Machine Learning Project 2016-2017 (Feb)

Classifying Domain Names

1 Problem

There are 1.6 million domain names registered in the .be zone. Around 80% of these domain names have a website. In this project, we are interested to know how many of these websites are used by private individuals or by companies. Further, we are interested to know how many contain a webshop, a holding page or blog, etc. This information is used to generate reports on internet usage in Belgium [2] but might also be used to direct users to correct websites or to detect anomalies.

Of course, it might take a long time to label all domain names and it is difficult to keep this list of labels up-to-date when domains are transferred or created (>250k/year). Automating this task is a promising path but it is non-trivial to write down an algorithm that does this automatically and this program also needs to be updated continuously. Therefore, this is an ideal task to apply machine learning where we manually label a small subset and incrementally feed corrections when we observe incorrect labels.

This project is in collaboration with DNS Belgium vzw. DNS Belgium is a not-for-profit organisation established in 1999 by ISPA Belgium (Internet Service Providers Association), Agoria (federation for the technology industry) and BELTUG (Belgium's communication technology and services user Group). Their aim is to register domain names, to make the internet more accessible, and to support its usage.

2 Approach

The problem will be tackled by designing a machine learning pipeline that given a domain name outputs the class label (or a set of weighted class labels). This pipeline will need to consist out of following steps:

1. Download information available at the domain name. Typically, the front page is sufficient. For example, when given kuleuven.be, you can download the html of the front page using a tool like curl or PhantomJs (to also execute Javascript).
2. Translate the domain contents into a feature vector (derived from text, topics, structure of site, links, colors, loaded libraries, . . .).
3. Feed the feature vector into a machine learning model.
4. Report the most likely class(es).

The data set will be a list of domain names and an associated label. You will be first given a set of domain names to label yourself. Afterwards this information is merged with a data set available from DNS.be and made available on Toledo. Finally, after submitting your report and code you will be given a list of unlabelled domain names that you label using your submitted setup.

It is not allowed to share, distribute or reuse the given data set. This data set is property of DNS Belgium and may only be used in the context of the Machine Learning Project.

Category	Description	Examples	#labels
company	Commercial website, website of a company.	http://www.telenet.be http://www.microsoft.be	36
holding page company	Calling card of a company OR default page of a hosting provider if it is clear the purpose of the website is commercial.	http://www.webjaseur.be	18
non-commercial	Website of a non-commercial organization.	http://www.nema.be	9
holding page non-commercial	Calling card of a non-commercial organization OR default page of a hosting provider if it is clear the purpose of the website is non-commercial.	http://vmarc.be	7
error	Page returns an error or doesn't exist.	http://www.klossebak.be	16
pay-per-click	Page with sponsored links. Choose this category only if the page is not for sale.	http://www.onehoney.be	2
personal-family-blog	Personal website.	http://www.familie-goelen.be	3
web-shop	Main purpose of the website is to sell products online.	http://www.topashop.be	3
portal/media	News sites, media, starting pages.	http://www.msn.be http://www.deredactie.be http://www.zita.be	2
for sale	Domain name is for sale. Choose this category if the page is for sale as well as pay-per-click.	http://www.rheumatology.be	2
password protected	Most part of the content is password protected.	http://polaroidfiction.be	2
~~porn~~	We ignore this category. It should not appear in your code.		

Table 1: The available labels are [2, page 7].

3 Tasks

3.1 Form Groups **Before Feb 24th, 23:59**

Mail wannes.meert@cs.kuleuven.be whether you work alone on this project or in a team of two. In the latter case, mention both team member names.

3.2 Literature Study and Experimental Setup

Familiarize yourself with machine learning techniques used to classify webpages and texts [3, 1]. Get a feeling for what has been done before in this context by searching for relevant terms[1]. Divide the work between the two people in the team.

3.3 First Report and Data Collection **Before March 24th, 23:59**

In a first phase of the project you explain in a brief report how you will address the task at hand and label the dataset available on Toledo.

Report 1: Mail a report (PDF, ∼2 pages) to the aforementioned email address in which you:

- Describe what **literature** you have read and what you have learned from it. This is not an annotated bibliography but a critical analysis of the relationship among the work and your project[2].
- Describe the overall **pipeline** you have in mind.
- Compile a list of **research questions** that you want to answer in the final report and link them to your pipeline. These have to include:
  - Which data representation will you use (classes and features)?
  - Which machine learning model(s) and parameter settings (hyperparameters)?
  - How will you evaluate your classifier and hyperparameters choice (methodology and formal score functions)?

[1] scholar.google.com, www.semanticscholar.org, academic.research.microsoft.com

[2] http://www.writing.utoronto.ca/advice/specific-types-of-writing/literature-review

  - How will you evaluate which features have the most impact?
  - What is the computational and memory cost of preprocessing, learning and evaluating?
  - How will you incrementally update your classifier when additional domain names are labeled?
  - Can you include expert knowledge (e.g. a page with more than 1000 links is a scam)?
  - How will you handle outdated labels (e.g. fast changing categories like `pay-per-click`)?

The quality of this report influences your final score for the project, so try your best to come up with a good plan. After the reports are handed in, you will get feedback on your report, and, where necessary, we may give more concrete guidelines on how to proceed.

Data labels: Label a random selection in `domainlabels_empty.csv` and mail the document together with the report. For every category at least #labels in Table 1 should occur per person.

3.4 Final Report and Prototype **Before May 12th, 23:59**

Report 2: Write a report with your findings.

Mail your final report (PDF, ≤ 10 pages) describing details of your approach and the results obtained:

- Summarize your solution in an **abstract**.
- Clearly state the **problem statement**.
- Answer the **research questions** you formulated in the first report.
- Write out the **conclusions** you draw from your experiments together with a scientifically supported motivation for these conclusions.
- Be concrete and precise about methods, formulas and numbers throughout the text. A scientific text should allow for **reproducibility**.
- Report the total **time you spent** on the project, and how it was divided over the different tasks mentioned.

Prototype: Include the code you used to perform the machine learning and experiments with the final report (also if the final application runs in the cloud). Running and experimenting your application is part of the evaluation. You are free to use any mainstream programming language and any publicly available toolbox. Make sure that:

- The code can be executed from the command line using (include a README for details): `./classifydotbe domainnames_in.csv domainnames_out.csv` (or `classifydotbe.cmd`). You can print extra information to the terminal or add extra columns to the output csv-file.
- The default setting is to use the model and parameters you have found to perform the best.
- Your code is self-contained and includes all dependencies.
- Your code is print-friendly (e.g., max 80 columns).

3.5 Peer assessment **Before May 12th, 23:59, individually**

Send by email a peer-assessment of your partner's efforts. This should be done on a scale from 0-4 where 0 means "I did all the work", 2 means "I and my partner did about the same effort", and 4 means "My partner did all the work". Add a short motivation to clarify your score. This information is used only by the professor and his assistants and is not communicated further.

3.6 Discussion **Week of May 22th**

There will be an oral discussion of your project report where you will also get the chance to demo your prototype. The instructors may try a certain set of actions to assess the adaptivity of your system.

3.7 What is (not) allowed?

The goal of this project is for the students to obtain hands-on experience with machine learning, and to deepen their insight in some of the topics taught in the machine learning course. The goal of the evaluation is to assess to what extent this is the case, for each individual student. The evaluation of the project is based on the report, the prototype, the score on the test set compared to other teams, the peer assessment, and the oral discussion. From this point of view, the following should be obvious, but we state it nevertheless.

Collaboration is only allowed *within* teams. Each team must work independently. You are not allowed to share source code with other teams. Within teams, the contributions of each team member must be represented fairly.

The authorship of each piece of the source code and the report must be clear and unambiguous. If parts of the code have been taken from elsewhere (e.g., copied from the internet), this must be indicated very clearly in the code. The report must provide a clear view on what has been copied from elsewhere, and what is your own work. The report itself must adhere to general scientific standards of source attribution. This means: if an idea is based on someone else's work, refer to that work. If a diagram, picture, text, ... is copied from elsewhere, refer to the source and clearly mention what the relationship with the original source is ("copied from ...", "based on ...", "inspired by ..."). More details on this can be found at http://eng.kuleuven.be/studenten/masterproef/masterproef/workshops/7november2016eng.pdf.

Any misrepresentation of your contribution to the project, deliberately or not, may be cause for sanctions, some of which can be severe.

Questions

Please direct any questions that you may have about the project to the Toledo forum or the classroom discussion moments.

Good luck!

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

[2] DNS Belgium. *Annual Report*. 2015. URL: https://dnsbelgium.be/sites/default/files/generated/files/documents/cijfers_DNS_2015_DEF2_NL[1].pdf.

[3] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of Tricks for Efficient Text Classification". In: *preprint arXiv:1607.01759* (2016).