ML Project: Intro

Hendrik Blockeel
Wannes Meert

21 Feb 2017

KU LEUVEN

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

KU LEUVEN

Classifying Domain Names

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

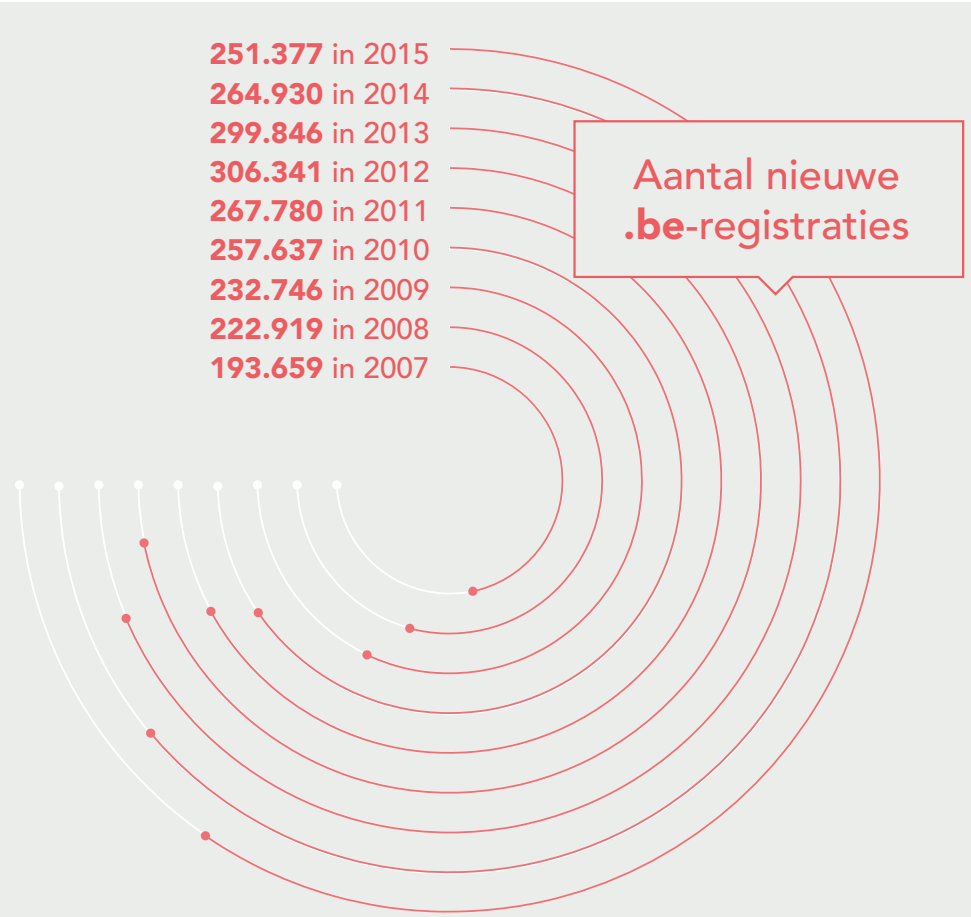
.be-websites

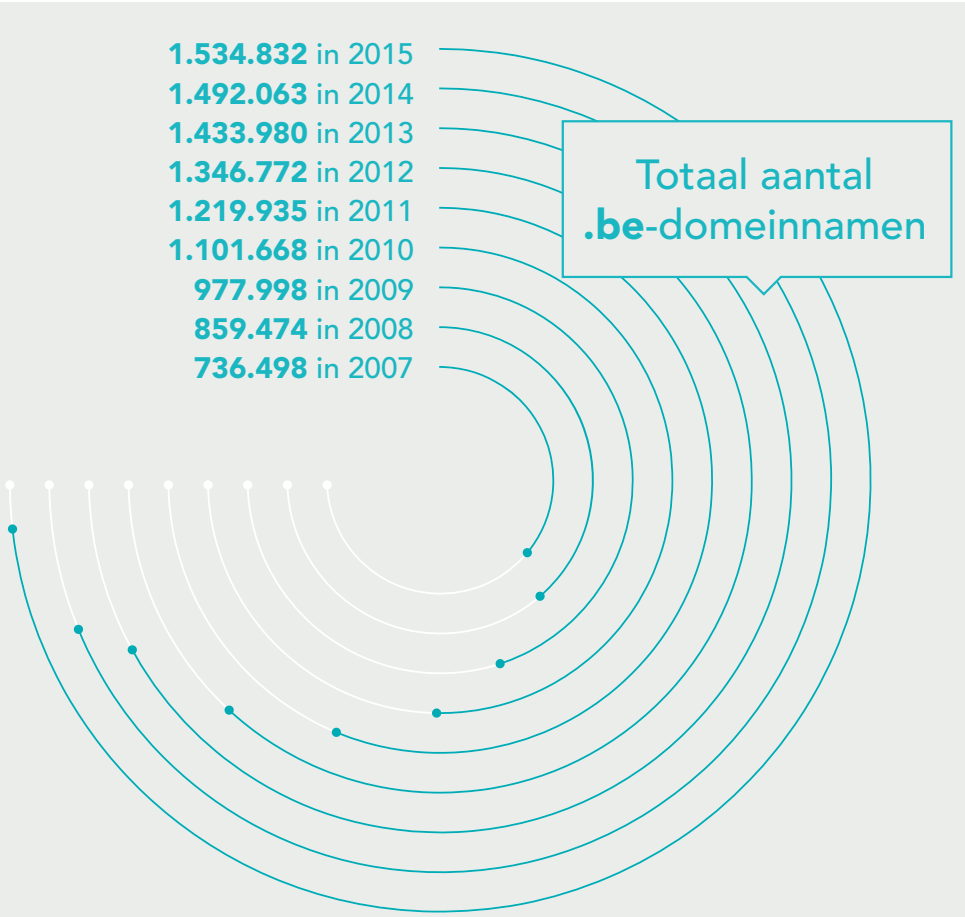
CIJFERS IN %	2012	2013	2014	2015	EVOLUTIE 2014-2015
Business site	37,52	32,72	32,42	33,81	+1,39
Company holding page	10,90	14,18	15,59	20,81	+5,22
Non-commercial site	7,28	8,82	8,68	8,94	+0,26
Non-commercial holding page	12,30	12,90	9,93	9,36	-0,57
Error reporting	7,55	7,40	15,39	16,13	+0,74
Pay-per-click	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





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

dnsbelgium

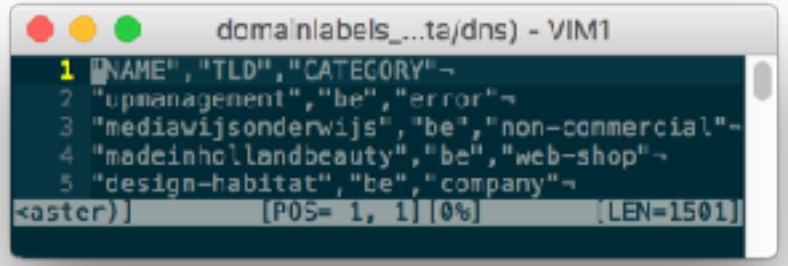
KU LEUVEN

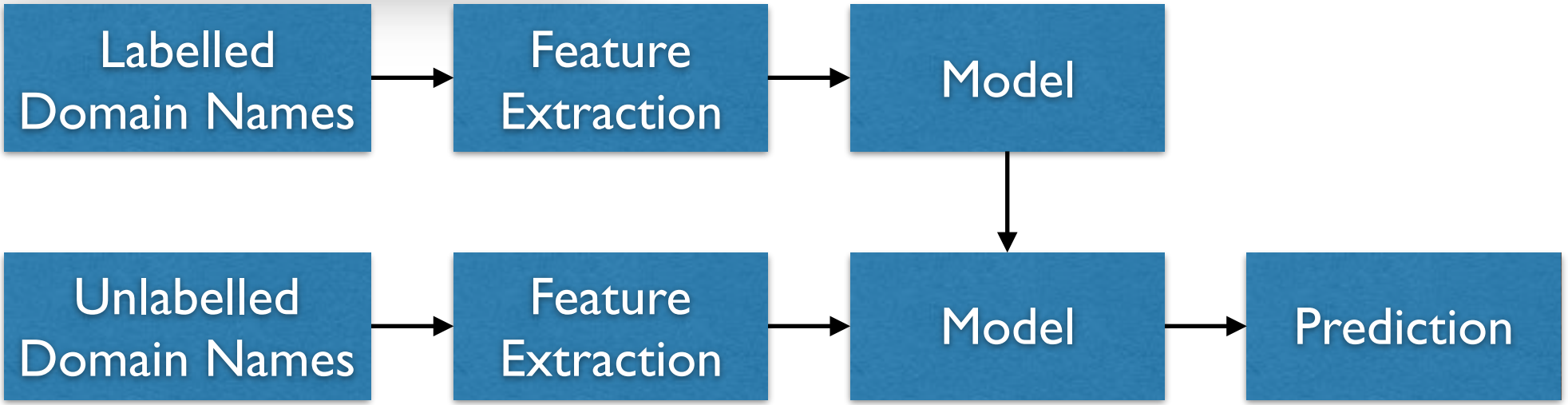
Feature Extraction

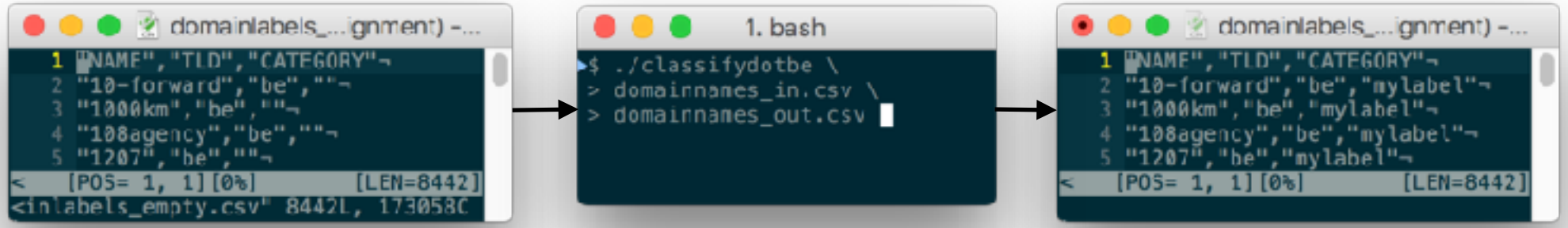
- There are no restrictions on the type of features you can use.
- You can only connect to internet resources while scraping the website, not while computing the feature values.
- Possible types:
  - f(words)
    Using synonym or dictionary databases is allowed (as internet resource or locally available)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

KU LEUVEN

Setup







KU LEUVEN

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.