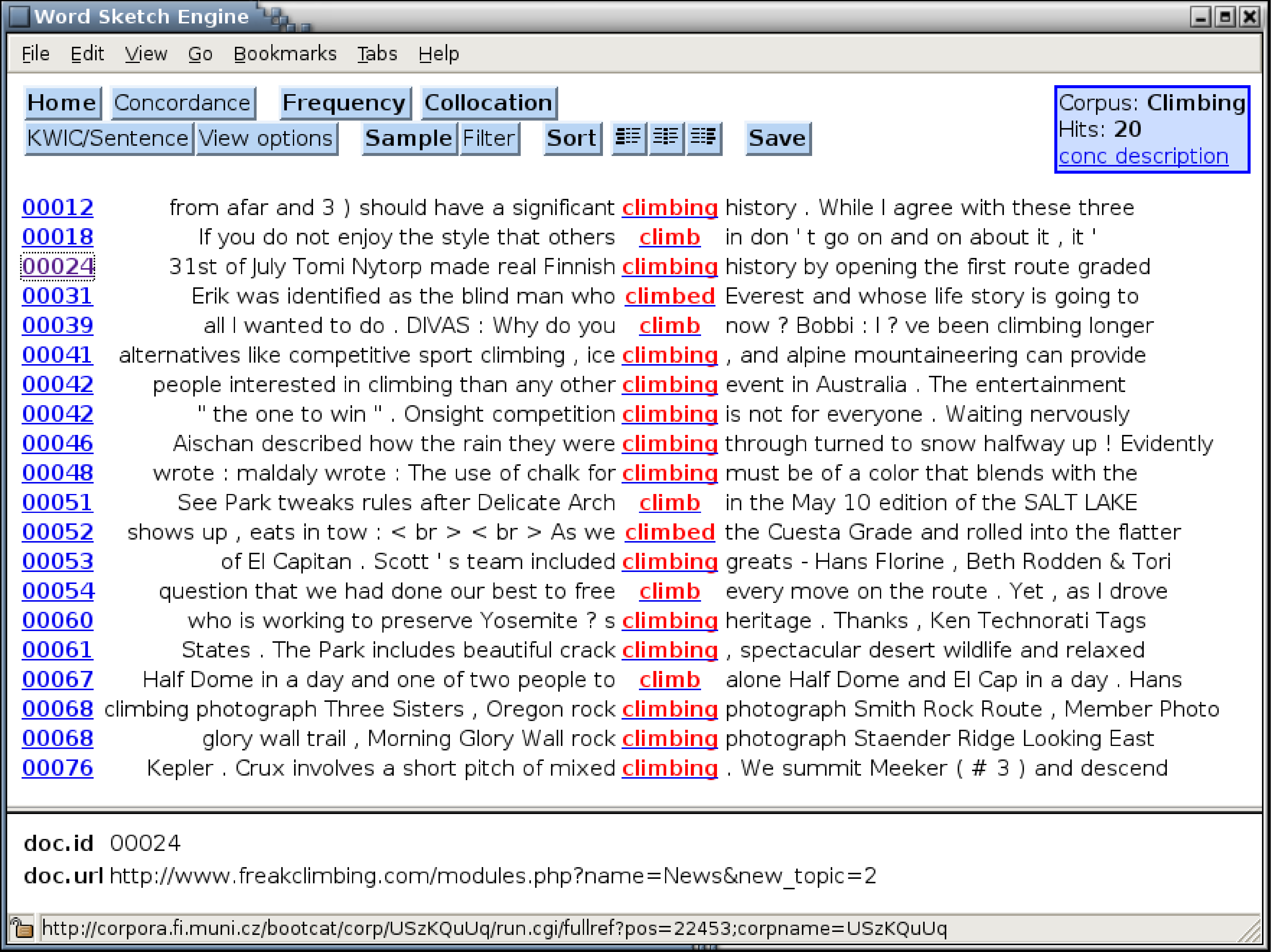
- getting statistically significant data about natural languages
- main areas
  - lexicography
  - speech recognition
  - machine translation
  - language learning/teaching

Size of Text Corpora

- DESAM – 1 million words
- BNC – 100 million words
- itWaC – 2 billion words

Corpus Manager

- software tool for working with corpora
- fast searching
- powerful query language
- statistics
  - words/bigrams/n-grams frequencies
  - collocations
- advanced features
  - word sketches
  - statistical thesaurus



Corpus: **Climbing**
Hits: **20**
conc description

```
00012        from afar and 3 ) should have a significant climbing history . While I agree with these three
00018          If you do not enjoy the style that others  climb   in don ' t go on and on about it , it '
00024        31st of July Tomi Nytorp made real Finnish climbing history by opening the first route graded
00031           Erik was identified as the blind man who climbed  Everest and whose life story is going to
00039            all I wanted to do . DIVAS : Why do you  climb   now ? Bobbi : I ? ve been climbing longer
00041  alternatives like competitive sport climbing , ice climbing , and alpine mountaineering can provide
00042     people interested in climbing than any other climbing event in Australia . The entertainment
00042          " the one to win " . Onsight competition climbing is not for everyone . Waiting nervously
00046        Aischan described how the rain they were climbing through turned to snow halfway up ! Evidently
00048     wrote : maldaly wrote : The use of chalk for climbing must be of a color that blends with the
00051        See Park tweaks rules after Delicate Arch  climb   in the May 10 edition of the SALT LAKE
00052   shows up , eats in tow : < br > < br > As we climbed  the Cuesta Grade and rolled into the flatter
00053             of El Capitan . Scott ' s team included climbing greats - Hans Florine , Beth Rodden & Tori
00054      question that we had done our best to free  climb   every move on the route . Yet , as I drove
00060       who is working to preserve Yosemite ? s climbing heritage . Thanks , Ken Technorati Tags
00061        States . The Park includes beautiful crack climbing , spectacular desert wildlife and relaxed
00067    Half Dome in a day and one of two people to  climb   alone Half Dome and El Cap in a day . Hans
00068 climbing photograph Three Sisters , Oregon rock climbing photograph Smith Rock Route , Member Photo
00068        glory wall trail , Morning Glory Wall rock climbing photograph Staender Ridge Looking East
00076      Kepler . Crux involves a short pitch of mixed climbing . We summit Meeker ( # 3 ) and descend
```

doc.id 00024
doc.url http://www.freakclimbing.com/modules.php?name=News&new_topic=2

http://corpora.fi.muni.cz/bootcat/corp/USzKQuUq/run.cgi/fullref?pos=22453;corpname=USzKQuUq

Motivation

- common corpora
  - expensive
  - limited electronic resources
  - printed resources have to be used
  - building is time consuming
  - copyright issues

- web corpora
  - cheap
  - almost unlimited resources
  - building is fast (can be automated)

Web as Corpus

- WWW is a very rich source of textual data (January 2005: 11.5 billion web pages)
- the data is available to everyone
- errors in texts – problem?

Why not just use Google?

- query language too weak
- texts not annotated
- no restrictions on domains
- for statistics we need local data

Using web as corpus

- pre-create
  - crawl web
  - download web pages
  - clean data
  - annotate
  - output = large ballanced web corpus (itWaC, deWac)

- advantages
  - huge corpora can be build
- disadvantages
  - time consuming
  - computer experts required

Using web as corpus (2)

- on-the-fly
  - input = query
  - search engine
  - download web pages
  - (annotate)
  - output = concordance lines

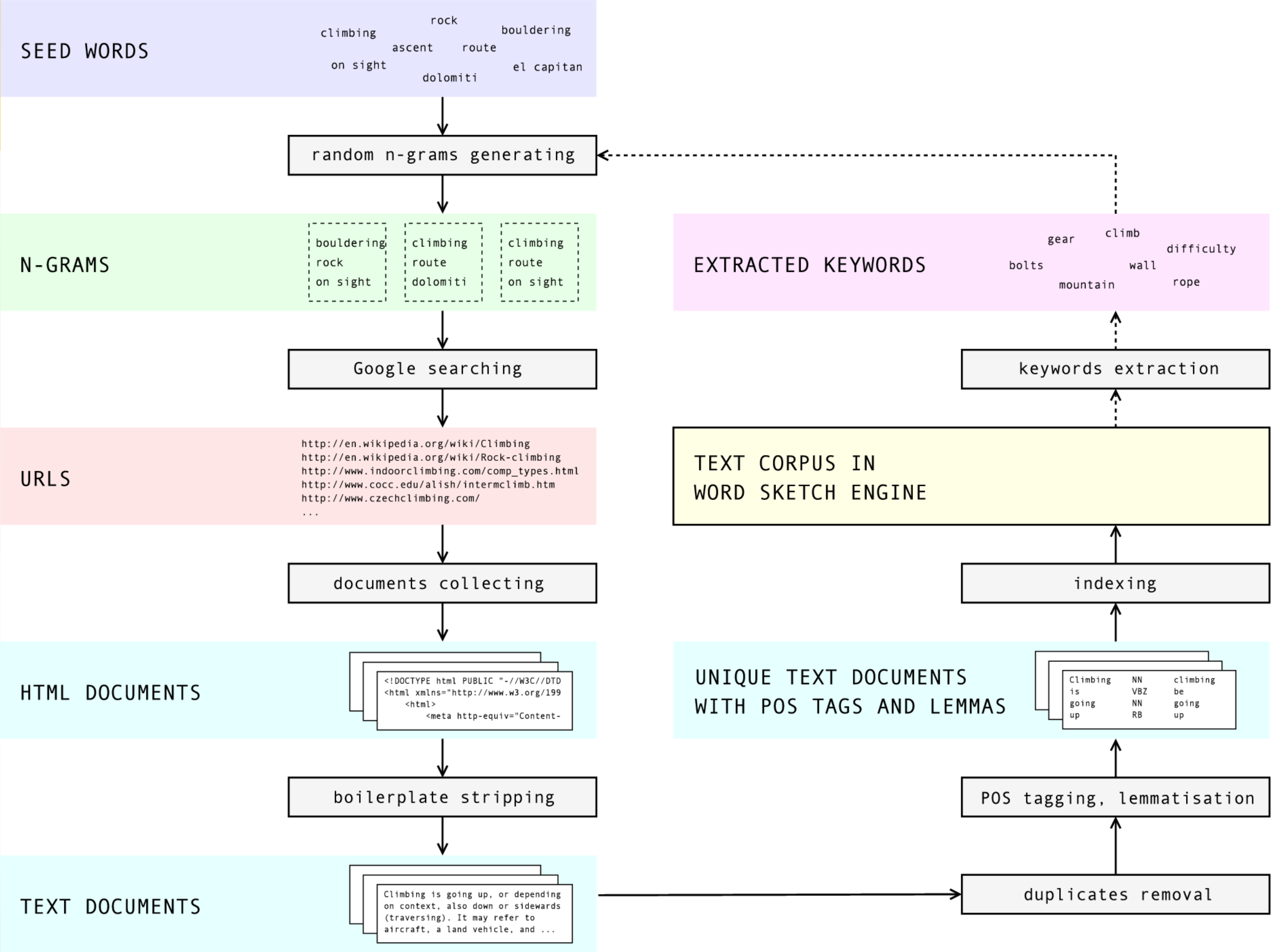
- disadvantages
  - limited query language
  - slow

WebBootCaT

- BootCaT = Simple Utilities to Bootstrap Corpora and Terms from the Web
  - Marco Baroni et al (University of Bologna)
- medium size domain specific corpora
  - ca 1 million words
- input = seed words + options
- output = annotated domain specific corpus loaded into Bonito

Domain specific corpora

- lexicography, speech recognition, machine translation
  - often also domain specific
- less data is sufficient than for general corpora


SEED WORDS
rock
climbing
bouldering
ascent
route
on sight
el capitan
dolomiti
random n-grams generating
N-GRAMS
bouldering
rock
on sight
climbing
route
dolomiti
climbing
route
on sight
Google searching
URLS
http://en.wikipedia.org/wiki/Climbing
http://en.wikipedia.org/wiki/Rock-climbing
http://www.indoorclimbing.com/comp_types.html
http://www.cocc.edu/alish/intermclimb.htm
http://www.czechclimbing.com/
...
documents collecting
HTML DOCUMENTS
<!DOCTYPE html PUBLIC "-//W3C//DTD
<html xmlns="http://www.w3.org/199
<html>
<meta http-equiv="Content-
boilerplate stripping
TEXT DOCUMENTS
Climbing is going up, or depending on context, also down or sidewards (traversing). It may refer to aircraft, a land vehicle, and ...
duplicates removal
POS tagging, lemmatisation
UNIQUE TEXT DOCUMENTS
WITH POS TAGS AND LEMMAS
Climbing NN climbing
is VBZ be
going NN going
up RB up
indexing
TEXT CORPUS IN
WORD SKETCH ENGINE
keywords extraction
EXTRACTED KEYWORDS
gear
climb
difficulty
bolts
wall
mountain
rope

WebBootCaT (2)

- n-grams generating
- Google search (Google API)
- download web pages
- boilerplate stripping
  - strip tag heavy parts
- duplicates removal
  - Text::DeDuper (CPAN)
  - n-gram based

WebBootCaT (3)

- POS-tagging, lemmatisation
  - TreeTagger
    - English, German, French, Italian, Spanish, Bulgarian
  - Czech tagging coming soon
- Indexing
  - manatee, Bonito (Sketch Engine)

Keywords extraction

- reference corpora
  - large web corpora (ca 500 million words)
- compare relative frequencies of words

word	WBC corpus	reference corpus
rope	$1.5 * 10^{-1}$ %	$8.3 * 10^{-4}$ %
wall	$1.1 * 10^{-1}$ %	$67.1 * 10^{-4}$ %
Yosemite	$1.2 * 10^{-1}$ %	$0.7 * 10^{-4}$ %
mountain	$0.7 * 10^{-1}$ %	$31.1 * 10^{-4}$ %

- multi-word expressions

KW extraction – problems

Kittyhawk: USS Kittyhawk calling. Request you alter course. Over and out.

Radio: Message received. Mission such we cannot alter cours. We request you alter course.

Kittyhawk: We are an aircraft carrier of the US Navy. We demand you alter course soonest to avoid collision.

Radio: We are unable to implement your request. We recommend you take avoiding action immediately.

Kittyhawk: If you continue to ignore our order we will open fire.

Radio: We are a lighthouse – your call!

Average reduced frequency

- look at the word distribution in the corpus
- the less uniform distribution the higher frequency reduction

What the future holds

- Near future
  - tokenisation for languages which do not use spaces between words
  - POS-tagging for more languages
  - keywords extraction for more languages
- Future
  - building parallel corpora
  - avoiding problems with Google API

WebBootCaT – Summary

- building domain specific corpora
- fast (1 million words in 10 to 20 minutes)
- easy to use
- web based
- keywords extraction
- powerful corpus manager