swisscom

<insert title> II

Projet d'approfondissement (PA)

Lucy Linder
01.05.2018

MSE | MASTER OF SCIENCE IN ENGINEERING

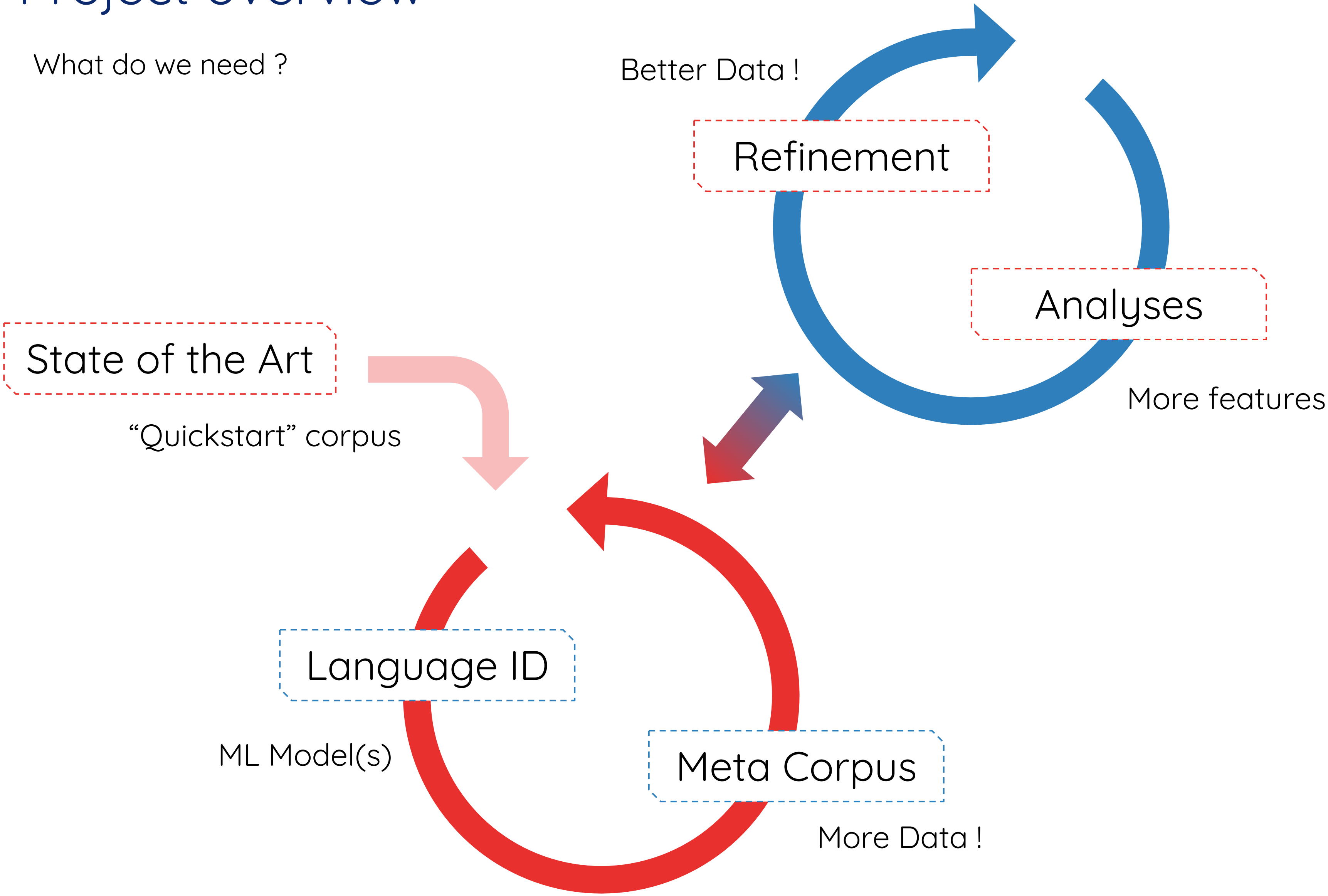
HEIA-FR
HTA-FR

Summary

(it’s been a while…)

Project overview

What do we need ?



State of the Art

“Quickstart” corpus

Language ID

ML Model(s)

Meta Corpus

More Data !

Refinement

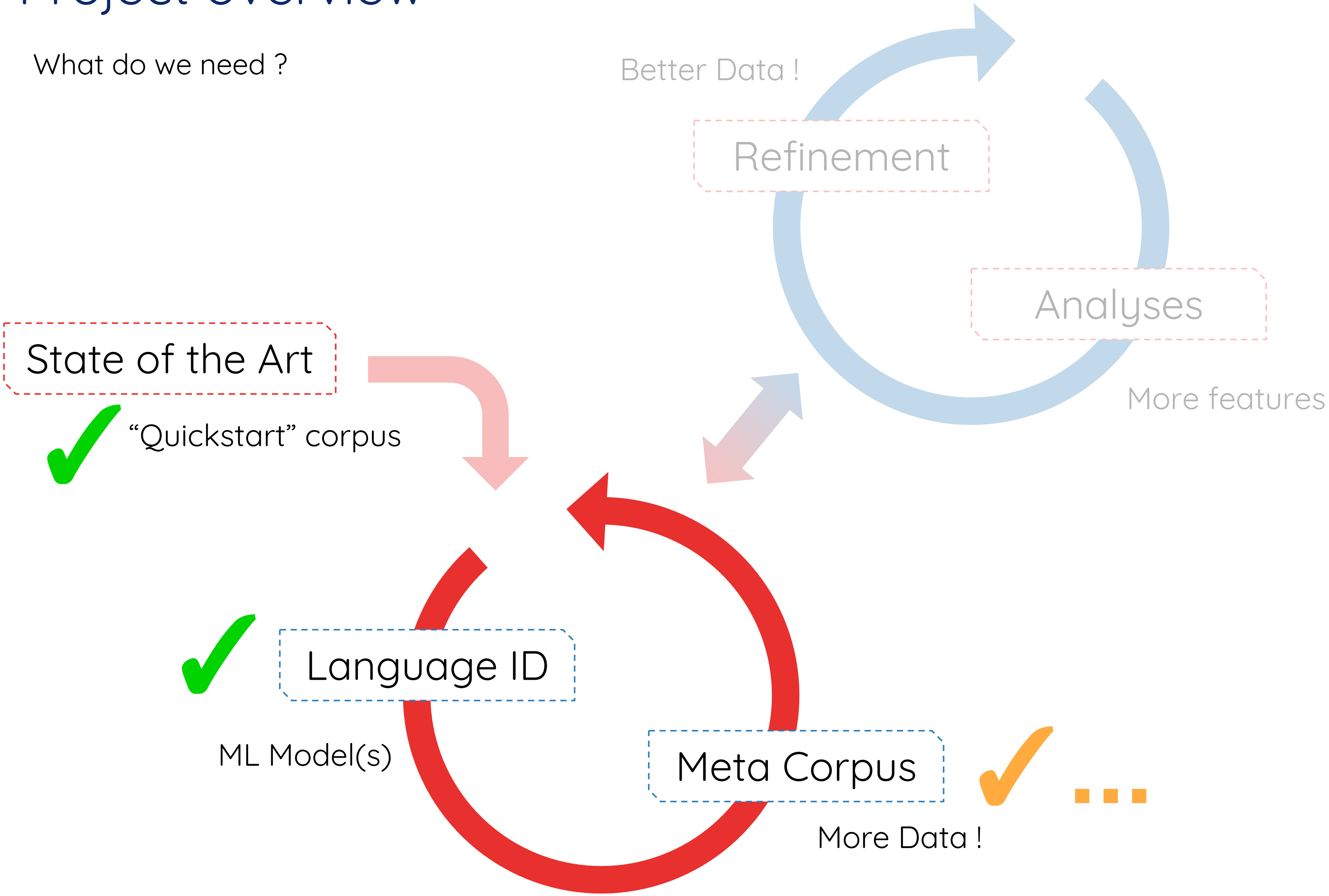
Better Data !

Analyses

More features

Project overview

What do we need ?



State of the Art

✓ “Quickstart” corpus

Better Data !

Refinement

Analyses

More features

✓ Language ID

ML Model(s)

Meta Corpus ✓ ...

More Data !

Language Identification

Dataset

Quickstart dataset - train+test set

FR, DE, IT, EN

→ Leipzig corpora: http://wortschatz.uni-leipzig.de/en/download/

→ Wikipedia sentences between 2010-2016, 10K

SG

→ NOAH corpus

→ 7’431 sentences (114+ empty)

about

7K

Sentences per lang.

Validation set

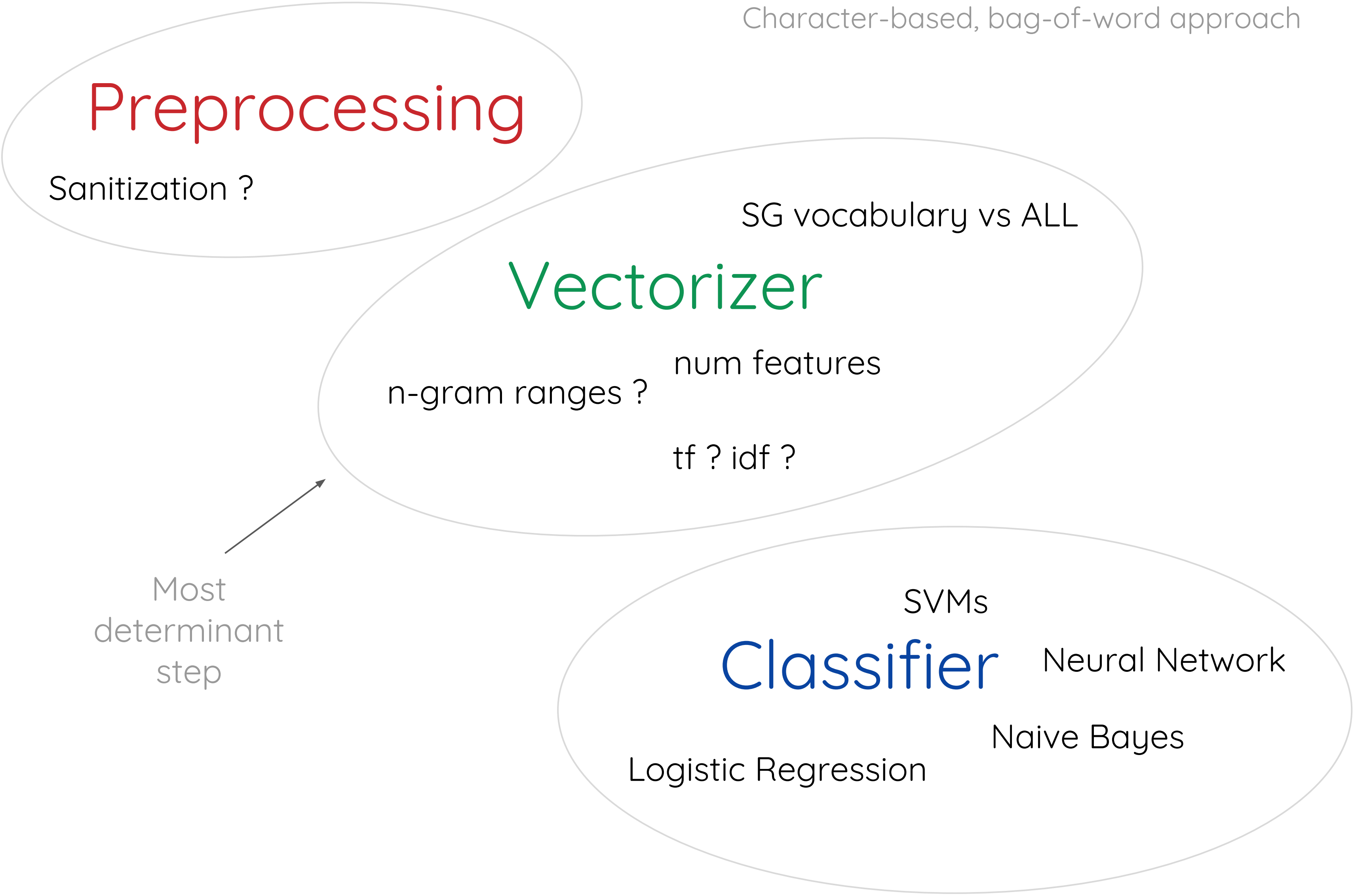
SG

→ sms4sciences, testing mostly the recall

→ 10’706 sentences

Models landscape

Character-based, bag-of-word approach



Preprocessing

Sanitization ?

SG vocabulary vs ALL

Vectorizer

num features

n-gram ranges ?

tf ? idf ?

Most determinant step

SVMs

Classifier

Neural Network

Naive Bayes

Logistic Regression

Feature extraction

⚠ !! biases !!

Vectorizer

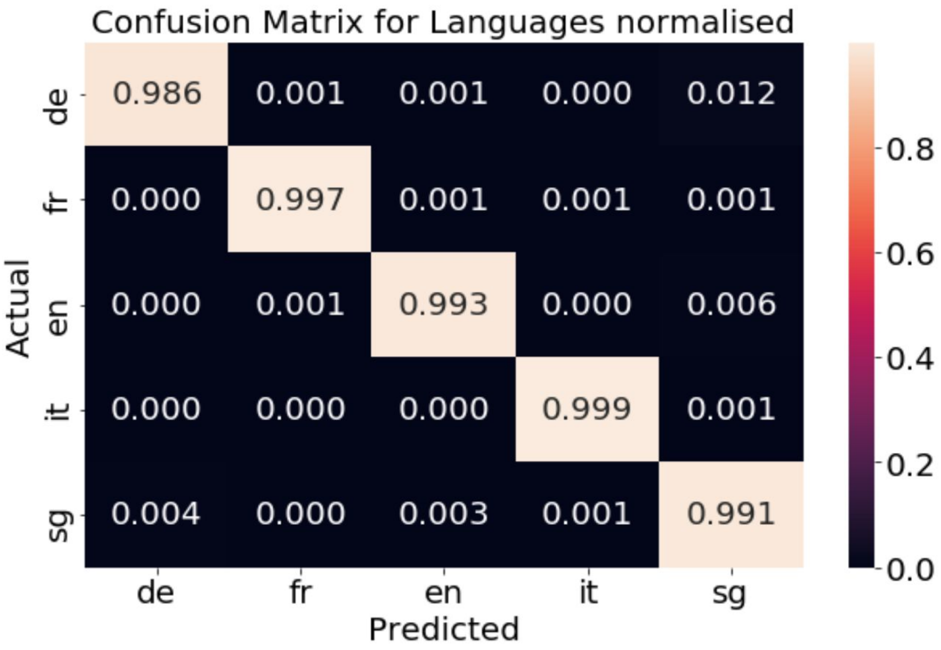
Using GridSearchCV + LogReg:

Best score (accuracy): **0.989**
Best parameters set:
- max_features: 10'000
- ngram_range: (3, 3)
- use_sanitizer: True
- sg_only: False
- sublinear_tf: True
- use_idf: True

Tested

sg_only: True, False
sanitizer: None, np_sanitize
max_features: 4000, 6000, 10000
ngram_range: (3,3), (4,4), (3,5)
use_idf: True, False
sublinear_tf: True, False

…



Confusion Matrix for Languages normalised

Actual \ Predicted	de	fr	en	it	sg
de	0.986	0.001	0.001	0.000	0.012
fr	0.000	0.997	0.001	0.001	0.001
en	0.000	0.001	0.993	0.000	0.006
it	0.000	0.000	0.000	0.999	0.001
sg	0.004	0.000	0.003	0.001	0.991

```
SMS samples: 10706, errors: 56 (0.52%)
--------------------------------------
other languages detected:
    de   48,   fr  3
    en    1,   it  4
```

Prediction

accuracy
90%+
For all but NN

Classifiers

Naive Bayes

→ just for fun, as a training ;)

Logistic Regression

→ easy and fast

→ efficient: 0.99+ accuracy

✓

SVM

→ training very slow, hard to fine-tune + never converges !

→ best: 0.99, linear kernel, C=1

✓

Neural Networks

→ tested with 1 hidden layer only

→ not enough [good] data for good results... (?)

Testing tools and evaluation

Scraping WebApp

Data gathering

Main hypothesis: still more SG to find on the web

IDEA 1: .ch domains

1’367’215 .ch domains

$ viewdns.ch domain list $

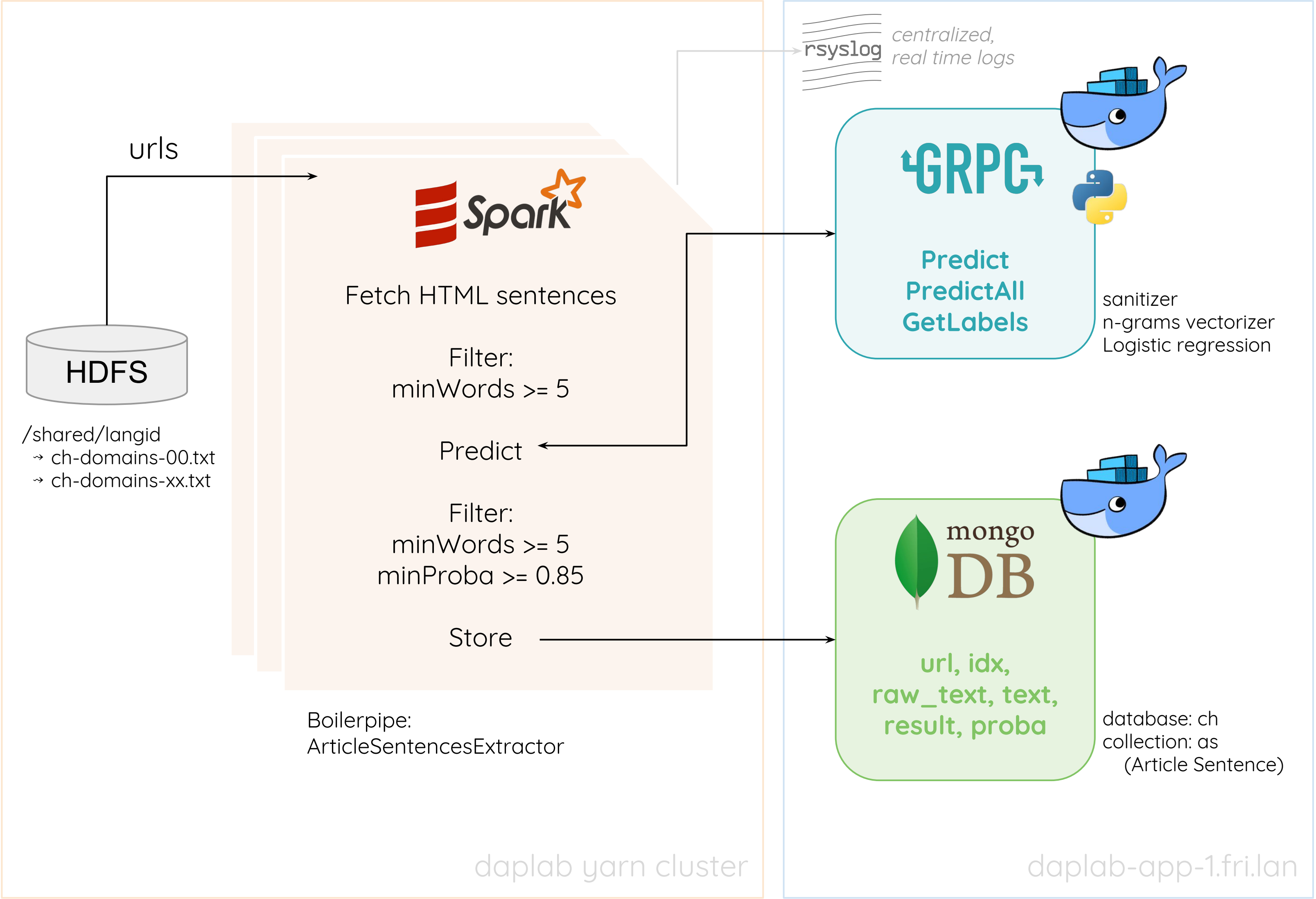
⚠ Not all hosting websites

lots of URLs …

→ analyse only the *landing page*

→ use a *distributed* pipeline

Crawling pipeline



urls

HDFS

/shared/langid
→ ch-domains-00.txt
→ ch-domains-xx.txt

Spark

Fetch HTML sentences

Filter:
minWords >= 5

Predict

Filter:
minWords >= 5
minProba >= 0.85

Store

Boilerpipe:
ArticleSentencesExtractor

daplab yarn cluster

rsyslog *centralized, real time logs*

GRPC

Predict
PredictAll
GetLabels

sanitizer
n-grams vectorizer
Logistic regression

mongo DB

url, idx,
raw_text, text,
result, proba

database: ch
collection: as
(Article Sentence)

daplab-app-1.fri.lan

Version 1.0-SNAPSHOT, 09.04.2018 @ Lucy Linder | slides

Crawling pipeline

Data format

*“**as**” table*

```
{
    "_id": "1-ASE|1700875192-0",
    "domain": "0713.ch",
    "url": "http://0713.ch",
    "idx": 0,
    "raw_text": "Wenn zom Fänschter use luegsch :)",
    "text": "wenn zom fänschter use luegsch",
    "result": "sg",
    "proba": [ ... ],
    "extractor_name": "ASE",
    "version_number": 1,
    "version_description": "ng3-5_sg_f6k_lreg",
    "when" : ISODate("2018-04-20T13:23:34Z")
}
```

*“**log**” table*

```
{
  "_id": ObjectId("..."),
  "url": "http://0-1.ch",
  "sg": 14,
  "count": 200,
  "ex": "",
  "when": ISODate("2018-04-20T13:23:19Z"),
  "model_version": 1,
  "model_version_descr": "ng3-5_sg_f6k_lreg",
  "extractor": "ASE"
}
```

Difficulties

→ Time about 45 minutes for 1’000 URLs … 42+ days ! (less using multiple processes)

→ Aleas lost nodes, OutOfMemoryError, …

→ Charset, Scala, Logging, …

Results

proba	≥ 0.85	≥ 0.90	≥ 0.95
count	30’452	7’969	1’517

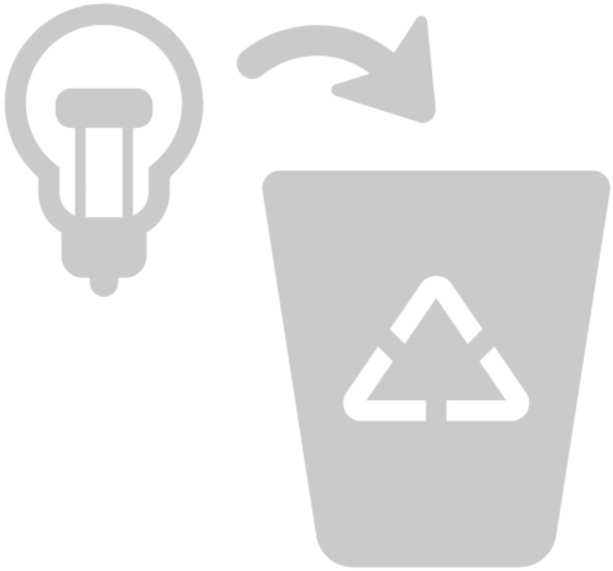
out of
1’150’975
URLs

227’547+
unreachable

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...	✓

93%	a........................... ad......................... adj...................... ADJ.................. ADJ.................. [...]	✗
86%	րդ դարի երրորդ միւռոնը եւ կոչուեցաւ յաղթութեան միւռոն	
92%	CD sRössli Hü bim König CD sRössli Hü bim König	
91%	o u r m i s s i o n	
86%	ception lockedfalse priority namemedium list wlsdexception lockedfalse priority namemedium list wlsdexception lockedfalse priority namemedium grid wlsdexception lockedfalse priority nam [...]	

IDEA 1: conclusion



IDEA 2: “search Google Approach”

Hypothesis:

“Swiss German is mostly used in informal contexts,

such as forums, golden books, etc.”

IDEA 2: “search Google Approach”

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
```

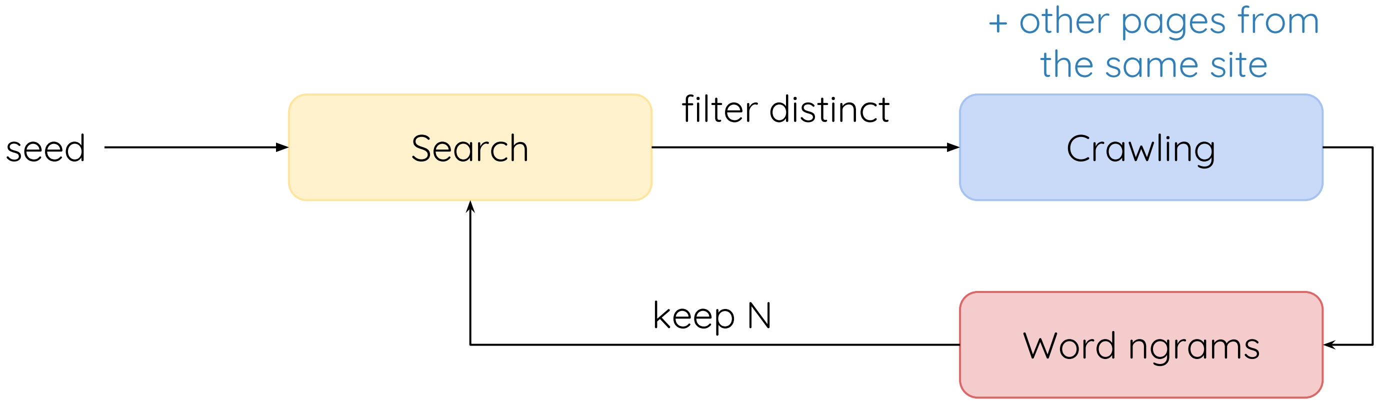
```
text (characters):
   avg:      198
   min:       16
   max:    3'657 (one at 553'307)

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

Processing time: 3 minutes

about 8’000 new sentences !

IDEA 2: “search Google Approach”

Proposition



Difficulties / questions

→ how to distribute ?

→ search engine limitations

→ dealing with duplicates …

→ …

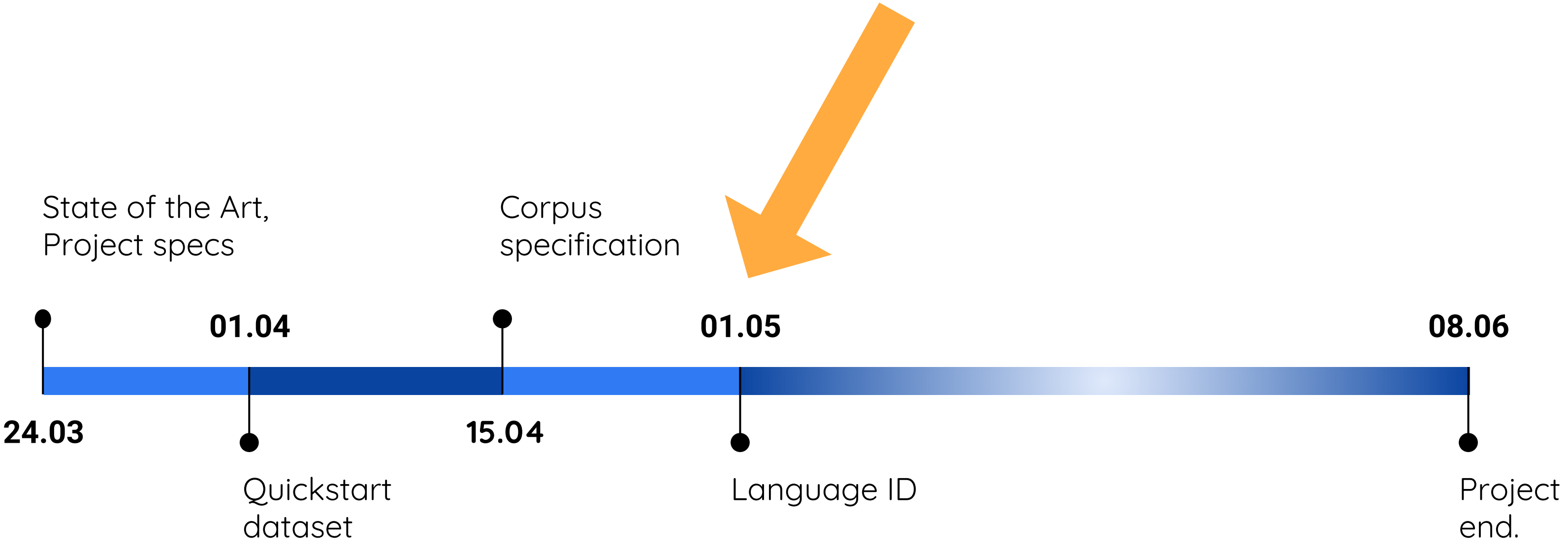
Summary and administration

Calendar

Agile methodology

Major milestones

We are here



State of the Art,
Project specs

24.03

01.04

Quickstart
dataset

Corpus
specification

15.04

01.05

Language ID

08.06

Project
end.

2-4 weeks for the report … It leaves us less than 2 weeks !

Open points

What’s next ?

→ human validation ?

→ new langid models ?

→ “Google Search” implementation ?

→ …

Source code and database ?

→ technology stack …

→ ~~Spark~~ ?

Merci Vilmal