WEB CORPUS IN ONE CLICK

Jan Pomikálek, Vít Suchomel
NLP Centre

Masaryk University
13 December 2011

OUTLINE

- Text corpora
- Sketch Engine
- Creating web corpora
    - Web crawling
    - Character encoding detection
    - Language detection
    - Removing junk
    - De-duplication
- Results

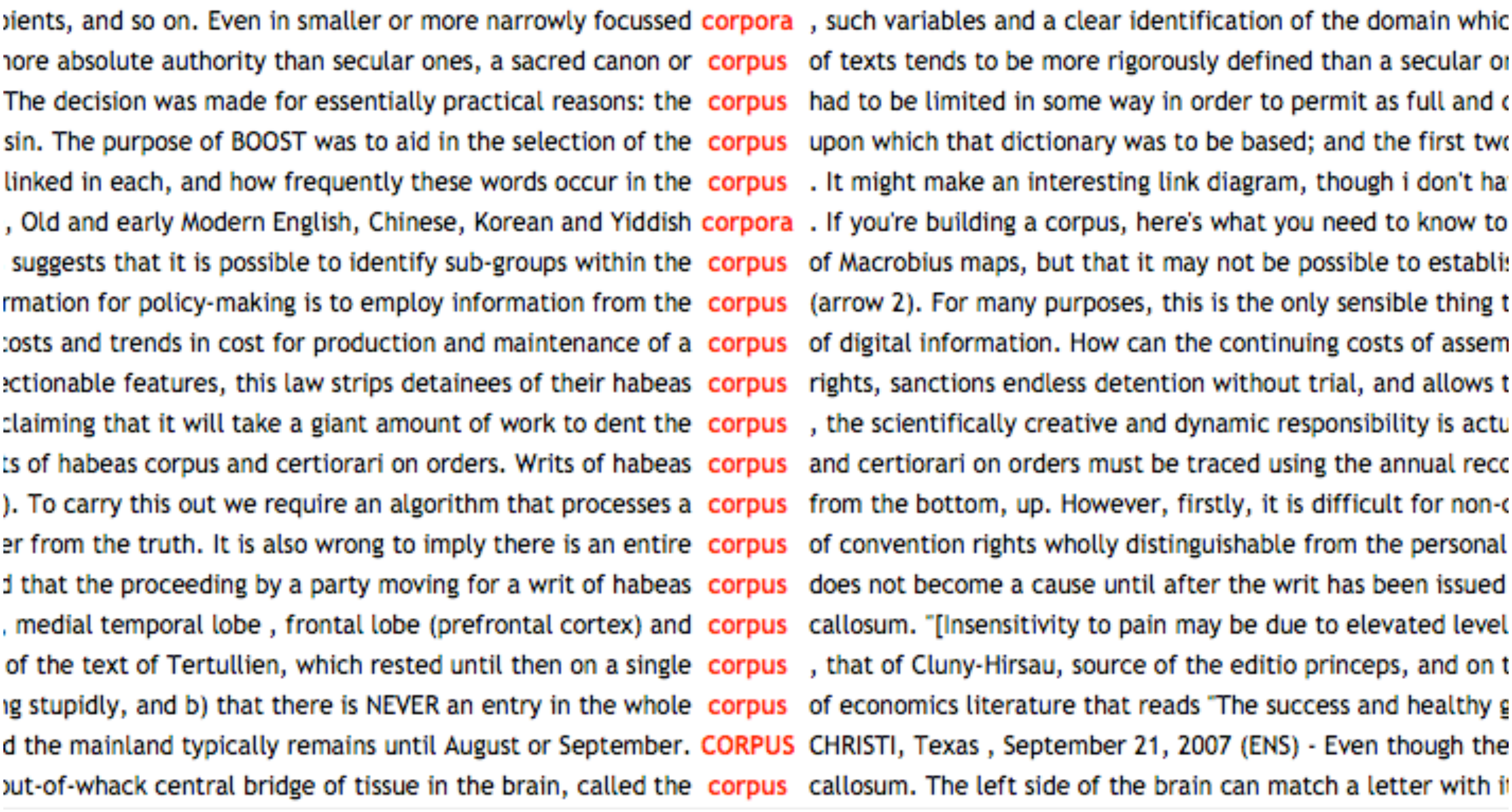
TEXT CORPORA

- Large collections of texts
- Typically monolingual
- Linguistically annotated (part-of-speech tags, lemmata)
- Wide range of applications
  - Speech recognition
  - Machine translation
  - Language teaching and learning
  - Lexicography (creating dictionaries)

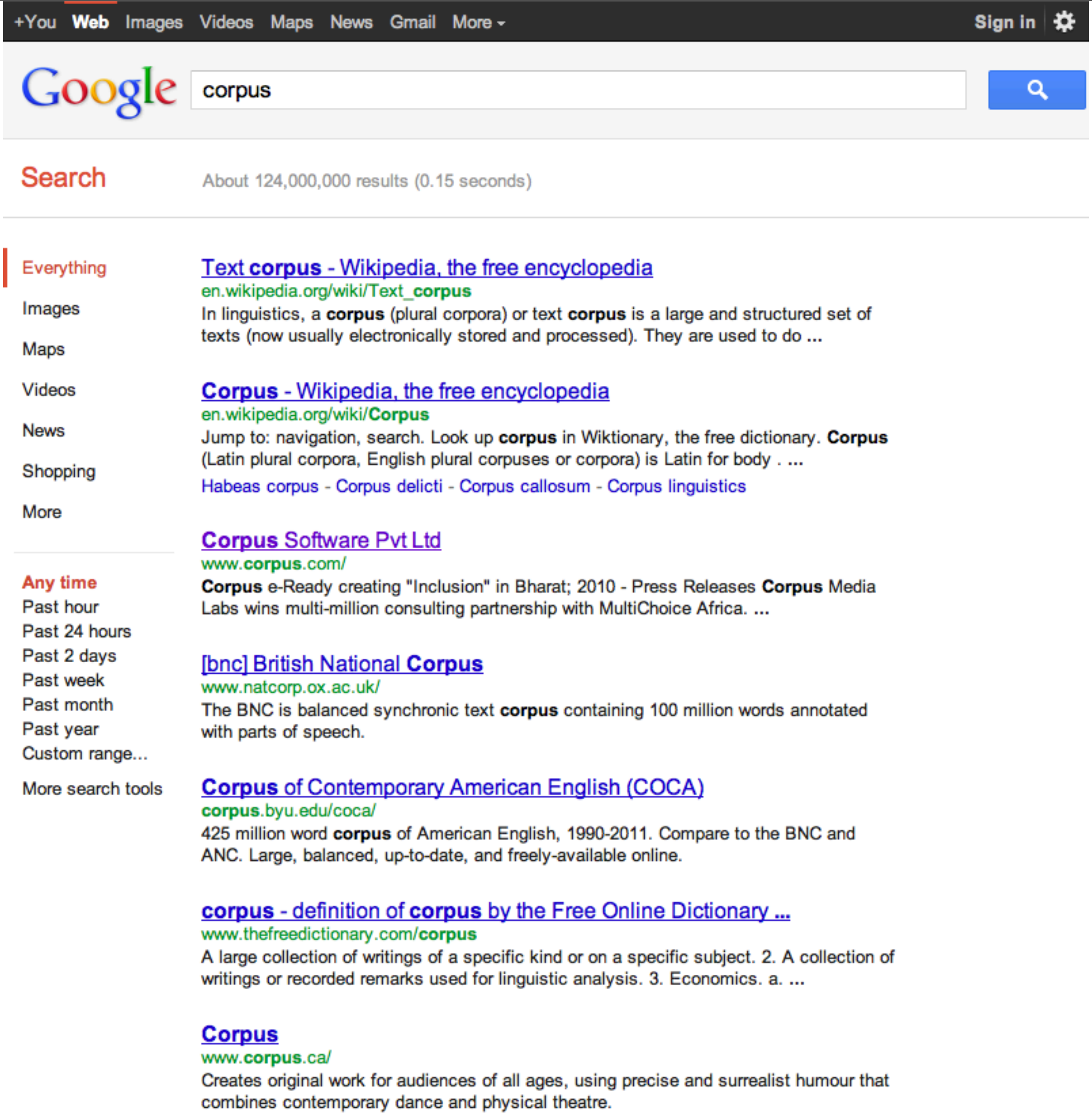
USE CASE EXAMPLES

- Looking up words

CONCORDANCE LINES FOR “CORPUS”



```
ients, and so on. Even in smaller or more narrowly focussed corpora , such variables and a clear identification of the domain whic
ore absolute authority than secular ones, a sacred canon or  corpus  of texts tends to be more rigorously defined than a secular o
The decision was made for essentially practical reasons: the  corpus  had to be limited in some way in order to permit as full and
sin. The purpose of BOOST was to aid in the selection of the  corpus  upon which that dictionary was to be based; and the first tw
linked in each, and how frequently these words occur in the  corpus  . It might make an interesting link diagram, though i don't ha
, Old and early Modern English, Chinese, Korean and Yiddish corpora . If you're building a corpus, here's what you need to know to
 suggests that it is possible to identify sub-groups within the  corpus  of Macrobius maps, but that it may not be possible to establi
rmation for policy-making is to employ information from the  corpus  (arrow 2). For many purposes, this is the only sensible thing t
osts and trends in cost for production and maintenance of a  corpus  of digital information. How can the continuing costs of assem
ctionable features, this law strips detainees of their habeas  corpus  rights, sanctions endless detention without trial, and allows t
laiming that it will take a giant amount of work to dent the  corpus  , the scientifically creative and dynamic responsibility is actu
s of habeas corpus and certiorari on orders. Writs of habeas  corpus  and certiorari on orders must be traced using the annual rec
). To carry this out we require an algorithm that processes a  corpus  from the bottom, up. However, firstly, it is difficult for non-
r from the truth. It is also wrong to imply there is an entire  corpus  of convention rights wholly distinguishable from the personal
 that the proceeding by a party moving for a writ of habeas  corpus  does not become a cause until after the writ has been issued
, medial temporal lobe , frontal lobe (prefrontal cortex) and  corpus  callosum. "[Insensitivity to pain may be due to elevated level
of the text of Tertullien, which rested until then on a single  corpus  , that of Cluny-Hirsau, source of the editio princeps, and on t
g stupidly, and b) that there is NEVER an entry in the whole  corpus  of economics literature that reads "The success and healthy
d the mainland typically remains until August or September. CORPUS CHRISTI, Texas , September 21, 2007 (ENS) - Even though the
ut-of-whack central bridge of tissue in the brain, called the  corpus  callosum. The left side of the brain can match a letter with i
```



About 124,000,000 results (0.15 seconds)

Text corpus - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Text_**corpus**
In linguistics, a **corpus** (plural corpora) or text **corpus** is a large and structured set of texts (now usually electronically stored and processed). They are used to do ...

Corpus - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Corpus**
Jump to: navigation, search. Look up **corpus** in Wiktionary, the free dictionary. **Corpus** (Latin plural corpora, English plural corpuses or corpora) is Latin for body . ...
Habeas corpus - Corpus delicti - Corpus callosum - Corpus linguistics

Corpus Software Pvt Ltd
www.**corpus**.com/
Corpus e-Ready creating "Inclusion" in Bharat; 2010 - Press Releases **Corpus** Media Labs wins multi-million consulting partnership with MultiChoice Africa. ...

[bnc] British National **Corpus**
www.natcorp.ox.ac.uk/
The BNC is balanced synchronic text **corpus** containing 100 million words annotated with parts of speech.

Corpus of Contemporary American English (COCA)
corpus.byu.edu/coca/
425 million word **corpus** of American English, 1990-2011. Compare to the BNC and ANC. Large, balanced, up-to-date, and freely-available online.

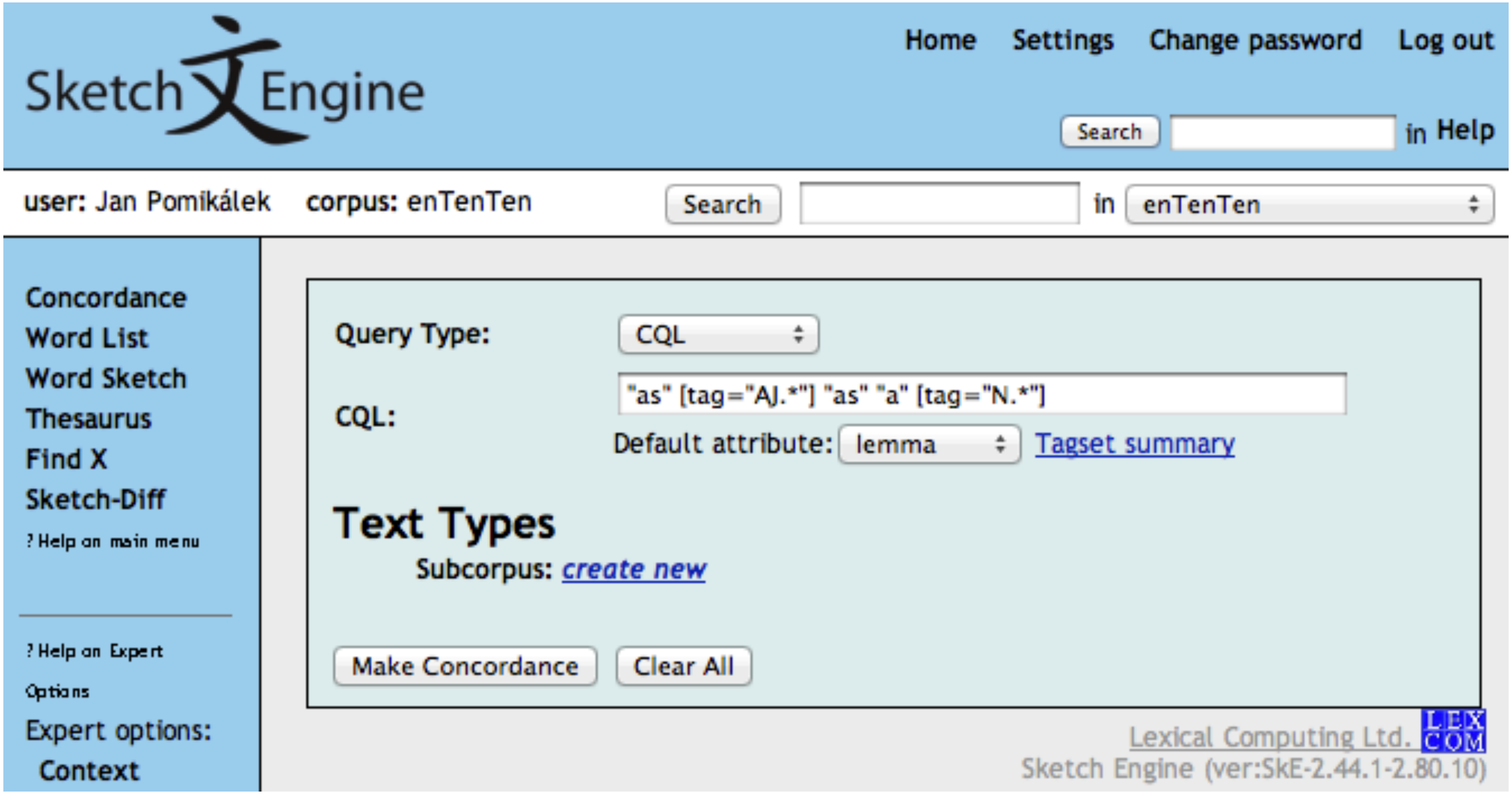
corpus - definition of **corpus** by the Free Online Dictionary ...
www.thefreedictionary.com/**corpus**
A large collection of writings of a specific kind or on a specific subject. 2. A collection of writings or recorded remarks used for linguistic analysis. 3. Economics. a. ...

Corpus
www.**corpus**.ca/
Creates original work for audiences of all ages, using precise and surrealist humour that combines contemporary dance and physical theatre.

USE CASE EXAMPLES

- Looking up words
- Looking up patterns

CQL: “AS <ADJECTIVE> AS A <NOUN>”



user: Jan Pomikálek corpus: enTenTen

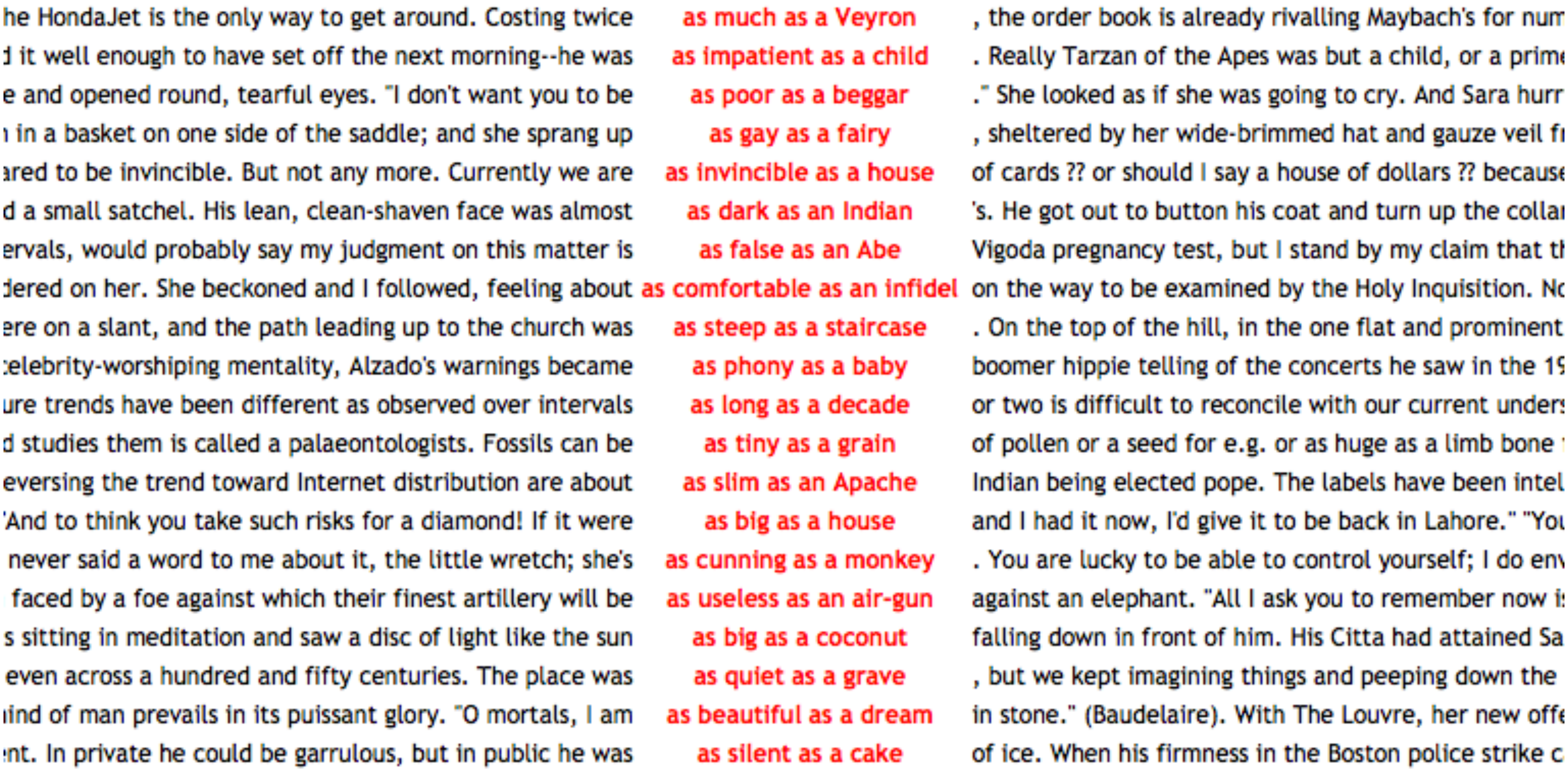
Query Type: CQL

CQL: "as" [tag="AJ.*"] "as" "a" [tag="N.*"]

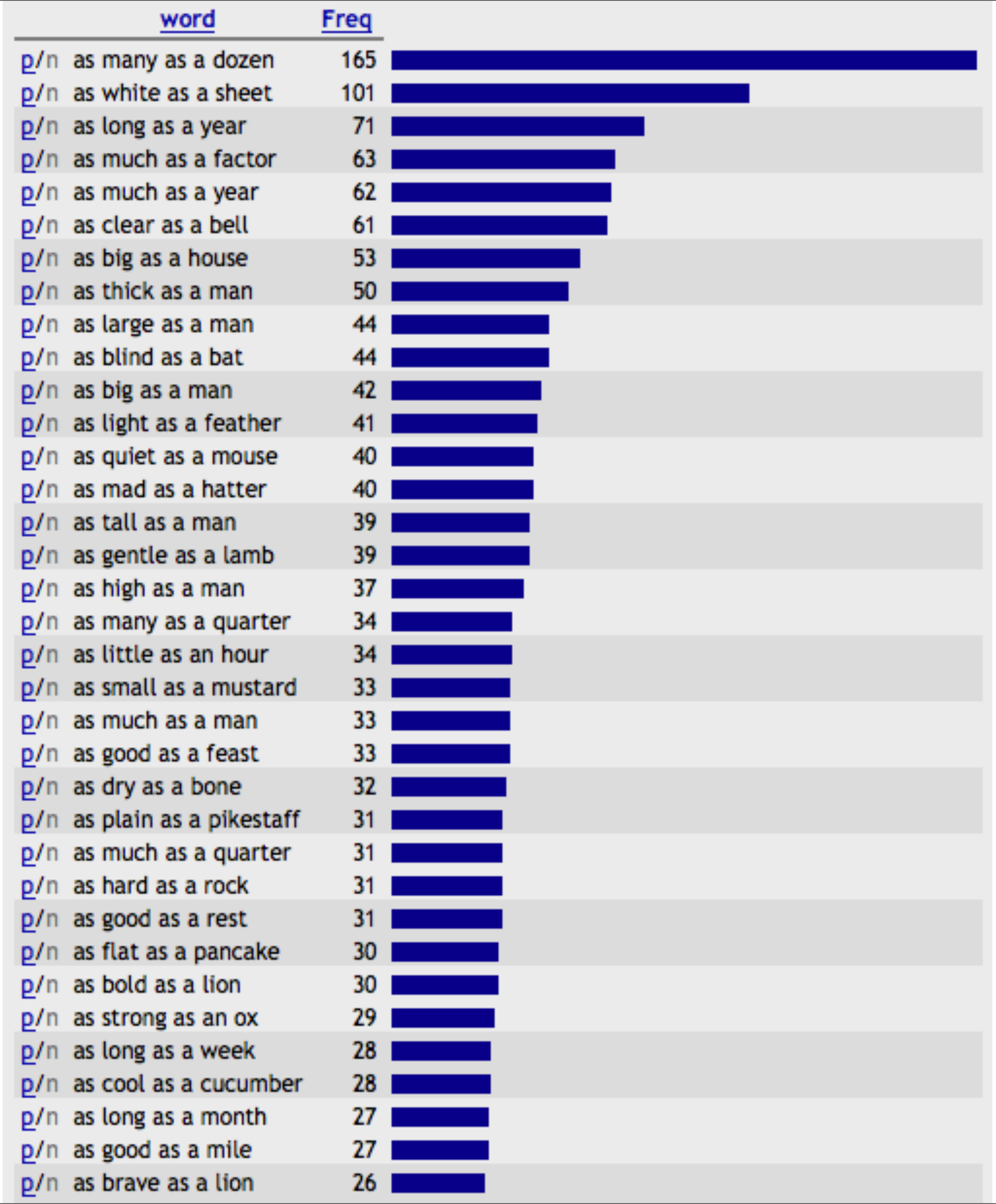
Default attribute: lemma

Subcorpus: *create new*

“AS <ADJECTIVE> AS A <NOUN>”



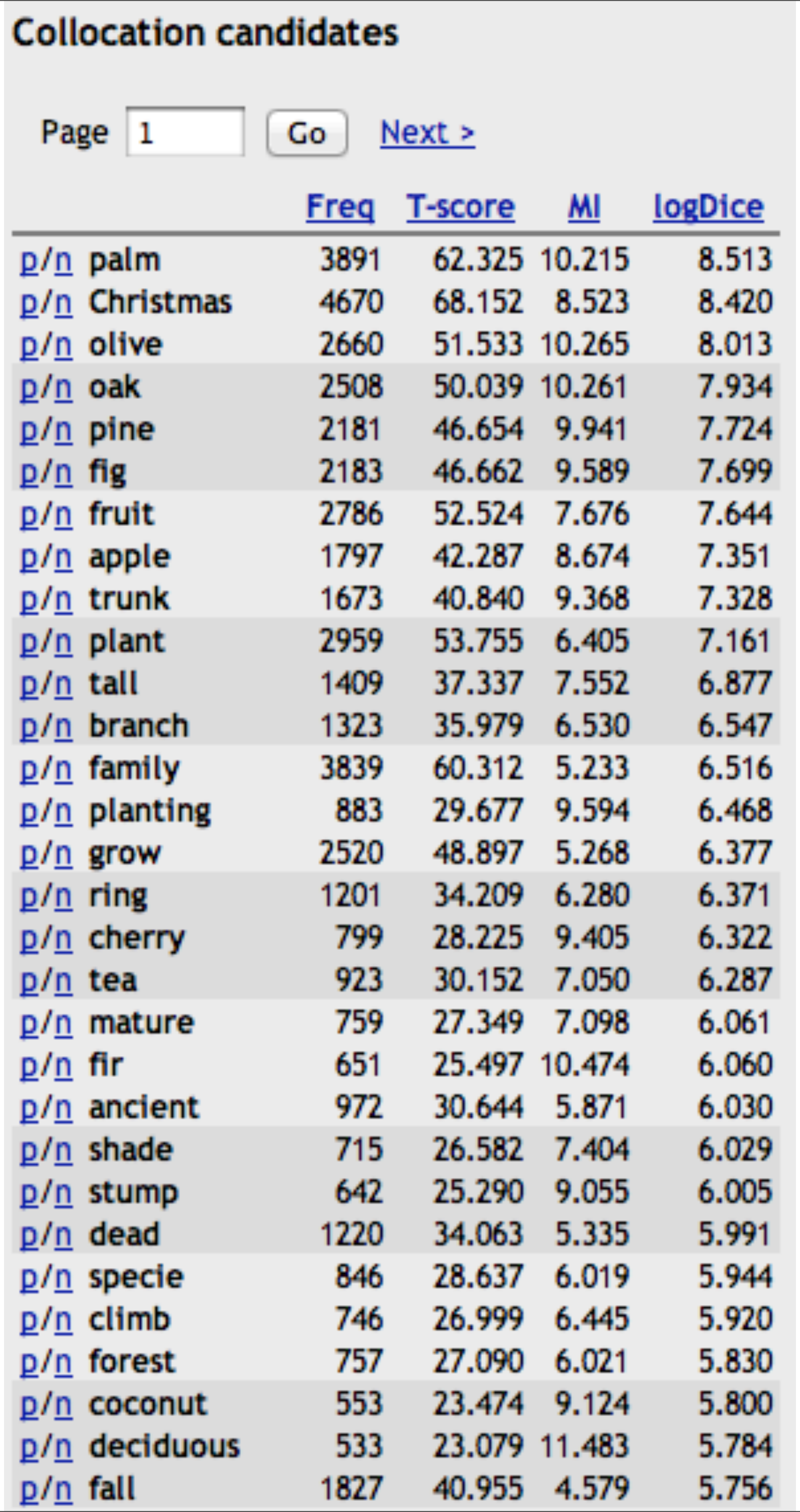
he HondaJet is the only way to get around. Costing twice	as much as a Veyron	, the order book is already rivalling Maybach's for num
d it well enough to have set off the next morning--he was	as impatient as a child	. Really Tarzan of the Apes was but a child, or a prime
e and opened round, tearful eyes. "I don't want you to be	as poor as a beggar	." She looked as if she was going to cry. And Sara hurr
ı in a basket on one side of the saddle; and she sprang up	as gay as a fairy	, sheltered by her wide-brimmed hat and gauze veil fı
ared to be invincible. But not any more. Currently we are	as invincible as a house	of cards ?? or should I say a house of dollars ?? because
d a small satchel. His lean, clean-shaven face was almost	as dark as an Indian	's. He got out to button his coat and turn up the collaı
ervals, would probably say my judgment on this matter is	as false as an Abe	Vigoda pregnancy test, but I stand by my claim that tl
dered on her. She beckoned and I followed, feeling about	as comfortable as an infidel	on the way to be examined by the Holy Inquisition. Nc
ere on a slant, and the path leading up to the church was	as steep as a staircase	. On the top of the hill, in the one flat and prominent
celebrity-worshiping mentality, Alzado's warnings became	as phony as a baby	boomer hippie telling of the concerts he saw in the 19
ure trends have been different as observed over intervals	as long as a decade	or two is difficult to reconcile with our current unders
d studies them is called a palaeontologists. Fossils can be	as tiny as a grain	of pollen or a seed for e.g. or as huge as a limb bone
eversing the trend toward Internet distribution are about	as slim as an Apache	Indian being elected pope. The labels have been intel
'And to think you take such risks for a diamond! If it were	as big as a house	and I had it now, I'd give it to be back in Lahore." "You
never said a word to me about it, the little wretch; she's	as cunning as a monkey	. You are lucky to be able to control yourself; I do env
faced by a foe against which their finest artillery will be	as useless as an air-gun	against an elephant. "All I ask you to remember now is
s sitting in meditation and saw a disc of light like the sun	as big as a coconut	falling down in front of him. His Citta had attained Sa
even across a hundred and fifty centuries. The place was	as quiet as a grave	, but we kept imagining things and peeping down the
ind of man prevails in its puissant glory. "O mortals, I am	as beautiful as a dream	in stone." (Baudelaire). With The Louvre, her new offe
nt. In private he could be garrulous, but in public he was	as silent as a cake	of ice. When his firmness in the Boston police strike c



	word	Freq
p/n	as many as a dozen	165
p/n	as white as a sheet	101
p/n	as long as a year	71
p/n	as much as a factor	63
p/n	as much as a year	62
p/n	as clear as a bell	61
p/n	as big as a house	53
p/n	as thick as a man	50
p/n	as large as a man	44
p/n	as blind as a bat	44
p/n	as big as a man	42
p/n	as light as a feather	41
p/n	as quiet as a mouse	40
p/n	as mad as a hatter	40
p/n	as tall as a man	39
p/n	as gentle as a lamb	39
p/n	as high as a man	37
p/n	as many as a quarter	34
p/n	as little as an hour	34
p/n	as small as a mustard	33
p/n	as much as a man	33
p/n	as good as a feast	33
p/n	as dry as a bone	32
p/n	as plain as a pikestaff	31
p/n	as much as a quarter	31
p/n	as hard as a rock	31
p/n	as good as a rest	31
p/n	as flat as a pancake	30
p/n	as bold as a lion	30
p/n	as strong as an ox	29
p/n	as long as a week	28
p/n	as cool as a cucumber	28
p/n	as long as a month	27
p/n	as good as a mile	27
p/n	as brave as a lion	26

USE CASE EXAMPLES

- Looking up words
- Looking up patterns
- Frequencies of words and patterns
- Collocations

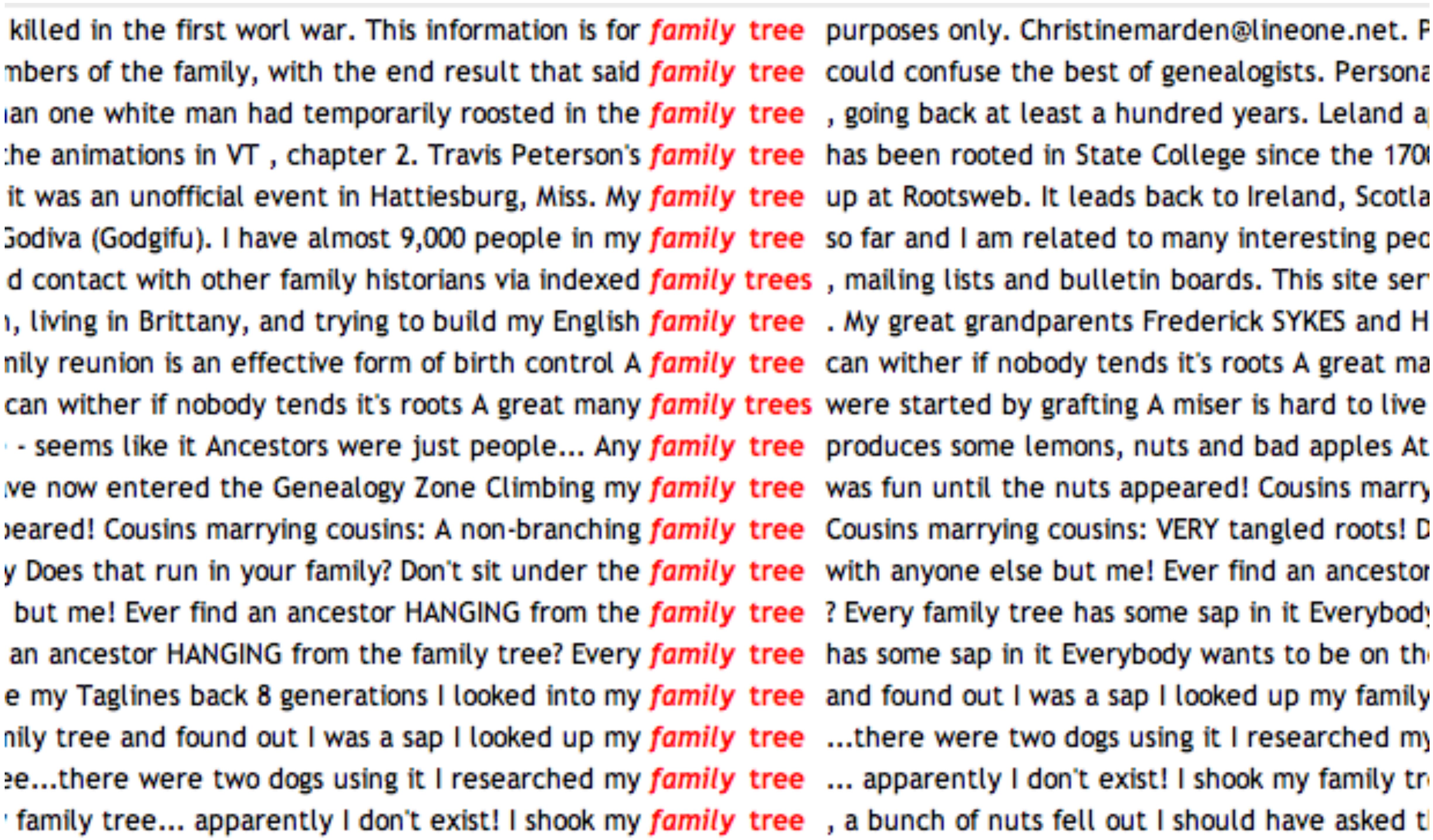


Collocation candidates

Page 1 Go Next >

		Freq	T-score	MI	logDice
p/n	palm	3891	62.325	10.215	8.513
p/n	Christmas	4670	68.152	8.523	8.420
p/n	olive	2660	51.533	10.265	8.013
p/n	oak	2508	50.039	10.261	7.934
p/n	pine	2181	46.654	9.941	7.724
p/n	fig	2183	46.662	9.589	7.699
p/n	fruit	2786	52.524	7.676	7.644
p/n	apple	1797	42.287	8.674	7.351
p/n	trunk	1673	40.840	9.368	7.328
p/n	plant	2959	53.755	6.405	7.161
p/n	tall	1409	37.337	7.552	6.877
p/n	branch	1323	35.979	6.530	6.547
p/n	family	3839	60.312	5.233	6.516
p/n	planting	883	29.677	9.594	6.468
p/n	grow	2520	48.897	5.268	6.377
p/n	ring	1201	34.209	6.280	6.371
p/n	cherry	799	28.225	9.405	6.322
p/n	tea	923	30.152	7.050	6.287
p/n	mature	759	27.349	7.098	6.061
p/n	fir	651	25.497	10.474	6.060
p/n	ancient	972	30.644	5.871	6.030
p/n	shade	715	26.582	7.404	6.029
p/n	stump	642	25.290	9.055	6.005
p/n	dead	1220	34.063	5.335	5.991
p/n	specie	846	28.637	6.019	5.944
p/n	climb	746	26.999	6.445	5.920
p/n	forest	757	27.090	6.021	5.830
p/n	coconut	553	23.474	9.124	5.800
p/n	deciduous	533	23.079	11.483	5.784
p/n	fall	1827	40.955	4.579	5.756

“FAMILY TREE”

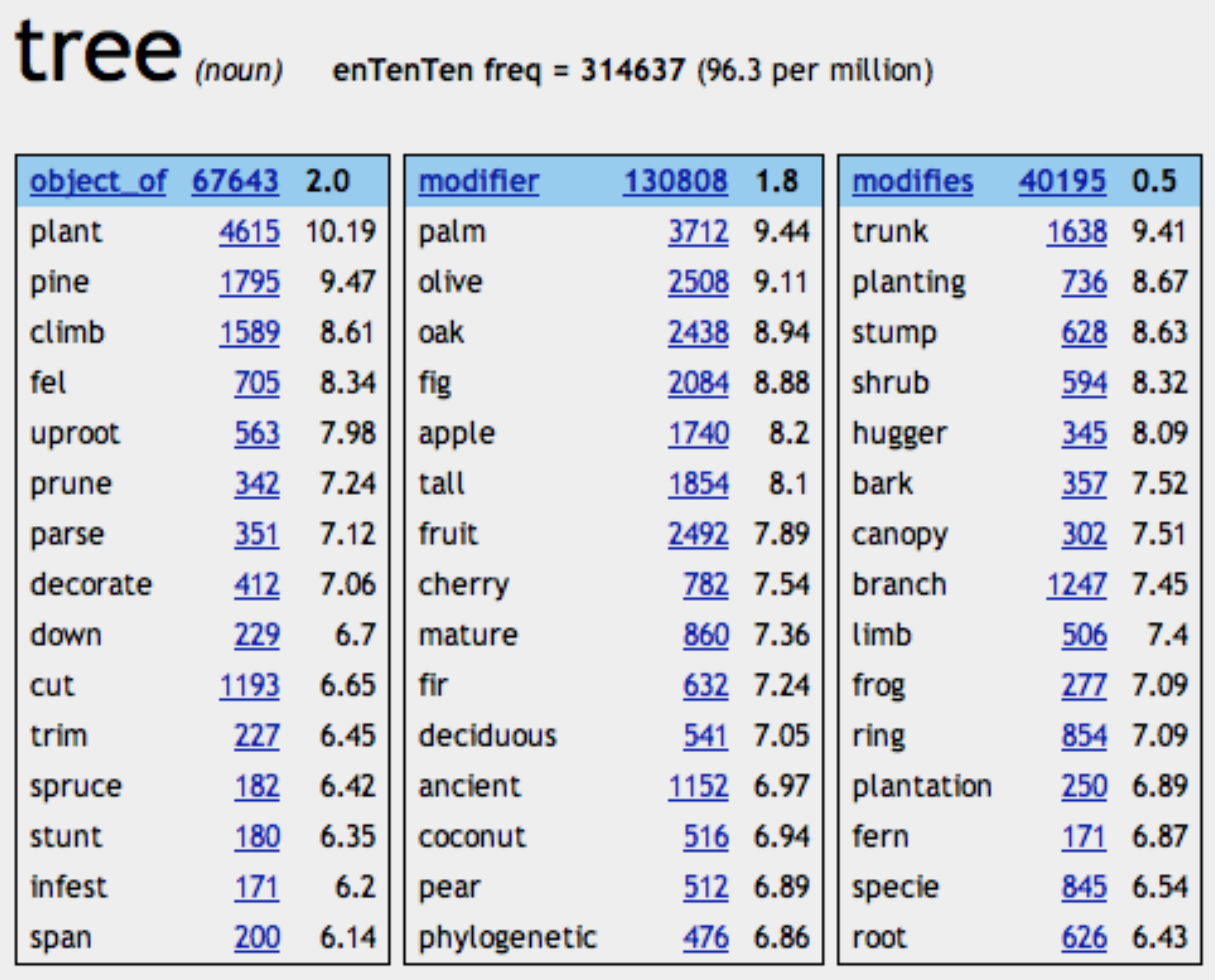


```
killed in the first worl war. This information is for family tree  purposes only. Christinemarden@lineone.net. F
mbers of the family, with the end result that said family tree  could confuse the best of genealogists. Persona
an one white man had temporarily roosted in the family tree  , going back at least a hundred years. Leland a
he animations in VT , chapter 2. Travis Peterson's family tree  has been rooted in State College since the 170
it was an unofficial event in Hattiesburg, Miss. My family tree  up at Rootsweb. It leads back to Ireland, Scotla
Godiva (Godgifu). I have almost 9,000 people in my family tree  so far and I am related to many interesting peo
d contact with other family historians via indexed family trees , mailing lists and bulletin boards. This site ser
, living in Brittany, and trying to build my English family tree  . My great grandparents Frederick SYKES and H
nily reunion is an effective form of birth control A family tree  can wither if nobody tends it's roots A great ma
can wither if nobody tends it's roots A great many family trees were started by grafting A miser is hard to live
- seems like it Ancestors were just people... Any family tree  produces some lemons, nuts and bad apples At
ve now entered the Genealogy Zone Climbing my family tree  was fun until the nuts appeared! Cousins marry
eared! Cousins marrying cousins: A non-branching family tree  Cousins marrying cousins: VERY tangled roots! D
y Does that run in your family? Don't sit under the family tree  with anyone else but me! Ever find an ancestor
 but me! Ever find an ancestor HANGING from the family tree  ? Every family tree has some sap in it Everybody
an ancestor HANGING from the family tree? Every family tree  has some sap in it Everybody wants to be on th
e my Taglines back 8 generations I looked into my family tree  and found out I was a sap I looked up my family
nily tree and found out I was a sap I looked up my family tree  ...there were two dogs using it I researched my
ee...there were two dogs using it I researched my family tree  ... apparently I don't exist! I shook my family tr
 family tree... apparently I don't exist! I shook my family tree  , a bunch of nuts fell out I should have asked t
```

USE CASE EXAMPLES

- Looking up words
- Looking up patterns
- Frequencies of words and patterns
- Collocations
- Word sketches

WORD SKETCHES FOR "TREE"



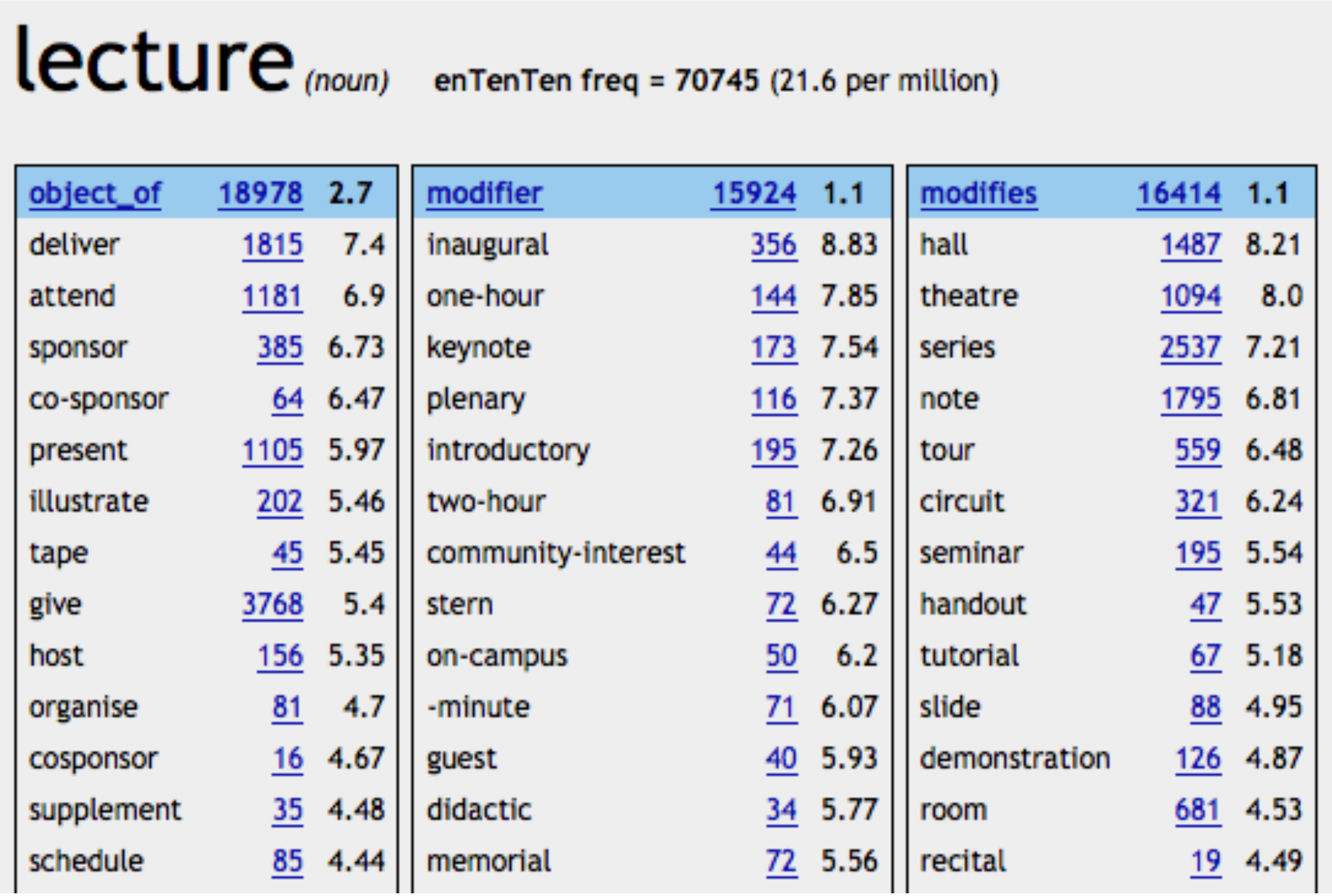
tree *(noun)* **enTenTen freq = 314637** (96.3 per million)

object_of	67643	2.0
plant	4615	10.19
pine	1795	9.47
climb	1589	8.61
fel	705	8.34
uproot	563	7.98
prune	342	7.24
parse	351	7.12
decorate	412	7.06
down	229	6.7
cut	1193	6.65
trim	227	6.45
spruce	182	6.42
stunt	180	6.35
infest	171	6.2
span	200	6.14

modifier	130808	1.8
palm	3712	9.44
olive	2508	9.11
oak	2438	8.94
fig	2084	8.88
apple	1740	8.2
tall	1854	8.1
fruit	2492	7.89
cherry	782	7.54
mature	860	7.36
fir	632	7.24
deciduous	541	7.05
ancient	1152	6.97
coconut	516	6.94
pear	512	6.89
phylogenetic	476	6.86

modifies	40195	0.5
trunk	1638	9.41
planting	736	8.67
stump	628	8.63
shrub	594	8.32
hugger	345	8.09
bark	357	7.52
canopy	302	7.51
branch	1247	7.45
limb	506	7.4
frog	277	7.09
ring	854	7.09
plantation	250	6.89
fern	171	6.87
specie	845	6.54
root	626	6.43

WORD SKETCHES FOR “LECTURE (NOUN)”



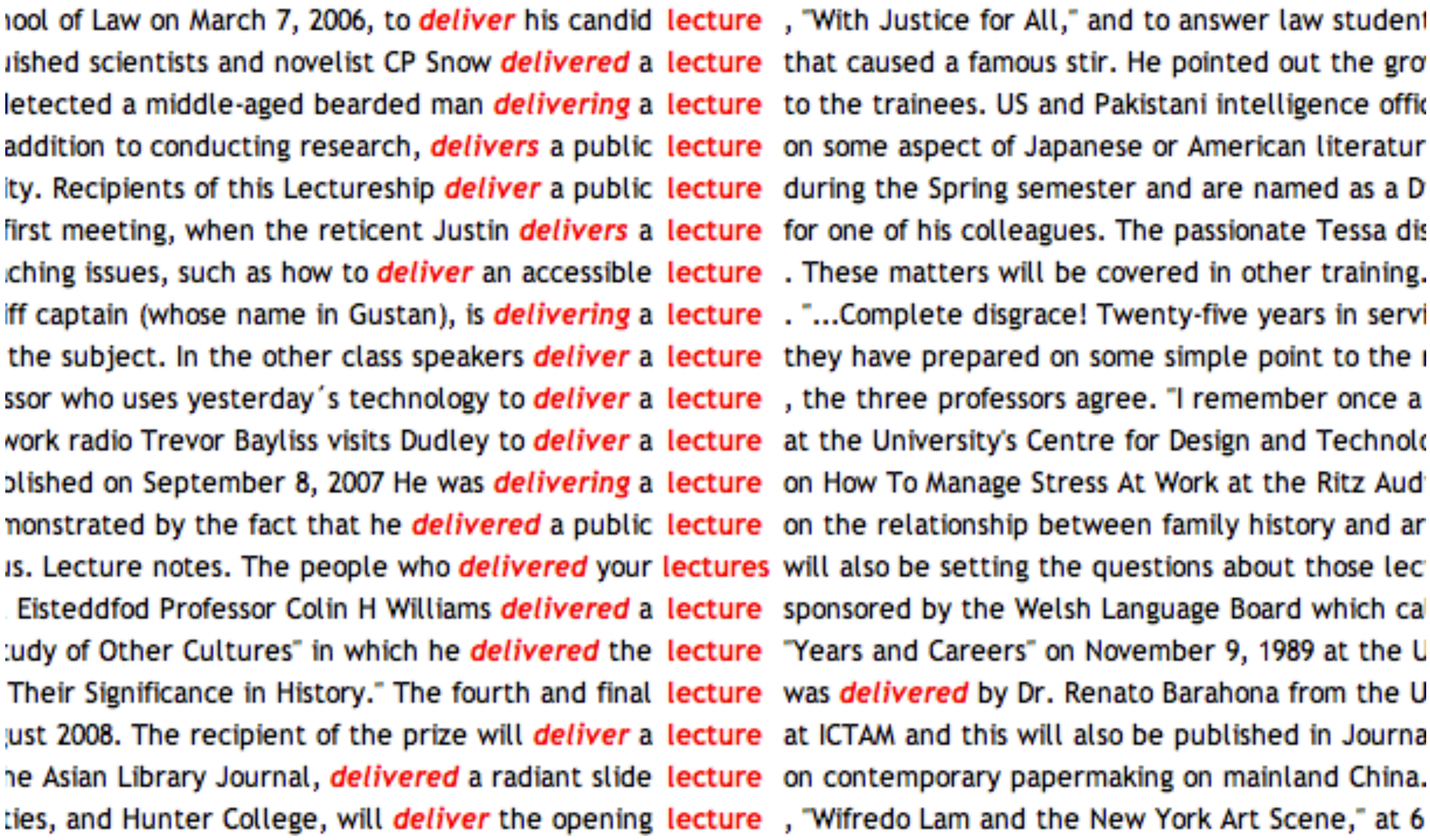
lecture *(noun)* enTenTen freq = 70745 (21.6 per million)

object_of	18978	2.7
deliver	1815	7.4
attend	1181	6.9
sponsor	385	6.73
co-sponsor	64	6.47
present	1105	5.97
illustrate	202	5.46
tape	45	5.45
give	3768	5.4
host	156	5.35
organise	81	4.7
cosponsor	16	4.67
supplement	35	4.48
schedule	85	4.44

modifier	15924	1.1
inaugural	356	8.83
one-hour	144	7.85
keynote	173	7.54
plenary	116	7.37
introductory	195	7.26
two-hour	81	6.91
community-interest	44	6.5
stern	72	6.27
on-campus	50	6.2
-minute	71	6.07
guest	40	5.93
didactic	34	5.77
memorial	72	5.56

modifies	16414	1.1
hall	1487	8.21
theatre	1094	8.0
series	2537	7.21
note	1795	6.81
tour	559	6.48
circuit	321	6.24
seminar	195	5.54
handout	47	5.53
tutorial	67	5.18
slide	88	4.95
demonstration	126	4.87
room	681	4.53
recital	19	4.49

“TO DELIVER A LECTURE”



hool of Law on March 7, 2006, to deliver his candid lecture , "With Justice for All," and to answer law student

uished scientists and novelist CP Snow delivered a lecture that caused a famous stir. He pointed out the gro

detected a middle-aged bearded man delivering a lecture to the trainees. US and Pakistani intelligence offic

addition to conducting research, delivers a public lecture on some aspect of Japanese or American literatur

ity. Recipients of this Lectureship deliver a public lecture during the Spring semester and are named as a D

first meeting, when the reticent Justin delivers a lecture for one of his colleagues. The passionate Tessa dis

aching issues, such as how to deliver an accessible lecture . These matters will be covered in other training.

iff captain (whose name in Gustan), is delivering a lecture . "...Complete disgrace! Twenty-five years in servi

the subject. In the other class speakers deliver a lecture they have prepared on some simple point to the

ssor who uses yesterday´s technology to deliver a lecture , the three professors agree. "I remember once a

work radio Trevor Bayliss visits Dudley to deliver a lecture at the University's Centre for Design and Technolo

blished on September 8, 2007 He was delivering a lecture on How To Manage Stress At Work at the Ritz Aud

monstrated by the fact that he delivered a public lecture on the relationship between family history and ar

us. Lecture notes. The people who delivered your lectures will also be setting the questions about those lec

Eisteddfod Professor Colin H Williams delivered a lecture sponsored by the Welsh Language Board which ca

tudy of Other Cultures" in which he delivered the lecture "Years and Careers" on November 9, 1989 at the U

Their Significance in History." The fourth and final lecture was delivered by Dr. Renato Barahona from the U

gust 2008. The recipient of the prize will deliver a lecture at ICTAM and this will also be published in Journa

he Asian Library Journal, delivered a radiant slide lecture on contemporary papermaking on mainland China.

ties, and Hunter College, will deliver the opening lecture , "Wifredo Lam and the New York Art Scene," at 6

SKETCH ENGINE

- Corpus query system
- 170 text corpora, 46 languages
- Developed by Lexical Computing Ltd in cooperation with NLP Centre MU
- Freely available to MU students and staff
  - http://ske.fi.muni.cz/
- Available to public for a fee
  - http://the.sketchengine.co.uk/
- Open source version (no word sketches)
  - NoSketch Engine
  - http://nlp.fi.muni.cz/trac/noske

SkE users at MU

- 370 registered users, 130 active (at least 1 access in the last month)
- 73,876 page views in November 2011
- Teaching (foreign languages, linguistics)
  - James Thomas (Faculty of Arts)
  - Jarmila Fictumová (Faculty of Arts)
  - Radomíra Bednářová (Faculty of Science)
- Research projects
  - Corpus Pattern Analysis (NLP Centre, Patrick Hanks)
  - Verbalex (NLP Centre)

SkE users world-wide

- ca. 100 organisations, 4000 users
- Dictionary publishing houses
  - Oxford University Press (UK)
  - Cambridge University Press (UK)
  - Harper Collins (UK)
  - Amebis (Slovenia)
- Research institutions
  - Institut voor Nederlandse Lexicologie (Netherlands)
  - Institute of the Estonian Language (Estonia)
  - University Pompeu Fabre (Spain)
  - Institut Libre Marie Haps (Belgium)

TRADITIONAL CORPORA

- From printed materials
  - Books, newspaper, magazines
  - Scanning, OCR
- Controlled
  - We know what kind of texts are inside
  - Contents based on sociologic studies
- British National Corpus (100M words, 1994)
- Czech National Corpus (1300M words, 2010)
- Disadvantages
  - Expensive
  - Time consuming
  - Limited size

WEB CORPORA

- Uncontrolled
  - We are not sure what kinds of texts we are downloading and at which amounts
- Cheap (provided we have required technology)
- Can be made really big (at least for some languages)
- The only option for under-resourced languages

TRADITIONAL VS. WEB CORPORA

- Existing studies suggest that they do not differ by much
- Serge Sharrof, 2006: Creating general-purpose corpora using automated search engine queries.
  - English web corpus vs. BNC
  - Similar word frequency lists
  - Comparison of text genres

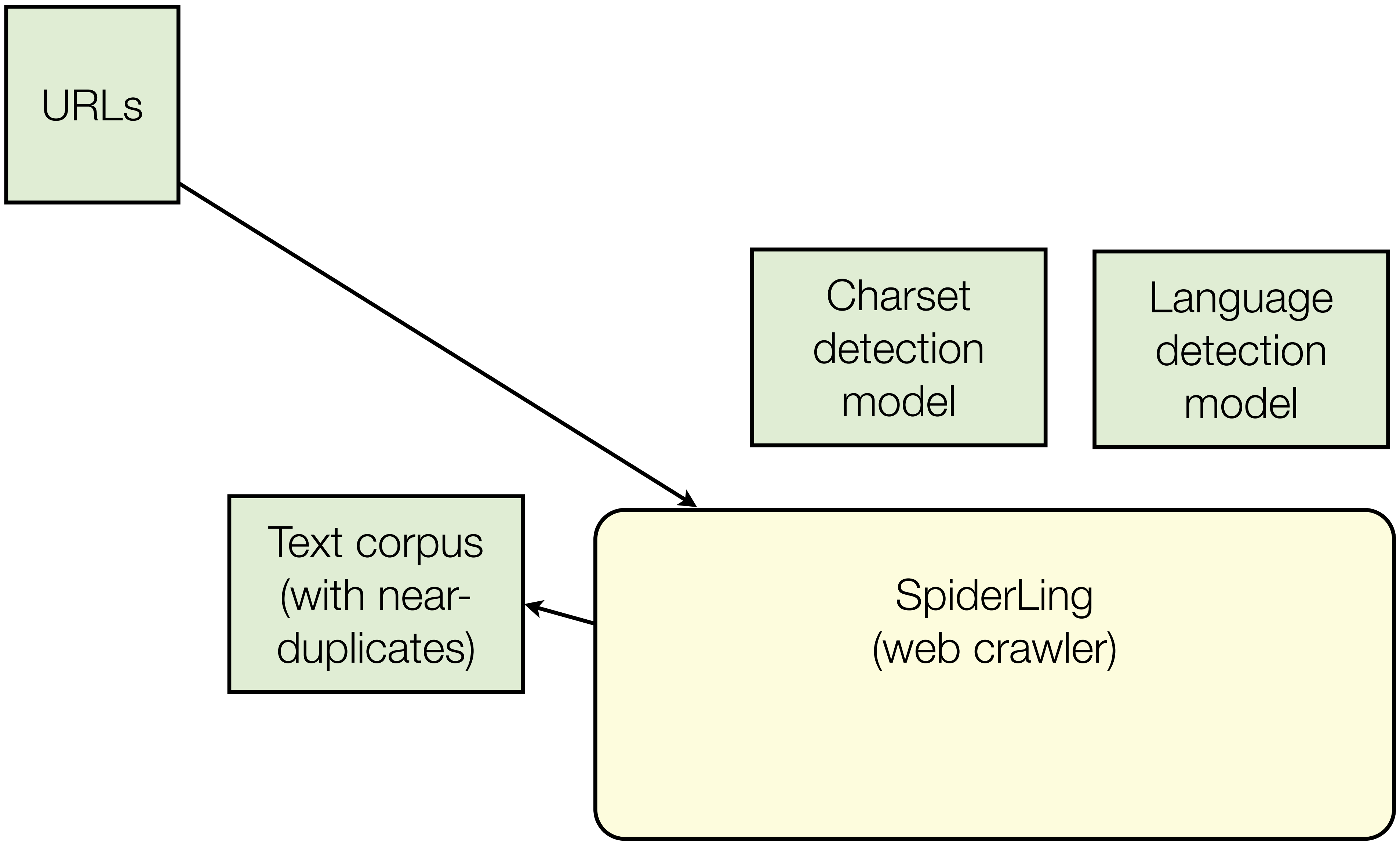
	BNC	I-EN (web corpus)
Life	27%	14%
Politics	19%	12%
Business	8%	13%
Natsci	4%	3%
Appsci	7%	**29%**
Socsci	17%	16%
Arts	7%	2%
Leisure	11%	11%

MORE DATA = BETTER DATA

- Rare language phenomena
- More data = more occurrences of rare items

corpus name	BNC	ukWaC	ClueWeb09 (7%)
size [tokens]	112 M	1,565 M	8,391 M
nascent	95	1 344	12 283
effort	7 576	106 262	805 142
unbridled	75	814	7 228
hedonism	63	594	4 061
nascent effort	0	1	22
unbridled hedonism	0	2	14

Building web corpora (with one click)



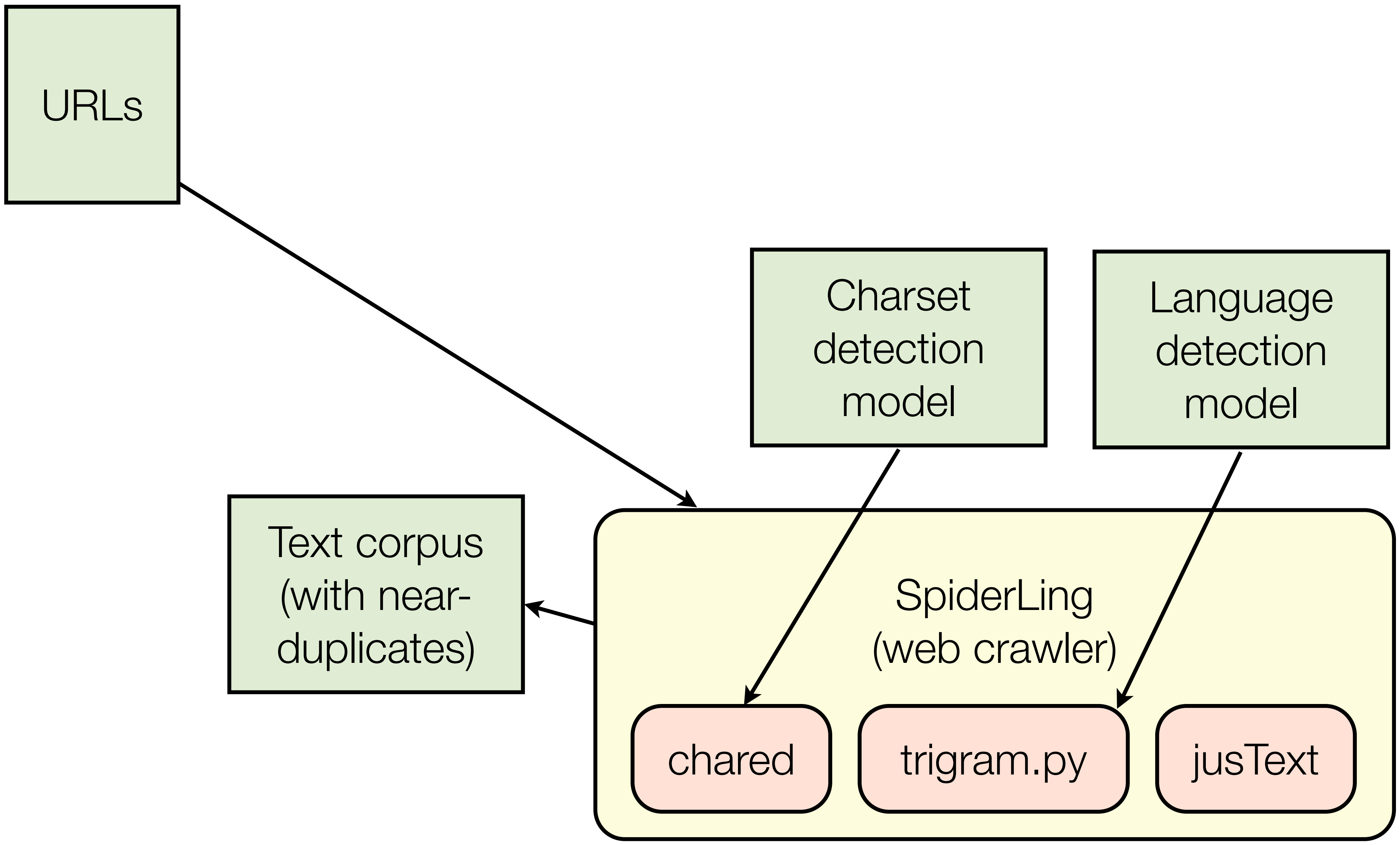
URLs

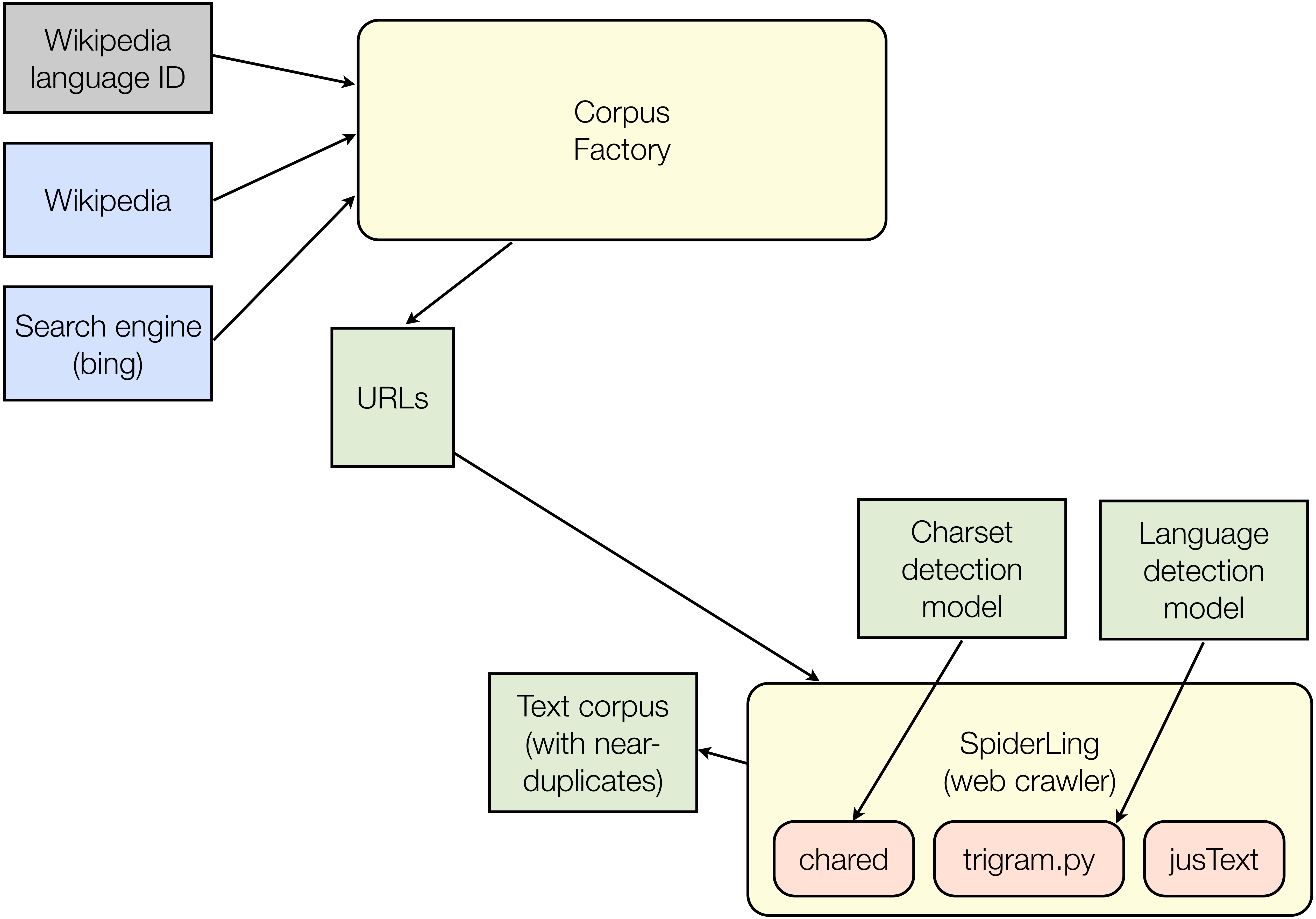
Charset detection model

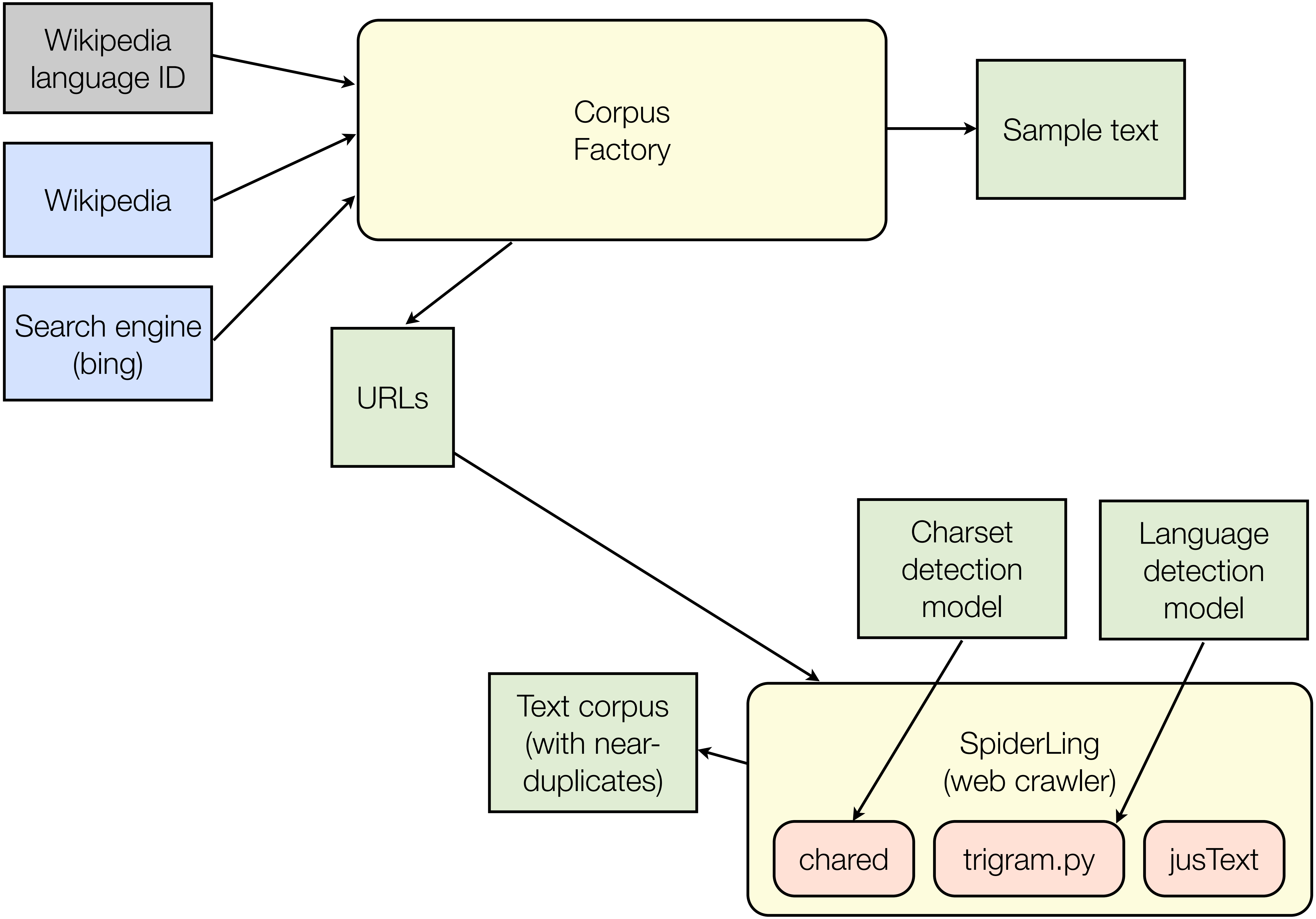
Language detection model

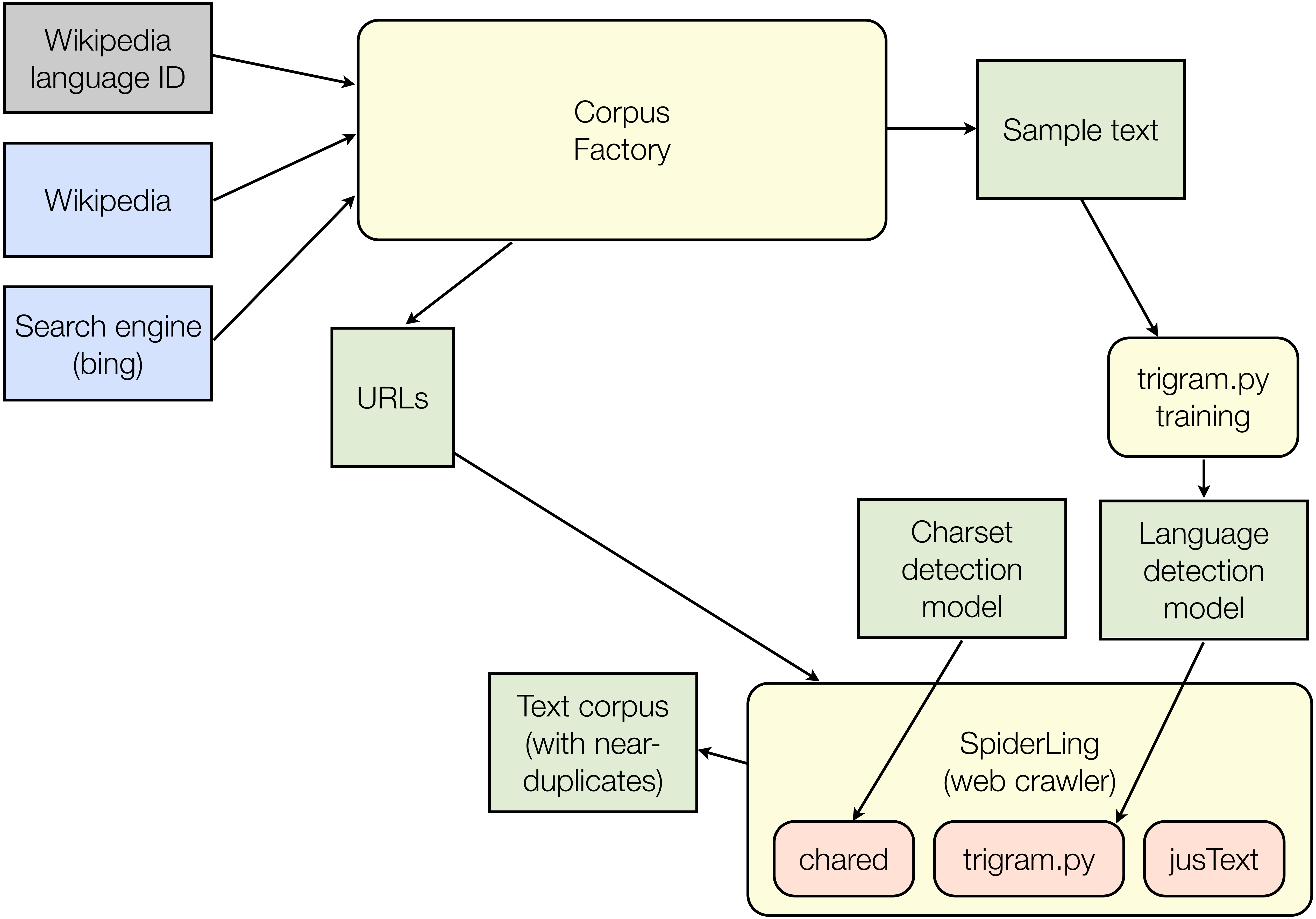
Text corpus (with near-duplicates)

SpiderLing (web crawler)


URLs
Charset detection model
Language detection model
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
URLs
Charset detection model
Language detection model
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
Charset detection model
Language detection model
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText



Wikipedia language ID

Wikipedia

Search engine (bing)

Corpus Factory

Sample text

URLs

trigram.py training

Charset detection model

Language detection model

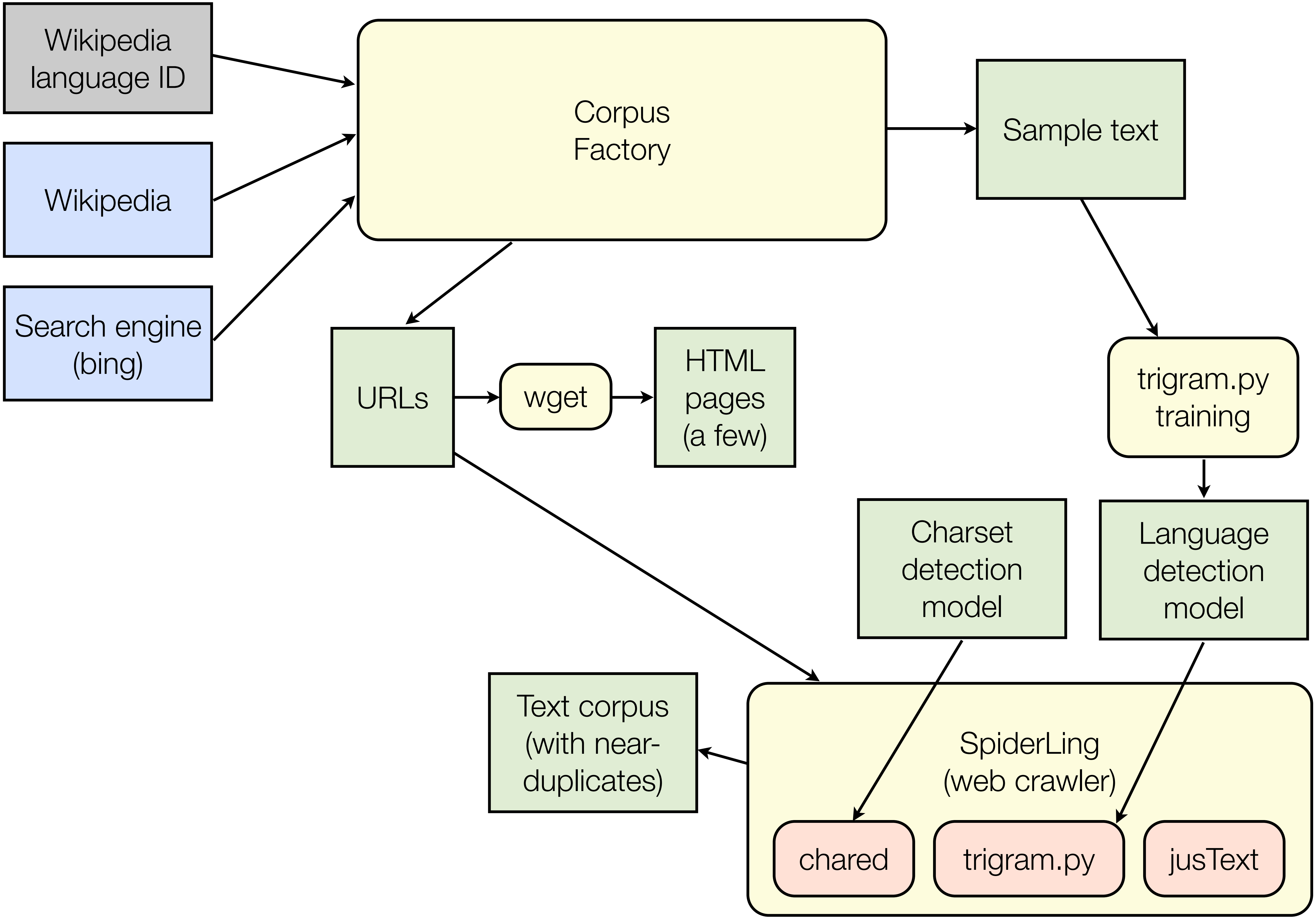
Text corpus (with near-duplicates)

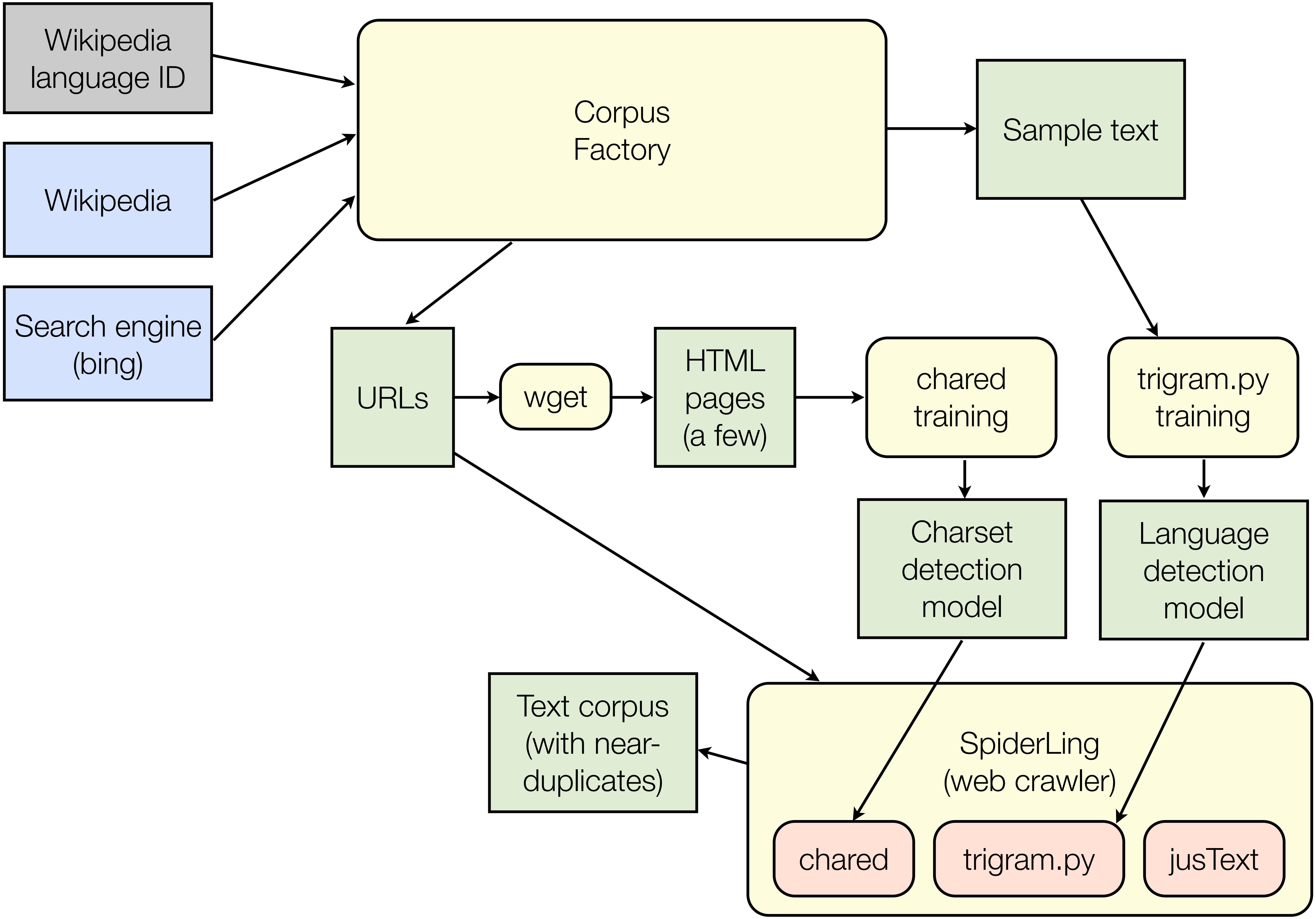
SpiderLing (web crawler)

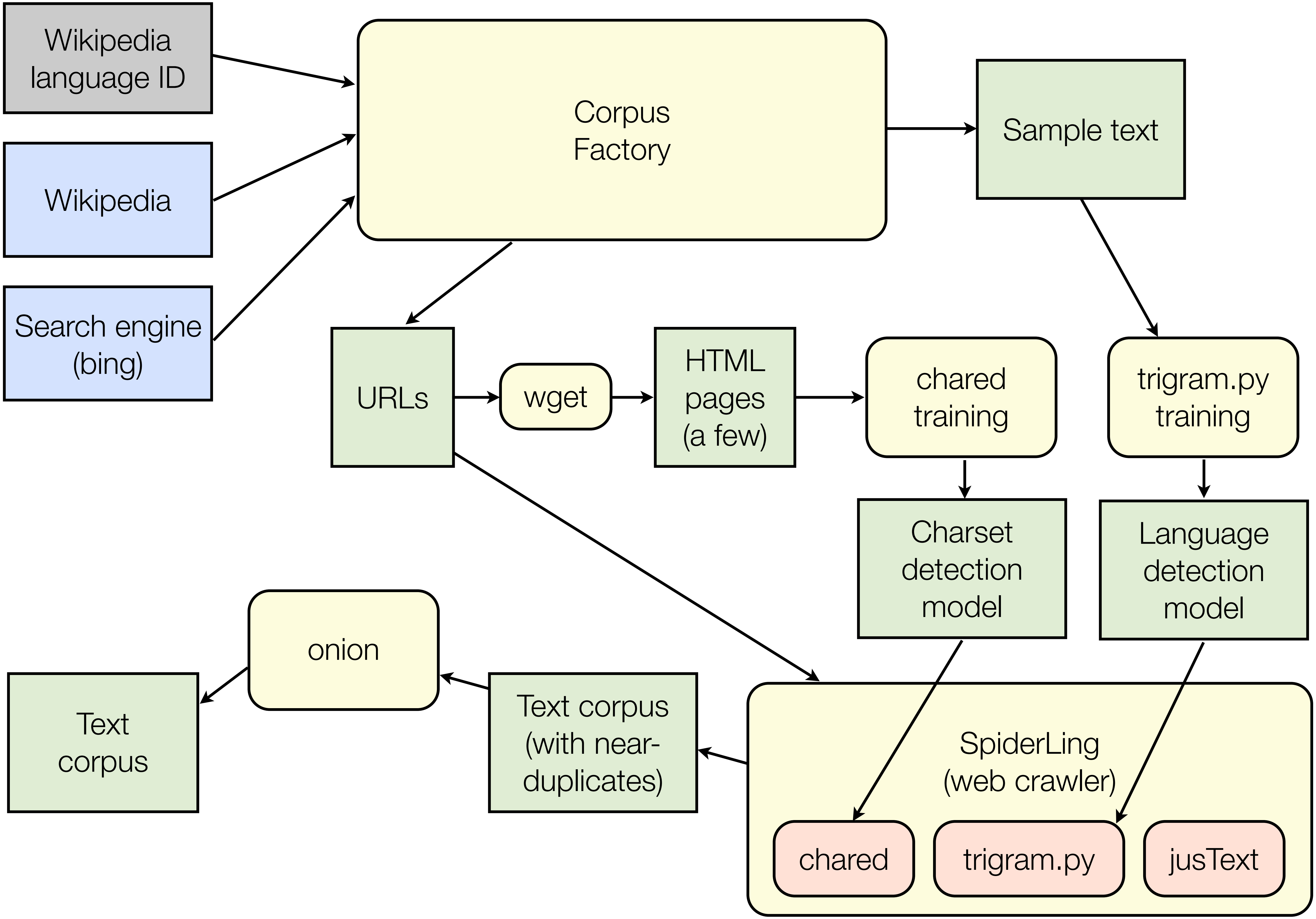
chared

trigram.py

jusText


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
trigram.py training
Charset detection model
Language detection model
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
chared training
trigram.py training
Charset detection model
Language detection model
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText



Wikipedia language ID

Wikipedia

Search engine (bing)

Corpus Factory

Sample text

URLs

wget

HTML pages (a few)

chared training

trigram.py training

Charset detection model

Language detection model

onion

Text corpus

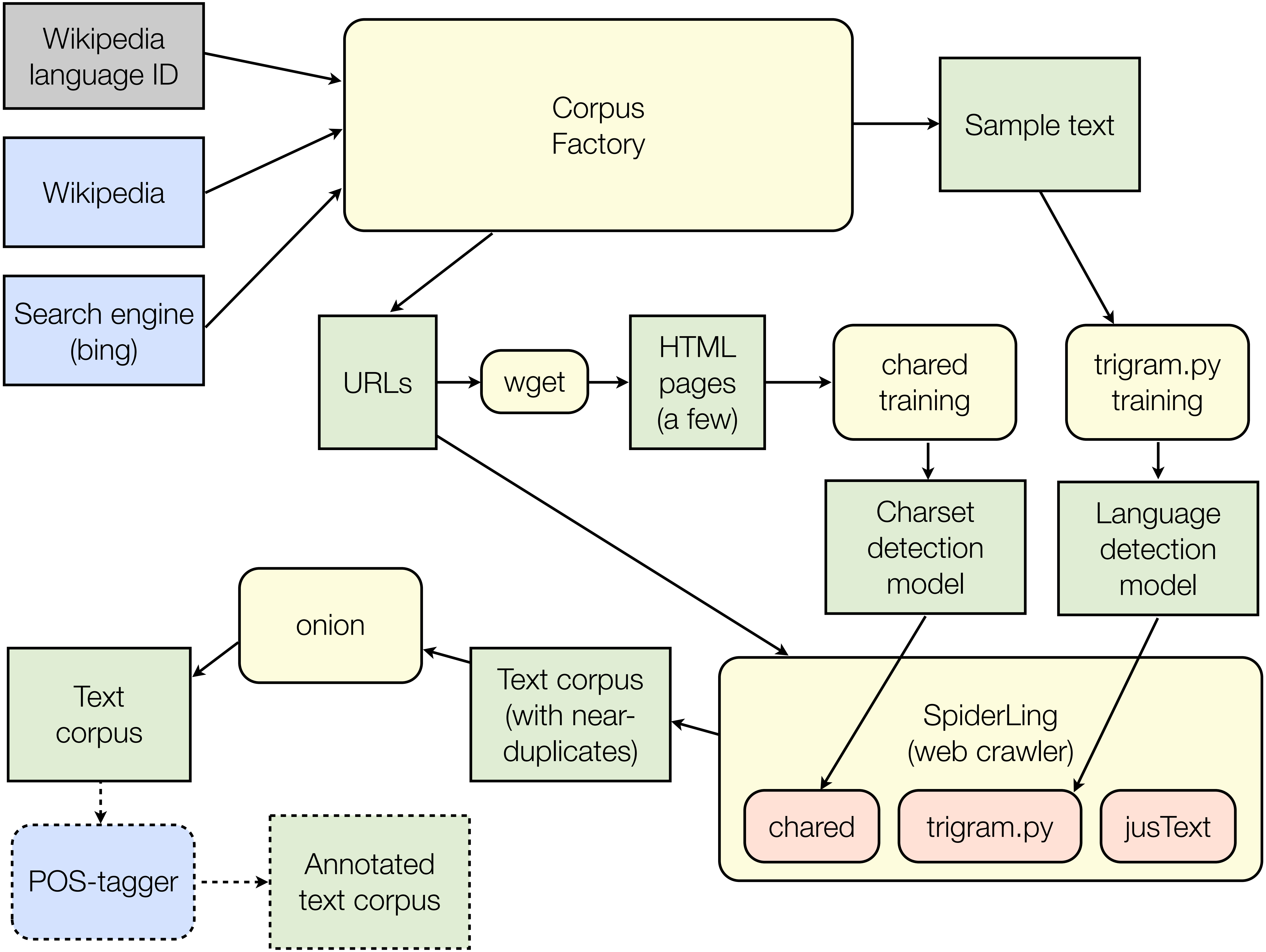
Text corpus (with near-duplicates)

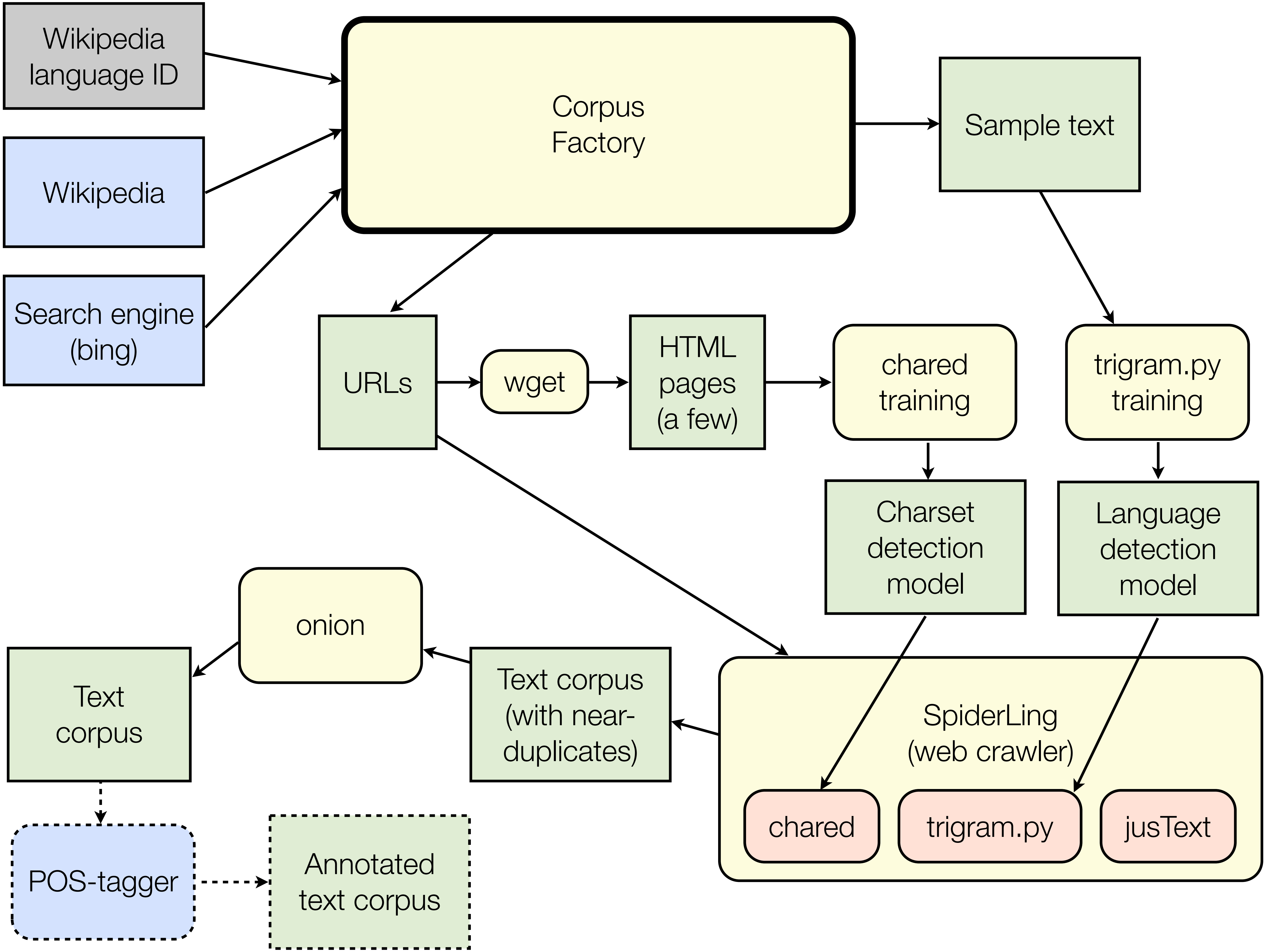
SpiderLing (web crawler)

chared

trigram.py

jusText


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
chared training
trigram.py training
Charset detection model
Language detection model
onion
Text corpus
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText
POS-tagger
Annotated text corpus



Wikipedia language ID

Wikipedia

Search engine (bing)

Corpus Factory

Sample text

URLs

wget

HTML pages (a few)

chared training

trigram.py training

Charset detection model

Language detection model

onion

Text corpus

Text corpus (with near-duplicates)

SpiderLing (web crawler)

chared

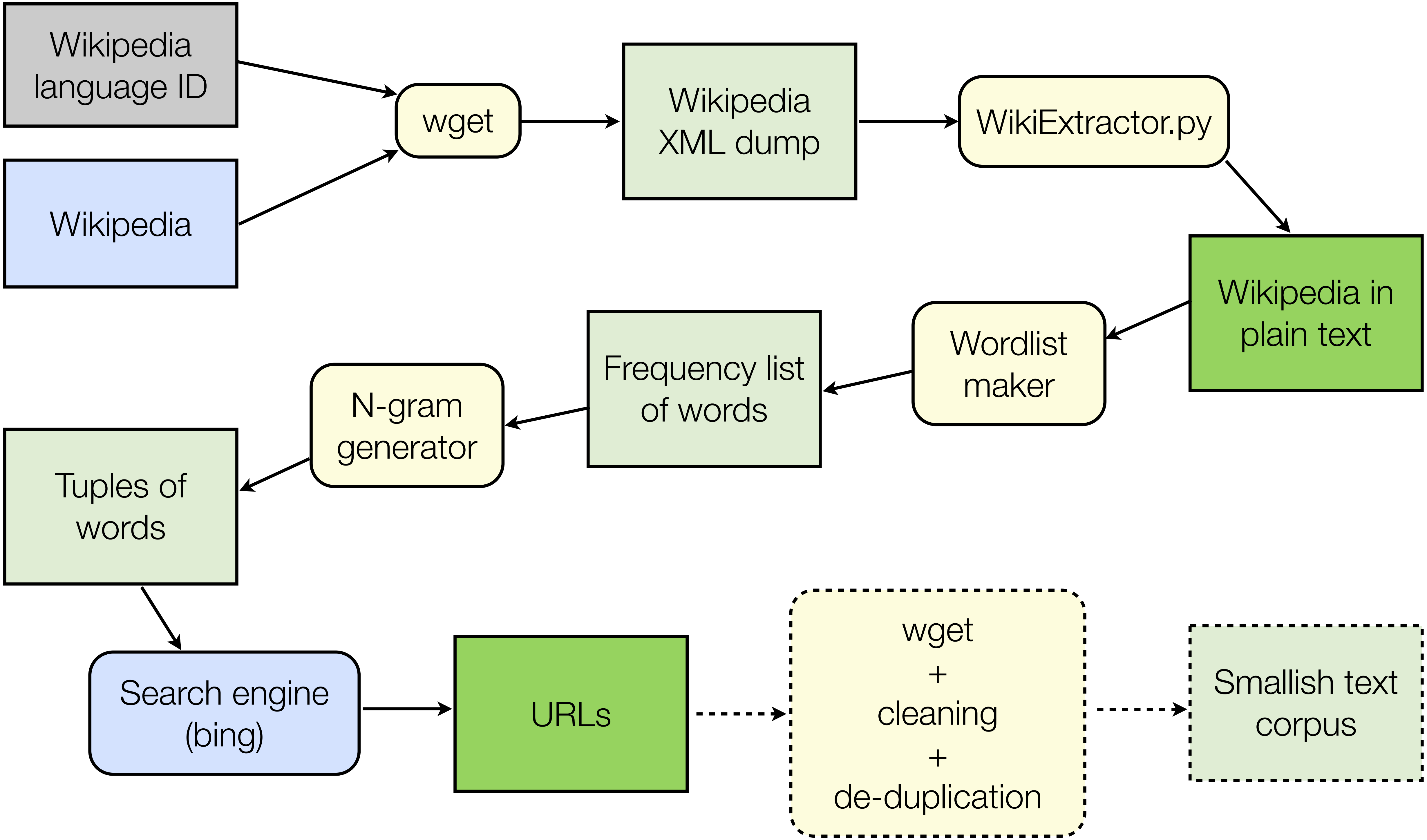
trigram.py

jusText

POS-tagger

Annotated text corpus

Corpus Factory



Wikipedia language ID

Wikipedia

wget

Wikipedia XML dump

WikiExtractor.py

Wikipedia in plain text

Wordlist maker

Frequency list of words

N-gram generator

Tuples of words

Search engine (bing)

URLs

wget + cleaning + de-duplication

Smallish text corpus

1. WIKIPEDIA XML DUMP

```
  <page>
    <title>Astronomie</title>
    <id>10</id>
    <revision>
      <id>5866929</id>
      <timestamp>2010-09-22T23:08:37Z</timestamp>
      <contributor>
        <username>ArthurBot</username>
        <id>34408</id>
      </contributor>
      <minor />
      <comment>Bot: [[en:Astronomy]] is a good article</comment>
       <text xml:space="preserve">[[Soubor:USA.NM.VeryLargeArray.02.jpg|thumb|Mezi zařízení, která se používají k astronomickým pozorováním, patří i [[radioteleskop]]y.]]

'''Astronomie''', [[řečtina|řecky]] αστρονομία z άστρον (astron) hvězda a νόμος (nomos) zákon, [[čeština|česky]] též '''hvězdářství''', je [[věda]], která se zabývá jevy za hranicemi [[Atmosféra Země|zemské atmosféry]]. Zvláště tedy výzkumem [[vesmír|vesmírných]] [[těleso|těles]], jejich soustav, různých dějů ve vesmíru i vesmírem jako celkem.

== Historie astronomie ==
Astronomie se podobně jako další vědy začala rozvíjet ve [[starověk]]u. První se z astronomie rozvíjela [[astrometrie]], zabývající se měřením [[poloha|poloh]] [[hvězda|hvězd]] a [[planeta|planet]] na obloze. Tato oblast astronomie měla velký význam pro [[navigace|navigaci]]. Podstatnou částí astrometrie je [[sférická astronomie]] sloužící k popisu poloh objektů na [[nebeská sféra|nebeské sféře]], zavádí [[souřadnice]] a popisuje významné [[křivka|křivky]] a [[bod]]y na nebeské sféře. Pojmy ze sférické astronomie se také používají při [[měření času]].

Další oblastí astronomie, která se rozvinula, byla [[nebeská mechanika]]. Zabývá se [[Mechanický pohyb|pohybem]] [[těleso|těles]] v [[gravitační pole|gravitačním poli]], například [[planeta|planet]] ve [[Sluneční soustava|sluneční soustavě]]. Základem nebeské mechaniky jsou práce [[Johannes Kepler|Keplera]] a [[Isaac Newton|Newtona]].
```

2. WIKIPEDIA IN PLAIN TEXT

Astronomie.

Astronomie, řecky αστρονομία z άστρον (astron) hvězda a νόμος (nomos) zákon, česky též hvězdářství, je věda, která se zabývá jevy za hranicemi zemské atmosféry. Zvláště tedy výzkumem vesmírných těles, jejich soustav, různých dějů ve vesmíru i vesmírem jako celkem.

Historie astronomie.

Astronomie se podobně jako další vědy začala rozvíjet ve starověku. První se z astronomie rozvíjela astrometrie, zabývající se měřením poloh hvězd a planet na obloze. Tato oblast astronomie měla velký význam pro navigaci. Podstatnou částí astrometrie je sférická astronomie sloužící k popisu poloh objektů na nebeské sféře, zavádí souřadnice a popisuje významné křivky a body na nebeské sféře. Pojmy ze sférické astronomie se také používají při měření času.

Další oblastí astronomie, která se rozvinula, byla nebeská mechanika. Zabývá se pohybem těles v gravitačním poli, například planet ve sluneční soustavě. Základem nebeské mechaniky jsou práce Keplera a Newtona.

Aristotelés ve svém díle "O nebi" z roku 340 př. n. l. dokázal, že tvar Země musí být kulatý, jelikož stín Země na Měsíci je při zatmění vždy kulatý, což by při plochém tvaru Země nebylo možné. Řekové také zjistili, že pokud sledujeme Polárku z jižnějšího místa na Zemi, jeví se nám níže nad obzorem než pro pozorovatele ze severu, kterému se bude její poloha na obloze jevit výše. Aristotelés dále určil poloměr Země, který ale odhadl na dvojnásobek skutečného poloměru. V aristotelovském modelu Země stojí a Měsíc se Sluncem a hvězdami krouží kolem ní, a to po kruhových drahách.

Myšlenky Aristotelovy rozvinul ve 2. století našeho letopočtu Ptolemaios, který také stavěl Zemi do středu a další objekty nechal obíhat kolem ní ve sférách, první byla sféra Měsíce, dále sféry Merkuru, Venuše, Slunce, Marsu, Jupitera, Saturna a sféra stálic (hvězd, jež byly považovány za nehybné, jak to plyne z názvu, měly se pohybovat jen společně s oblohou). Tento model poměrně vyhovoval polohám těles na obloze.

Roku 1514 navrhl Mikuláš Koperník nový model, ve kterém bylo ve středu soustavy Slunce a planety obíhaly kolem něj po kruhových drahách, setkal se ale s problémy při pozorováních, objekty se nenacházely na správných souřadnicích.

Roku 1609 zkonstruoval Galileo Galilei dalekohled, s jehož pomocí objevil čtyři měsíce obíhající kolem planety Jupiter, a tím dokázal Koperníkovu teorii o Slunci ve středu a planetách kroužících kolem. Johannes Kepler zaměnil kruhové dráhy planet za eliptické, čímž bylo dosaženo souladu s pozorovanými polohami těles.

V roce 1687 vydal sir Isaac Newton knihu Philosophiae Naturalis Principia Mathematica o poloze těses v prostoru a čase a zákon obecné přitažlivosti, podle něhož jsou k sobě tělesa vázana gravitací, která závisí na hmotnosti těles a na jejich vzdálenosti. Z gravitačního zákona vychází eliptický pohyb planet.

Roku 1929 studoval Edwin Hubble daleké galaxie, zjistil rudý posuv, který se zvětšuje se vzdáleností, to byl důkaz o rozpínání vesmíru. Fakt, že se od sebe objekty vzdalují, naznačuje, že někdy v minulosti byly objekty velmi blízko od sebe, tím se zrodily myšlenky o velkém třesku, místě a čase, kdy byl vesmír nekonečně malý a hustý.

V letech 1905–1915 napsal Albert Einstein teorii relativity – speciální, ve které zavedl konečnou rychlost světla a obecnou relativitu o gravitaci, čase a prostoru ve velkých rozměrech. Na začátku 20. století vznikla kvantová teorie o chování elementárních částic.

3. Frequency list of words

1	a	1188065
2	v	907365
3	se	751213
4	na	578619
5	je	502439
6	z	288445
7	s	256238
8	do	248581
9	V	222782
10	byl	198371
11	roce	181944
12	ve	177070
13	i	170934
14	jako	165863
15	o	165078
...		

1000	informace	3238
1001	státy	3233
1002	vzhledem	3229
1003	minulosti	3218
1004	největším	3215
1005	vychází	3213
1006	podobně	3213
1007	řešení	3206
...		
5993	tabulky	649
5994	Bill	649
5995	živočichů	649
5996	vrcholy	649
5997	(za	649
5998	farnosti	649
5999	vítězem	649

4. TUPLES OF WORDS

```
místem určené veřejné vojáci
Enterprise provozu. přinesla teprve
hodnocení kvalitní považováno vystupoval
Nejstarší hlavu pohlaví ukončení
francouzskou procesu scény snadno
nearabské náboženství). pravý přechodu
Francii mužů náměstí. závody
Oblast doprava. proběhl zahrál
Alois kříže příběhu ruce
konstrukce letounu rekonstrukce rekord
1907 Vznik povýšen zemí.
dřívější miliónů nepříliš výšky
1902 I., verzí členové
hvězdná studia vyslal °C
budovu grafické kolonie počasí
```

5. URLs

místem určené veřejné vojáci

http://sporkova.spytech.cz/ApDownload.php?id=10318

http://bunkrr.blog.cz/0807

http://www.nacesty.cz/last-minute/ZIEL=SSA&FFT=1

http://vrchovsky.blog.cz/

http://www.servingthenations.org/article.asp?ArticleID=152

http://druidova.mysteria.cz/

http://radoskydancer.wordpress.com/

http://nwo.corpx.eu/

http://www.aifp.cz/cz/clanky.php?kat=10

...

Enterprise provozu. přinesla teprve

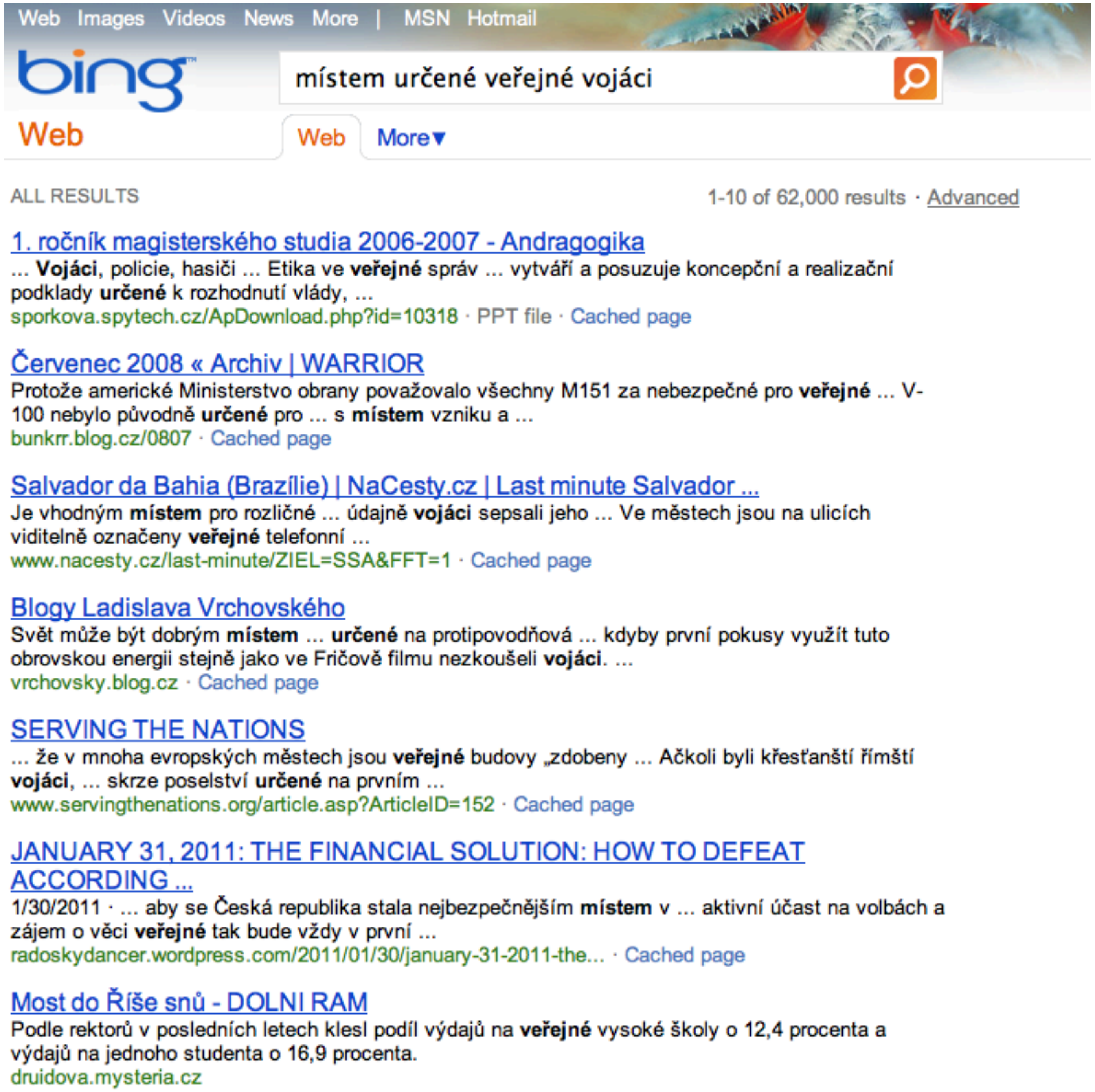
http://alave.cz/sitove-prvky-bezdratove:c:668

http://www.cs.hukol.net/themenreihe.p?c=Firmy

http://www.cs.hukol.net/themenreihe.p?c=Syst%C3%A9mov%C3%BD%20software

http://www.root.cz/clanky/stalo-se-tyden-13-04/

...



místem určené veřejné vojáci

ALL RESULTS

1-10 of 62,000 results · Advanced

1. ročník magisterského studia 2006-2007 - Andragogika
... **Vojáci**, policie, hasiči ... Etika ve **veřejné** správ ... vytváří a posuzuje koncepční a realizační podklady **určené** k rozhodnutí vlády, ...
sporkova.spytech.cz/ApDownload.php?id=10318 · PPT file · Cached page

Červenec 2008 « Archiv | WARRIOR
Protože americké Ministerstvo obrany považovalo všechny M151 za nebezpečné pro **veřejné** ... V-100 nebylo původně **určené** pro ... s **místem** vzniku a ...
bunkrr.blog.cz/0807 · Cached page

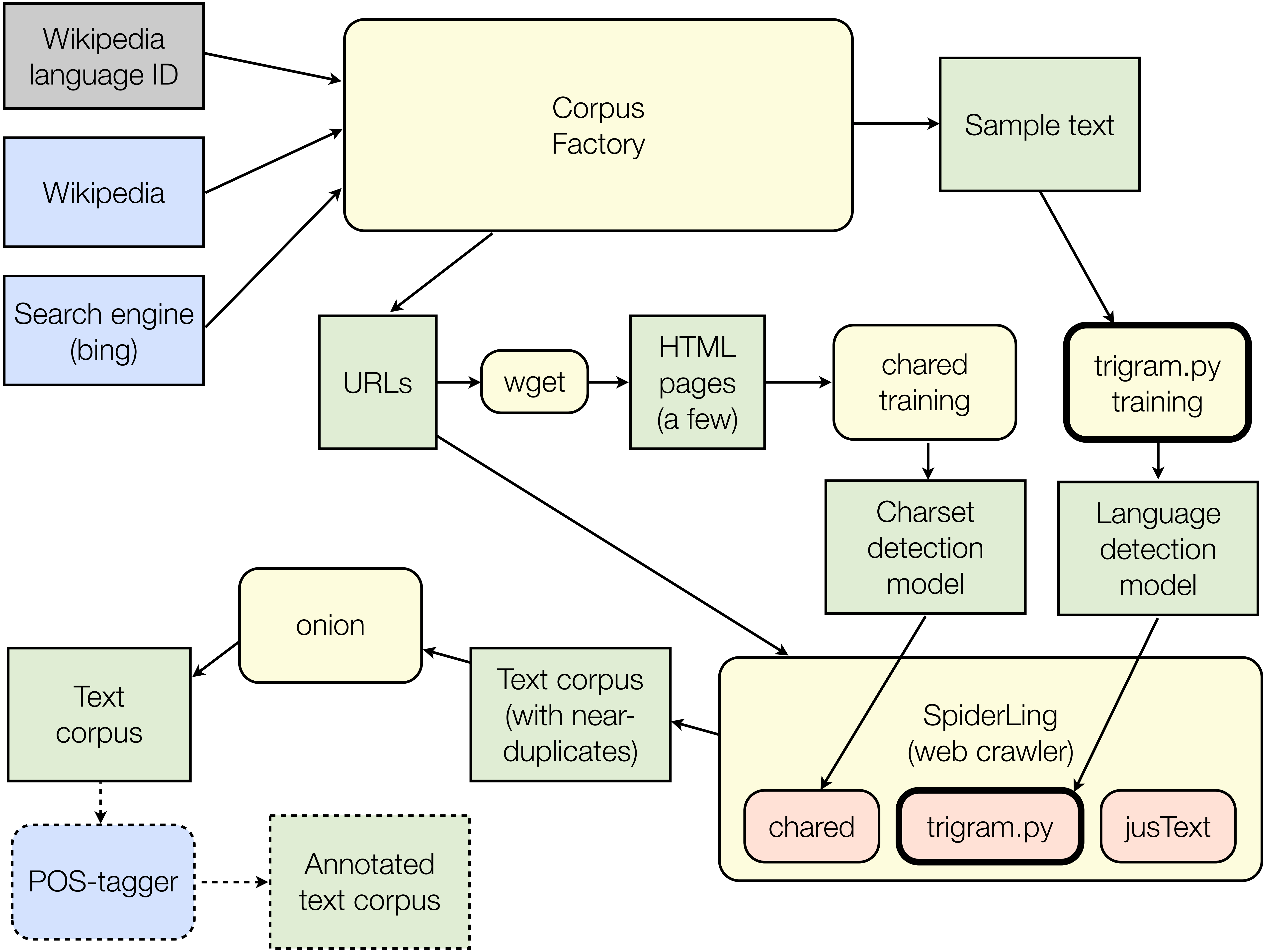
Salvador da Bahia (Brazílie) | NaCesty.cz | Last minute Salvador ...
Je vhodným **místem** pro rozličné ... údajně **vojáci** sepsali jeho ... Ve městech jsou na ulicích viditelně označeny **veřejné** telefonní ...
www.nacesty.cz/last-minute/ZIEL=SSA&FFT=1 · Cached page

Blogy Ladislava Vrchovského
Svět může být dobrým **místem** ... **určené** na protipovodňová ... kdyby první pokusy využít tuto obrovskou energii stejně jako ve Fričově filmu nezkoušeli **vojáci**. ...
vrchovsky.blog.cz · Cached page

SERVING THE NATIONS
... že v mnoha evropských městech jsou **veřejné** budovy „zdobeny ... Ačkoli byli křesťanští římští **vojáci**, ... skrze poselství **určené** na prvním ...
www.servingthenations.org/article.asp?ArticleID=152 · Cached page

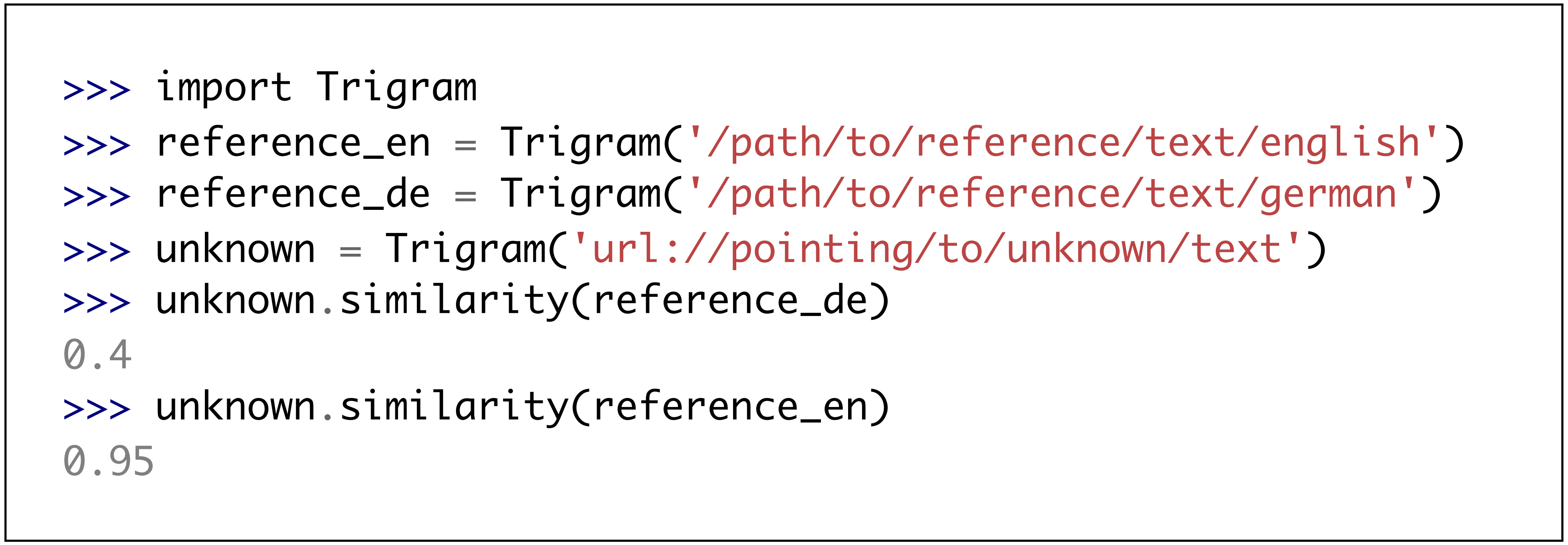
JANUARY 31, 2011: THE FINANCIAL SOLUTION: HOW TO DEFEAT ACCORDING ...
1/30/2011 · ... aby se Česká republika stala nejbezpečnějším **místem** v ... aktivní účast na volbách a zájem o věci **veřejné** tak bude vždy v první ...
radoskydancer.wordpress.com/2011/01/30/january-31-2011-the... · Cached page

Most do Říše snů - DOLNI RAM
Podle rektorů v posledních letech klesl podíl výdajů na **veřejné** vysoké školy o 12,4 procenta a výdajů na jednoho studenta o 16,9 procenta.
druidova.mysteria.cz

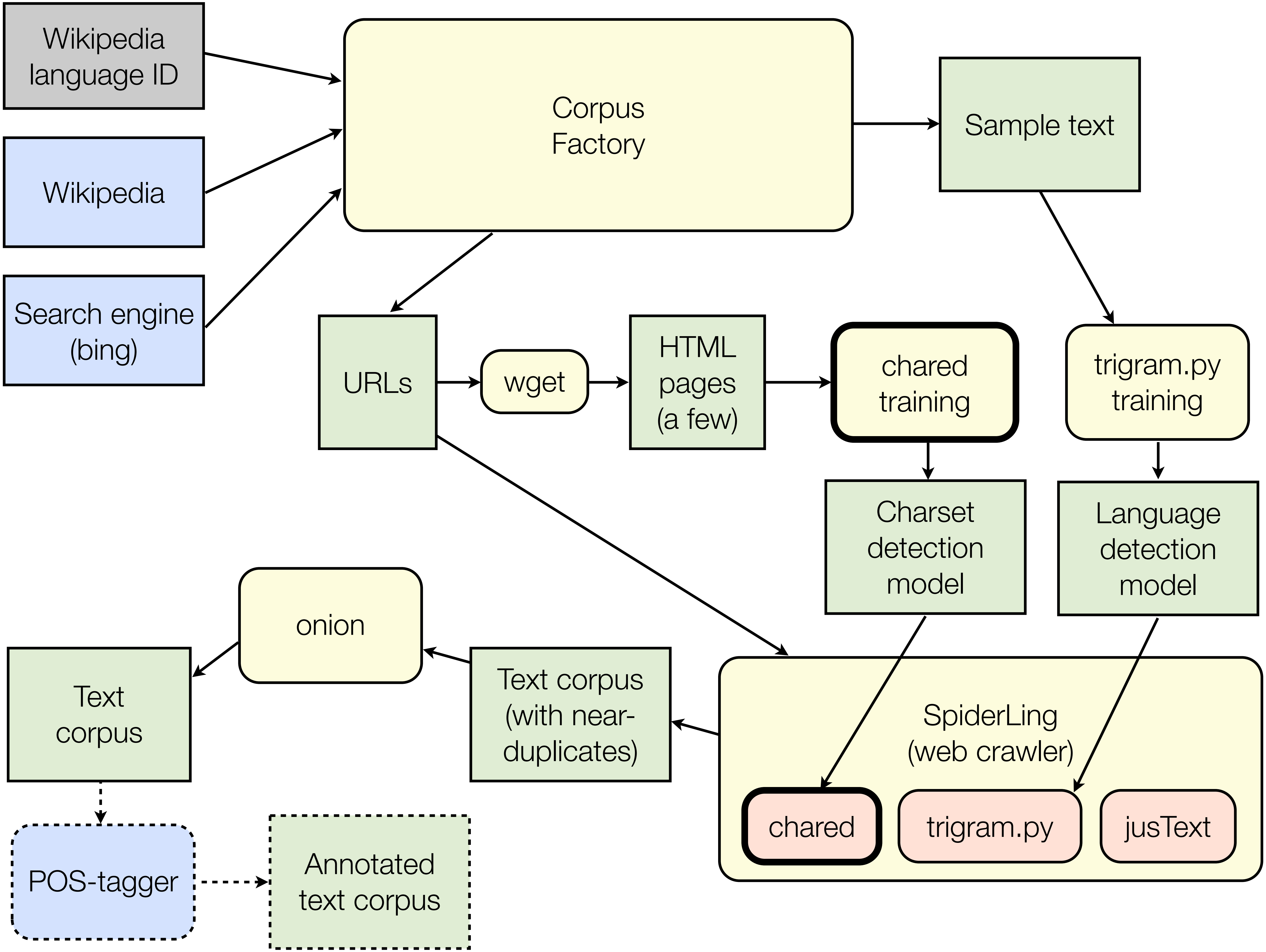

Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
chared training
trigram.py training
Charset detection model
Language detection model
onion
Text corpus
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText
POS-tagger
Annotated text corpus

LANGUAGE DETECTION / FILTERING

- Trigram class
  - http://code.activestate.com/recipes/326576-language-detection-using-character-trigrams/
- Similarity score based on frequencies of 3-grams of characters



```
>>> import Trigram
>>> reference_en = Trigram('/path/to/reference/text/english')
>>> reference_de = Trigram('/path/to/reference/text/german')
>>> unknown = Trigram('url://pointing/to/unknown/text')
>>> unknown.similarity(reference_de)
0.4
>>> unknown.similarity(reference_en)
0.95
```


Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
chared training
trigram.py training
Charset detection model
Language detection model
onion
Text corpus
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText
POS-tagger
Annotated text corpus

Character encoding detection

- Bytes
  - 70 c5 99 c3 ad 6c 69 c5 a1 20 c5 be 6c 75 c5 a5 6f 75 c4 8d 6b c3 bd 20 6b c5 af c5 88 20 70 c4 9b 6c 20 c4 8f c3 a1 62 65 6c 73 6b c3 a9 20 c3 b3 64 79
- In windows-1250
  - pĹ™Ă-liĹˇ ĹľluĹĄouÄŤkĂ˝ kĹŻĹ ĂşpÄ›l ÄŹĂˇbelskĂ© Ăłdy
- In iso-8859-2
  - pĹĂ-liĹĄ ĹžluĹĽouÄkĂ˝ kĹŻĹ ĂşpÄl ÄĂĄbelskĂŠ Ăłdy
- In utf-8
  - příliš žluťoučký kůň úpěl ďábelské ódy

WEB PAGE ENCODING SPECIFICATION

- Meta tags

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
```

- HTTP protocol

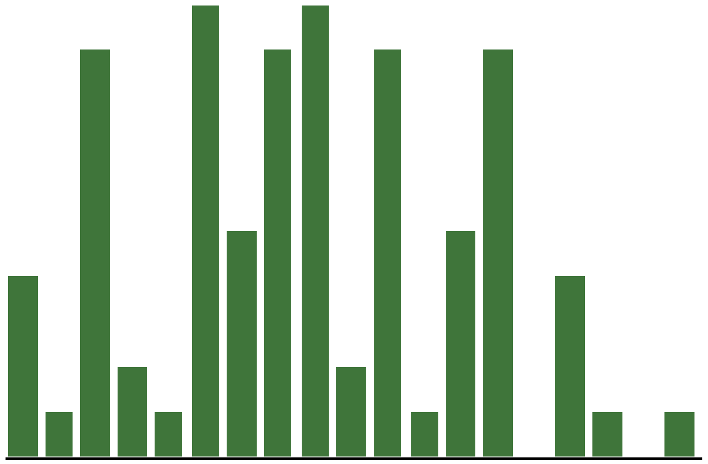
```
200 OK

Content-Type: text/html; charset=UTF-8
```

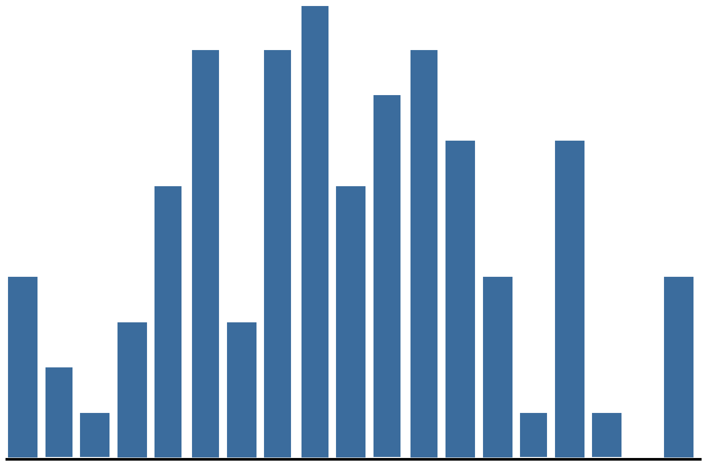
- Not always available
- Not always correct
- Guessing from text is more reliable

AUTOMATIC ENCODING DETECTION

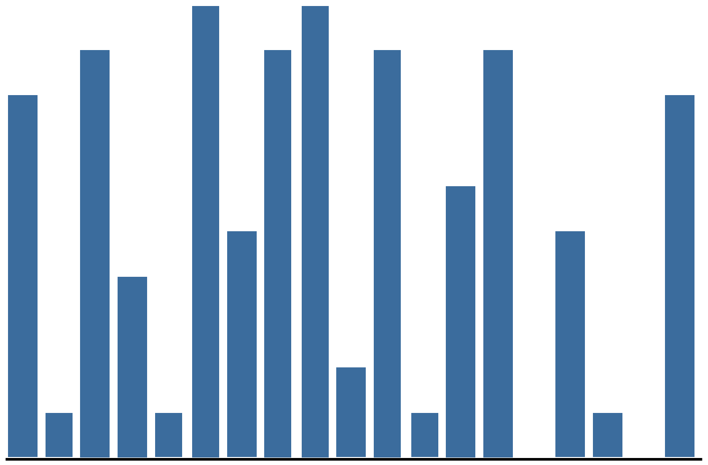
- Byte frequency vector for the input text
  - 3-grams of bytes



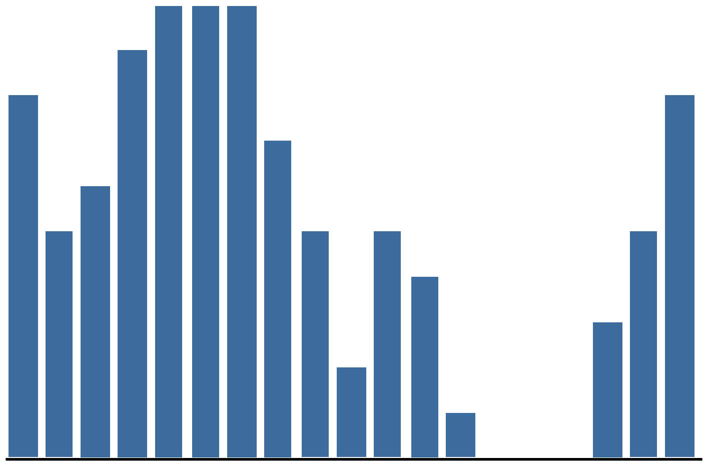
- Compare with model vectors (scalar product)



iso-8859-1



koi8-r



utf-8

TRAINING DATA

- Take ca. 1000 web pages with texts in the target language (Corpus Factory)
- Extract encoding information from meta tags
  - Mostly correct; errors “cancelled out” by statistical processing
  - Discard pages for which encoding cannot be determined
- Usage frequency of encodings for the target language
  - E.g. for Czech: **utf-8**: 60.2%, **windows-1250**: 32.2%, **iso-8859-2**: 6.0%
  - Ignore encodings with freq < 0.5%
- Convert all pages to all frequently used encodings
  - To balance training data
- Create models

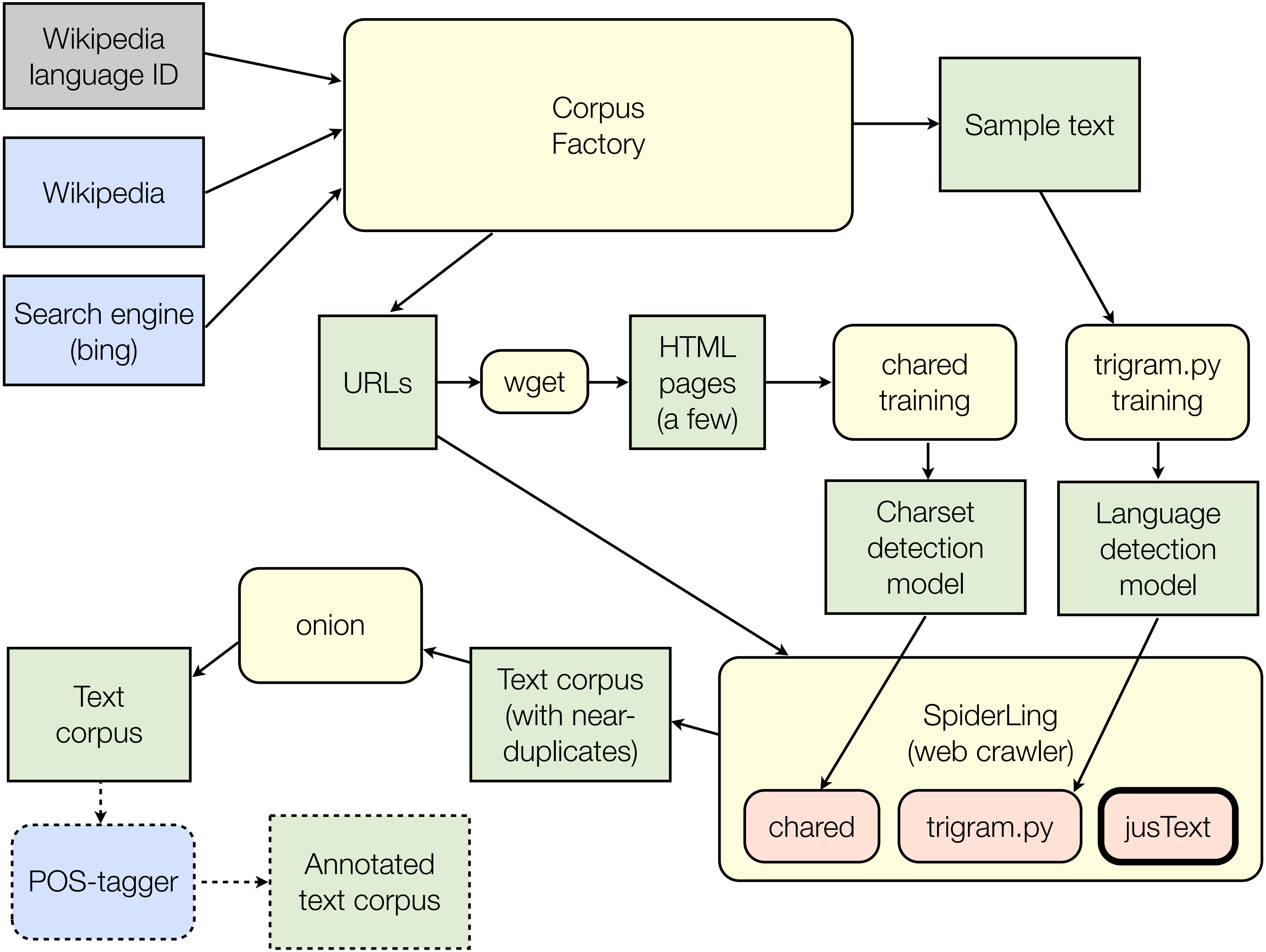
EVALUATION

- 5-fold cross-validation on training data

	Czech		English		German		Greek		Italian		Norwegian	
	freq	accuracy	freq	accuracy	freq	accuracy	freq	accuracy	freq	accuracy	freq	accuracy
utf-8	60,2%	100,0%	56,9%	95,8%	54,6%	100,0%	68,5%	100,0%	54,2%	100,0%	63,0%	100,0%
windows-1250	32,2%	100,0%	0,3%	n/a	0,1%	n/a	0,2%	n/a	0,0%	n/a	0,1%	n/a
windows-1252	0,4%	n/a	9,4%	97,5%	6,5%	97,3%	3,1%	75,8%	7,1%	95,7%	7,0%	97,4%
windows-1253	0,0%	n/a	0,0%	n/a	0,0%	n/a	14,3%	99,3%	0,0%	n/a	0,0%	n/a
iso-8859-1	1,0%	89,5%	32,8%	90,9%	37,1%	85,8%	1,7%	71,2%	37,9%	85,1%	29,3%	88,2%
iso-8859-2	6,0%	99,6%	0,0%	n/a	0,1%	n/a	0,0%	n/a	0,1%	n/a	0,1%	n/a
iso-8859-7	0,0%	n/a	0,0%	n/a	0,0%	n/a	12,0%	97,2%	0,0%	n/a	0,0%	n/a
iso-8859-15	0,0%	n/a	0,0%	n/a	1,2%	85,6%	0,0%	n/a	0,0%	n/a	0,4%	n/a
training docs	801		668		773		879		771		740	
w. avg accuracy	99,2%		93,5%		93,7%		97,9%		93,3%		95,7%	

IMPLEMENTATION

- In Python
- Open source (BSD License)
  - Available from: http://code.google.com/p/chared/
  - Online demo: http://nlp.fi.muni.cz/projects/chared/
- Currently supports 51 languages
  - http://code.google.com/p/chared/source/browse/#svn%2Ftrunk%2Fchared%2Fmodels



Wikipedia language ID

Wikipedia

Search engine (bing)

Corpus Factory

Sample text

URLs

wget

HTML pages (a few)

chared training

trigram.py training

Charset detection model

Language detection model

onion

Text corpus

Text corpus (with near-duplicates)

SpiderLing (web crawler)

chared

trigram.py

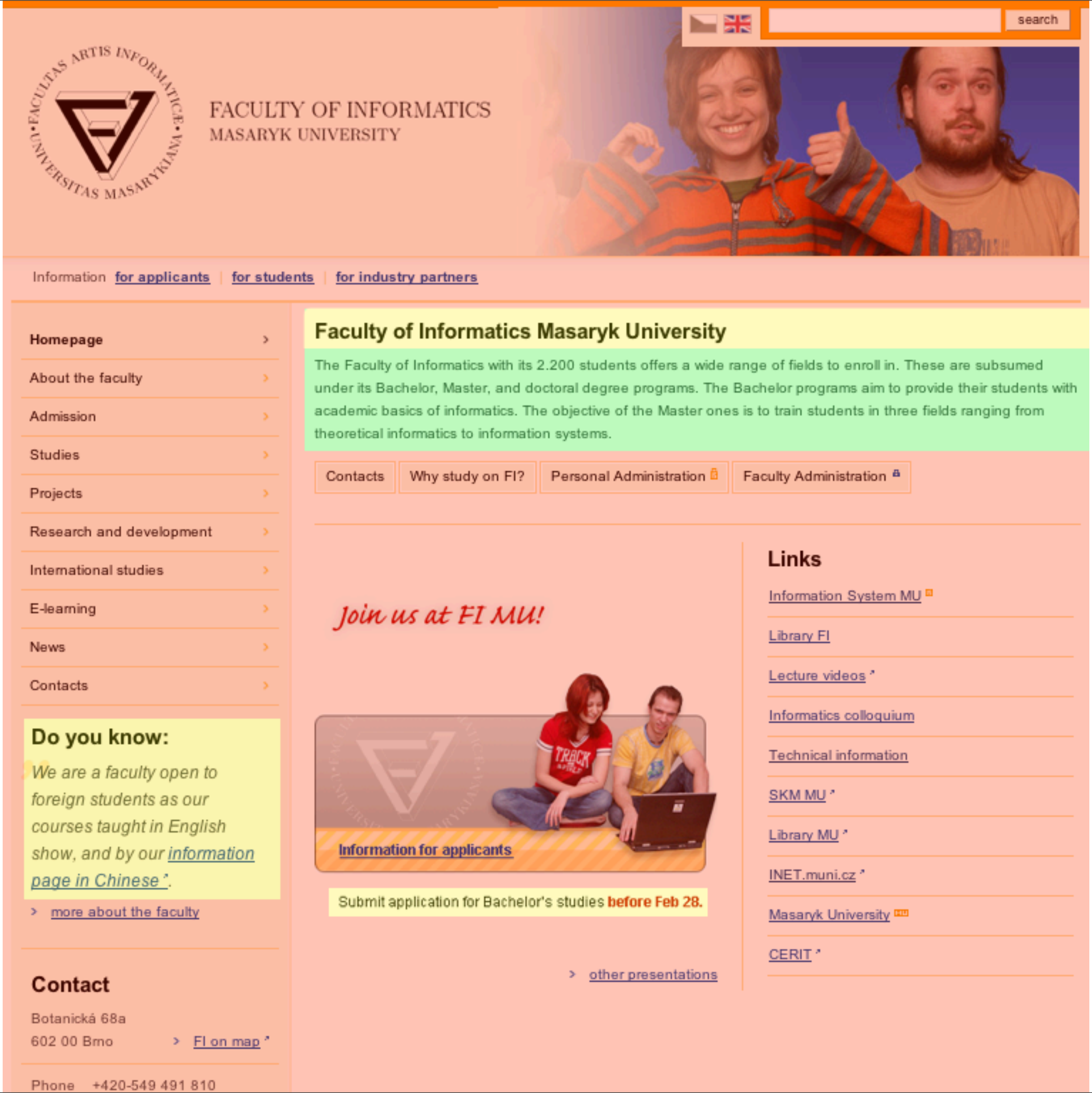
jusText

POS-tagger

Annotated text corpus

BOILERPLATE

- The content outside of the main body of a page, e.g.
  - Headers, footers
  - Navigation links
  - Copyright notices
  - Advertisements
- Does not contain full sentences
- Mostly noise (for text corpora)
- Inflates frequency of some terms, such as *home, search, print*



FACULTY OF INFORMATICS
MASARYK UNIVERSITY

Information for applicants | for students | for industry partners

- Homepage ›
- About the faculty ›
- Admission ›
- Studies ›
- Projects ›
- Research and development ›
- International studies ›
- E-learning ›
- News ›
- Contacts ›

Faculty of Informatics Masaryk University

The Faculty of Informatics with its 2.200 students offers a wide range of fields to enroll in. These are subsumed under its Bachelor, Master, and doctoral degree programs. The Bachelor programs aim to provide their students with academic basics of informatics. The objective of the Master ones is to train students in three fields ranging from theoretical informatics to information systems.

Contacts | Why study on FI? | Personal Administration | Faculty Administration

Join us at FI MU!

Information for applicants

Submit application for Bachelor's studies **before Feb 28.**

› other presentations

Links

- Information System MU
- Library FI
- Lecture videos
- Informatics colloquium
- Technical information
- SKM MU
- Library MU
- INET.muni.cz
- Masaryk University
- CERIT

Do you know:

We are a faculty open to foreign students as our courses taught in English show, and by our information page in Chinese.

› more about the faculty

Contact

Botanická 68a
602 00 Brno › FI on map

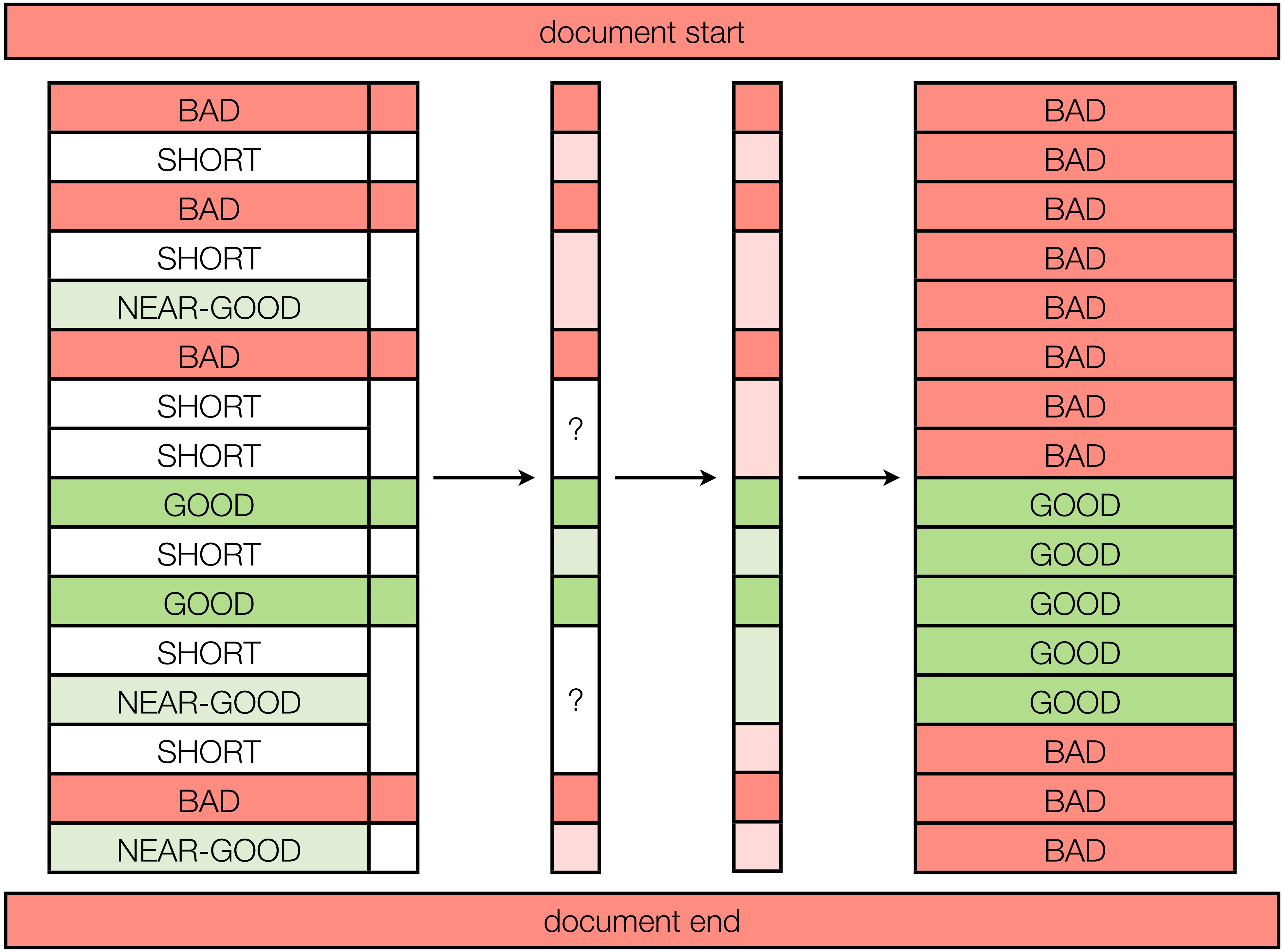
Phone +420-549 491 810

JUSTEXT

- Boilerplate cleaning algorithm
- Operates in 3 basic steps:

1. Segmentation - a page is split into text blocks (segments)
2. Context-free classification - preliminary class assigned to each segment (**good**, **bad**, **near-good**, **short**)
    - Length
    - Number of hyperlinks
    - Number of function (stoplist) words
3. Context-sensitive classification - final class assigned (**good**, **bad**)

JUSTEXT: TEXT BLOCK CLASSIFICATION



document start

BAD	BAD
SHORT	BAD
BAD	BAD
SHORT	BAD
NEAR-GOOD	BAD
BAD	BAD
SHORT	BAD
SHORT	BAD
GOOD	GOOD
SHORT	GOOD
GOOD	GOOD
SHORT	GOOD
NEAR-GOOD	GOOD
SHORT	BAD
BAD	BAD
NEAR-GOOD	BAD

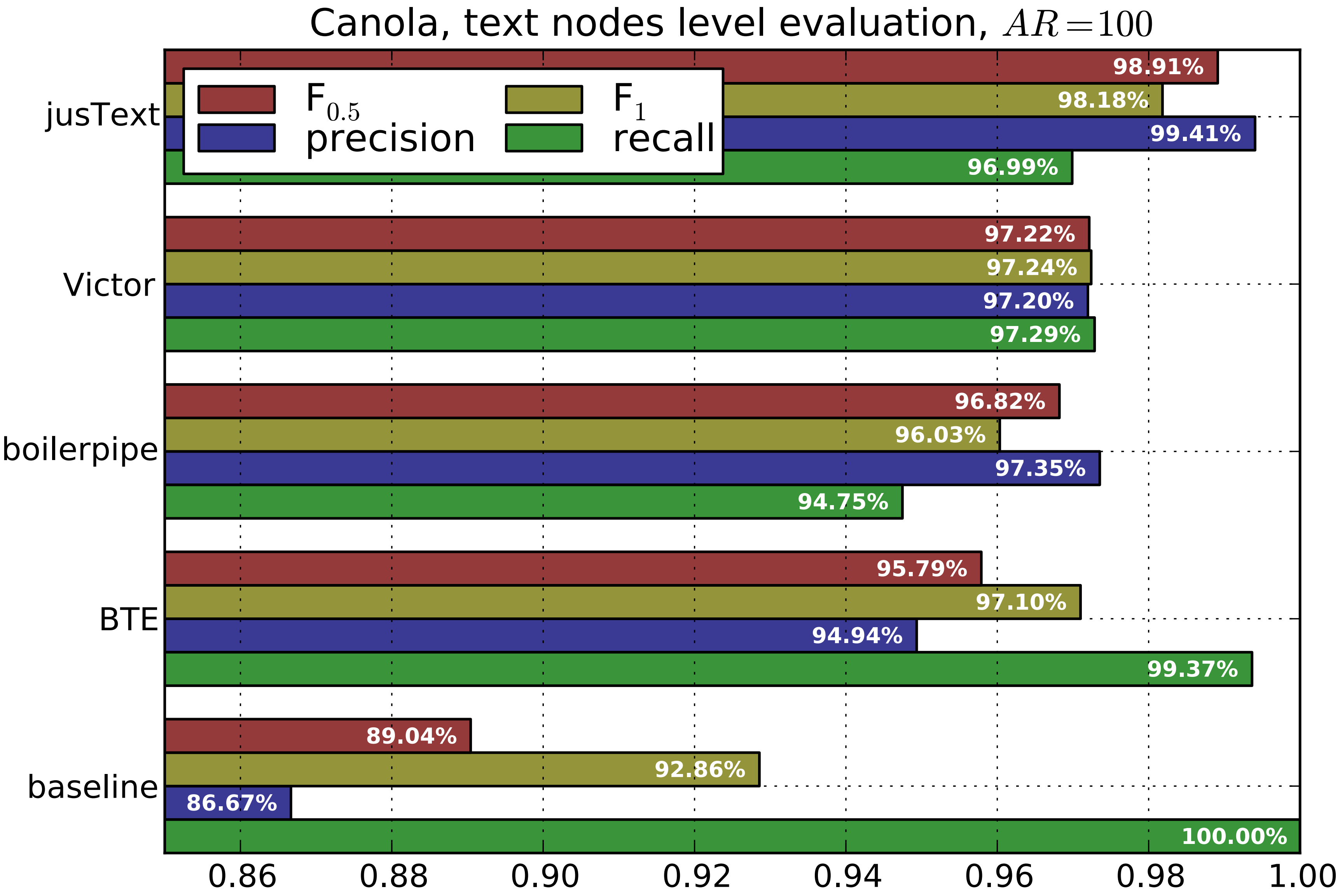
document end

JUSTEXT EVALUATION

- Data collections
  - Canola (KrdWrd team, http://krdwrd.org/)
  - CleanEval
  - L3S-GN1 (C. Kohlschütter, news articles)
- Algorithms
  - Victor (CRF classifier; Marek, Pecina, Spousta; CleanEval winner)
  - NCLEANER / StupidOS (n-gram language model; S. Evert)
  - boilerpipe (C4.8-based decision trees; C. Kohlschütter)
  - BTE (tag density; Finn et al; better than Victor on CleanEval)

JUSTEXT: EVALUATION RESULTS

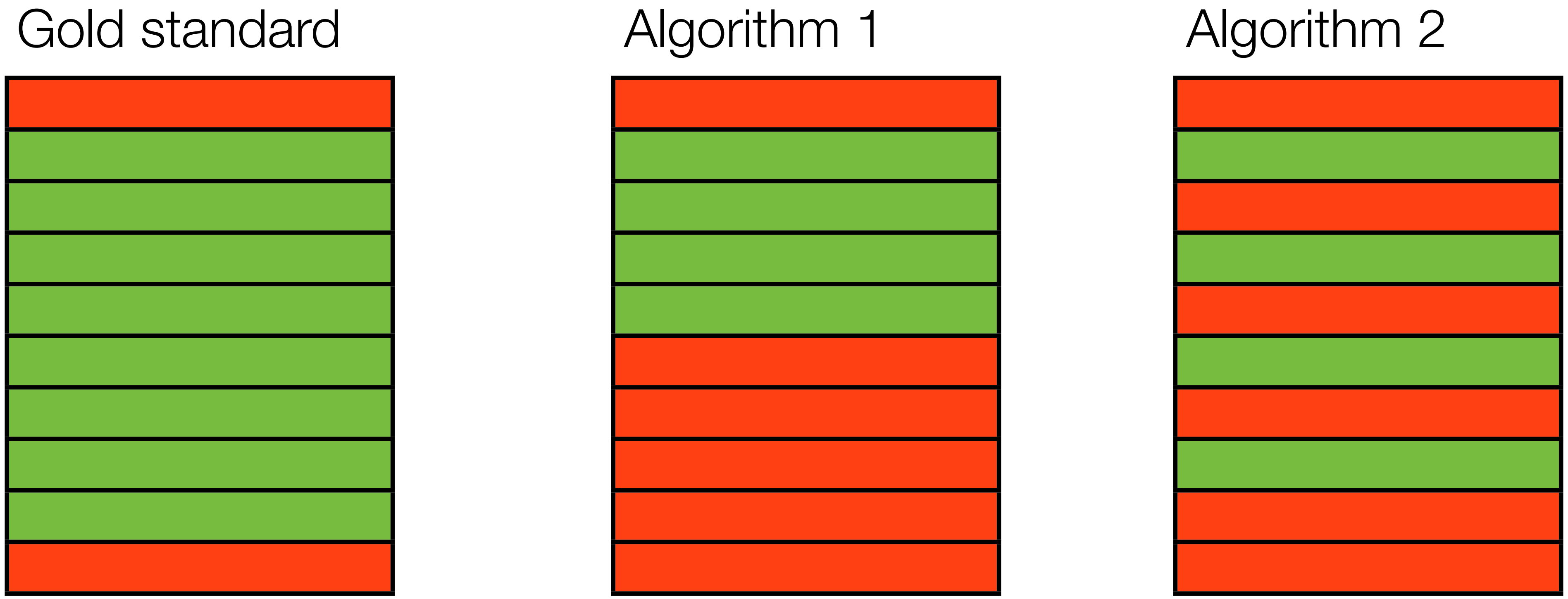
- On all collections on par with the best algorithms or even slightly better



DOCUMENT FRAGMENTATION

- Boilerplate cleaning algorithms with too fine-grained segmentation tend to have problems with fragmented output
- Example - Victor:

1. *A few days ago, I mentioned that I'd begun playing on one of the older text- based MMOGs, Gemstone IV. Many of you*
2. *commented on the Loading forums*
3. *about your own old experiences with text and how they compared with my own. ...*



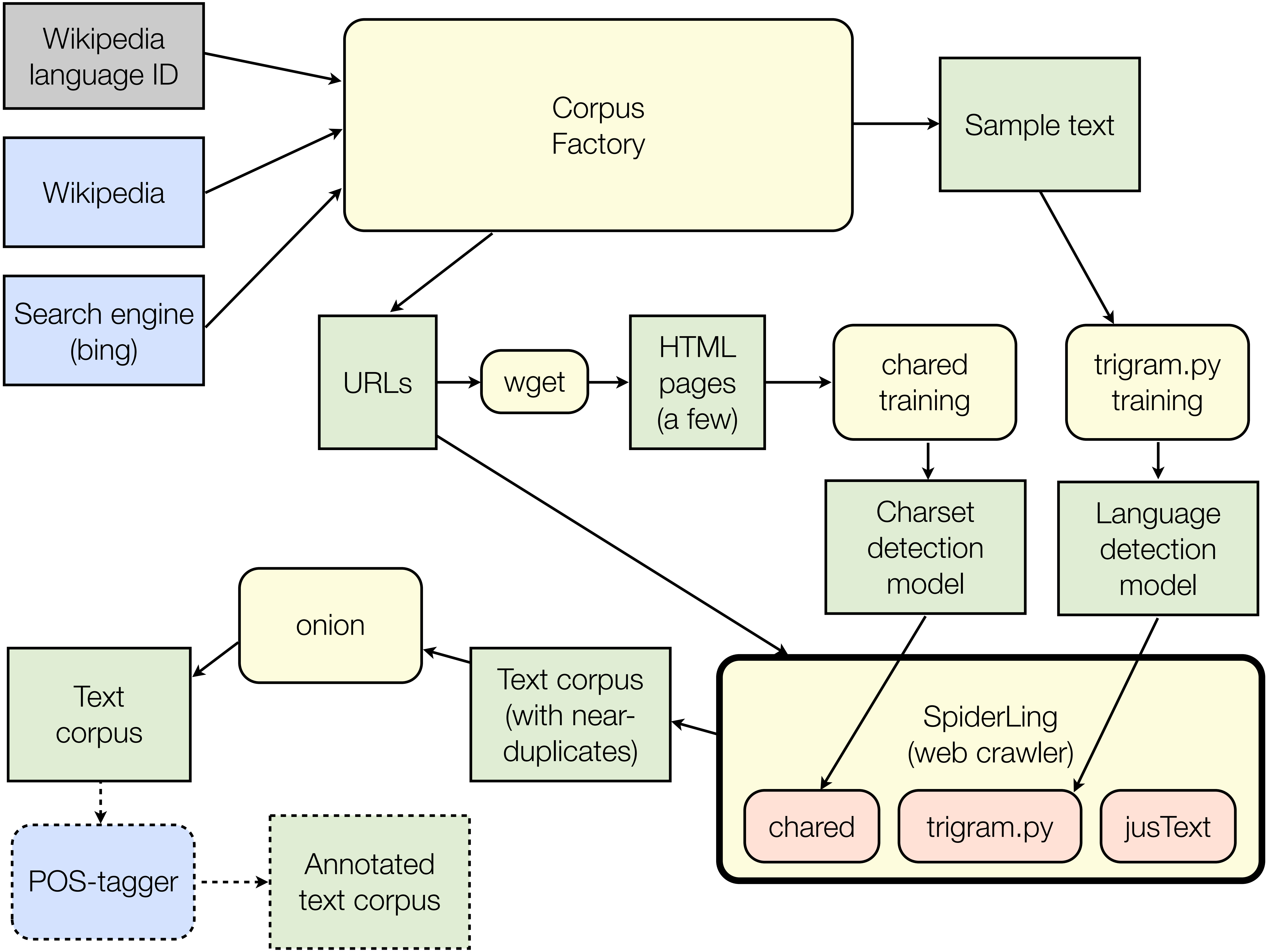
Gold standard	Algorithm 1	Algorithm 2
🟥	🟥	🟥
🟩	🟩	🟩
🟩	🟩	🟥
🟩	🟩	🟩
🟩	🟩	🟥
🟩	🟥	🟩
🟩	🟥	🟥
🟩	🟥	🟩
🟩	🟥	🟥
🟥	🟥	🟥

Evaluation of fragmentation on Canola

	avg fragment length	median fragment length	avg fragments per document
perfect cleaning	1315,1	279	6,98
BTE	11095,6	7611	1,00
boilerpipe	671,2	68	13,81
Victor	637,9	126	14,13
jusText	2304,3	794	3,88

AVAILABILITY OF JUSTEXT

- In Python
- Open source (BSD License)
- http://code.google.com/p/justext/
- 563 downloads since March 2011
- Online demo: http://nlp.fi.muni.cz/projects/justext/


Wikipedia language ID

Wikipedia

Search engine (bing)

Corpus Factory

Sample text

URLs

wget

HTML pages (a few)

chared training

trigram.py training

Charset detection model

Language detection model

onion

Text corpus

Text corpus (with near-duplicates)

SpiderLing (web crawler)

chared

trigram.py

jusText

POS-tagger

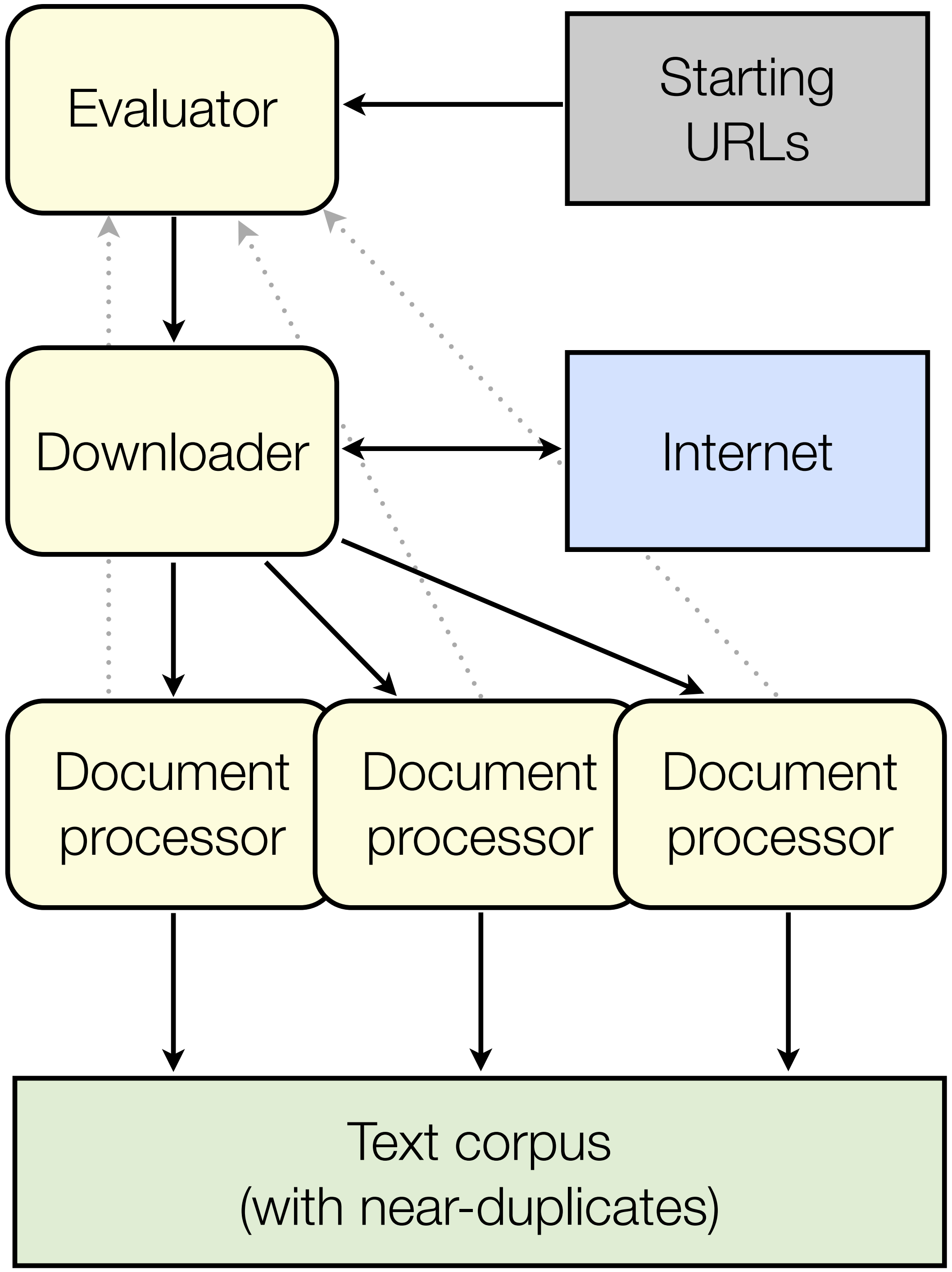
Annotated text corpus

SpiderLing

- In-house created web crawler for text corpora
- Why not use existing software?
  - Specific requirements
  - Robustness (must handle terabyte downloads)
  - Simple crawlers
    - Not robust enough
  - Complex robust crawlers (e.g. Heritrix)
    - Difficult to customize

SpiderLing

- **Web crawler which focuses on text-rich sources**
- Written in Python
- Emphasis on simplicity
- Asynchronous communication with servers
- Main modules (subprocesses)
    - Evaluator
    - Downloader
    - Data processors



DOCUMENT PROCESSORS

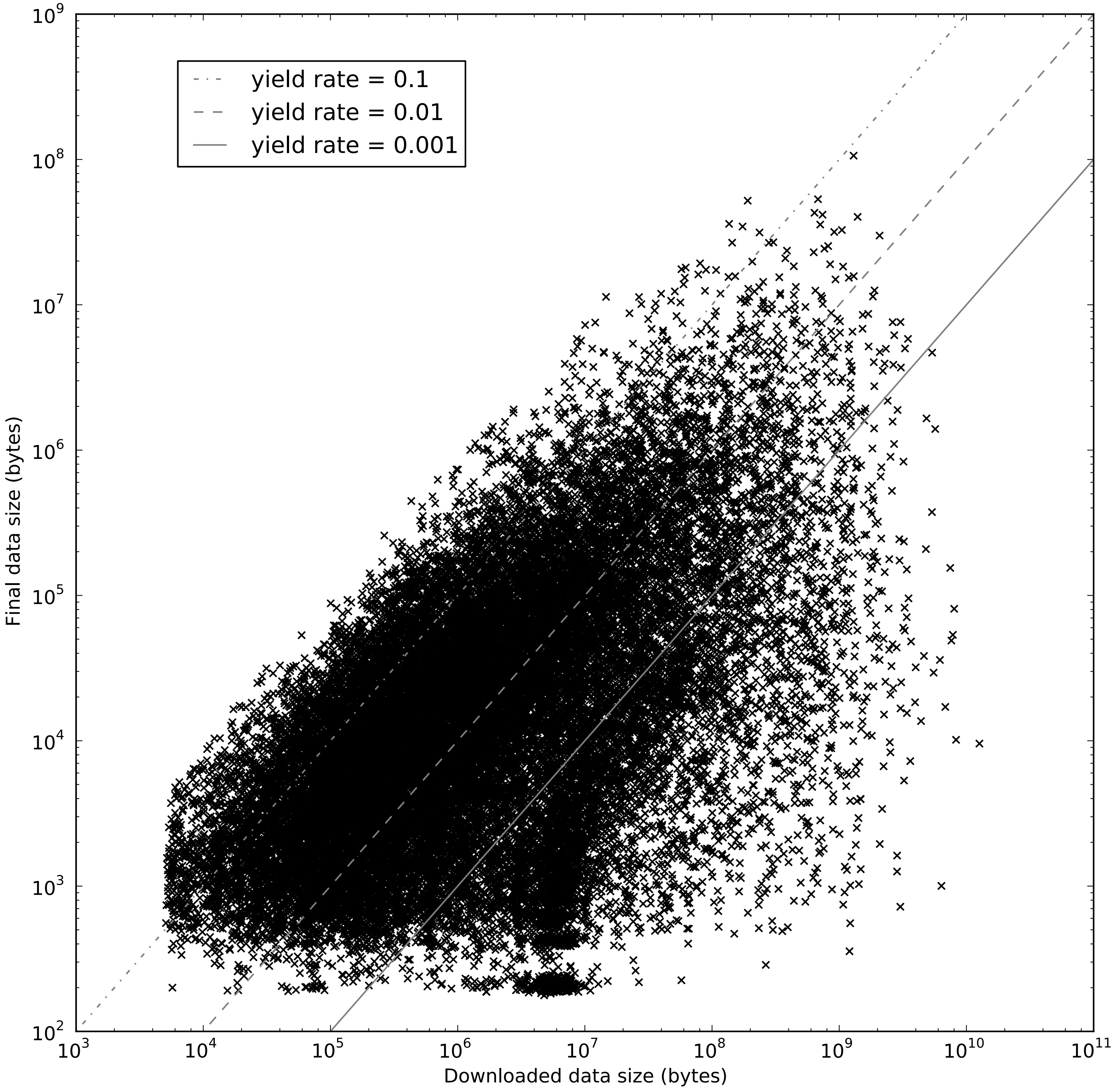
- Character encoding detection (chared)
- Language filters (trigram.py)
- Removing boilerplate (jusText)

YIELD RATE

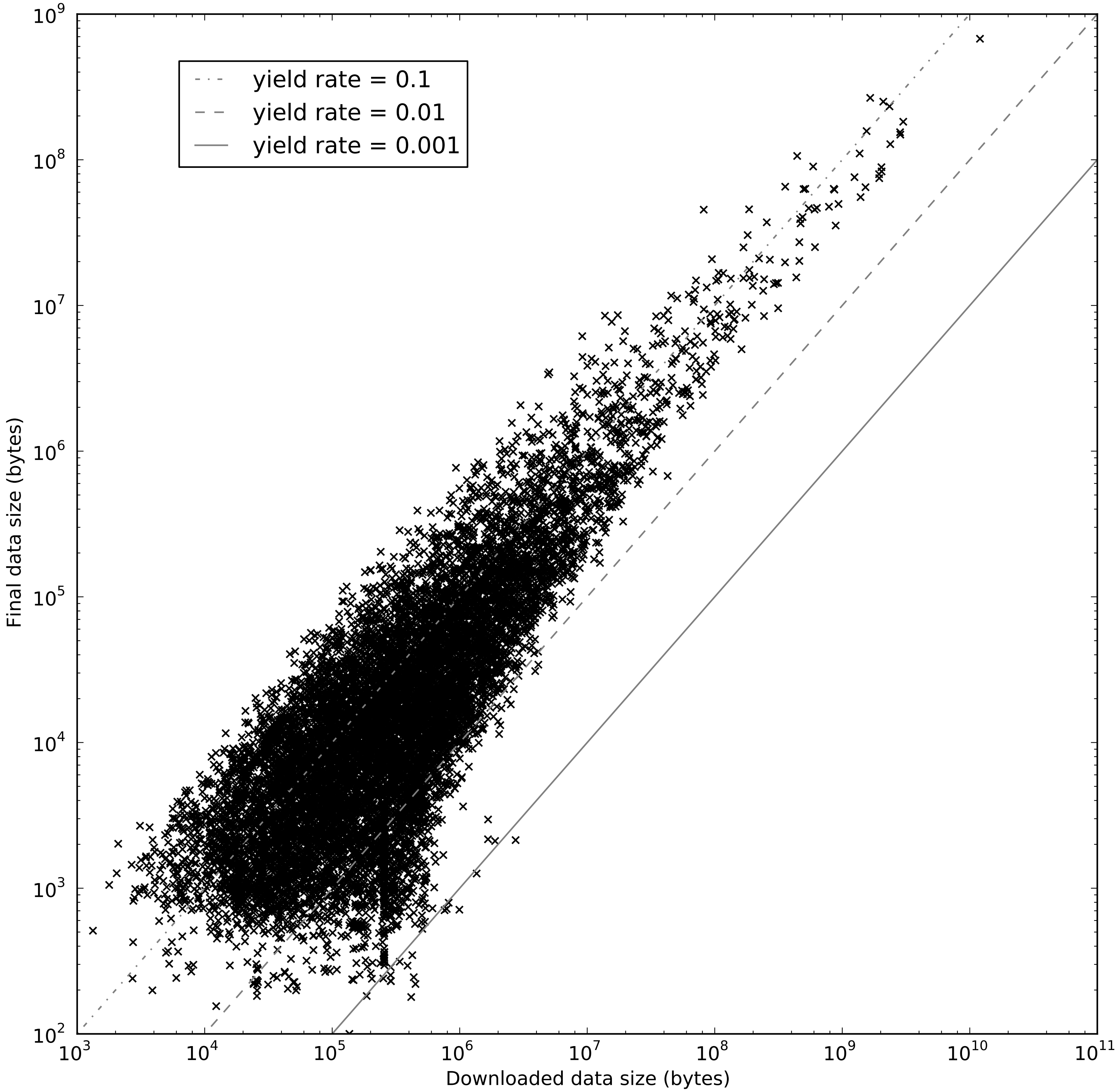
- yield rate = final corpus size / downloaded data size
- Yield rate stats for internet domains (on the fly)
    - Prune bad domains

domain	downloaded data	clean data	yield rate
www.prozakladnu.cz	198.7 MB	151.3 MB	76,4%
www.astrologie.cz	138.1 MB	97.8 MB	70,9%
slovnik.online-clanky.cz	17.2 MB	11.1 MB	64,5%
www.darius.cz	13.8 MB	8.5 MB	61,8%
www.pavlat-znalec.cz	12.7 MB	7.7 MB	60,8%
...			
stanpilot.rajce.idnes.cz	12.6 MB	0.2 MB	1,4%
spojene-arabske-emiraty.orbion.cz	10.6 MB	0.1 MB	1,2%
sof.rock.cz	11.2 MB	0.1 MB	1,1%

Yield rates for Heritrix crawled data



YIELD RATES FOR SPIDERLING CRAWLED DATA



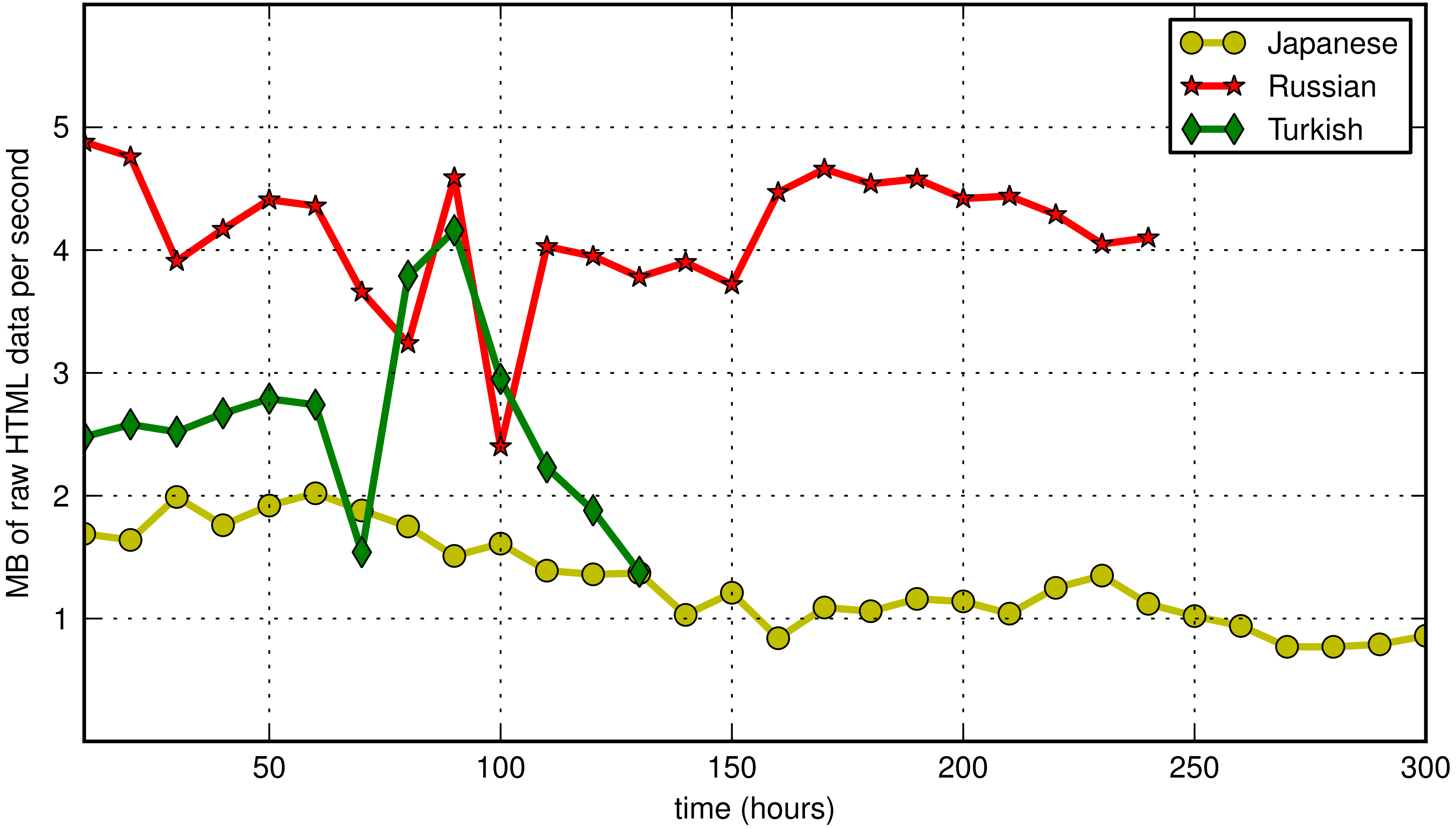
YIELD RATE THRESHOLD

- Yield rate threshold is a function of the number of downloaded documents:

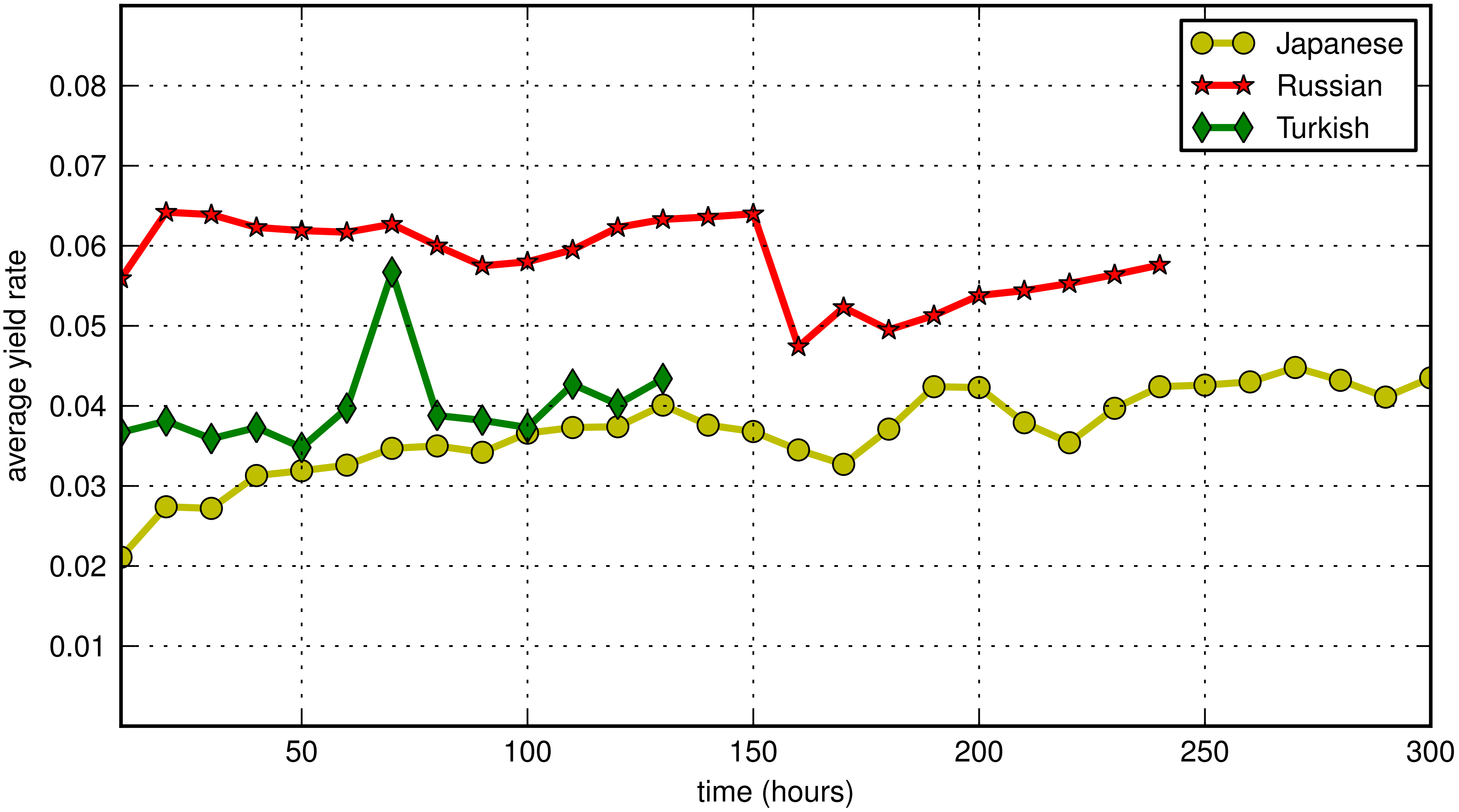
$$t(n) = 0.01 \cdot (\log_{10}(n) - 1)$$

# of documents	yr threshold
10	0%
100	1%
1000	2%
10000	3%

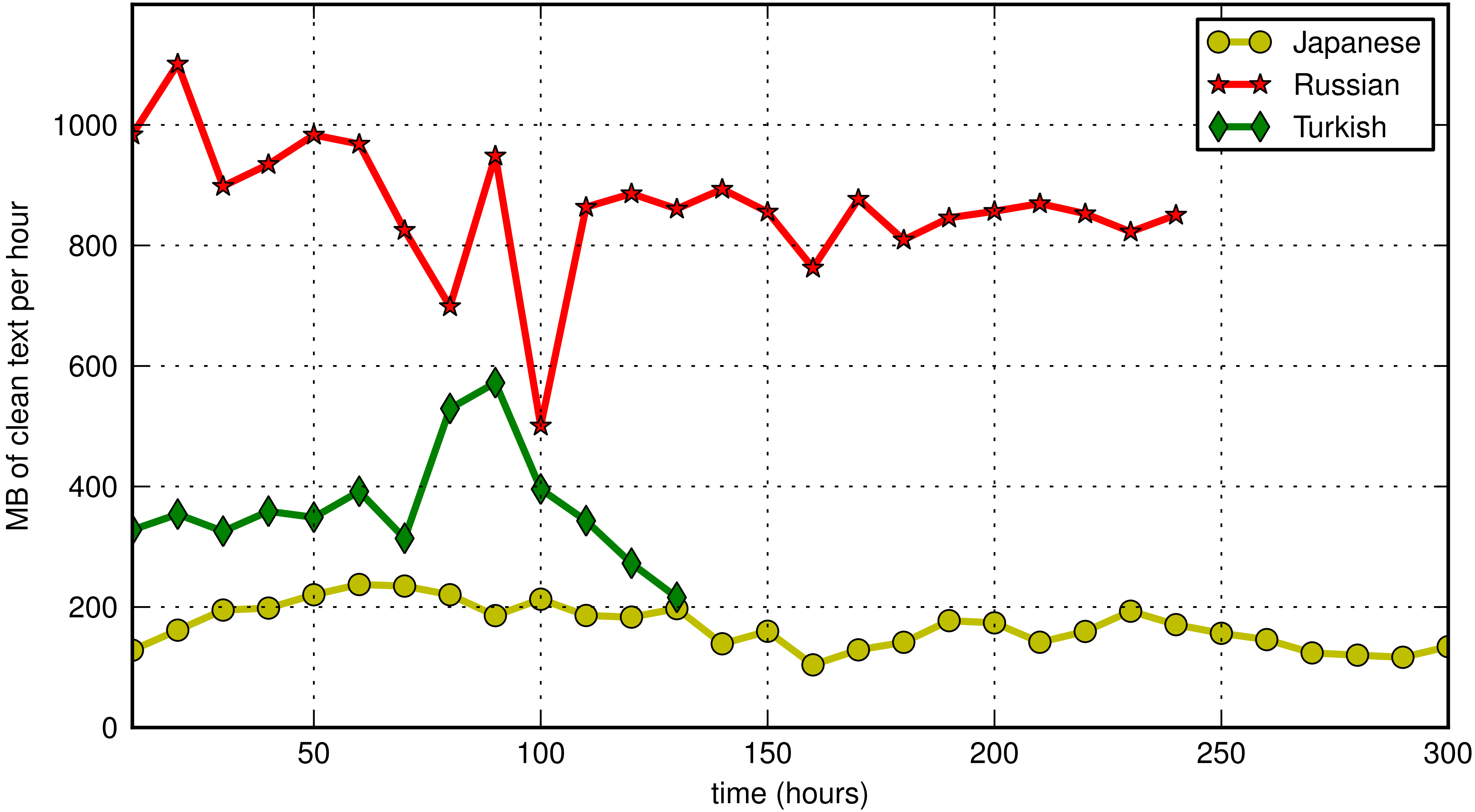
CRAWLING SPEED (RAW HTML DATA)

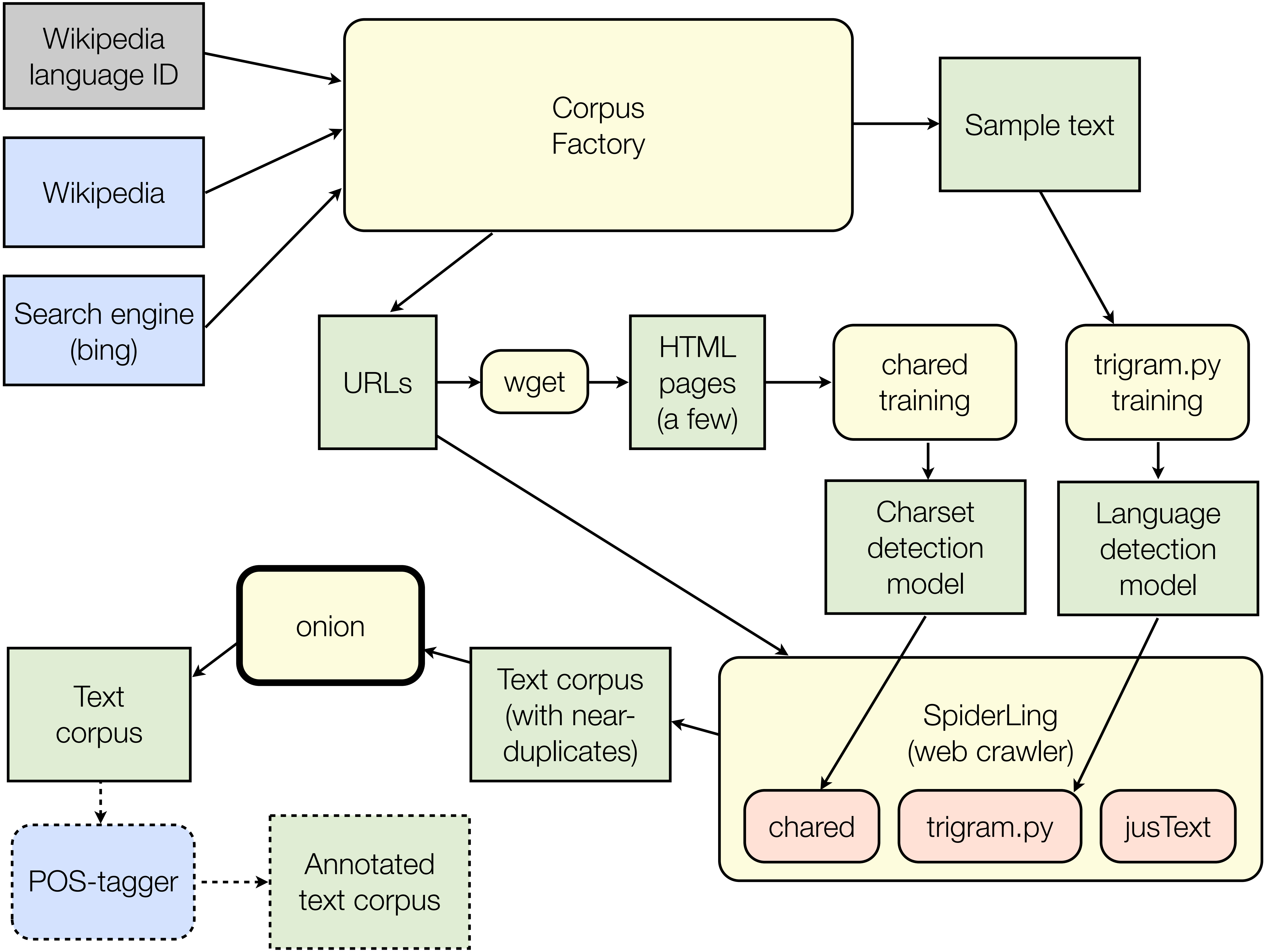


YIELD RATE DEVELOPMENT



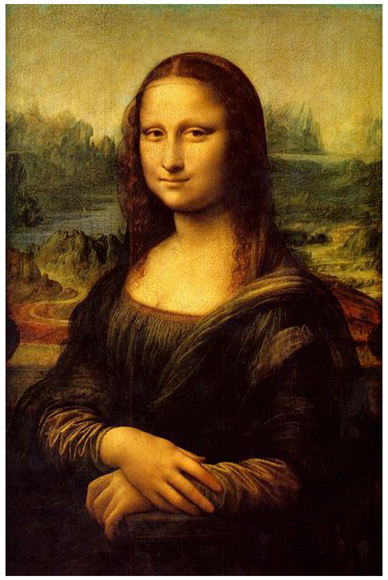
CRAWLING SPEED (CLEAN DATA)



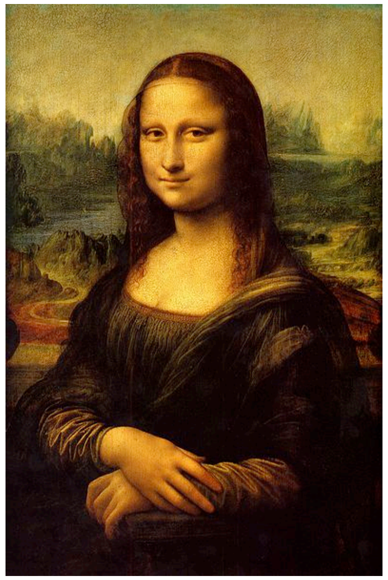

Wikipedia language ID
Wikipedia
Search engine (bing)
Corpus Factory
Sample text
URLs
wget
HTML pages (a few)
chared training
trigram.py training
Charset detection model
Language detection model
onion
Text corpus
Text corpus (with near-duplicates)
SpiderLing (web crawler)
chared
trigram.py
jusText
POS-tagger
Annotated text corpus

REMOVING DUPLICATE AND NEAR-DUPLICATE DATA

- Exact duplicates - easy

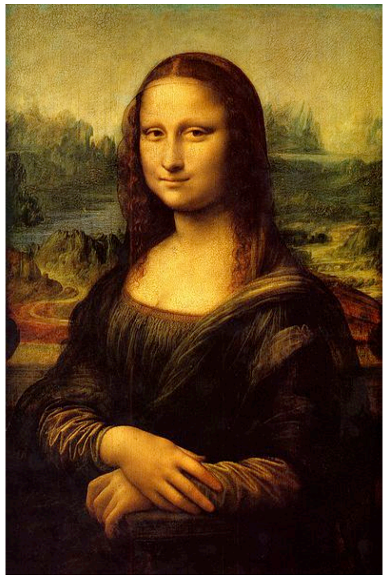


aa7d66733dbe

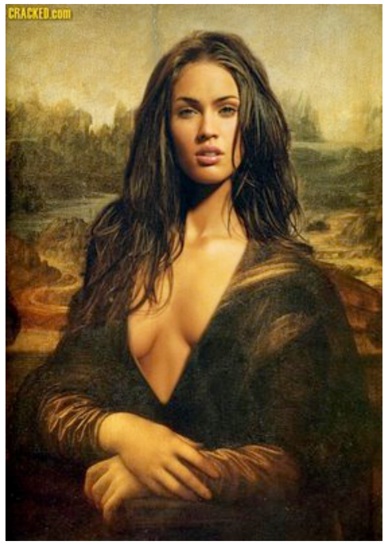


aa7d66733dbe

- Near duplicates - difficult



aa7d66733dbe

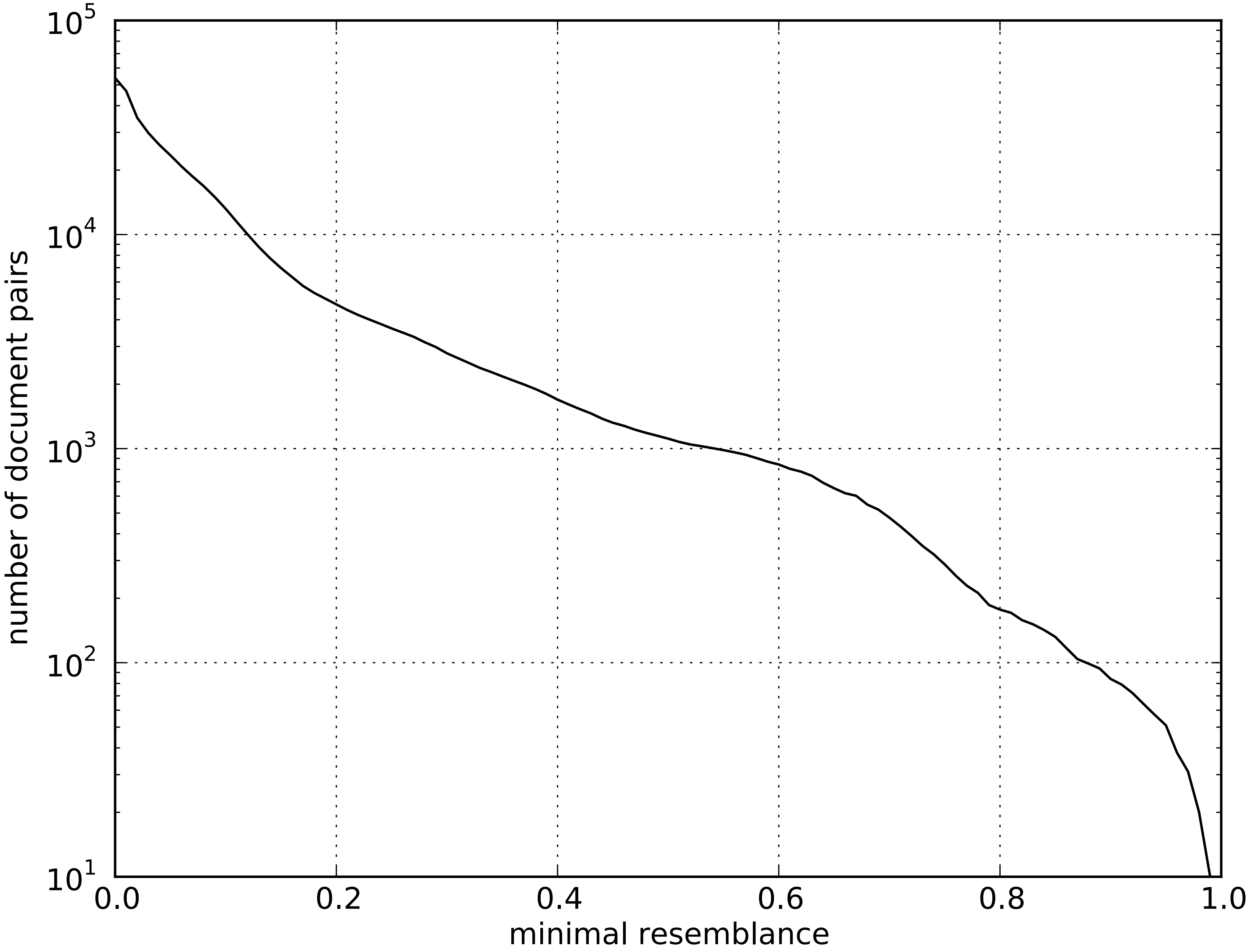


d87a79bfe197

KNOWN ALGORITHMS

- Mostly from IR field (web search engines)
- Broder’s shingling algorithm
- Charikar’s algorithm
- Fail to detect similarities at intermediate level (say 50-80%)
  - Not a problem for search engines (a feature rather than a bug)
  - For text corpora, all duplicates are problematic
- It seems that in web collections, document pairs at an intermediate level of similarity are much more frequent than document pairs at a high level of similarity

SIMILAR DOCUMENTS IN CLUEWEB09



Experiment performed on a small sample of ClueWeb09: 21,776 Web pages from 2,722 different domains.

ONION (DE-DUPLICATION PROGRAM)

- N-gram based
- We don’t need to know which pairs of documents are near-duplicates

6-grams for “what can we do with a drunken sailor”:
(what, can, we, do, with, a)
(can, we, do, with, a, drunken)
(we, do, with, a, drunken sailor)

- It suffices to make sure that the text we are adding to a corpus is not already there
- Keep a set of n-grams already present in a corpus
- A new document is added to the corpus only if it doesn’t contain too many n-grams already contained in the corpus
- The set of n-grams may grow out of RAM capacity
- Precompute list of duplicate n-grams (with 2 or more occurrences in the whole corpus); usually less than 10% of all n-grams (n >= 7)
- Prune unique n-grams

ONION: TECHNICAL DETAILS

- Finding duplicate n-grams -- standard external sort
- Storing 64-bit hashes of n-grams (rather than raw n-grams)
- Hashes stored in a Judy array -- a complex memory efficient associative array data structure
    - Judy1 - integer->Boolean (for representing sets)
    - RAM requirements as low as 6 bytes per hash

ONION: EVALUATION

corpus name	**enTenTen**	**itTenTen**	**deTenTen**	**ClueWeb09 (7%)**
language	English	Italian	German	English
words before de-duplication*	4.09 bil.	4.22 bil.	4.13 bil.	8.98 bil.
words after de-duplication	3.15 bil. (76.9%)	2.59 bil. (61.5%)	2.44 bil. (59.1%)	7.28 bil. (81.0%)
dupl. 10-grams before de-dupl.	528 mil.	662 mil.	625 mil.	913 mil.
dupl. 10-grams after de-dupl.	41 mil. (7.7%)	66 mil. (10.0%)	36 mil. (5.8%)	97 mil. (10.6%)

* after language filtering, removing boilerplate and exact duplicates

ONION: SCALABILITY

- Used for de-duplication of English ClueWeb09 (1bn web pages)
- Input size 920 GB (more than 100bn words)
- 72bn words after de-duplication
- aura.fi.muni.cz (8x 8-core Intel Xeon 2.27GHz, 440 GB RAM)
- ca 5 days on a single CPU
- Required 148 GB RAM

ONION: AVAILABILITY

- In C
- Open source (BSD License)
- http://code.google.com/p/onion/

RESULTS

Language	Tokens	Time
Czech	5.8G	?
Tajik	32.5M	3.4 days
Russian	20.2G	12.5 days
Japan	12-18G	22.5 days (+)
Turkish	5-10G	6.5 days (+)

FUTURE WORK

- Distinguishing between similar languages (e.g. Czech vs. Slovak)
- What kind of data is there in the web corpora?
    - Document clustering
    - Manual inspection of the clusters
- Corpus evaluation

Thank you!
Questions?