Wikipidea Articles Classification

Nolì Manzoni

October 15, 2020

Overview

Introduction

The goal of this project is to develop :

- A scraper for Wikipedia articles (10 categories)
- A model for classification
- Flask API (scraping/training and model querying)
- Docker container to expose the API and Jupyter notebook

Scraping

Wikipedia classes from main categories [1]

Scraping gets:

- Page paragraphs (content)
- Page title
- Page categories

Wikipedia's categories follow a hierarchy

[1] https://en.wikipedia.org/wiki/Category:Main_topic_classifications

Scraping

Tools:

- **Requests** HTTP library
- **Beautiful Soup** extract data out of HTML files

Slow process ... 11.07 seconds for 10 articles → multithreading [2]

Scraping 10 articles now takes 2.8 seconds!

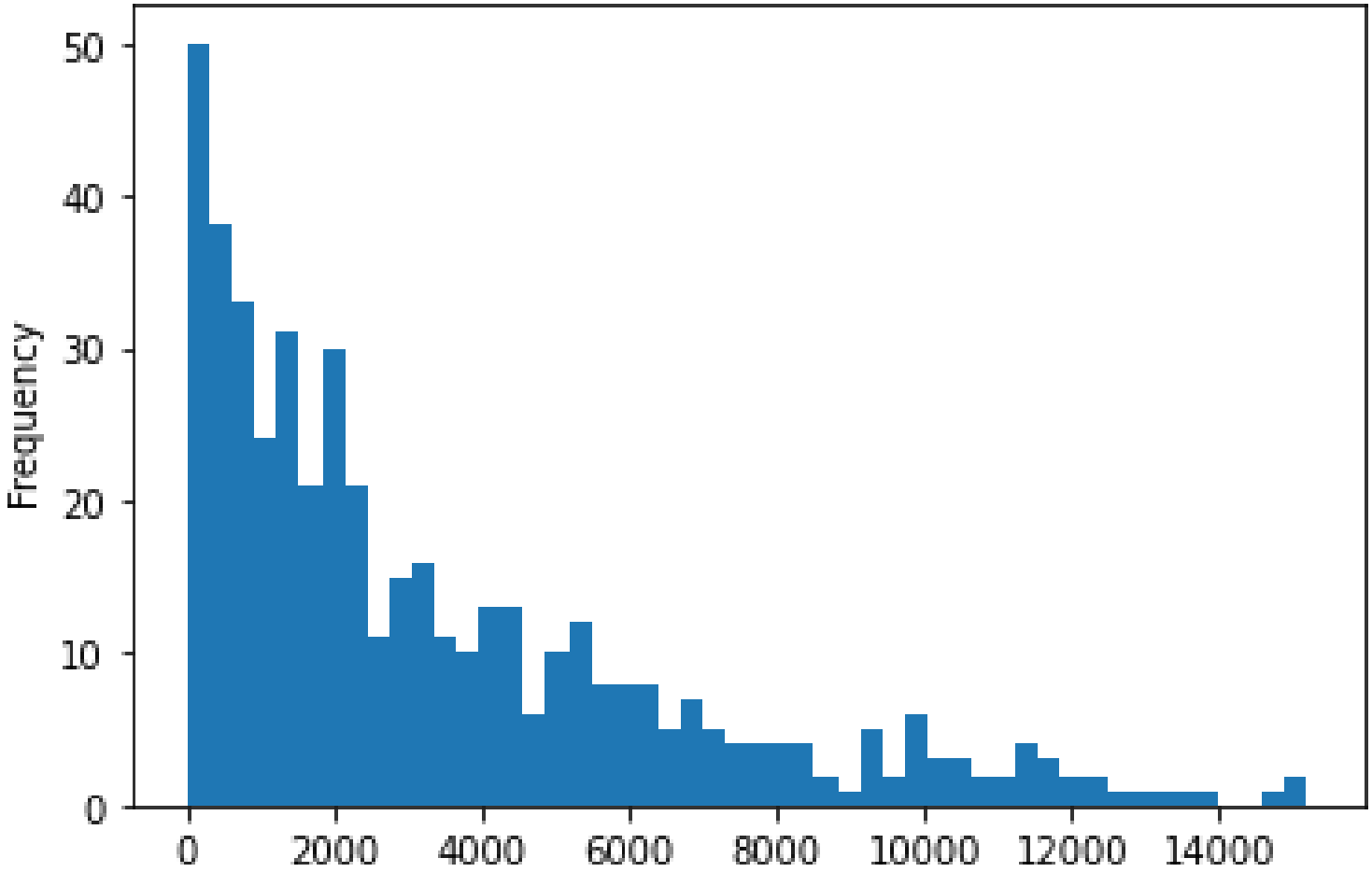
Scraping 50 articles for 10 categories takes ≈ **20 minutes** (467 articles)

[2] with ThreadPoolExecutor is an Executor subclass that uses a pool of threads to execute calls asynchronously.

Data Analysis

Wikipedia articles are quite long ... a mean of 3420 words

82% of the scraped articles contains more than 500 words



Model

BERT [3] Pytorch transformers for fine-tuning

BERT is limited to 512 tokens ... the dataset has a mean of 3420 words...

Sun et al. [4] → fine tuning for big text classification

Simple truncation is more powerful than more complex approaches

Method	IMDb	Sogou
head-only	5.63	2.58
tail-only	5.44	3.17
head+tail	**5.42**	**2.43**
hier. mean	5.89	2.83
hier. max	5.71	2.47
hier. self-attention	5.49	2.65

[3] https://arxiv.org/abs/1810.04805

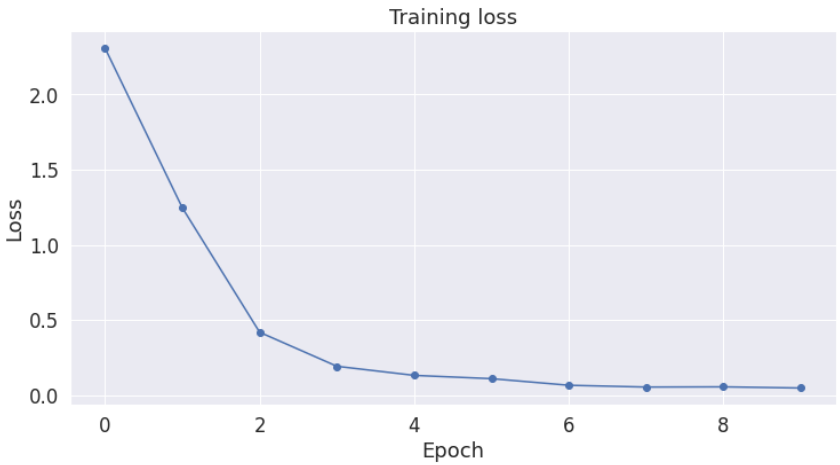
[4] https://arxiv.org/abs/1905.05583

Results

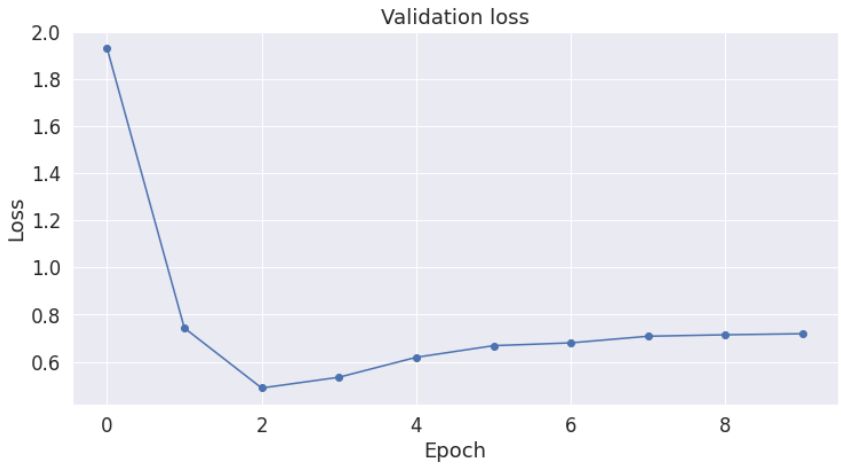
Dataset splitted in train (64%), validation (16%) and test (20%)

Model is fast to train ≈ 5 minutes

Validation accuracy of **88%**



(a) Train loss



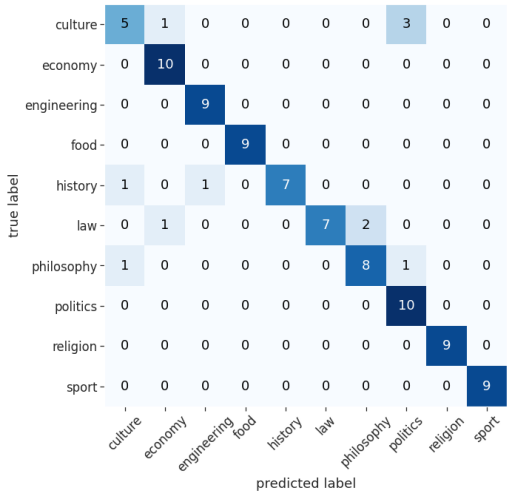
(b) Validation loss

Results

Test accuracy of **88%**

	precision	recall	f1-score	support
culture	0.71	0.56	0.63	9
economy	0.83	1.00	0.91	10
engineering	0.90	1.00	0.95	9
food	1.00	1.00	1.00	9
history	1.00	0.78	0.88	9
law	1.00	0.70	0.82	10
philosophy	0.80	0.80	0.80	10
politics	0.71	1.00	0.83	10
religion	1.00	1.00	1.00	9
sport	1.00	1.00	1.00	9
accuracy			0.88	94
macro avg	0.90	0.88	0.88	94
weighted avg	0.89	0.88	0.88	94

(a) Test set classification report



(b) Test set confusion matrix

Good performance and errors are expected
→ culture and politics could be similar

Questions

Questions?