Introduction to Machine Learning
NPFL 054

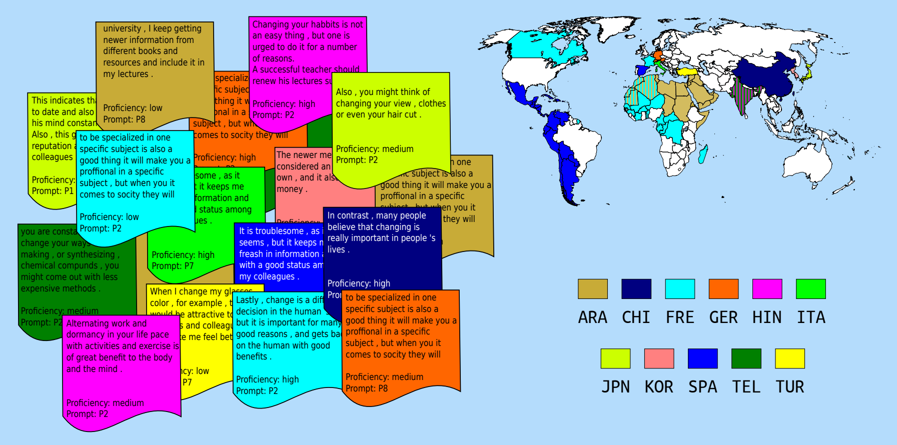
http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub**

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Native language identification task (NLI)



university , I keep getting newer information from different books and resources and include it in my lectures .

Proficiency: low
Prompt: P8

Changing your habbits is not an easy thing , but one is urged to do it for a number of reasons.
A successful teacher should renew his lectures su

Proficiency: high
Prompt: P2

specializ
fic subjec
thing it w
onal in a
subject , but wh
comes to socity they will

Proficiency: high

Also , you might think of changing your view , clothes or even your hair cut .

Proficiency: medium
Prompt: P2

This indicates th
to date and also
his mind constan
Also , this g
reputation a
colleagues

Proficiency:
Prompt: P1

to be specialized in one specific subject is also a good thing it will make you a proffional in a specific subject , but when you it comes to socity they will

Proficiency: low
Prompt: P2

some , as it
t it keeps me
formation and
l status among
ues .

Proficiency: high
Prompt: P7

The newer me
considered an
own , and it als
money .

Proficiency:

n one
ic subject is also a
good thing it will make you a
proffional in a specific
subject , but when you it
they will

you are consta
change your ways
making , or synthesizing ,
chemical compunds , you
might come out with less
expensive methods .

Proficiency: medium
Prompt: P2

It is troublesome , as i
seems , but it keeps n
freash in information a
with a good status am
my colleagues .

In contrast , many people believe that changing is really important in people 's lives .

Proficiency: high
Pro

When I change my glasses
color , for example , t
would be attractive to
s and colleagu
e me feel bet

cy: low
P7

Alternating work and dormancy in your life pace with activities and exercise is of great benefit to the body and the mind .

Proficiency: medium
Prompt: P2

Lastly , change is a diff
decision in the human
but it is important for many
good reasons , and gets ba
on the human with good
benefits .

Proficiency: high
Prompt: P2

to be specialized in one specific subject is also a good thing it will make you a proffional in a specific subject , but when you it comes to socity they will

Proficiency: medium
Prompt: P8

NLI

Identifying the native language (L1) of a writer based on a sample of their writing in a second language (L2)

Our data

- **L1s**: Arabic (ARA), Chinese (ZHO), French(FRA), German (DEU) Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), Turkish (TUR)
- **L2**: English
- **Real-world objects**: For each L1, 1,000 texts in L2 from The ETS Corpus of Non-Native Written English (former TOEFL11), i.e. $\mathit{Train} \cup \mathit{DevTest}$
- **Target class:** L1

More detailed info is available at the course website.

NLI

Topic
Most advertisements make products seem much better than they really are

Sample text
now a days the publisity is the best way to promoved a produt and if you wanth to sale a product you should bring some information that makes , that the people who is seeing the advertisements make sure that the product very good and in the future this person could buy it .

L1 = Spanish

tf-idf

Term **F**requency-**I**nverse **D**ocument **F**requency

- How important a word is to a document D in a collection C ($|C| = N$)?
- term frequency
  $\mathrm{tf}_{t,D}$ = the number of times a term t occurs in D (other possibilities exist)
- document frequency
  $\mathrm{df}_{t,D}$ = the number of documents in C in which a term t occurs, i.e., $|\{D \in C : t \in D\}|$
- inverse document frequency
  $\mathrm{idf}_{t,D} = \log N/\mathrm{df}_{\mathrm{t,D}}$

$$\mathrm{tfidf}_{t,D,C} = \mathrm{tf}_{t,D} \cdot \mathrm{idf}_{t,C}$$

Other variants of $\mathrm{tf}_{t,D}$ and $\mathrm{idf}_{t,D}$ exist.