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

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
- 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, term importance
  $\mathrm{idf}_{t,D} = \log \frac{N}{\mathrm{df}_{\mathrm{t,D}}}$

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