swisscom

SwigSpot

Creation of a Swiss German Dataset

Projet d'approfondissement (PA)

Lucy Linder

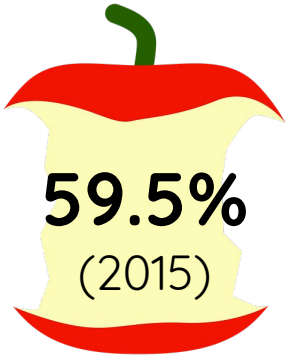
26.06.2018

MSE | MASTER OF SCIENCE IN ENGINEERING

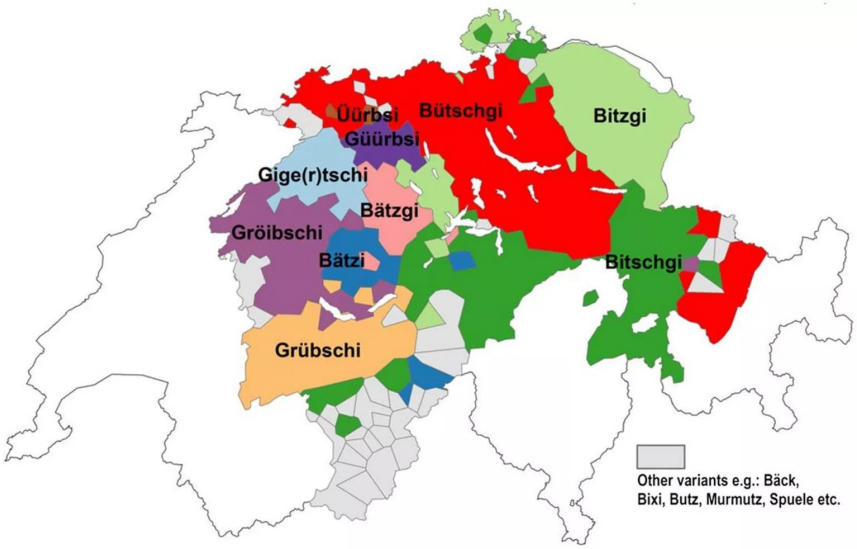
HEIA-FR
HTA-FR

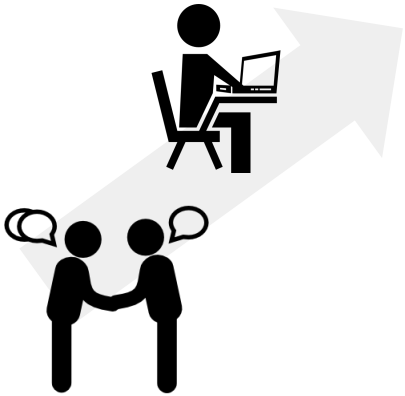
Context

Swiss German Dialects...



0.066% worldwide





"If you talk to a man in a language he understands, that goes to his head.
If you talk to him in his language, that goes to his heart."

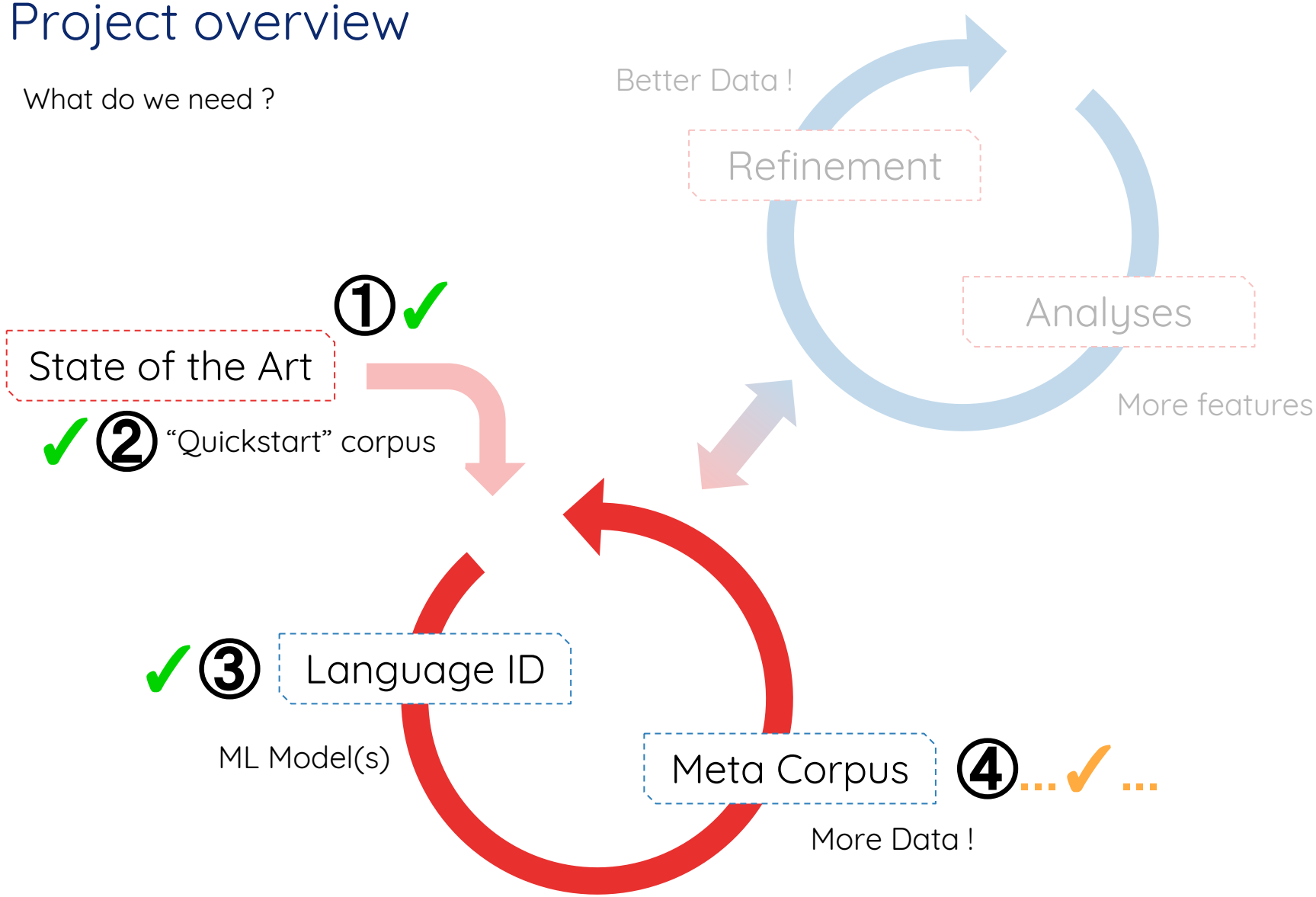
Nelson Mandela

Project Outline

The goal of the SwigSpot project is to **gather Swiss German resources** into a well designed corpus available to researchers.

Project overview

What do we need ?



①State of the Art and ②*Quickstart* Dataset

Existing corpora

ArchiMob corpus

34 XML transcripts

NOAH's Corpus

XML from 5 kind of sources

- Sms4science

Leipzig Corpora Collection

136 languages
Source: Web, wikipedia, ...
Already processed

- An Crúbadán
- SB-CH corpus

Quickstart dataset

TRAINING SET

7'387 samples per language

`de.txt`
`fr.txt`
`it.txt`
`en.txt`
`sg.txt`

NOAH + LEIPZIG

```
./get_quickstart_dataset.py
```

VALIDATION SETS

10'692 Swiss German SMS

`sms-sg.txt`

Swiss SMS corpus

```
./get_sms4science.py -l sg
```

200 SMS per language

`sms-any.txt`

Swiss SMS corpus

```
for lang in de fr en it sg; do
./get_sms4science.py \
    -l $lang -y -n 200
done
```

+

2'613 samples per language

`valid_de.txt`
`valid_fr.txt`
`valid_it.txt`
`valid_en.txt`
`valid_sg.txt`

Remaining LEIPZIG sentences

Added afterwards

③Language ID

Using machine learning

Models landscape

LID as a multi-class supervised ML classification task using N-grams as feature set

Letters + spaces ?

Preprocessing

Raw sentences ?

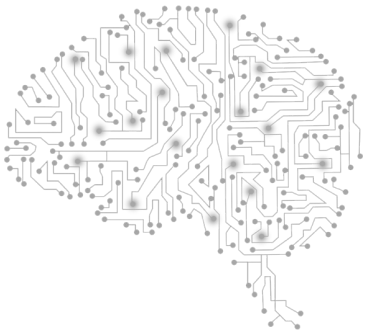
SG vocabulary vs ALL

Feature extraction

number of features

fixed vs variable-size N-grams

logfreq, TF-IDF ?



Neural Network

Classification

SVMs

Logistic Regression

Naive Bayes

Technologies & implementation

~ 20 notebooks

Best models

- using sanitized data: letters and spaces only
- after hyperparameters tuning

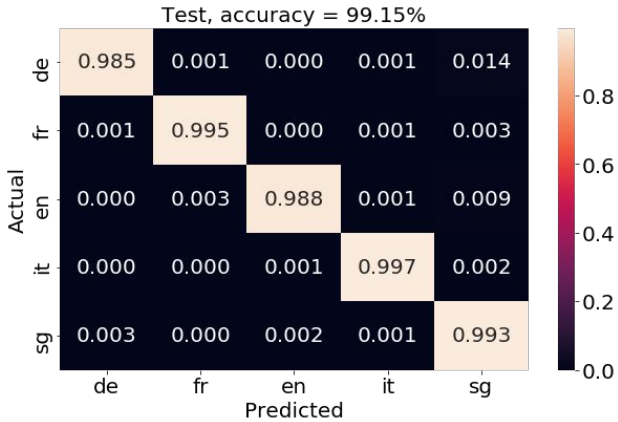
Model	Weights	N-grams	Type	#features
LogisticRegression	log-TF, IDF scaling.	trigrams	char.	10'000.
SVM	log-TF, IDF scaling.	trigrams	char.	10'000
Multinomial NB	Raw frequencies.	1-3 grams	char.	10'000
NaiveIdentifier	Raw frequencies.	1-3 grams	words	3'000/lang
NeuralNetwork	Raw frequencies.	1-3 grams	char.	3'000

Results

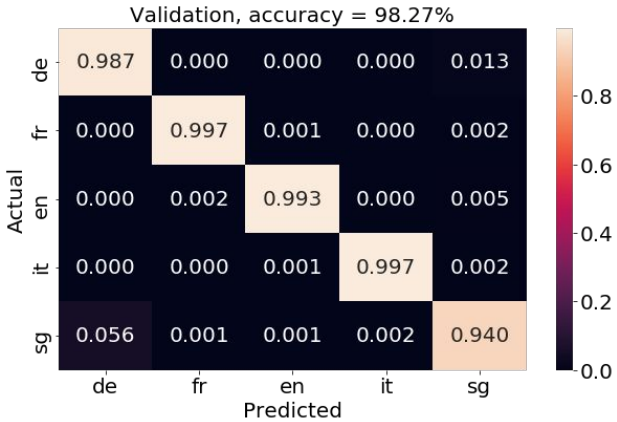
Performances

Classifier	Accuracy			SG (Valid.)			SG SMS err.	
	Test	SMS	Valid.	Prec.	Recall	F1	Count	%
LogisticRegression	99.40	85.57	98.55	98.37	94.57	96.43	63	0.59
SVM	99.45	87.06	98.61	99.11	94.07	96.52	85	0.79
MultinomialNB	98.55	95.22	98.29	99.47	93.03	96.14	386	3.61
NaiveIdentifier	98.16	92.64	98.38	99.75	92.61	96.05	460	4.30
Neural Network	98.01	88.46	97.20	96.96	89.21	92.92	389	3.64

Scraping model Logistic regression, 6000 features, 3-5 grams




```
SMS RECALL
==========
total samples 10692
total errors  57 (0.53%)
------------------------
languages detected
    de        47
    fr         1
    en         3
    it         6
    sg     10635
```



Results

Performances

Classifier	Accuracy			SG (Valid.)			SG SMS err.	
	Test	SMS	Valid.	Prec.	Recall	F1	Count	%
LogisticRegression	99.40	85.57	98.55	98.37	94.57	96.43	63	0.59
SVM	99.45	87.06	98.61	99.11	94.07	96.52	85	0.79
MultinomialNB	98.55	95.22	98.29	99.47	93.03	96.14	386	3.61
NaiveIdentifier	98.16	92.64	98.38	99.75	92.61	96.05	460	4.30
Neural Network	98.01	88.46	97.20	96.96	89.21	92.92	389	3.64

Scraping model

Logistic regression, 6000 features, 3-5 grams

④Data gathering

Main hypothesis: still more SG to find on the web

Warming up: SG webapp

SG Crawler Lang ID

SG Language detection

Query

Enter an url with potential Swiss German sentences:

URL

http://example.com

Extractor

ArticleExtractor

Model

MultinomialNB, CountVectorizer(1-3 ngrams, 10000 features)

Min. words: 5

Display raw sentences

Go!

Results from http://www.martinfrank.ch

Labels: de fr en it sg displayed: 6/6 >= min proba Show colors SG only

I write novels stories poems magazine articles film scripts and plays I write what I want to read myself If you happen to have the same taste Welcome to our world

I write in Swiss German English and the Swiss form of written German In my head I think when I think probably half of the time in Swiss German and half of the time in broken English

Am April habe ich bei der Buchvernissage von Dominic Oppligers acht schtumpfo züri empfernt eine Ansprache gehalten Ansprachen liegen mir nicht besser als ich Klavier oder Gitarre oder Flöte oder Geige spiele

Ort war der Helsinki Club an der Geroldstrasse Zürich Der Club sah schlimm aus die Leute vom Club sahen ebenso schlimm aus doch sie waren freundlich und fröhlich und das Publikum auch Was mich an dieser Szene beeindruckt ist dass viele einschliesslich Dominic Oppliger schon viel mehr geleistet haben als sie aussehen

the heaviness of the sighs on the endless nights i m in pain from missing you i can no long

don t let me go don t leave me only you i believe in ah to warm my lips my emotions make my hea

Le paysage devint accidenté abrupt le train s arrêta à une petite gare entre deux montagnes

fr with probability 1.00
de: 0.000
fr: 1.000
en: 0.000
it: 0.000
sg: 0.000

i hoken ufter schtange forter kasse slouft äiäm seiling luegen uf tur schhaubi achti xene bue ufter a ite for pan länt are sülen ei fuess ufem gumiramp for pan ter anger azüle gschtemt luegp mi a luegene gnau a erzäni füfzäni xe kli us wine pönk mizo rötleche schtachuhor träkigi auti blutschins nideri tenischschue äs häugäups tischört unes auz plutschins jäggli he peid häng ide hosesek luegpmer it ouge lue kli de skuter zue luegt wider zu mir übere risch chliner aus i u ender düm für si grössi risig fiu schwanz u ejer ide hose machpmi huere geil än arsch wine chline fuesspauer ter hoselade so haub off oder ter rissferschluss isch kabut luegen uf mini bei abe di schwarze läderhose töffschtifu mi schwanz ut ejer ide hose xetno geil us

boilerpipe

python webapp
dockerized

.ch domains approach

IDEA 1

hypothesis: SG spoken only in Switzerland

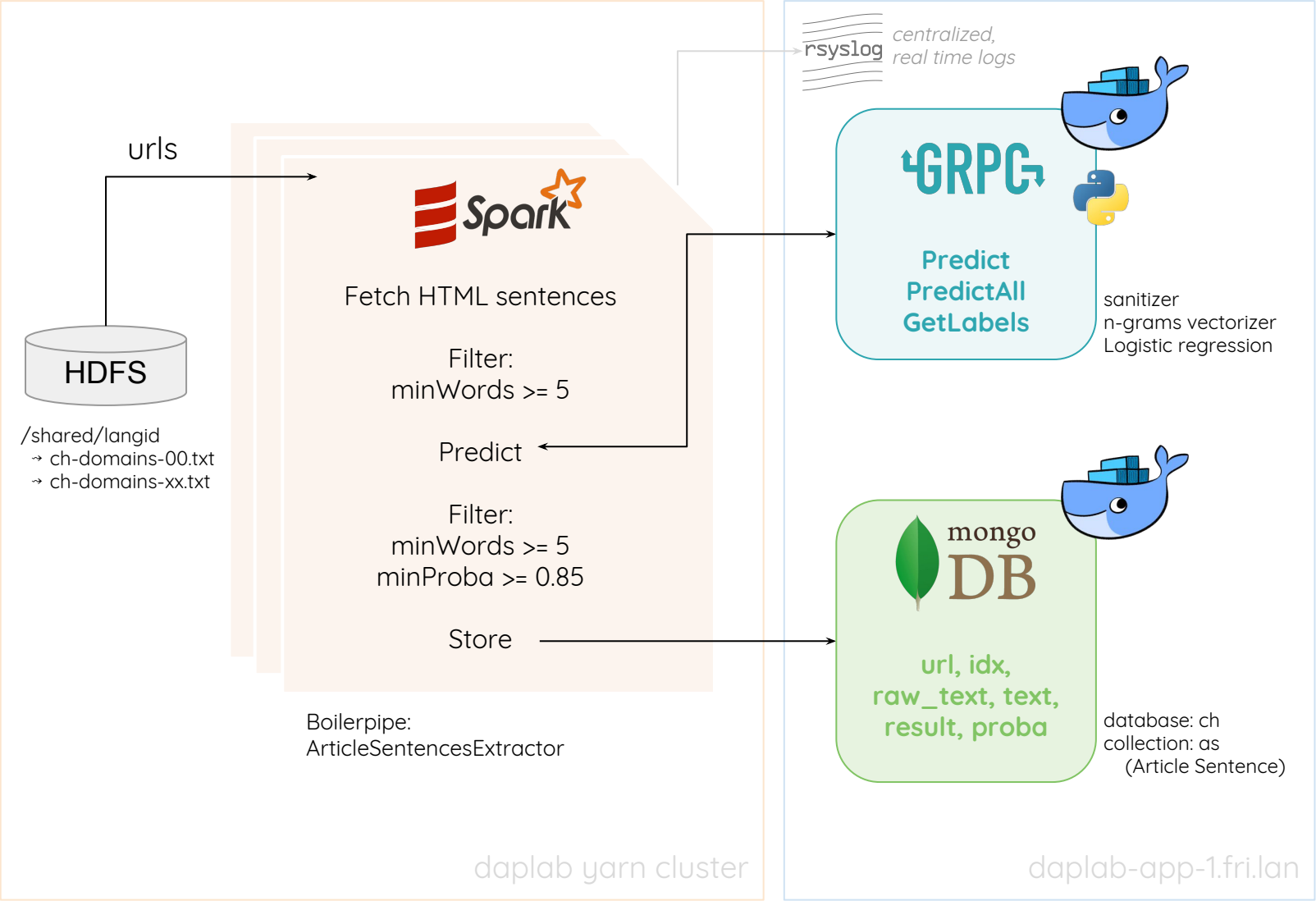
→ scrape the entire .ch domain

1'367'215 domains, that's a lot of crawling...

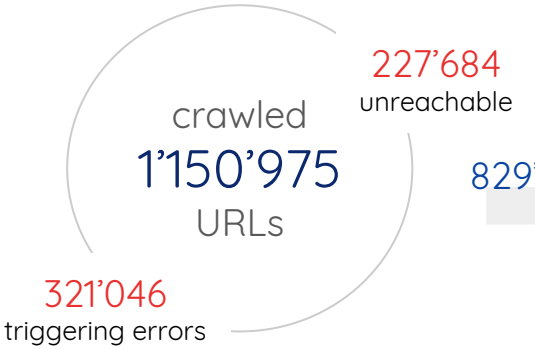
→ analyse only the *landing page*

→ use a *distributed* pipeline

Crawling pipeline



Results



829'929 pages

Threshold	≥ 0.85	≥ 0.90	≥ 0.95
Number of results	$30'452$	$7'969$	$1'517$

Correct sentences:

97% s gliche isch mitem stromnetz und de wasserversorgig i new york. all die leitige und versorgigsinfrastruktur isch extrem alt, und drum isches nid sälte dass es mal n komplette stromuusfall git. glaubs im summer isch de letschti riesä shutdown xi, [...]

95% än wichtigä teil vo dä päge isch di umfangriichi galerie

96% merci thömu, jetzt isch zzwänzgi abe gheit

90% Itz si mer o über Facebook derbi...

Incorrect sentences:

93% a..... ad..... adj..... ADJ..... ADJ..... [...]

86% ոդ դարի երրորդ միտոնը եւ կոչուեցաւ յաղթութեան միտոն

92% CD sRössli Hü bim König CD sRössli Hü bim König

91% o u r m i s s i o n

91% Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich

85% Telefonisch +49 1805 2455269 oder per Mail: welcome@billbox.com

86% saved from urlhttpinternet.email htmlhead titlemeine homepagetitle head hintergr. weiss, body bgcolorffffff linkdee vlinkdede bild über die ganze seitenbreite p aligncenterimg [...]

Lots of false positive:
Address, copyright, Armenian, ASCII art...

about 40 sites *actually* written in Swiss German...

Results

crawled
1'150'975
URLs

227'684

829'9[illegible]9 pages

321'046
triggering errors

Threshold	≥ 0.85	≥ 0.90	≥ 0.95
Number of results	30'452	7'969	1'517

Correct sentences:

97% s gliche isch mitem stromnetz und de wasserversorgig i new [illegible] versorgigsinfrastruktur isch extrem alt, und drum isches nid sälte[illegible] stromuusfall git. glaubs im summer isch de letschti riesä shutdow[illegible]

95% än wichtigä teil vo dä päge isch di umfangriichi galerie

96% merci thömu, jetzt isch zzwänzgi abe gheit

90% Itz si mer o über Facebook derbi...

Incorrect sentences:

93% a..... ad..... adj..... ADJ..... A[illegible]

86% ոդ դարի երրորդ միտոնը եւ կոչուեցաւ յաղթութեան միտոն

92% CD sRössli Hü bim König CD sRössli Hü bim König

91% o u r m i s s i o n

91% Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich Zürich

85% Telefonisch +49 1805 2455269 oder per Mail: welcome@billbox.com

86% saved from urlhttpinternet.email htmlhead titlemeine homepagetitle head hintergr. weiss, body bgcolorffffff linkdee vlinkdede bild über die ganze seitenbreite p aligncenterimg [...]

of false positive:
A[illegible]s, copyright, Armenian, ASCII art...

ut 40 sites *actually*
tten in Swiss German...

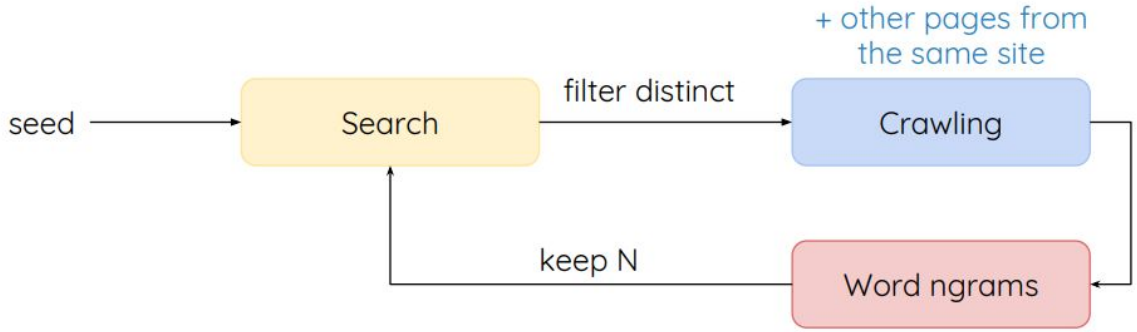
“Search Google” approach

IDEA 2

hypothesis: Swiss German is mostly used in informal contexts (forums, golden books, etc.)

→ use seeds in a Search Engine to gather interesting URLs

Potential *reinforcing loop*



Proof-of-concept

Using the first 100 results for 5 SG sentences:

"das isch sone seich", "das isch super", "weiss öpper", "het öpper", "wär chamer".

```
#URLs: 212
avg proba: 0.94

#sentences with proba:
   >= 0.85: 10289 (unique: 8555)
   >  0.90:  6556 (unique: 5504)
   >  0.95:  2197 (unique: 1883)

SG sentences per URL:
   avg:   68
   min:    1
   max: 1487
```

```
text (characters):
   avg:     198
   min:      16
   max:   3'657 (one at 553'307)

raw text (characters):
   avg:     202
   min:      16
   max:   3'472 (one at 561'059)
```

Processing time: **3 minutes**

about 8'000 new sentences !

Existing work

CorpusBuilder (2001)

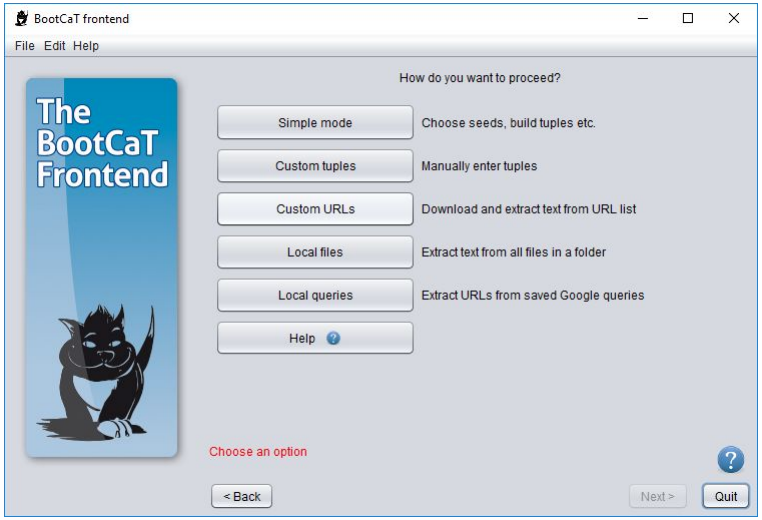
“automatically collecting documents in a minority language using Web queries”

The Leipzig Corpora collection

“Frequent terms [extracted from the Declaration of Human Rights] are combined to form Google search queries and retrieve the resulting URLs as a basis for the default download system”

The BootCaT toolkit

“one can build a relatively large quick-and-dirty corpus (about 80 texts, with default parameters and no manual quality checks) in less than half an hour”



We reinvented the ⊙wheel⊙…
But what a wheel !

Wrapping up

Summary

Objectives

SG Miner status unlocked.
Gather more than 10k sentences from the Web

Conclusion

There are a lot of Swiss German resources still to be discovered

Next steps

DIALECTS + SEARCH ENGINE APPROACH + HUMAN IN THE LOOP

Merci Vilmal

https://github.com/derlin/SwigSpot_Schwyzertuutsch-Spotting