



ML Project: Intro

Hendrik Blockeel
Wannes Meert

21 Feb 2017

ML Project

- Goal
 - Design and implement a complete ML pipeline, including: Training, Prediction, Evaluation
- Task
 - Classify .be domain names
 - In collaboration with **dnsbelgium**
- Timeline
 - 24 Feb: Form groups
 - 24 March: First report and data collection
 - 12 May: Final report and prototype + peer assessment
 - ~22 May: Oral discussion

Classifying Domain Names

Classify a .be domain name into one of the following categories:

.be-websites

| CIJFERS IN % | 2012 | 2013 | 2014 | 2015 | EVOLUTIE 2014-2015 |
|-----------------------------|-------|-------|-------|-------|-----------------------|
| Business site | 37,52 | 32,72 | 32,42 | 33,81 | +1,39 |
| Company holding page | 10,90 | 14,18 | 15,59 | 20,81 | +5,22 |
| Non-commercial site | 7,28 | 8,82 | 8,68 | 8,94 | +0,26 |
| Non-commercial holding page | 12,30 | 12,90 | 9,93 | 9,36 | -0,57 |
| Error reporting | 7,55 | 7,40 | 15,39 | 16,13 | +0,74 |
| Pay-per-click | 6,50 | 4,05 | 2,10 | 0,66 | -1,44 |
| Personal/family/blog | 6,15 | 4,72 | 4,83 | 3,29 | -1,54 |
| Webshop | 4,32 | 4,70 | 3,80 | 2,21 | -1,59 |
| Portal/media | 1,20 | 2,30 | 1,68 | 0,66 | -1,02 |
| Website for sale | 4,48 | 3,40 | 2,78 | 2,40 | -0,38 |
| Password-protected | 0,98 | 4,42 | 2,30 | 1,28 | -1,02 |
| Pornography | 0,82 | 0,38 | 0,50 | 0,46 | -0,04 |

dnsbelgium

KU LEUVEN

Motivation

251.377 in 2015
264.930 in 2014
299.846 in 2013
306.341 in 2012
267.780 in 2011
257.637 in 2010
232.746 in 2009
222.919 in 2008
193.659 in 2007

Aantal nieuwe
.be-registraties

1.534.832 in 2015
1.492.063 in 2014
1.433.980 in 2013
1.346.772 in 2012
1.219.935 in 2011
1.101.668 in 2010
977.998 in 2009
859.474 in 2008
736.498 in 2007

Totaal aantal
.be-domeinnamen

Difficult to keep up with manual labor.
Can we automate using machine learning?

dnsbelgium

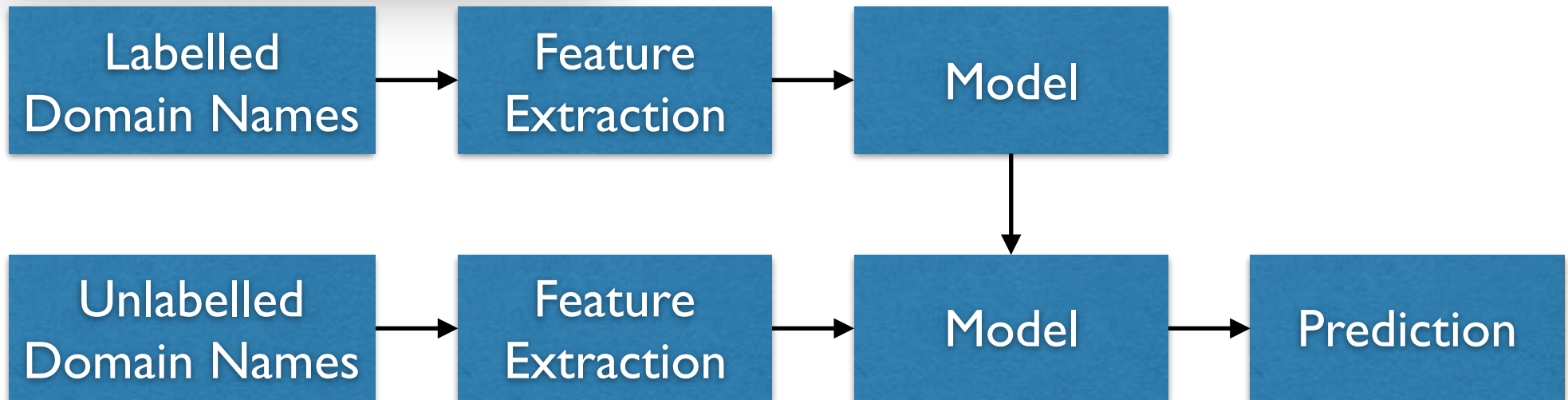
KU LEUVEN

Feature Extraction

- There are no restrictions on the type of features you can use.
- You can only connect to internet resources while scraping the website, not while computing the feature values.
- Possible types:
 - $f(\text{words})$
Using synonym or dictionary databases is allowed (as internet resource or locally available)
 - $f(\text{sentences})$
 - $f(\text{DOM})$
 - $f(\text{CSS})$
 - $f(\text{javascript})$
 - $f(\text{links})$
 - $f(\text{images})$
 - ...

Setup

```
domainlabels_...ta/dns) - VIM1
1 "NAME", "TLD", "CATEGORY"~
2 "upmanagement", "be", "error"~
3 "mediavijsonderwijs", "be", "non-commercial"~
4 "madeinhollandbeauty", "be", "web-shop"~
5 "design-habitat", "be", "company"~
<aster> [POS= 1, 1] [0%] [LEN=1501]
```



```
domainlabels_...ignment) -...
1 "NAME", "TLD", "CATEGORY"~
2 "10-forward", "be", ""~
3 "1000km", "be", ""~
4 "108agency", "be", ""~
5 "1207", "be", ""~
< [POS= 1, 1] [0%] [LEN=8442]
<inlabels_empty.csv" 8442L, 173058C
```

```
1. bash
$ ./classifydotbe \
> domainnames_in.csv \
> domainnames_out.csv
```

```
domainlabels_...ignment) -...
1 "NAME", "TLD", "CATEGORY"~
2 "10-forward", "be", "mylabel"~
3 "1000km", "be", "mylabel"~
4 "108agency", "be", "mylabel"~
5 "1207", "be", "mylabel"~
< [POS= 1, 1] [0%] [LEN=8442]
```

Evaluation

- You will be asked to label an unseen test set after submission of your code and before the oral discussion. Use the exact same setup as you submitted.
- At the oral discussion:
 - You might be asked to apply your setup to a domain name. Bring a laptop that is ready to *immediately* execute a prediction (eduroam wifi is available).
 - You might be asked questions about your pipeline, results and insights (or lack thereof) in the report, code, and general insights into ML models and methods.

What is (not) allowed?

- Data:
 - It is not allowed to share, distribute or reuse the given data set.
- Code and report:
 - Avoid a disqualification due to plagiarism!
 - Each team must work independently. Do not share code.
 - Clearly indicate your sources (ideas, concepts, code snippets).
 - You can discuss general issues on the Toledo forum or in the sessions.

Questions

- Use the Toledo Forum
- Use the sessions
 - Add your questions beforehand to the Toledo forum such that we can prepare
 - The sessions are not mandatory. You are only evaluated on the project results.