Natural Language Processing

Jiří Materna

Machine Learning College

@mlcollegecom

@mlcollegecom

#mlcollege

About me

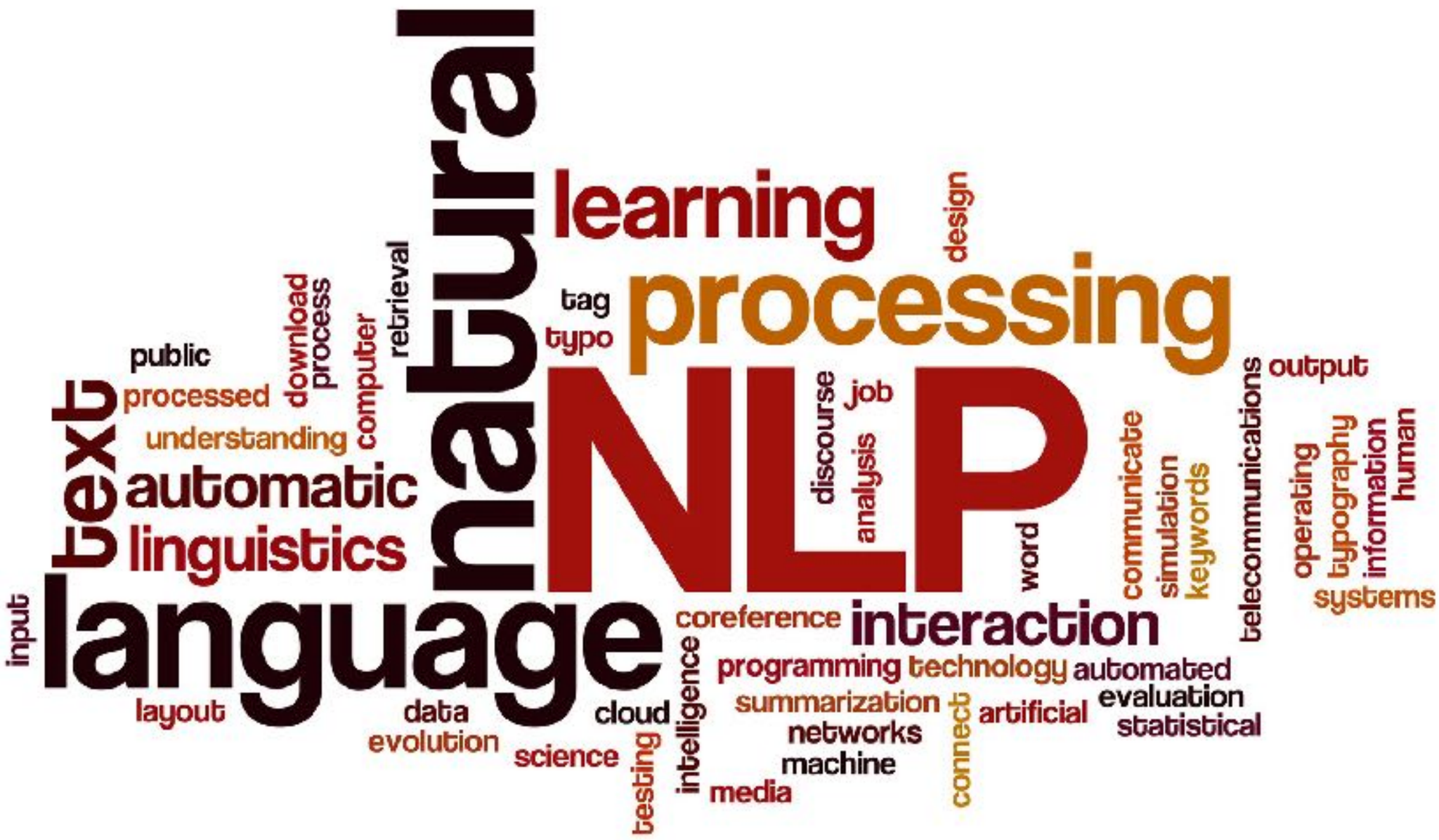
- Ph.D. in Natural Language Processing and Artificial Intelligence at Masaryk University
- 10 years at seznam.cz (last 8 years as Head Of Research)
- Founder and co-organiser of ML Prague
- Author of the ML Guru blog
- Mentor at StartupYard and Startup AI Incubator
- ML Freelacer and consultant

www.mlguru.com **www.mlprague.com** **www.mlcollege.com**

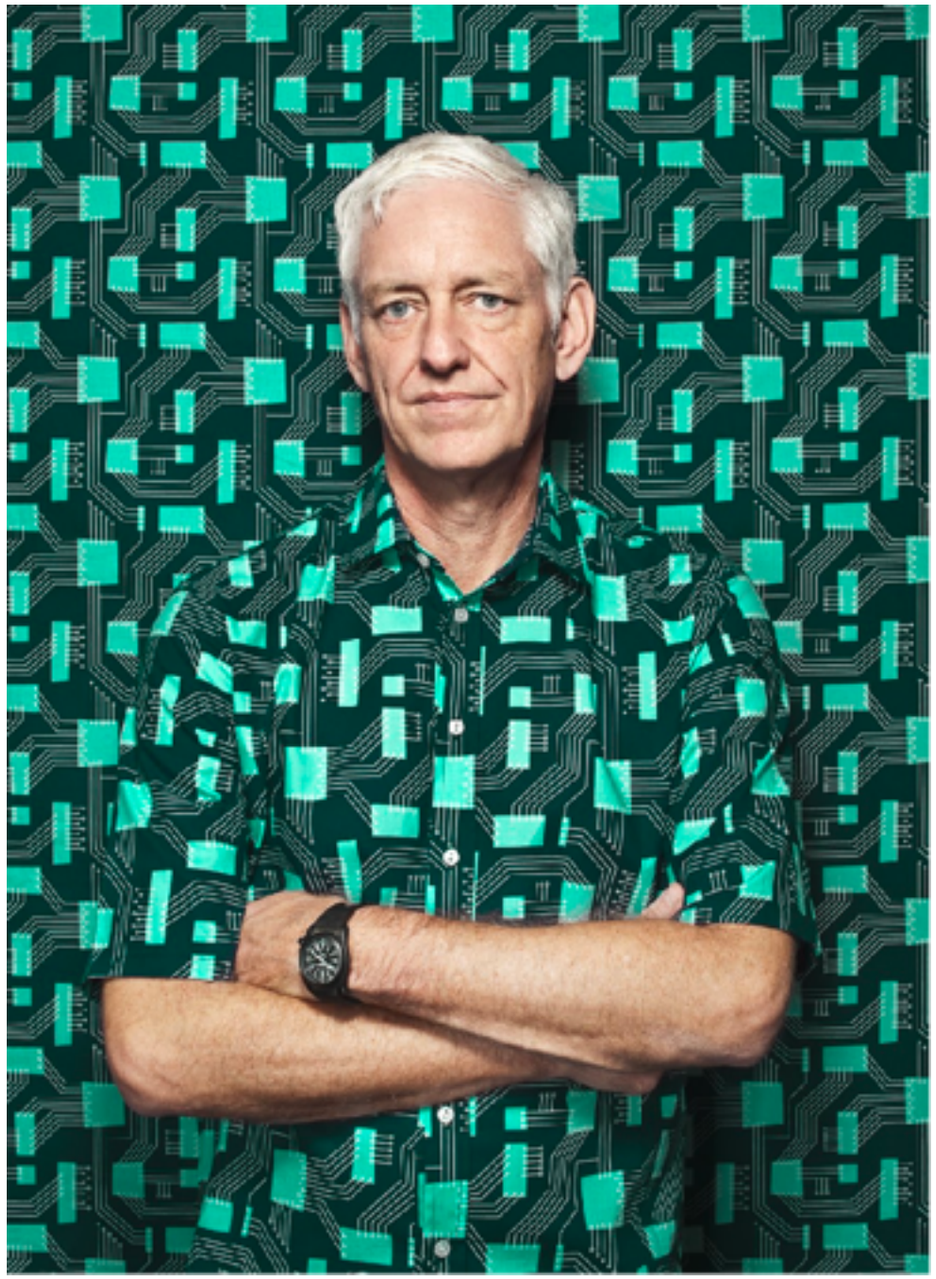
Outline

- Introduction to natural language processing
- Computational linguistics
- Text document vectorization
- Practical document classification task
- Language modeling
- Practical tasks on language modeling
- Word embeddings
- Text generating
- Practical tasks on language modeling

What is Natural Language Processing?



Norvig vs. Chomsky



source: https://www.commarts.com

source: https://citaty.net

Text corpus

Sketch Engine | natural language | British National Corpus

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Trends
Corpus info
My jobs
User guide

Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample

Query natural, language 255 (2.27 per million)

Page 1 of 13 Go Next | Last

J2K	nature of deixis (see Chapter 2 below) in	natural languages	, for sentences like (II) are true or false only
J2K	of the simple but immensely important fact that	natural languages	are primarily designed, so to speak, for use in
J2K	. The many facets of deixis are so pervasive in	natural languages	, and so deeply grammaticalized, that it is hard
J2K	the utterance, within the utterance itself.	Natural language	utterances are thus" anchored" directly to
J2K	semantics deals with certain	natural language	expressions. Suppose we identify the semantic
J2K	or self-referring expressions in	natural languages	, as in (12) and, arguably, in (13) (see Chapter 5
J2K	, is perhaps a philosophical red-herring.	Natural languages	, after all, just do have indexicals, and it is
J2K	. Semantics is then not concerned directly with	natural language	at all, but only with the abstract entities
J2K	to leave us with no term for all those aspects of	natural language	significance that are not in any way amenable to
J2K	of the deictic expressions that occur in	natural languages	, and we should now turn to consider linguistic
J2K	in familiar languages. Deictic systems in	natural languages	are not arbitrarily organized around the
J2K	. But this has the consequence, as we noted, that	natural languages	will only have a syntax and a pragmatics, and no
J2K	more or less directly on fragments of	natural language	(as initiated by Montague, 1974) would make
J2K	The semanticist who takes the other tack, that	natural language	senses are protean, sloppy and variable, is
J2K	offers a way out, for it allows one to claim that	natural language	expressions do tend to have simple, stable and
J2K	radical differences between logic and	natural language	seem to fade away. We shall explore this below
J2K	on what can be a possible lexical item in	natural languages	. Finally, the principles that generate
J0V	is meant any single document, or any stretch of	natural language	regarded as a self-contained unit for
J53	recognition and those that can understand	natural languages	, such as English, are known by the collective
HRK	through a dialogue, which approaches a	natural language	dialogue, or via a menu. In figure 6.2, the users

Page 1 of 13 Go Next | Last

Token & tokenization

This is a non-trivial English sentence: Ludolph's number is approx. 3.14.

Python library: http://www.nltk.org/

Stemming & lemmatization

Original	Stemming	Lemmatization
compensation	compens	compensation
compensations	compens	compensation
mouse	mous	mouse
mice	mice	mouse

Stemming & lemmatization

English:

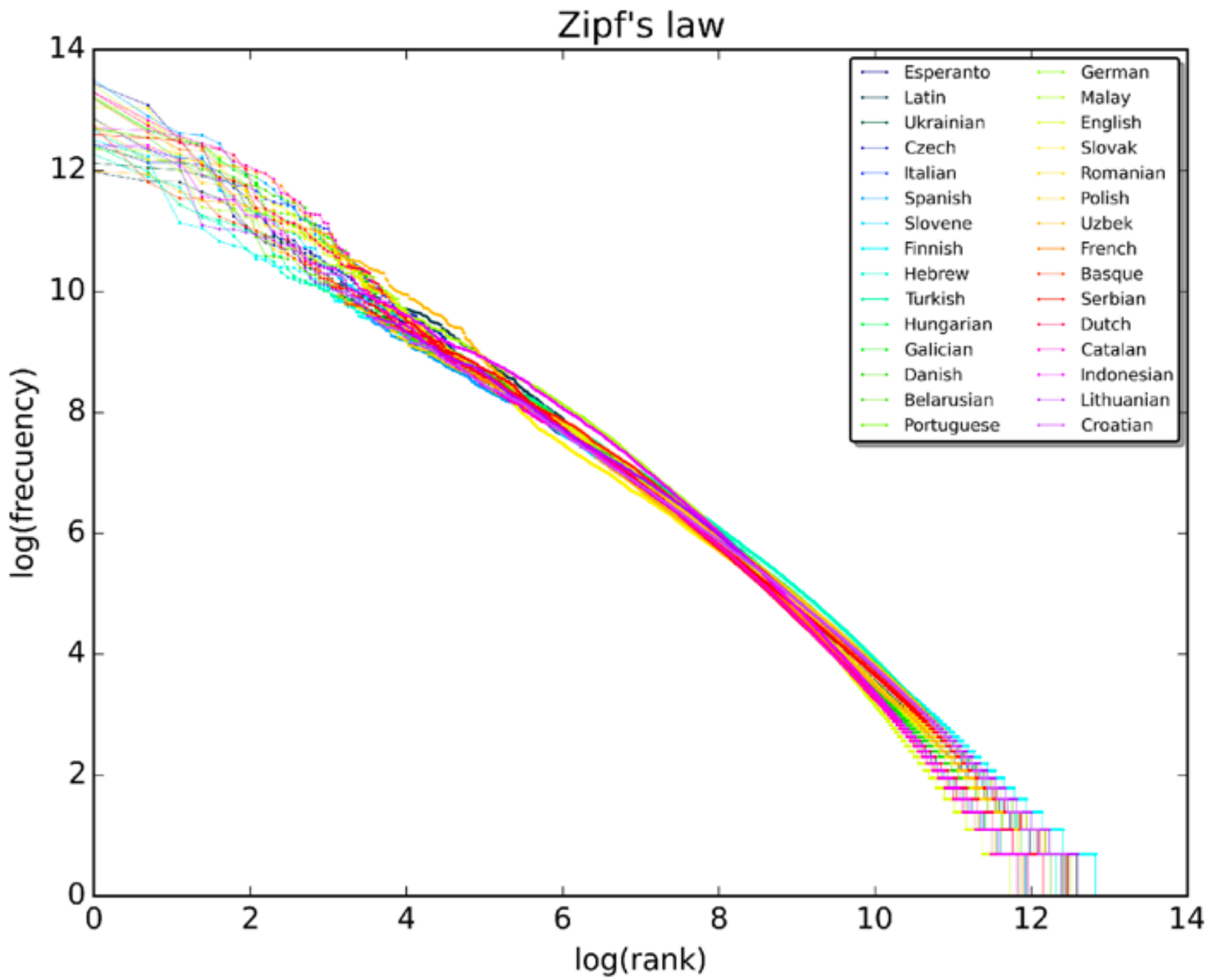
https://tartarus.org/martin/PorterStemmer/

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

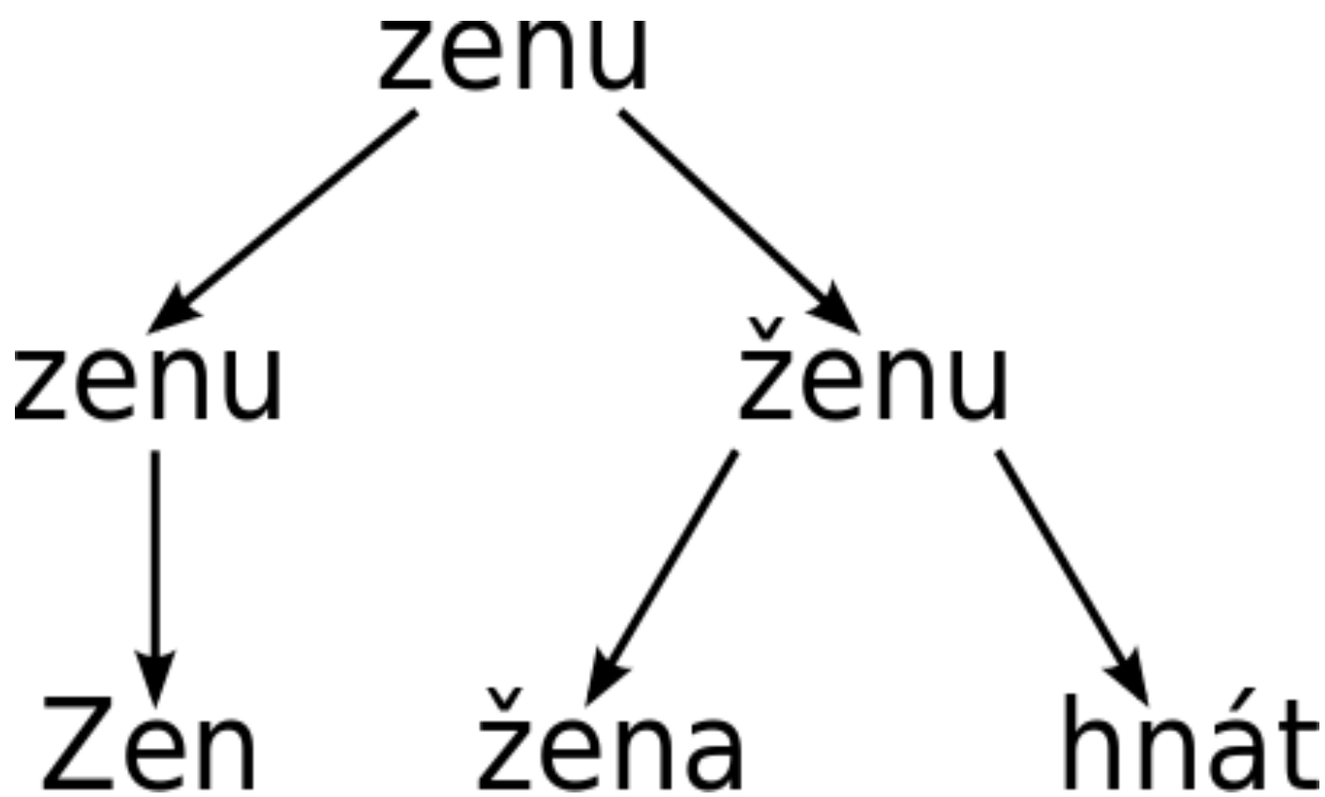
Czech:

http://ufal.mff.cuni.cz/morphodita

Zipf’s law & long tail

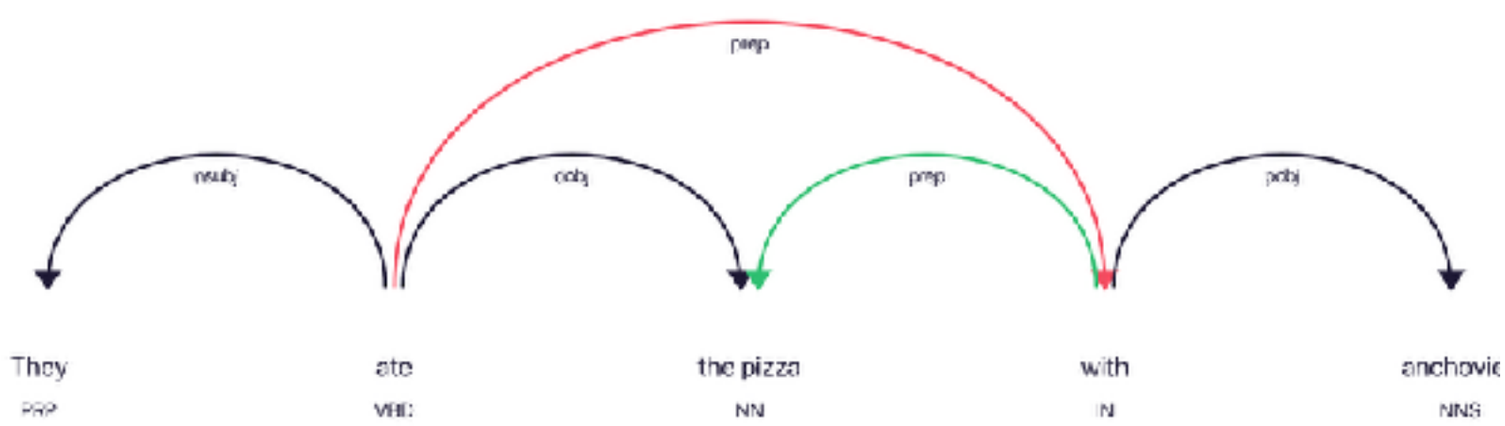


Ambiguity



Parsing

They ate the pizza with anchovies



source: https://explosion.ai/blog/syntaxnet-in-context

Publicly available corpora

British National Corpus: http://www.natcorp.ox.ac.uk/

Common Crawl: http://commoncrawl.org/the-data/get-started/

Wikipedia: https://dumps.wikimedia.org/

Feature extraction for NLP

1. *the man walked the dog*
2. *the man took the dog to the park*
3. *the dog went to the park*

[dog, man, park, the, to, took, walked, went]

1. [1, 1, 0, 1, 0, 0, 1, 0]
2. [1, 1, 1, 1, 1, 1, 0, 0]
3. [1, 0, 1, 1, 1, 0, 0, 1]

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

1. [1, 1, 0, 2, 0, 0, 1, 0]
2. [1, 1, 1, 3, 1, 1, 0, 0]
3. [1, 0, 1, 2, 1, 0, 0, 1]

1. [0, 0.18, 0, 0, 0, 0, 0.48, 0]
2. [0, 0.18, 0.18, 0, 0.18, 0.48, 0, 0]
3. [0, 0, 0.18, 0, 0.18, 0, 0, 0.48]

NLP Introduction task

http://localhost:9998/notebooks/01-text-classification-introduction.ipynb

Language models

- spell checking
- speech recognition
- machine translation
- …

n-gram models

$$P(w_1, w_2, \ldots w_n) = P(w_1)P(w_2|w_1)\ldots P(w_n|w_1, \ldots w_{n-1})$$

$$= \prod_i P(w_i|w_1, w_2 \ldots w_{i-1})$$

$$\approx \prod_i P(w_i|w_{i-k}, w_{i-k-1} \ldots w_{i-1})$$

$$P(w_i|w_{i-k}, w_{i-k-1} \ldots w_{i-1}) = \frac{count(w_{i-k}, w_{i-k-1} \ldots w_{i-1}, w_i)}{count(w_{i-k}, w_{i-k-1} \ldots w_{i-1})}$$

n-gram models — example

P(<s>, machine, learning, college, </s>) =

P(machine|<s>)P(learning | machine)P(college | learning).P(<s/>|college)

P(learning | machine) = count(machine, learning)/count(machine)

Language model smoothing

- Laplace smoothing (plus one)

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i) + 1}{count(w_{i-1}) + V}$$

- interpolation
- Good-Turing
- Witten-Bell
- …

Perplexity

$$PP(W) = P(w_1, w_2, \ldots, w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1, \ldots, w_{i-1})}}$$

$$= 2^{-\frac{1}{N} \sum_{i=1}^{N} log_2 P(w_i | w_1, \ldots, w_{i-1})}$$

Language detection using language models

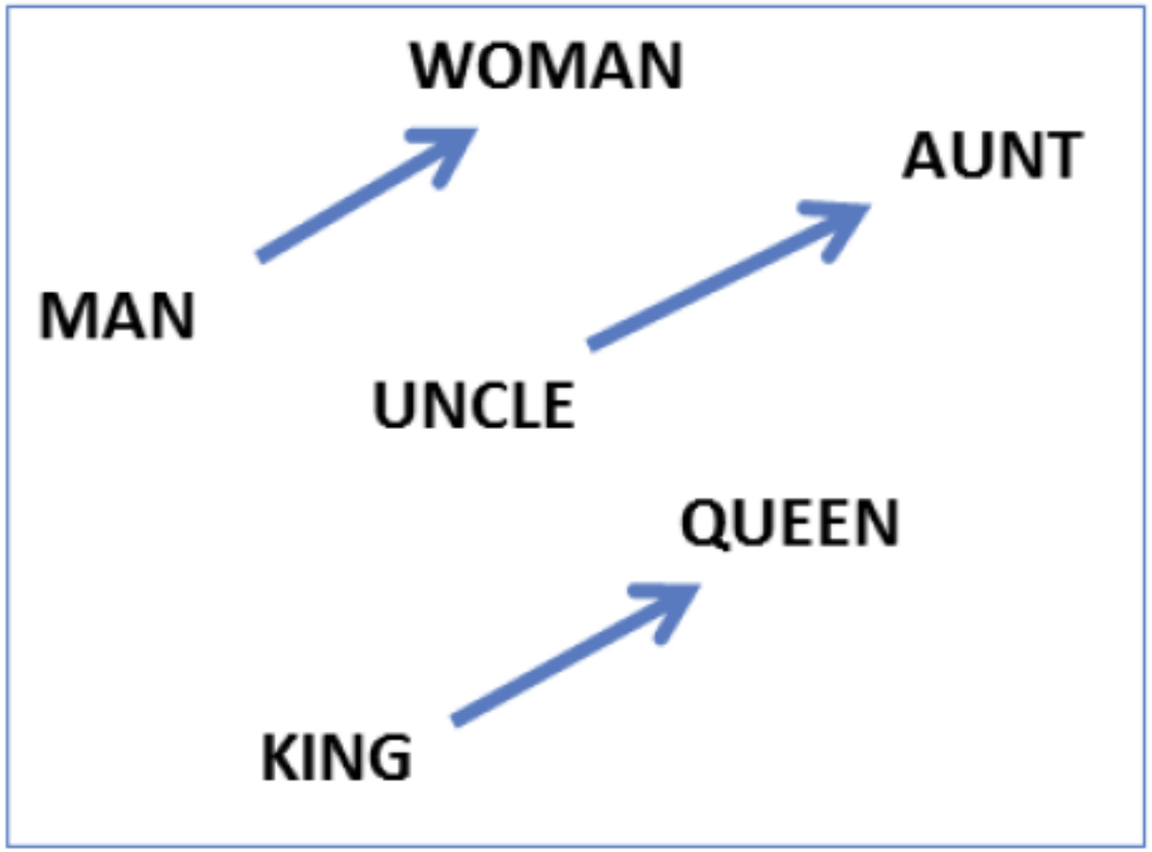
http://localhost:9998/notebooks/02-Language-detection-assignment.ipynb

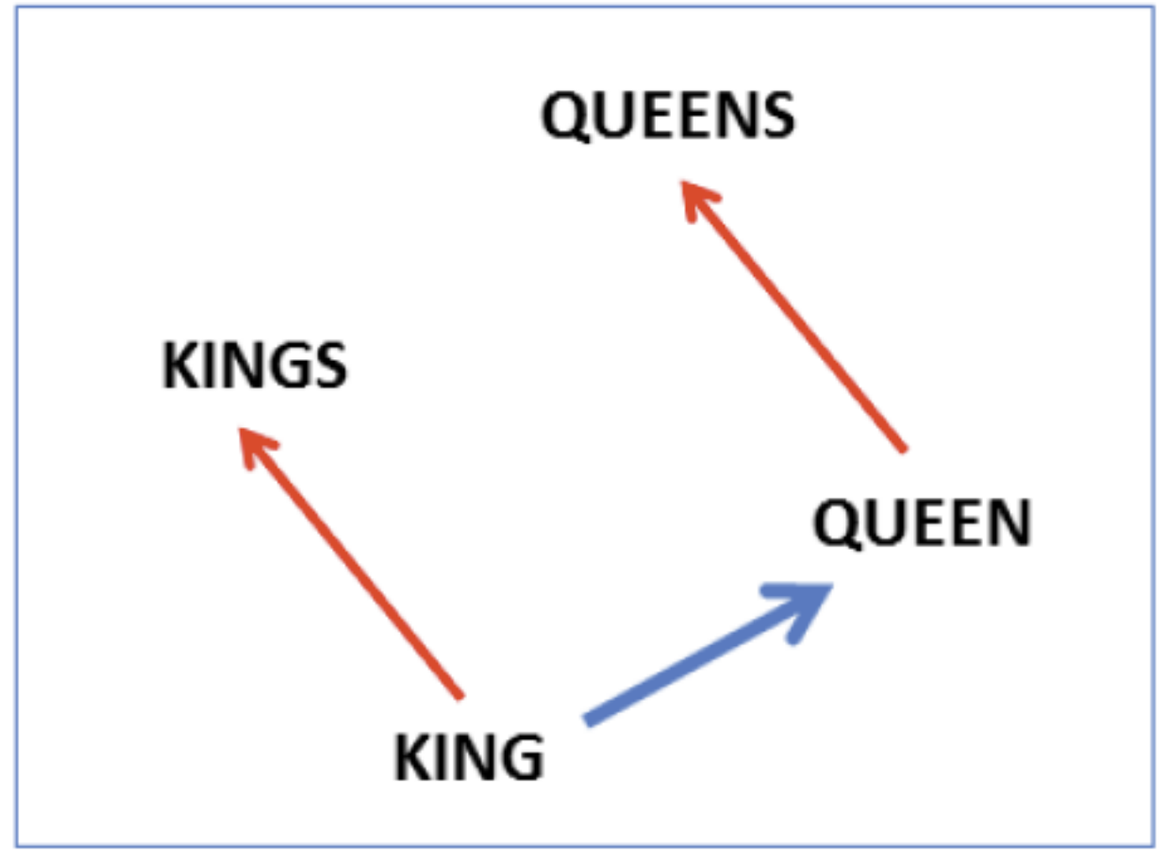
Travel agency review classification

http://localhost:9998/notebooks/03-Review-data-filtering-assignment.ipynb

http://localhost:9998/notebooks/04-Review-classification-assignment.ipynb

word2vec

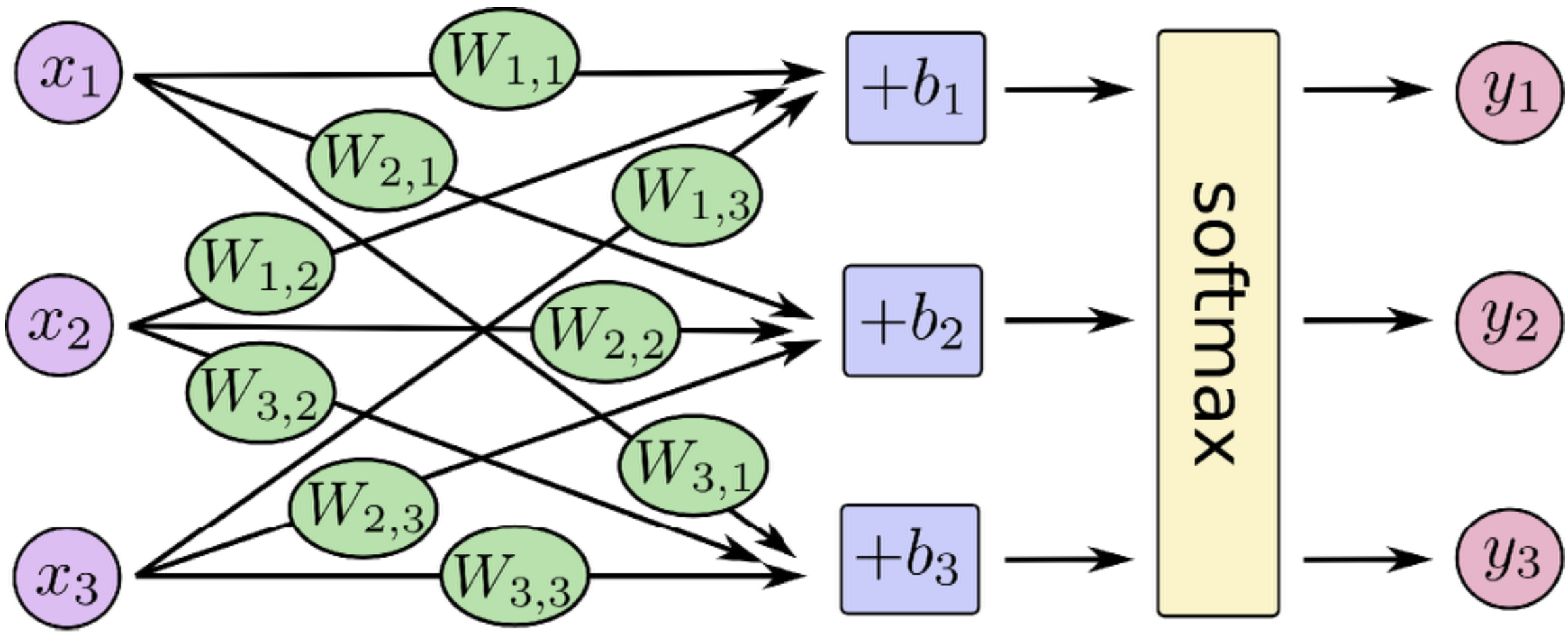




king is to **kings** as **queen** to ***?***.

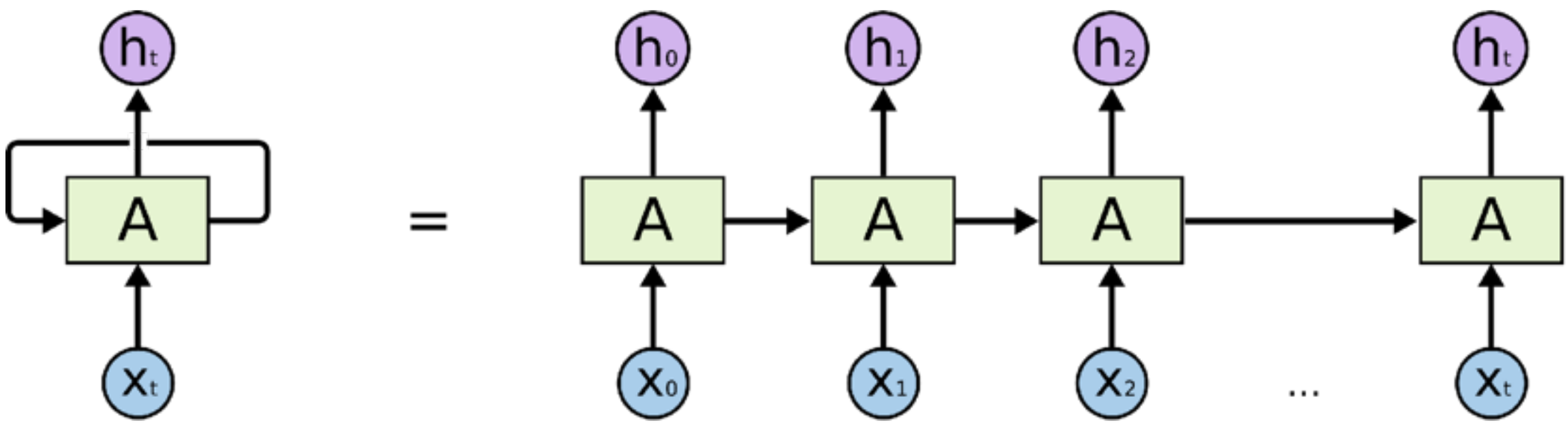
v(**kings**) - v(**king**) = v(**queens**) - v (**queen**)

Feed-Forward Neural Network



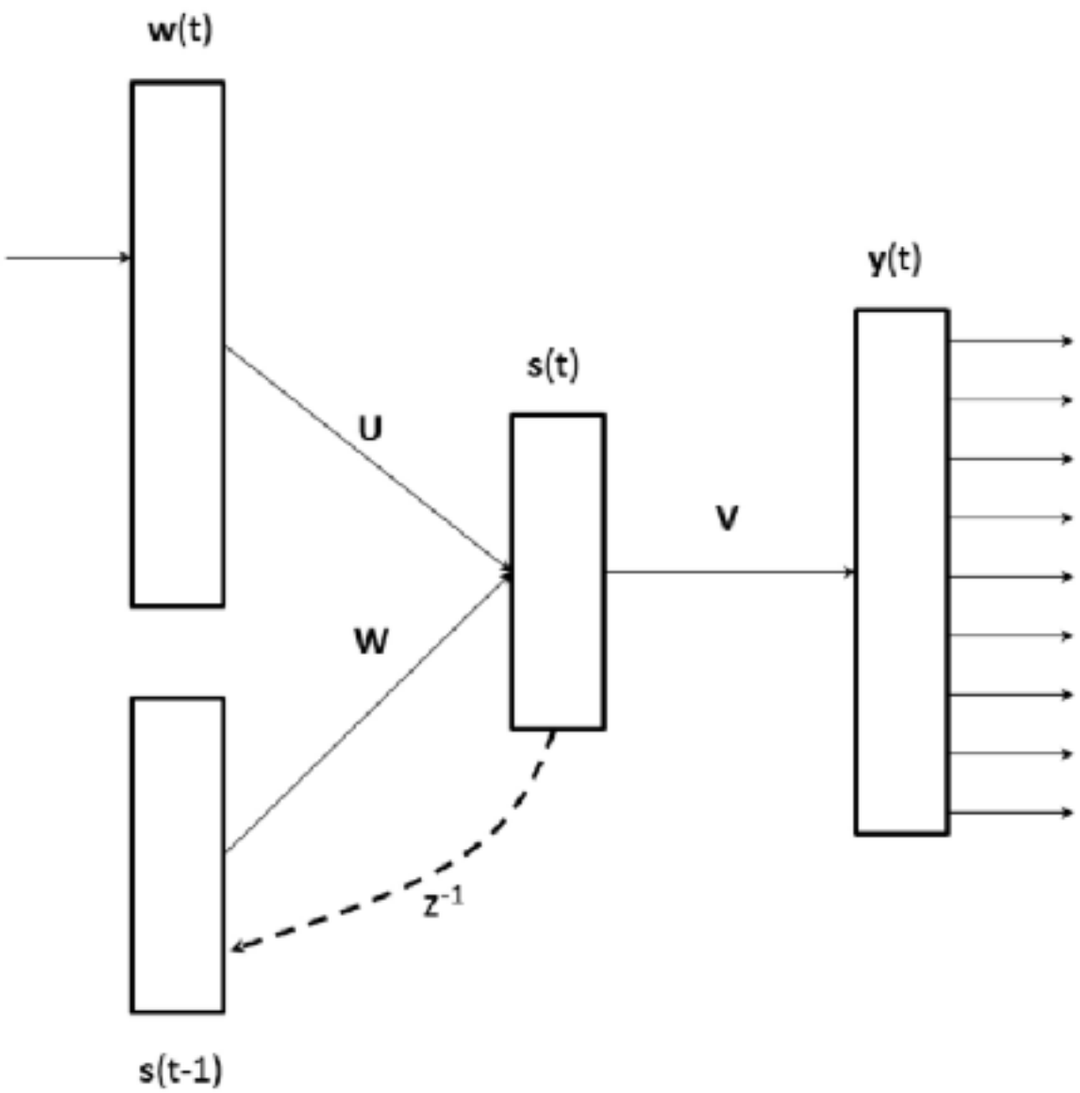
source: https://www.tensorflow.org

Recurrent Neural networks



source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent Neural Network Language Modeling Toolkit



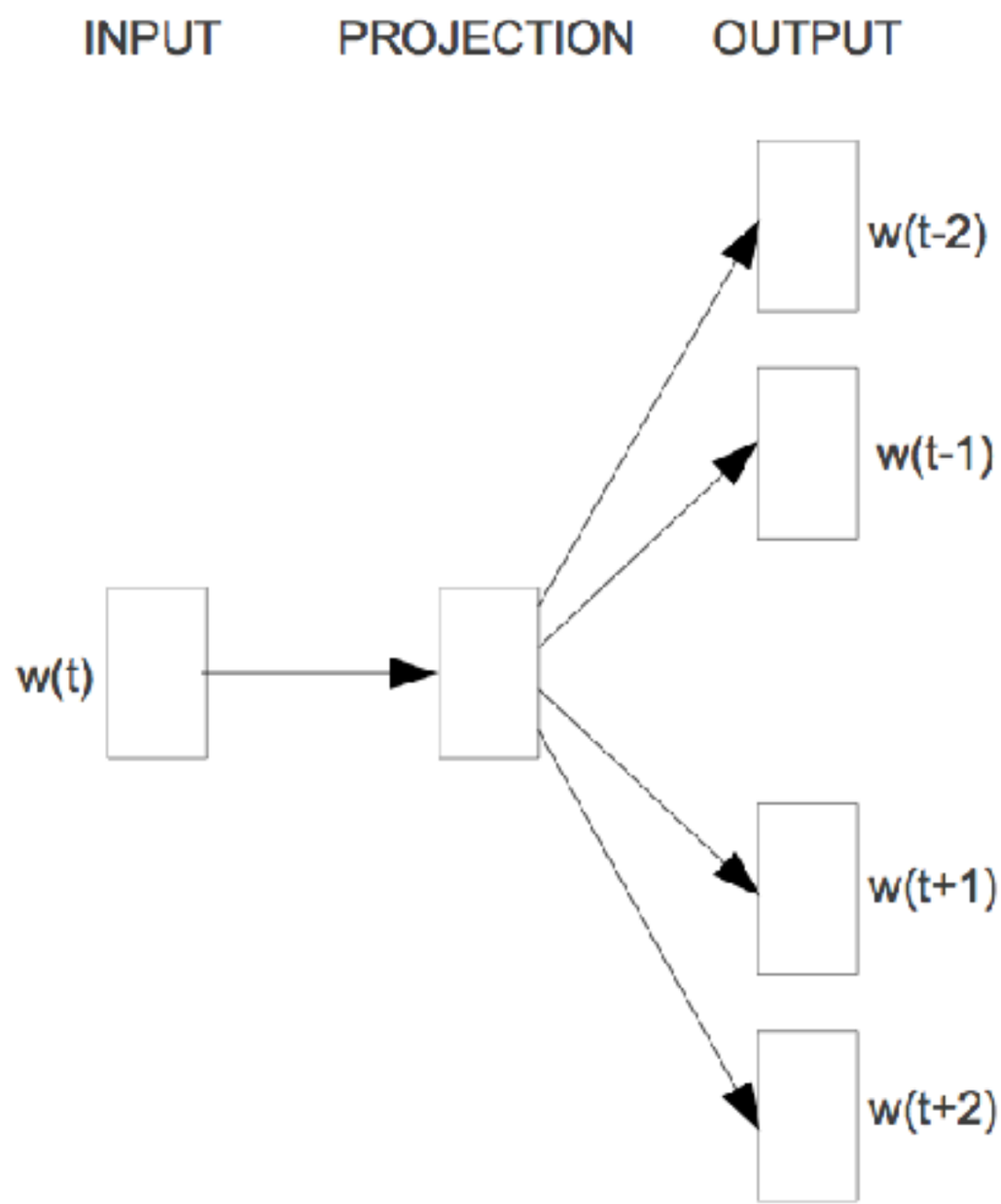
$$\mathbf{s}(t) = f\left(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1)\right)$$

$$\mathbf{y}(t) = g\left(\mathbf{V}\mathbf{s}(t)\right),$$

$$f(z) = \frac{1}{1+e^{-z}}, \quad g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}.$$

source: http://www.fit.vutbr.cz/~imikolov/rnnlm/

The skip-gram model



Experiments with word2vec

http://localhost:9998/notebooks/05-Word2vec-in-gensim.ipynb

http://localhost:9998/notebooks/06-Review-classification-w2v-assignment.ipynb

Language models for text generating

Nacházíte se: Úvod > Oddělení > Krásná literatura > Poezie > Česká a slovenská poezie > **Elektronická kniha Poezie umělého světa**

Ukázka pdf

Ukázka epub

Ukázka mobi

Like 180

Poezie umělého světa [E-kniha]

Jiří Materna

Hodnotilo 7 uživatelů, zatím žádné recenze, napsat vlastní recenzi

Popis: Elektronická kniha, 50 stran, bez zabezpečení DRM, ePUB, Mobi, PDF, česky - více

Stáhnout

Zdarma
K dispozici pro **okamžité** stáhnutí

Ke stažení

Anotace

Všechny básně v této knize byly automaticky vygenerovány počítačem za pomoci umělých neuronových sítích. Neuronová síť sama o sobě nic neumí a je třeba ji natrénovat pro činnost, kterou má vykonávat.

LISTOPAD

usínám, pláču, umírám, přemýšlím
co cítíš ty?
cítím tvou slabost
a whisky

SPRAVEDLNOST

na tvou dekadentní duši
ráno i v poledne
bůh má připravenou kuši

IMAGINACE

v pivu je poezie
jako jsou motýli v housenkách
popelník je pro prach
a strach

neboj se vidět a tvořit
spoutané srdce je hrob

Metafory

...tělo plné červánků...

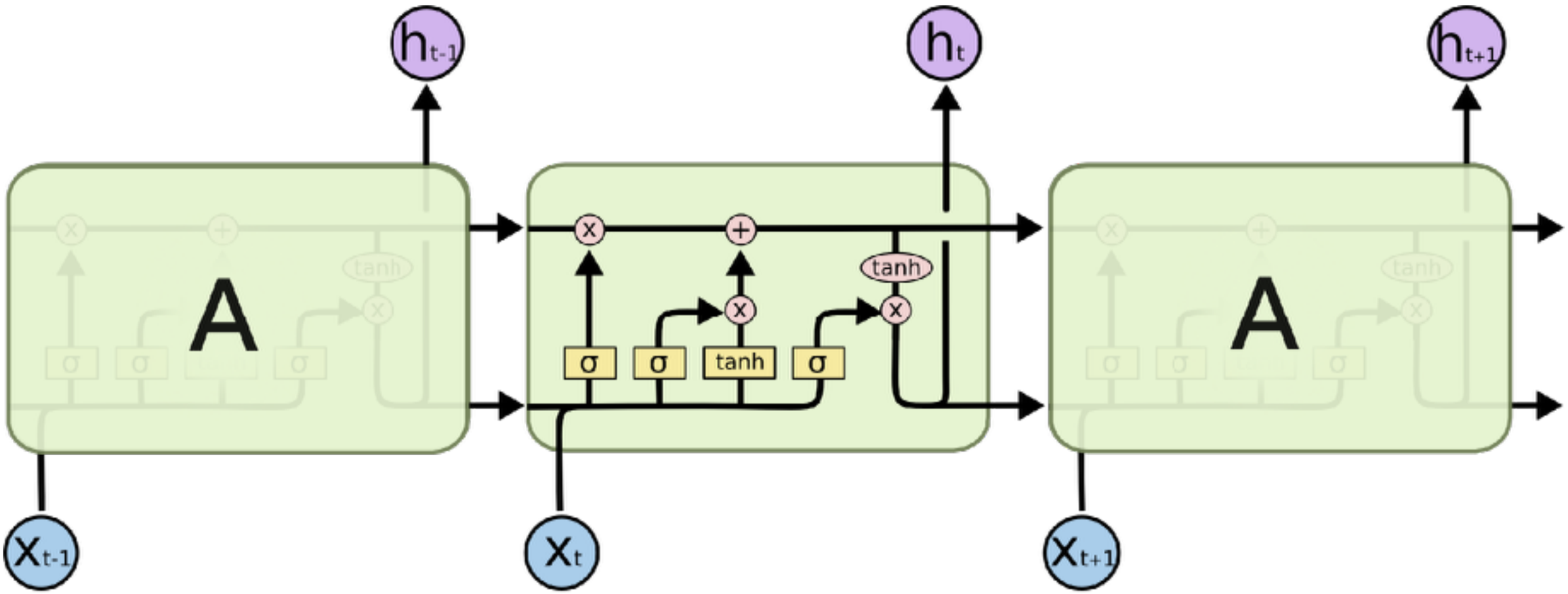
...tak vzácný jako listí...

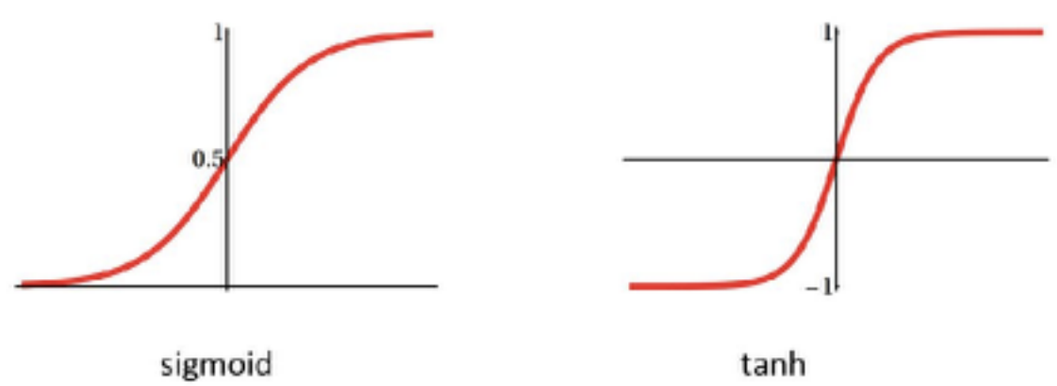
Language models for text generating

P(maso | máma, mele) = 0.5
P(Emu | máma, mele) = 0.3
P(tátu | máma, mele) = 0.2

```
t ~ Uniform(0, 1)
s = 0
for v in Vocabulary:
   s += v.prob
   if t < s:
      return v.word
```

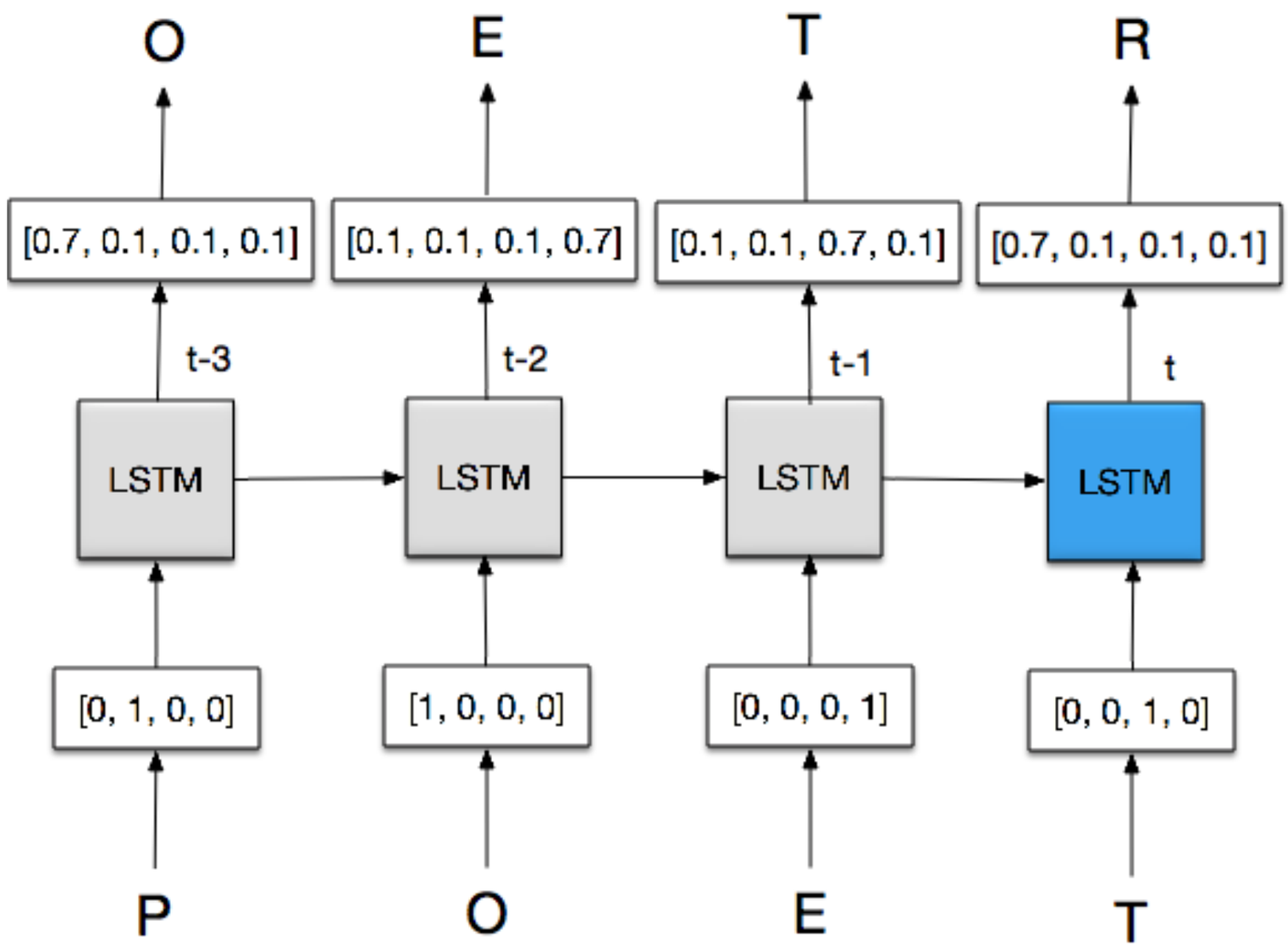
Long Short-Term Memory





Zdroj: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM language model



LSTM review generator

http://localhost:9998/notebooks/07-Review-generator.ipynb

What next?

Machine Learning Prague

MACHINE LEARNING meetups

http://www.mlguru.com

Thank you for your attention

e-mail: jiri@mlguru.com

Web: www.mlguru.com

Twitter: @JiriMaterna

Facebook: https://www.facebook.com/maternajiri

LinkedIn: https://www.linkedin.com/in/jirimaterna/