Presented and discussed at lab session on Friday, Apr 22

NPFL 054 – Homework #2 – assignment

A more realistic machine learning project

Barbora Hladká and Martin Holub
Charles University, 2022

1. Authorship recognition – a machine learning task

- Categorization task – given a ***passage*** of text, you should recognize (predict) its ***author***
- You will be given a training ***collection of passages***, which will be your training data set – Czech authors, Czech texts, around 100 years old (mainly novels)
- In fact we try to recognize the ***author's style***, and we suppose that it could be used also for the authorship identification
- Important: there could be *topics* that are characteristic of a particular author(s) – however, we should NOT learn *topic recognition*, as we want to focus on the *author's style independently of the topics*

  ⟶ we will work with ***delexicalized texts***

2. Experimental data set

- **Origin of the data** – selected Czech classical books
    - provided by Czech National Library
    - several well known Czech authors, several whole books by each one
- **NLP processing** – from the raw data to the xml structure
    - the starting point = scanned original books + OCR output
    - OCR output cleaning – some noise removed (titles, page headers, etc), hyphens at the end of the lines eliminated
    - UDPipe processing
    - delexicalization
    - segmentation
    - random splitting to train+dev+test subsets
    - simple xml structure of passages
- **Elementary statistics**
  The data set contains text by 6 authors, 5 books per author. Each book was segmented into passages. There are two versions of the data set that differ in the average number of tokens in passages.

Number of passages and sentences/tokens in the data sets

Passage size (avg)	train	devel	test
1000	1352 67590 / 1352204	431 21542 / 431470	431 21505 / 431204
200	6751 67590 / 1352204	2152 21542 / 431470	2152 21505 / 431204

3. Your job

- You should work ***only with the provided data sets*** and you should NOT use any other language data sources
- **Data analysis and ML experiments – developing your ML model**
    - common practice: start with a simple (baseline) model, then try to tune and improve it – better feature (sub)sets, hyperparameters tuning
    - your ***main goal*** is to build a ***model for authorship prediction*** that will generalize to the (blind) test set
- **Development evaluation, and estimation of the generalization error**
- **Error analysis** – are there differences among different authors?
- **Presentation** of your work and results in the written form

4. Organization

- **Deadline for submission is *hard – May 18 (Wednesday, midnight)***
- Further communication with teachers
    - the main contact teacher is Martin Holub – available all the time except one week (will be out from Saturday Apr 30 to Sunday May 8)
    - more details can be communicated and discussed at two remaining lab sessions with MH – Fridays Apr 29 and May 13
    - each student can meet the teacher in person and talk about the work – possibly 2x 30 minutes until the submission deadline

5. Warning – the job is a bit time consuming

- Your assignment involves
    - some programming – data processing, feature engineering
    - running ML experiments, and their evaluation
    - presentation in the form of written report – describe you experiments, your motivation and expectations, explain and interpret your results carefully, making nice tables and charts
- ***Do not delay your work***, start as soon as possible
    - do not hesitate to contact the teacher if you have troubles
    - ***written report*** is a bit demanding as well, and it matters (!)

6. Technical details

- available data will be posted on the course web page
- how to submit your homework
    - make one .zip file with “everything” – it should include your codes, your outputs, your written report in pdf, and all sources that you used to complete the report
    - send it by email to holub@ufal.mff.cuni.cz
      by Wednesday May 18

7. Developing your model(s) – obligatory details

- obligatory: ***n-gram feature extraction*** (token level, n = 1, 2, 3)
    - of course, you can extract any other features from the available xml data – however, be ready to do the same also with the tests examples!
    - using n-gram features is both recommended and obligatory
- obligatory: ***analyse the distribution*** of the n-gram features
- obligatory: simple ***model based on sentence length*** distribution
    - can you recognize the authors by the number of tokens in their sentences?
    - optionally (a hint): similarly you can experiment with punctuation (full stops, commas, periods, question marks, etc) – its distribution within sentences or whole passages may be informative – should be experimentally explored
- obligatory: using ***SVM method***
    - SVM model is both recommended and obligatory
    - compare the performance of the SVM model using different amounts of n-gram features (= different feature set sizes)
    - try different ways to calculate feature values
        - simple n-gram count = ***term frequency*** (in a given passage)
        - normalized n-gram count = ***relative term frequency***
          = term frequency divided by the number of all terms in a given passage
        - ***weighted term frequency*** = tf * idf
          where idf = log(N / df),
          N is the total number of passages in the collection,
          df is the number of passages that contain a given term
    - of course, you can experiment with different ML methods as well, or combine them
- obligatory: evaluation on development test sets with different average passage size

8. Remarks on provided data sets and tools

- You will download the data sets in a clear xml format. Passages of type ‘devel’ should be used for your development tests. Each student will use either smaller, or bigger passages for training, according to teacher’s individual instruction. However, each student will evaluate his/her models on both shorter and longer test passages.
- If you like, you can use the attached tools, which work under linux (written for bash and perl). They can help to filter a set of passages (and make their subsets), and to make lists of n-grams from a given (sub)set of passages.

9. Final evaluation of your model using the blind test data set

- At last, after you submit your work, your best model will be evaluated using the blind test data set.
    - Your best model should be trained on all available data that you have (train+devel)
    - Format of the test passages is exactly the same as the format of your train and devel set, but attribute “author” is missing

      Example: `<passage pid="1784" length="1020" target_length="1000" type="test">`
- Be ready to take a set of test passages and predict their author. You should develop and submit a procedure that will do that.

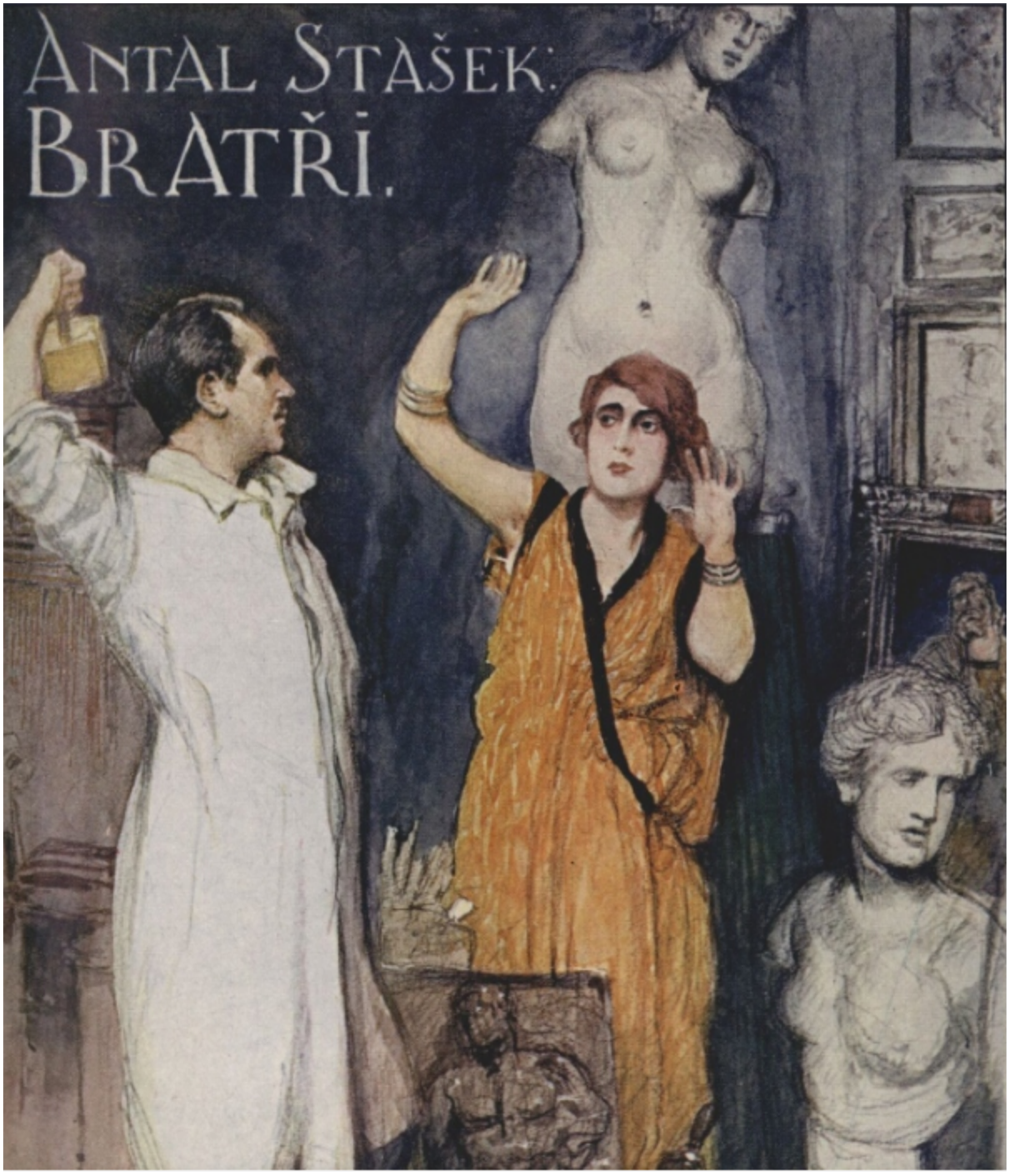
I WISH YOU GOOD LUCK!

M. H.

****

ILLUSTRATIONS

****



ANTAL STAŠEK

BRATŘI

II. VYDÁNÍ

Tázal se:

»Je to tajemství?«

»Ano, tajemství všech lidí: smrt!«

»Čí?«

»Mé dny jsou sečteny.«

»Matičko!« zvolal, a objav ji, jako by ji nechtěl dát smrti v plen, celoval ji po tvářích.

Umílený úsměv kmitl jasným mihem přes její tvrdý obličej. Řídké byly projevy lásky synovské, a proto světlem v jejích černých dnech.

»Ne, nezemřete; budete živa ještě mnoho, mnoho let.«

»Chci být dlouho živa jen v tvém srdci a ve tvých vzpomínkách. Chci mít i na tvých dílech, na tvých tvorbách podíl a v nich dále žít jako roditelka, jež zplodila nejen tělo, ale i tvou tvůrčí sílu a ukovala ji v pevný kov.«

Hluboce se zamyslil.

»Přemítáš o tom společném našem životě?«

»Ano . . . Ale nejvíce o tom tajemství, jež mne pronásleduje a visí nade mnou jako mlha všech mých dnů.«

»Vím to, vím.«

»Nepovíte mně ani nyní, když jsem dospěl v zralého člověka a kráčím vlastní drahou svou, kdo a kde je nemanželský otec můj?«

»Nepovím . . . povědět nesmím, nemohu.«

»Proč?«

»Mé tušení . . .«

Zalomil rukama:

»Zase tušení!«

»Býváme dálnovidkami my milující matky a zíráme jasně tam, kde pro jiné je čirá tma. Jsme neviditelnými nitkami s budoucností spojeny, našeptávající

One page OCR output

vedlo к cíli. Nebylo důkazů, stopy spáchaných činů byly setřeny. Ale dlouhá vyšetřovací vazba a stále stupňované rozčilování rozrušily mu nervy a to ostatní dokonali lékaři. V díle tom jim znamenitě pomáhal Polanův zápisník, jenž se jim dostal do rukou. Zapisoval do něho občas své příhody, ale ještě více své myšlenky. Bylo tam plno nejhoráznějších paradox a nejnebezpečnějších zásad, jaké se mohou vylíhnout jen v mozku nesmyslnými teoriemi sprazeném . . . Většina revolucionářů, at mladých, at starých, je přecpána ztřeštěnými nápady, jež se jim sypou z mozku jako drtiny z děravého pytle. Mívají neodolatelný pud, načmárat je na papír ... To bývá při stíhání velmi vydatnou pomůckou ... Ne socialisté, ale anarchisté, a mezi nimi i ti, kteří sami sebe nazývají anarchisty teoretickými a etickými, jsou největšími nepřáteli lidské společnosti, protože ze samých základů chtějí přetvořit její pořádky. Vyhlazovací boj proti nim je nejpřednější úlohou mezinárodní policie . . . Ano, ano, zápisníky« .,.
»Ale kde je Polanův?« tázal se Rudný.
»Tadyhle ho máme . . . vidíte . . . tuhle je . . . čtěte!«
Komisař vzal podávanou mu knížku, jež měla pozlacenou ořízku a byla vázána šagrenovou kůží, chvilku v ní obracel listy a pak spustil:

One page OCR output

Uchopil ji za ni a vtiskl na růžové prsty žhavý polibek.
Ano, ruku ... I o mou i o tvou jde.«
Nastaly snad prekážky?«
Nikoliv . . . Zůstává při tom, že máš v neděli dopoledne přijít a rodiče o mne požádat.«
»Proč má být tedy zle? ...«
Dověděla jsem se zrazeným tajemstvím, že ti budou položeny podmínky, které prý nebudeš ani chtít a snad ani moci vyplnit.«
Zarazil se, v chůzi se zastavil, sevřel pevněji Lidunčinu ruku, kterou dosud nepustil, a děl:
Ani moci, ani chtít? ... Má síla má své hranice; ale má vůle tebe dosáhnout je bez hranic. Záleželo-li by to jen na ní, budeš mon tak jistě, jako že slunečko právě zachází tamhle za borový háj.«
Dodal za chvíli:
»Marně přemýšlím o těch podmínkách. Matička ti neřekla?«
»Otec to tají i před matkou.. . Neboj se. Mně srdce říká, že konec bude vítězný. Ty jen přijď, a to dopoledne mezi jedenáctou a dvanáctou... Viď, že vezmeš cylindr a frak?«
»Nemohla bys mně ty věci odpustit?«
»Ne, miláčku, nemohu. Víš, jak otec dbá těch formalit. «
»Tak cylindr a frak mají naší lásce pomáhat?«
»Stejně jako kněz,« žertovala na rtech s risměvem.
»Unesu tě, nedají-li mně tě po dobrém,« vyrazil ze sebe zpola vážně, zpola žertovně.
»Kam?« zasmála se Lidunka a položila mu hlavu na rámě.
»Do krajin věčného jara, kde slunce hřeje žárněji, krev proudí rychleji, myšlenky pádí překotem, kde

UDPipe analysis – original output

Listopadové jitro rozbřesklo se nad Vídní. Rozbřesklo se?

```
<s>
	<w n="1" pos="ADJ" morph="AANS1----1A----"
msd="Case=Nom|Degree=Pos|Gender=Neut|Number=Sing|Polarity=Pos"
lemma="víistopadový">víistopadové</w>
	<w n="2" pos="NOUN" morph="NNNS1-----A----"
msd="Case=Nom|Gender=Neut|Number=Sing|Polarity=Pos" lemma="jitro">jitro</w>
	<w n="3" pos="VERB" morph="VpNS---XR-AA--1"
msd="Aspect=Perf|Gender=Neut|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act"
lemma="rozbřesknout">rozbřesklo</w>
	<w n="4" pos="PRON" morph="P7-X4----------"
msd="Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short" lemma="se">se</w>
	<w n="5" pos="ADP" morph="RR--7----------" msd="AdpType=Prep|Case=Ins"
lemma="nad">nad</w>
	<w n="6" pos="PROPN" morph="NNFS7-----A----"
msd="Case=Ins|Gender=Fem|NameType=Geo|Number=Sing|Polarity=Pos" lemma="Vídeň">Vídní</w>
	<pc n="7" pos="PUNCT" join="??" morph="Z:-------------" msd="" lemma=".">.</pc>
</s>
<s>
	<w n="1" pos="VERB" morph="VpNS---XR-AA--1"
msd="Aspect=Perf|Gender=Neut|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act"
lemma="rozbřesknout">Rozbřesklo</w>
	<w n="2" pos="PRON" morph="P7-X4----------"
msd="Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short" lemma="se">se</w>
	<pc n="3" pos="PUNCT" join="??" morph="Z:-------------" msd="" lemma="?">?</pc>
</s>
```

Segmentation into passages, sentences, and tokens
Simplified xml structure – full text, original word forms

Listopadové jitro rozbřesklo se nad Vídní. Rozbřesklo se?

```
<passage id="a-04.b-02.r-04.s-1000.p-0" length="1009" target_length="1000"
type="train">
        <s>
                <token>víistopadové</token>
                <token>jitro</token>
                <token>rozbřesklo</token>
                <token>se</token>
                <token>nad</token>
                <token>Vídní</token>
                <token>.</token>
        </s>
        <s>
                <token>Rozbřesklo</token>
                <token>se</token>
                <token>?</token>
        </s>
        <s>
                <token>Nevím</token>
                <token>,</token>
                <token>zda</token>
                <token>se</token>
                <token>snad</token>
                <token>na</token>
                <token>město</token>
                <token>nesvalilo</token>
                <token>,</token>
                <token>je</token>
                <token>tomu</token>
                <token>už</token>
                <token>dávno</token>
                <token>,</token>
                <token>skoro</token>
                <token>třicet</token>
                <token>let</token>
                <token>;</token>
                <token>ale</token>
```

Delexicalized xml – the working format

Listopadové jitro rozbřesklo se nad Vídní. Rozbřesklo se?

```
<passage id="a-04.b-02.r-08.s-1000.p-0" length="1009" target_length="1000"
type="train">
        <s>
                <token>POS_ADJ</token>
                <token>POS_NOUN</token>
                <token>POS_VERB</token>
                <token>se</token>
                <token>nad</token>
                <token>POS_PROPN</token>
                <token>.</token>
        </s>
        <s>
                <token>POS_VERB</token>
                <token>se</token>
                <token>?</token>
        </s>
        <s>
                <token>POS_VERB</token>
                <token>,</token>
                <token>zda</token>
                <token>se</token>
                <token>POS_ADV</token>
                <token>na</token>
                <token>POS_NOUN</token>
                <token>POS_VERB</token>
                <token>,</token>
                <token>POS_VERB</token>
                <token>ten</token>
                <token>POS_ADV</token>
                <token>POS_ADV</token>
                <token>,</token>
                <token>POS_ADV</token>
                <token>POS_NUM</token>
                <token>POS_NOUN</token>
                <token>;</token>
                <token>ale</token>
```