Website Classification

Optional Semester Project in Data Science COM-508

Michael Allemann

May 28, 2019

EPFL

Overview

- Preprocessing
- Preparation for classification
- Classification
- Postprocessing

Preprocessing

Json to csv

- Read json using Pandas.
- Drop where snippet or title is NaN.
- Merge snippet and title to text.
- Split args to first and drug.
- Use link as index.
- Keep display link.
- 510’192 rows.

Unique display link

- Index display link.
- Count occurences of display link and save as count.
- Sort by count.
- Keep last occurence of text.

Remove most popular websites

- Extract most frequently visited websites from Alexa.
- Remove display links containing one of the websites.
- Remove all duplicate rows.

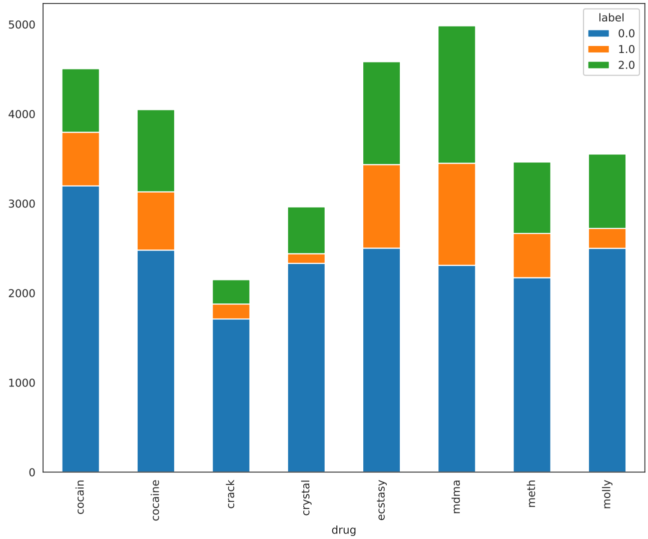
Labelling

- label 500 websites with highest count.
- Merge labelled column with original csv file.
- 0 → false positive
- 1 → true positive
- 2 → dead link

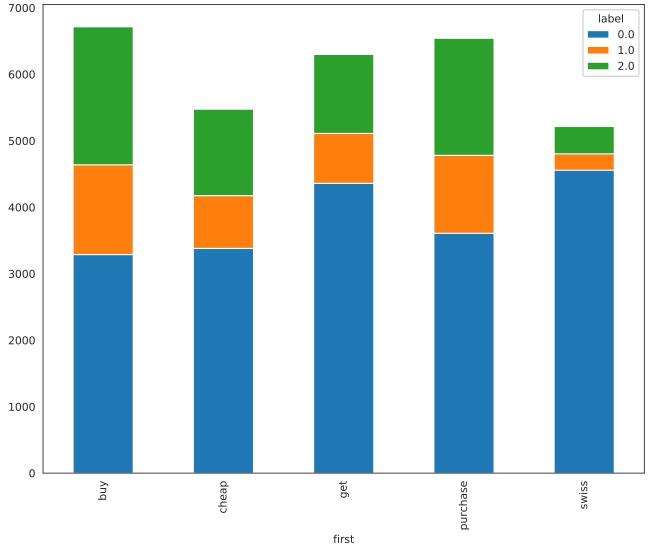
Expand labelling

- Merge labelled column with original csv file.
- Index link.
- Expanded labelled data to 101320.

Drug to label distribution



First to label distribution



Preparation for classification

Split into labelled and unlabelled data

- Drop labels $= 2$.
- 30246 labelled websites for learning.
- 71074 unlabelled for prediction.

Scikit-learn

- Split labelled data into train and test set.
- 80% train, 20% test.
- Extract labels to vector.
- Extract text from train, test and unlabelled.
- Remove stop words.
- Vectorize text using tf-idf transformation.
- Select 10’000 best features using F-Value.
- Transform to PyTorch tensor.

Classification

Multilayer Perceptron

- Input layer 10'000 Nodes.
- Linear layer 64 Nodes, ReLU, Dropout 0.5.
- Linear layer 64 Nodes, ReLU, Dropout 0.5.
- Linear layer 64 Nodes, ReLU, Dropout 0.5.
- Output layer 1 Nodes, Sigmoid.

Multilayer Perceptron

- Optimizer: Adam.
- Learning rate: 0.001.
- 100 Epochs.
- Binary Cross Entropy.

Results

- Length of test set: 4700.
- Test samples with label 0: 3802.
- Test samples with label 1: 898.
- Test accuracy: 99.28 %
- False positives: 15.
- False negatives: 19.

Postprocessing

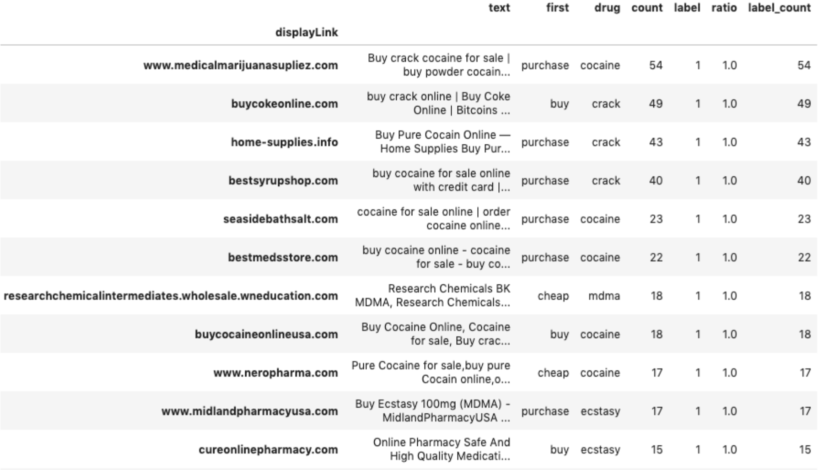
Unique display link

- Group by display link.
- Count labels.
- Count occurrences.
- Calculate ratio labels / occurrences.
- index display link.
- order by ratio and occurrences.

Predictions

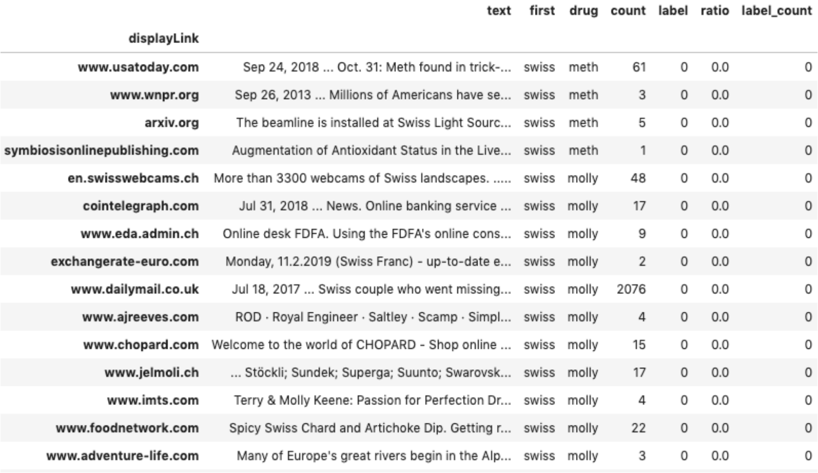
- Unlabelled display links: 14043.
- Predicted to be illegal web shop: 602.

Labelled as Webshop



displayLink	text	first	drug	count	label	ratio	label_count
www.medicalmarijuanasupliez.com	Buy crack cocaine for sale | buy powder cocain...	purchase	cocaine	54	1	1.0	54
buycokeonline.com	buy crack online | Buy Coke Online | Bitcoins ...	buy	crack	49	1	1.0	49
home-supplies.info	Buy Pure Cocain Online — Home Supplies Buy Pur...	purchase	crack	43	1	1.0	43
bestsyrupshop.com	buy cocaine for sale online with credit card |...	purchase	crack	40	1	1.0	40
seasidebathsalt.com	cocaine for sale online | order cocaine online...	purchase	cocaine	23	1	1.0	23
bestmedsstore.com	buy cocaine online - cocaine for sale - buy co...	purchase	cocaine	22	1	1.0	22
researchchemicalintermediates.wholesale.wneducation.com	Research Chemicals BK MDMA, Research Chemicals...	cheap	mdma	18	1	1.0	18
buycocaineonlineusa.com	Buy Cocaine Online, Cocaine for sale, Buy crac...	buy	cocaine	18	1	1.0	18
www.neropharma.com	Pure Cocaine for sale,buy pure Cocain online,o...	cheap	cocaine	17	1	1.0	17
www.midlandpharmacyusa.com	Buy Ecstasy 100mg (MDMA) - MidlandPharmacyUSA ...	purchase	ecstasy	17	1	1.0	17
cureonlinepharmacy.com	Online Pharmacy Safe And High Quality Medicati...	buy	ecstasy	15	1	1.0	15

Labelled as not Webshop



	text	first	drug	count	label	ratio	label_count
displayLink							
www.usatoday.com	Sep 24, 2018 ... Oct. 31: Meth found in trick-...	swiss	meth	61	0	0.0	0
www.wnpr.org	Sep 26, 2013 ... Millions of Americans have se...	swiss	meth	3	0	0.0	0
arxiv.org	The beamline is installed at Swiss Light Sourc...	swiss	meth	5	0	0.0	0
symbiosisonlinepublishing.com	Augmentation of Antioxidant Status in the Live...	swiss	meth	1	0	0.0	0
en.swisswebcams.ch	More than 3300 webcams of Swiss landscapes.	swiss	molly	48	0	0.0	0
cointelegraph.com	Jul 31, 2018 ... News. Online banking service ...	swiss	molly	17	0	0.0	0
www.eda.admin.ch	Online desk FDFA. Using the FDFA's online cons...	swiss	molly	9	0	0.0	0
exchangerate-euro.com	Monday, 11.2.2019 (Swiss Franc) - up-to-date e...	swiss	molly	2	0	0.0	0
www.dailymail.co.uk	Jul 18, 2017 ... Swiss couple who went missing...	swiss	molly	2076	0	0.0	0
www.ajreeves.com	ROD · Royal Engineer · Saltley · Scamp · Simpl...	swiss	molly	4	0	0.0	0
www.chopard.com	Welcome to the world of CHOPARD - Shop online ...	swiss	molly	15	0	0.0	0
www.jelmoli.ch	... Stöckli; Sundek; Superga; Suunto; Swarovsk...	swiss	molly	17	0	0.0	0
www.imts.com	Terry & Molly Keene: Passion for Perfection Dr...	swiss	molly	4	0	0.0	0
www.foodnetwork.com	Spicy Swiss Chard and Artichoke Dip. Getting r...	swiss	molly	22	0	0.0	0
www.adventure-life.com	Many of Europe's great rivers begin in the Alp...	swiss	molly	3	0	0.0	0