Website Classification

Optional Semester Project in Data Science COM-508

Michael Allemann

May 7, 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.
- 517'217 rows.

Unique display link

- Index display link.
- Count occurences of display link and save as count.
- Sort by count.
- Keep last occurence of text.
- 14586 rows.

Remove most popular websites

- Extract most frequently visited websites from Alexa.
- Remove display links containing one of the websites.
- 13330 rows.

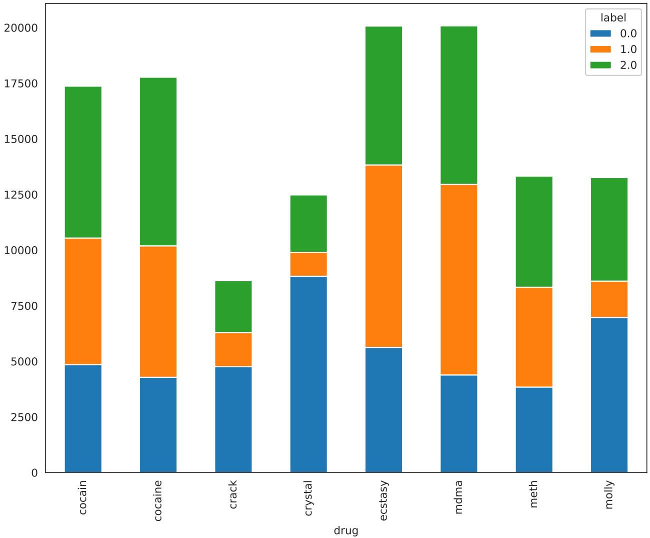
Labelling

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

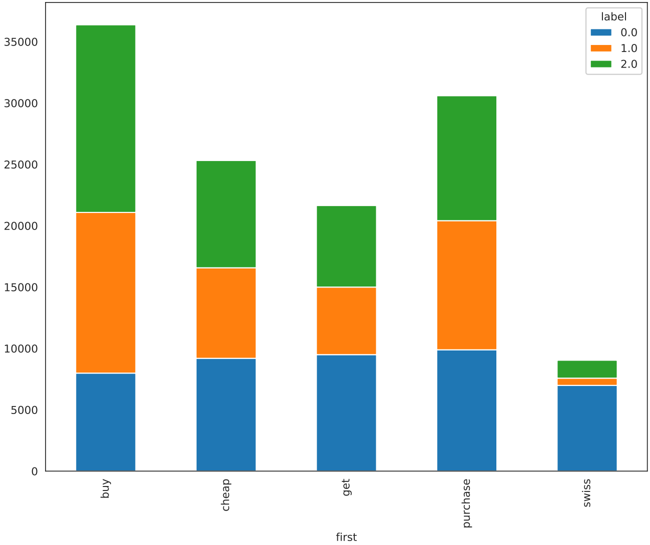
Expand labelling

- Merge labelled column with original csv file.
- Index link.
- Expanded labelled data to 122968.

Drug to label distribution



First to label distribution



Preparation for classification

Split into labelled and unlabelled data

- Drop labels $= 2$.
- 80638 labelled websites for learning.
- 387224 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.
- ReLU, Linear layer 32 Nodes, Dropout 0.2.
- ReLU, Linear layer 64 Nodes, Dropout 0.2.
- ReLU, Linear layer 32 Nodes, Dropout 0.2.
- Output layer 2 Nodes.

Multilayer Perceptron

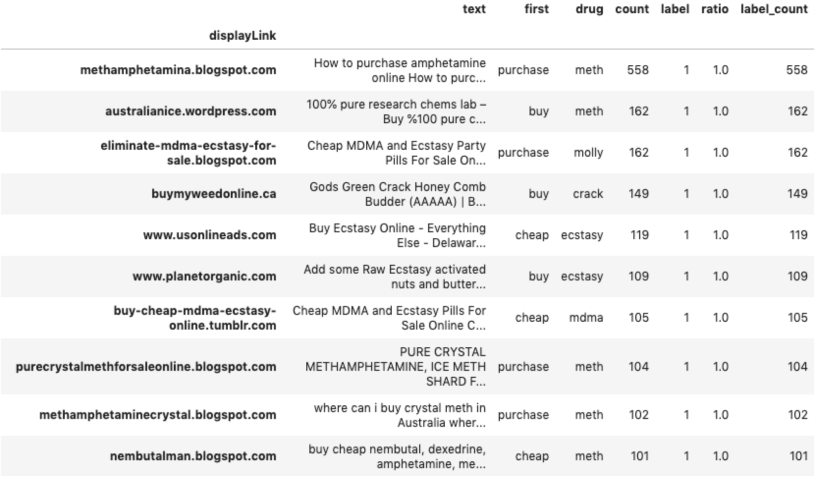
- Optimizer: Adam.
- Learning rate: 0.001.
- 100 Epochs.
- 6% test error.

Postprocessing

Unique display link

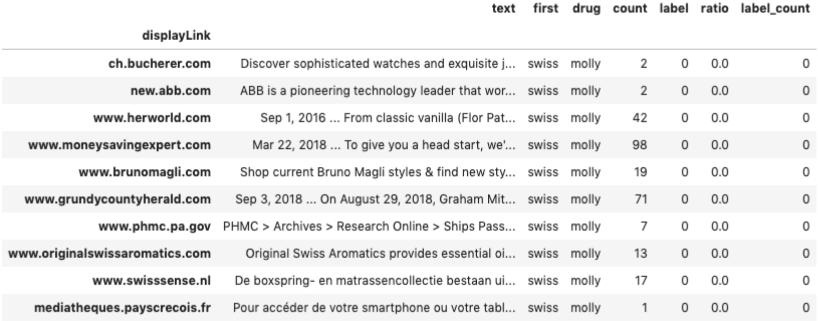
- Group by display link.
- Count labels.
- Count occurrences.
- Calculate ratio labels / occurrences.
- index display link.
- order by ratio and occurrences.

Labelled as Webshop



	text	first	drug	count	label	ratio	label_count
displayLink							
methamphetamina.blogspot.com	How to purchase amphetamine online How to purc...	purchase	meth	558	1	1.0	558
australianice.wordpress.com	100% pure research chems lab – Buy %100 pure c...	buy	meth	162	1	1.0	162
eliminate-mdma-ecstasy-for-sale.blogspot.com	Cheap MDMA and Ecstasy Party Pills For Sale On...	purchase	molly	162	1	1.0	162
buymyweedonline.ca	Gods Green Crack Honey Comb Budder (AAAAA) | B...	buy	crack	149	1	1.0	149
www.usonlineads.com	Buy Ecstasy Online - Everything Else - Delawar...	cheap	ecstasy	119	1	1.0	119
www.planetorganic.com	Add some Raw Ecstasy activated nuts and butter...	buy	ecstasy	109	1	1.0	109
buy-cheap-mdma-ecstasy-online.tumblr.com	Cheap MDMA and Ecstasy Pills For Sale Online C...	cheap	mdma	105	1	1.0	105
purecrystalmethforsaleonline.blogspot.com	PURE CRYSTAL METHAMPHETAMINE, ICE METH SHARD F...	purchase	meth	104	1	1.0	104
methamphetaminecrystal.blogspot.com	where can i buy crystal meth in Australia wher...	purchase	meth	102	1	1.0	102
nembutalman.blogspot.com	buy cheap nembutal, dexedrine, amphetamine, me...	cheap	meth	101	1	1.0	101

Labelled as not Webshop



	text	first	drug	count	label	ratio	label_count
displayLink							
ch.bucherer.com	Discover sophisticated watches and exquisite j...	swiss	molly	2	0	0.0	0
new.abb.com	ABB is a pioneering technology leader that wor...	swiss	molly	2	0	0.0	0
www.herworld.com	Sep 1, 2016 ... From classic vanilla (Flor Pat...	swiss	molly	42	0	0.0	0
www.moneysavingexpert.com	Mar 22, 2018 ... To give you a head start, we'...	swiss	molly	98	0	0.0	0
www.brunomagli.com	Shop current Bruno Magli styles & find new sty...	swiss	molly	19	0	0.0	0
www.grundycountyherald.com	Sep 3, 2018 ... On August 29, 2018, Graham Mit...	swiss	molly	71	0	0.0	0
www.phmc.pa.gov	PHMC > Archives > Research Online > Ships Pass...	swiss	molly	7	0	0.0	0
www.originalswissaromatics.com	Original Swiss Aromatics provides essential oi...	swiss	molly	13	0	0.0	0
www.swisssense.nl	De boxspring- en matrassencollectie bestaan ui...	swiss	molly	17	0	0.0	0
mediatheques.payscrecois.fr	Pour accéder de votre smartphone ou votre tabl...	swiss	molly	1	0	0.0	0