Introduction to Natural Language Processing and Time Series for Raiffeisenbank International

Jiří Materna

About me

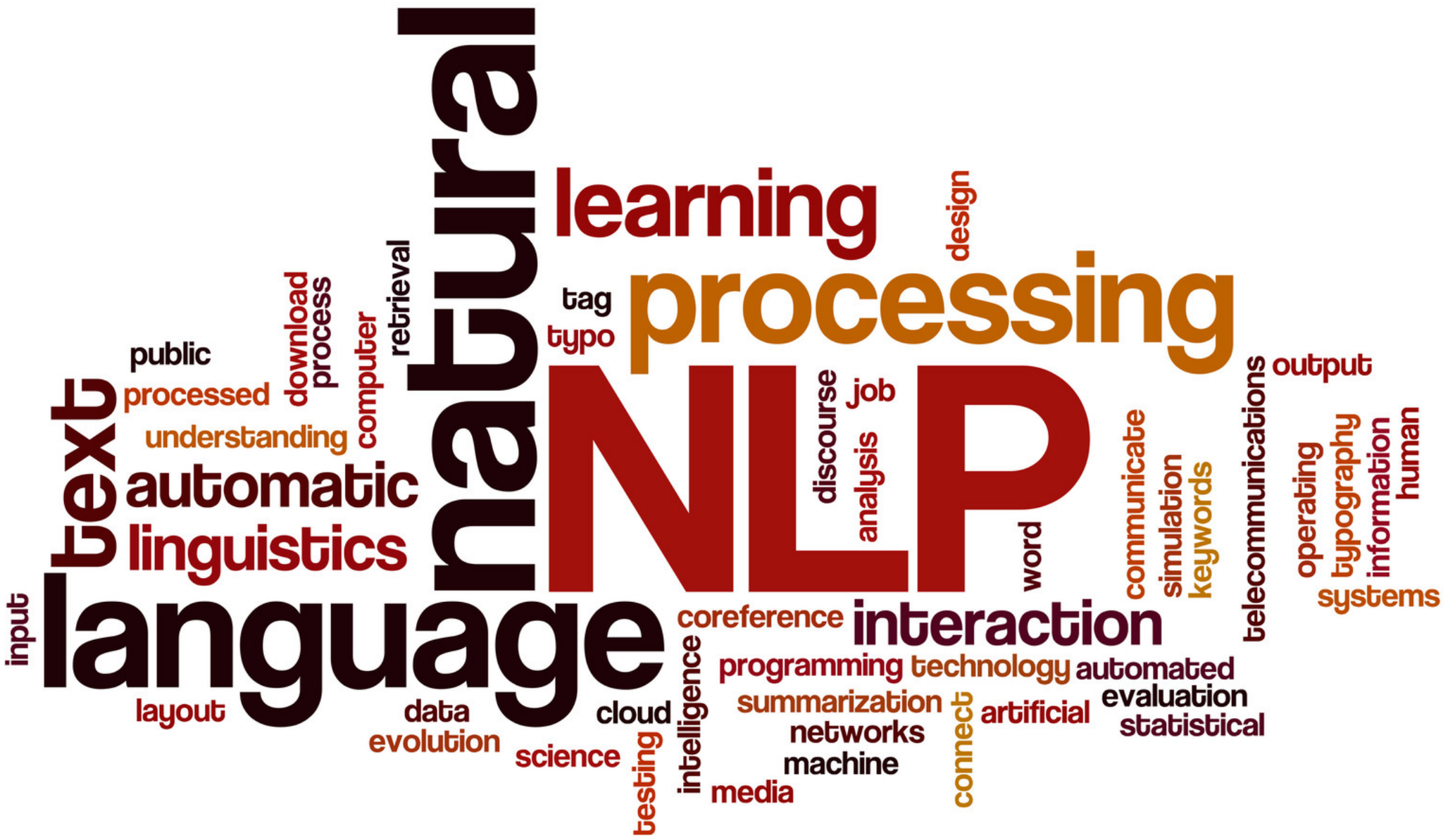
- Ph.D. in Natural Language Processing and Artificial Intelligence at Masaryk University
- 10 years at seznam.cz (last 8 years as Head Of Research)
- Founder and lecturer at ML College
- Founder and co-organizer of ML Prague
- ML Freelancer and consultant

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

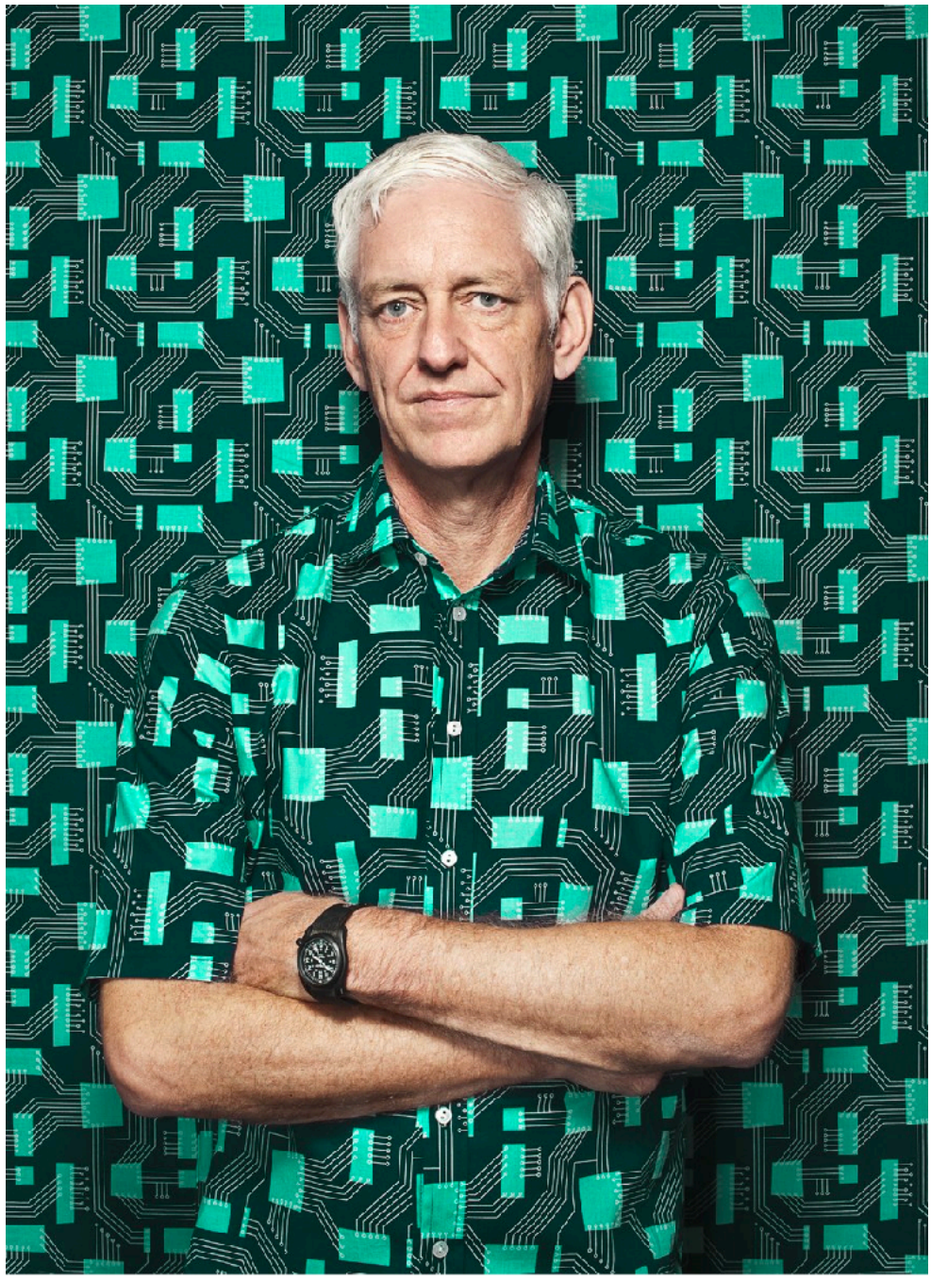
Outline

- Introduction to natural language processing
- Computational linguistics
- Text document vectorization
- Practical document classification task
- Language modeling
- Practical tasks on language modeling
- Word embeddings
- Transformers
- LSTM/GRU and time series prediction

What is Natural Language Processing?



Norvig vs. Chomsky



source: https://www.commarts.com

source: https://citaty.net

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/

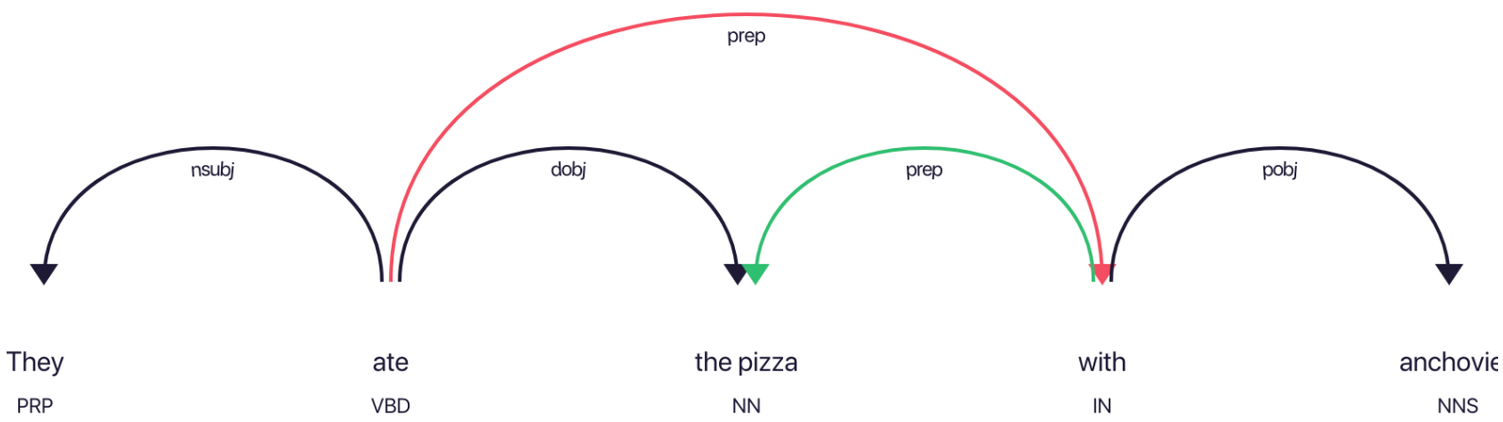
Czech:

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

Parsing

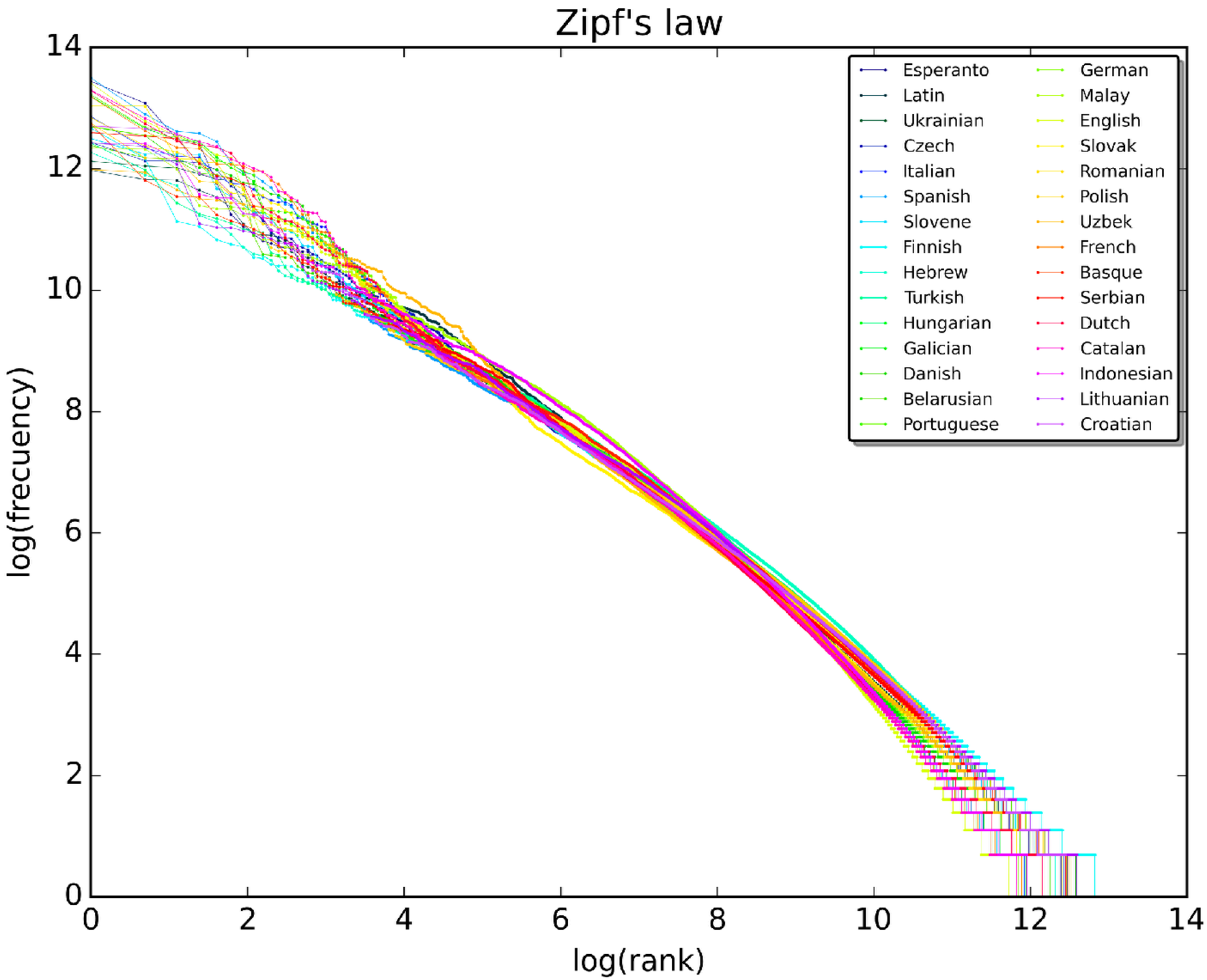
They ate the pizza with anchovies

Creative Commons Attribution-NonCommercial 2.5
James Constable, 2010



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

Zipf’s law & long tail



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]

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

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

01-text-classification-introduction.ipynb

Language models

- spell checking
- speech recognition
- machine translation
- …

n-gram models

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

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

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

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

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

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

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

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

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

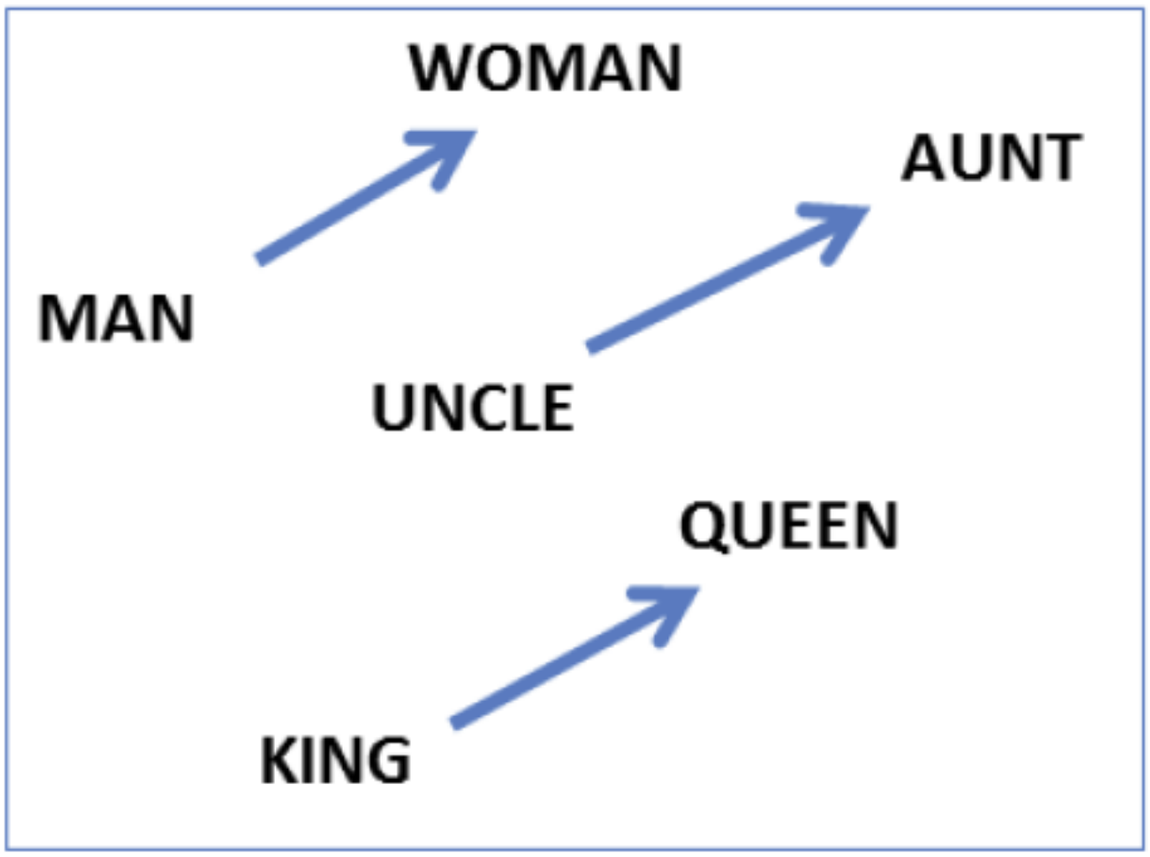
Language detection using language models

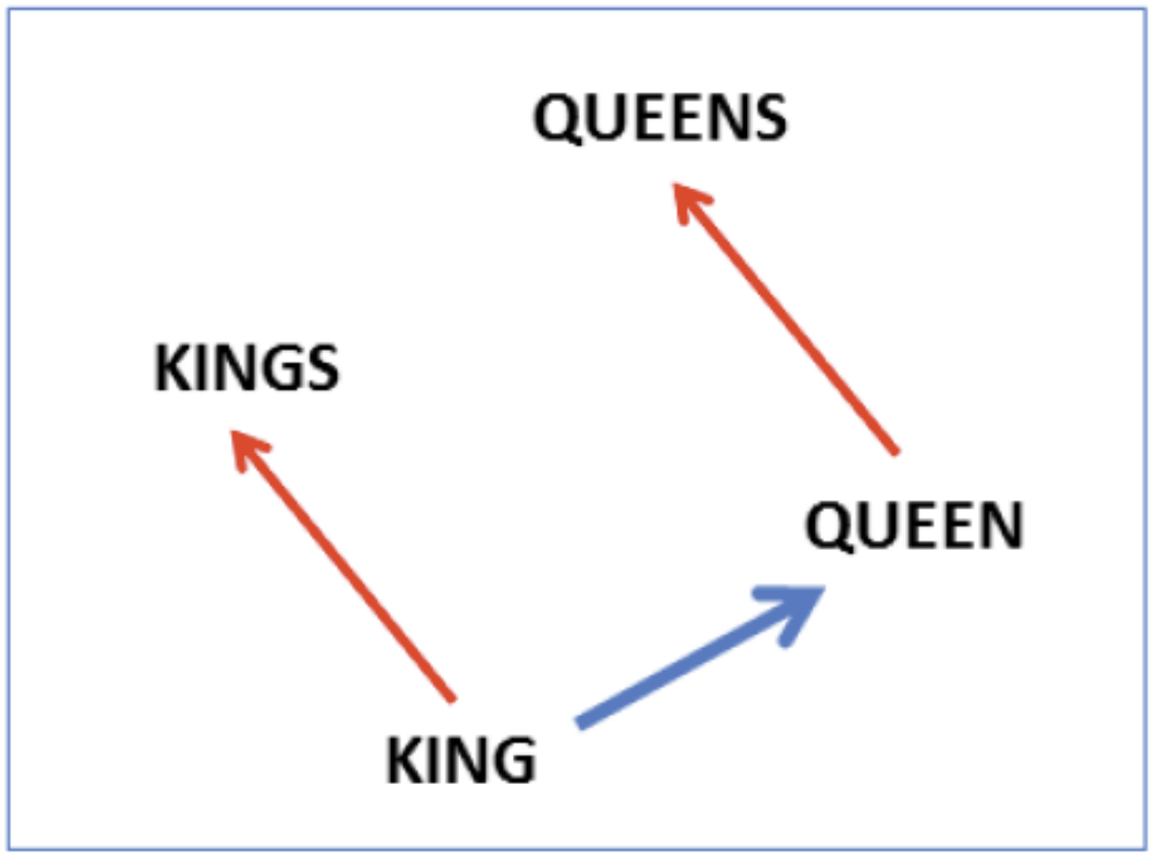
02-Language-detection-assignment.ipynb

Travel agency review classification

03-Review-classification-assignment.ipynb

word2vec

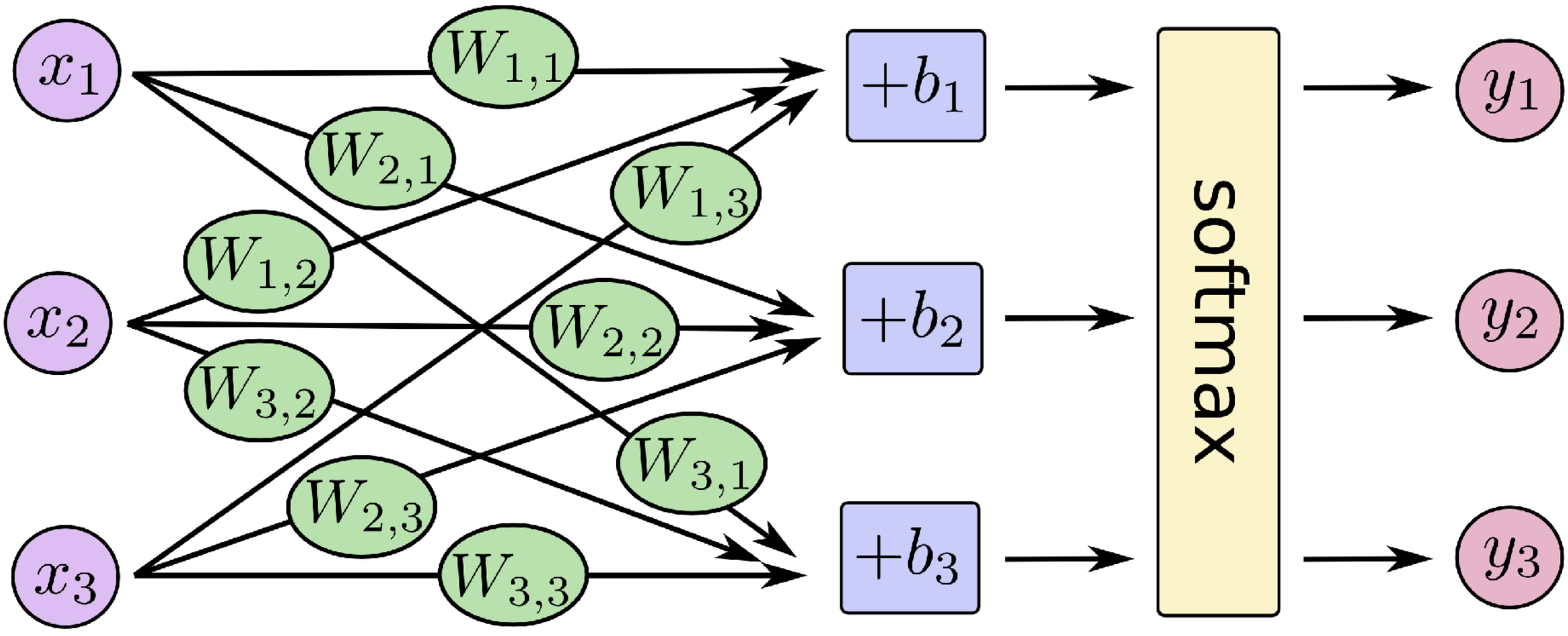




king is to **kings** as **queen** to ***?***.

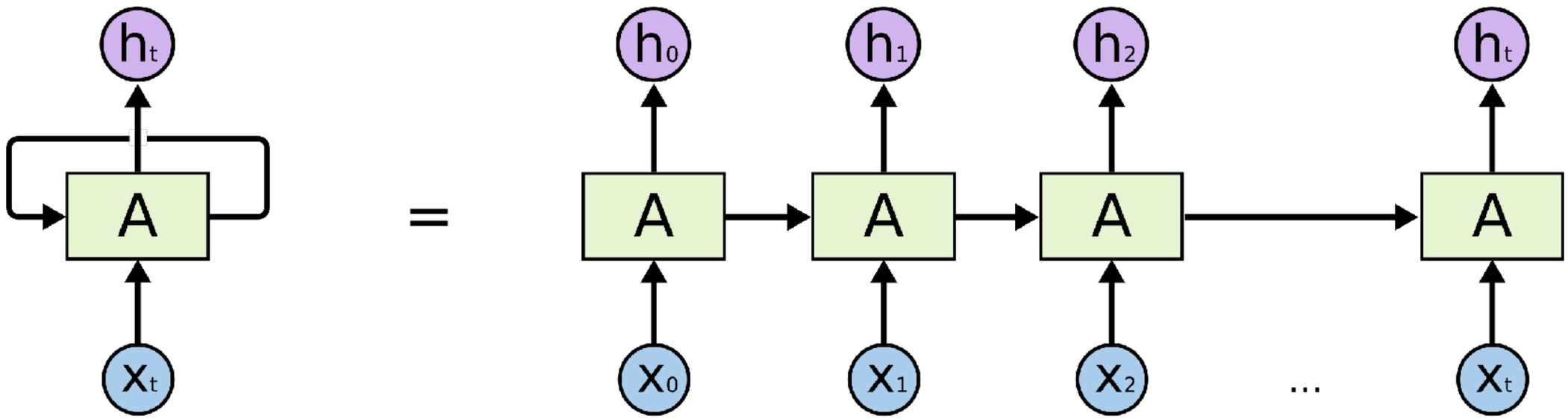
v(**kings**) - v(**king**) = v(**queens**) - v (**queen**)

Feed-Forward Neural Network



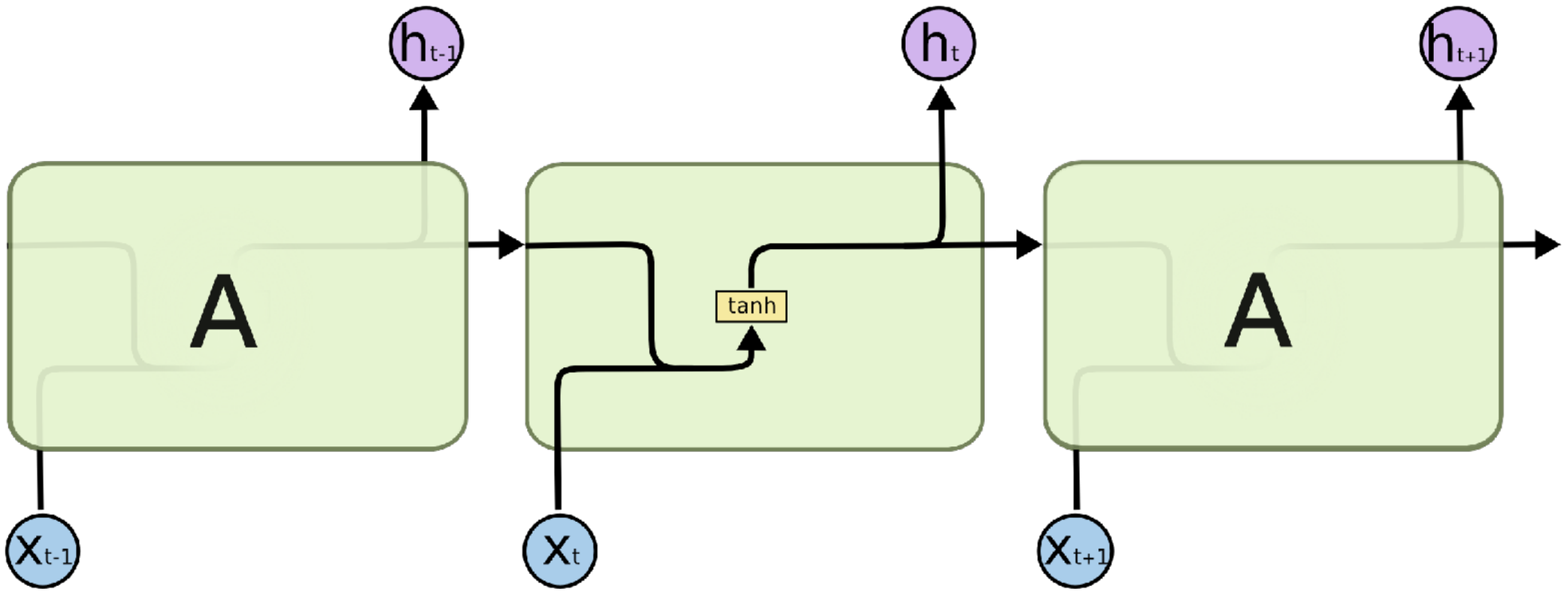
source: https://www.tensorflow.org

Recurrent Neural networks 1/2



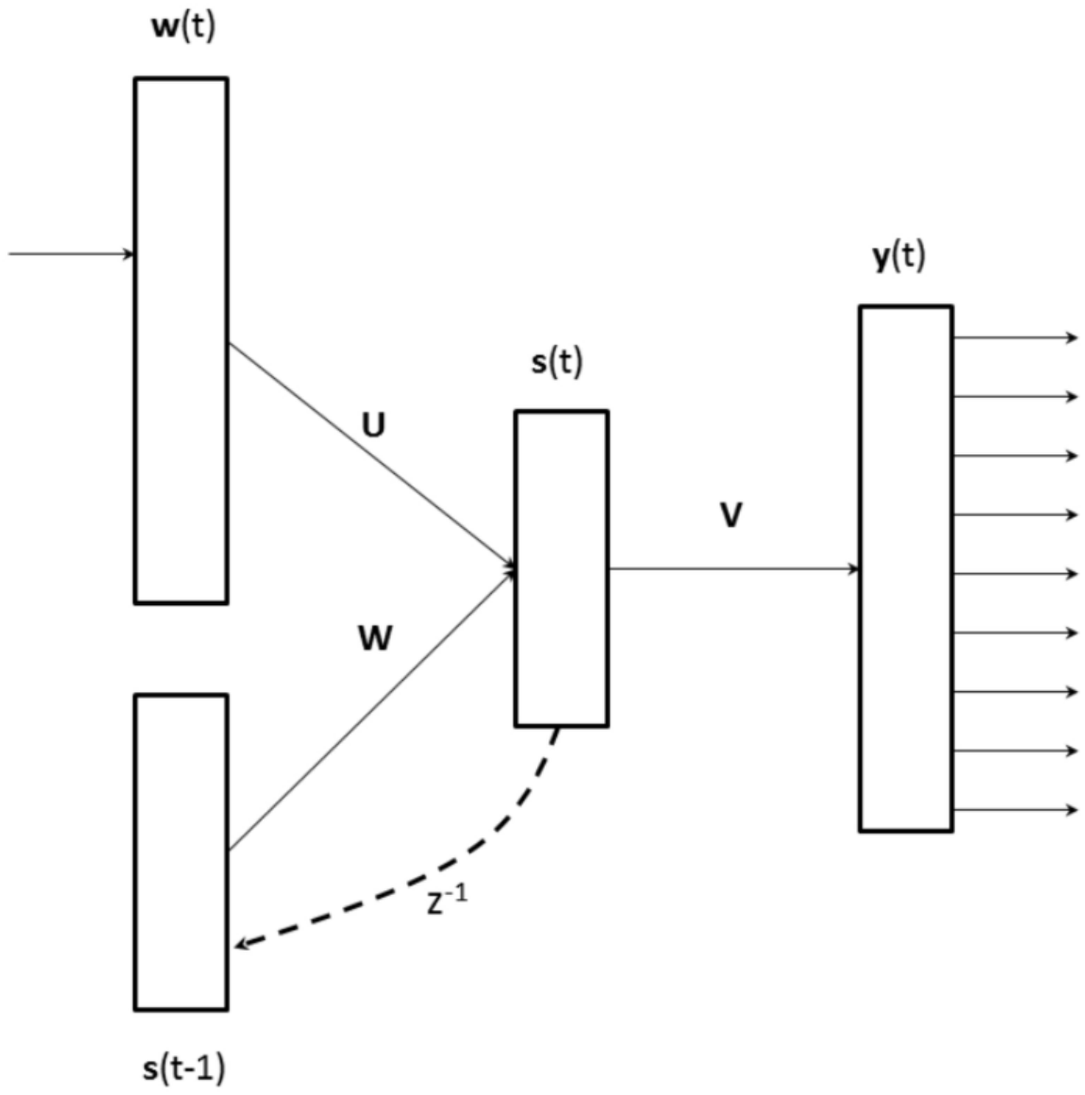
source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent Neural Networks 2/2



https://www.tensorflow.org

Recurrent Neural Network Language Modeling Toolkit



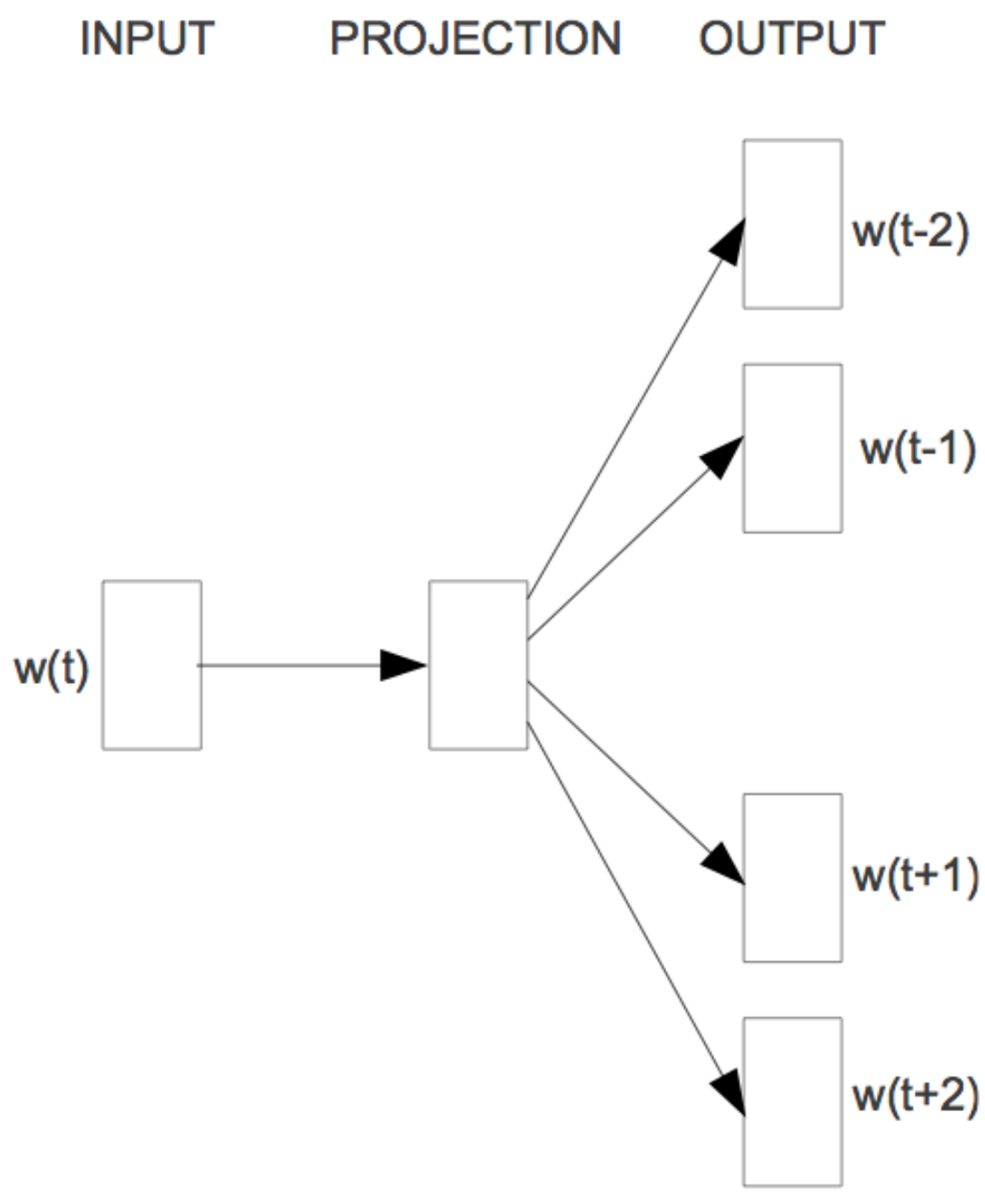
$$\mathbf{s}(t) = f\left(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1)\right)$$

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

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

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

The skip-gram model



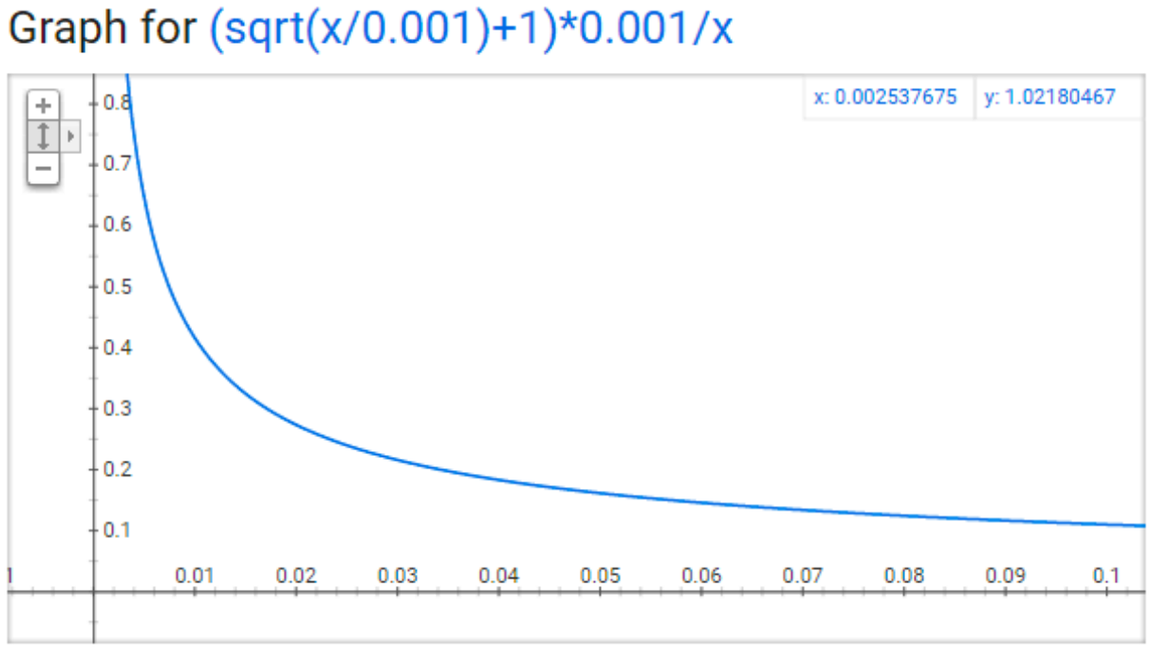
Skip-gram improvements

Subsampling frequent inputs

$$P(w_i) = (\sqrt{\frac{z(w_i)}{0.001}} + 1) \cdot \frac{0.001}{z(w_i)}$$

z(w) Relative frequency of word *w*

P(w) Probability of keeping word *w*



Negative sampling

We select only 5-20 negative samples in the loss function.
The probability of picking a word *w* is given by *z(w).*

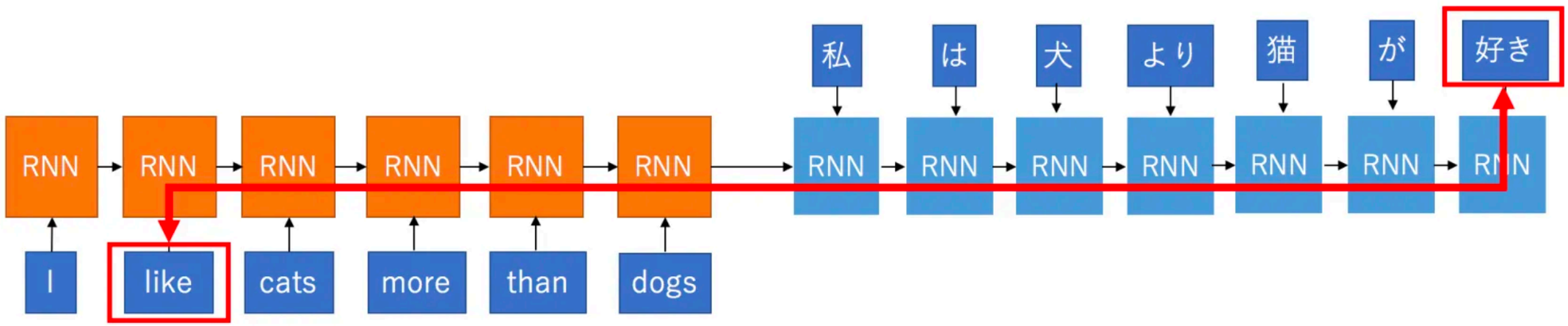
http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/

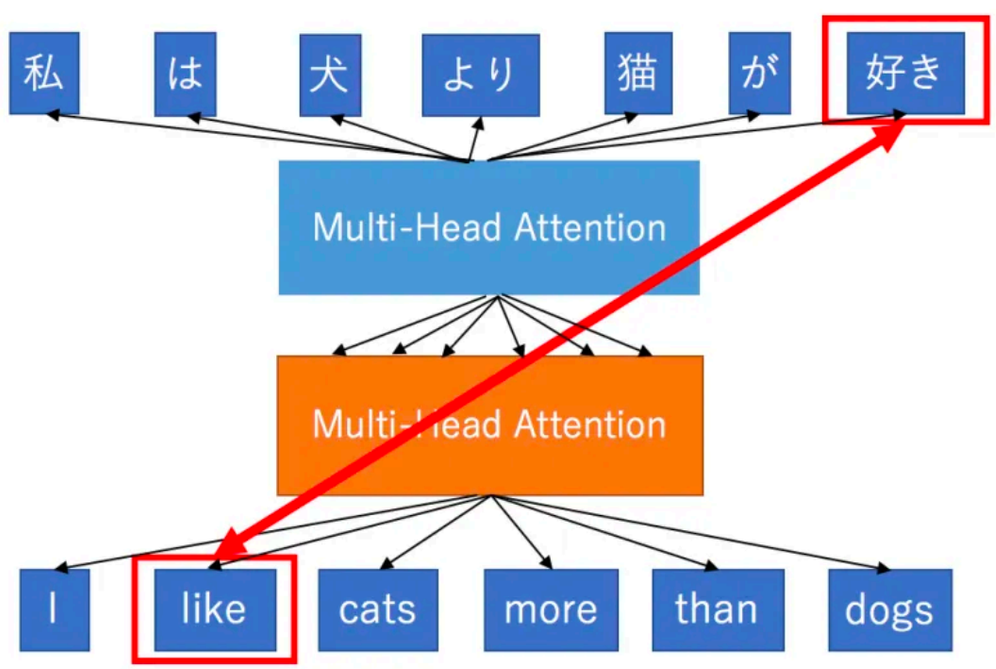
Experiments with word2vec

04-Word2vec-in-gensim.ipynb

05-Review-classification-w2v-assignment.ipynb

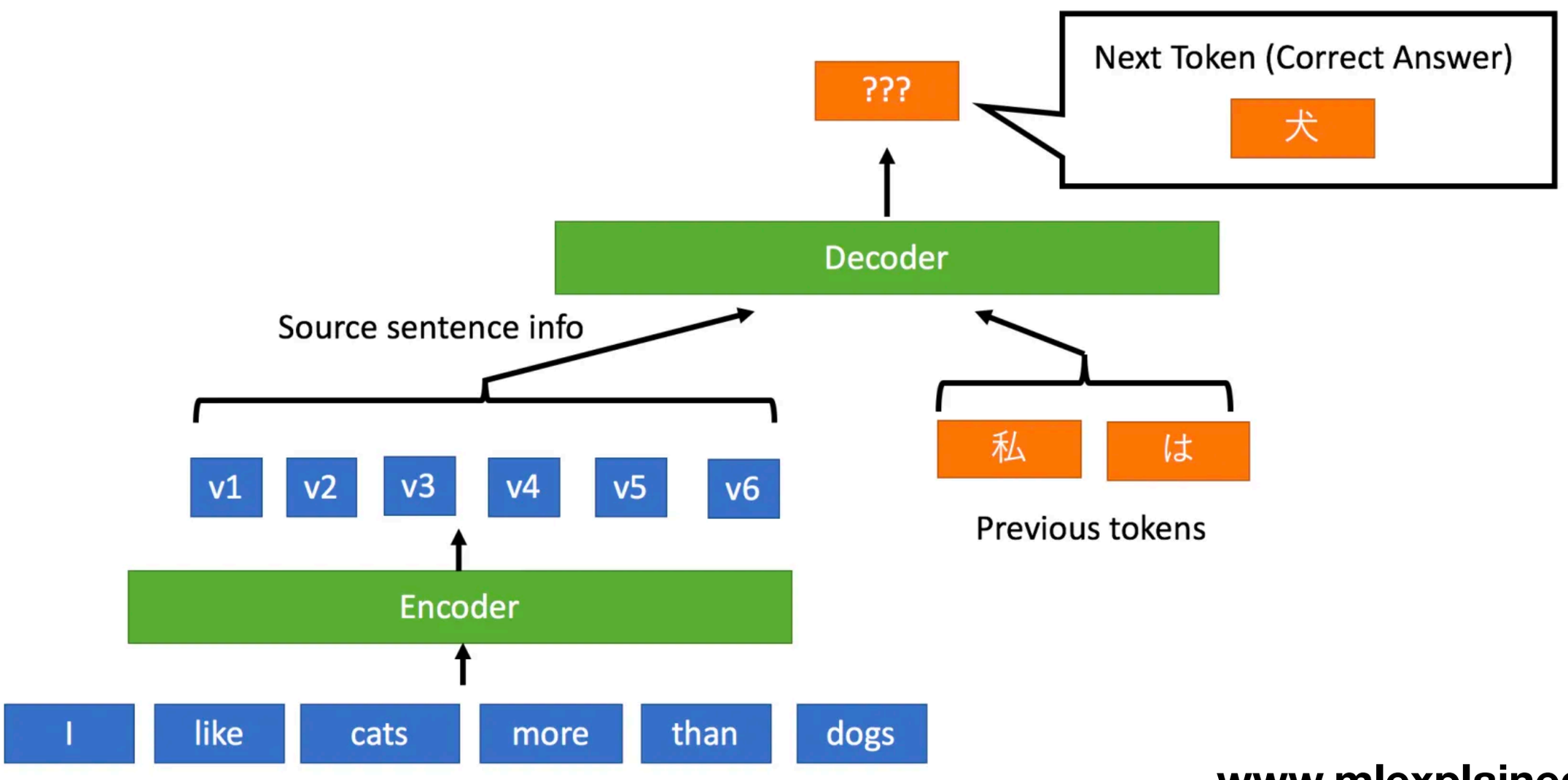
Transformer





source: www.mlexplained.com

Translation with Transformers



www.mlexplained.com

GPT-2 Language model

Donald Trump told...

GPT-3: **https://openai.com/blog/openai-api/**

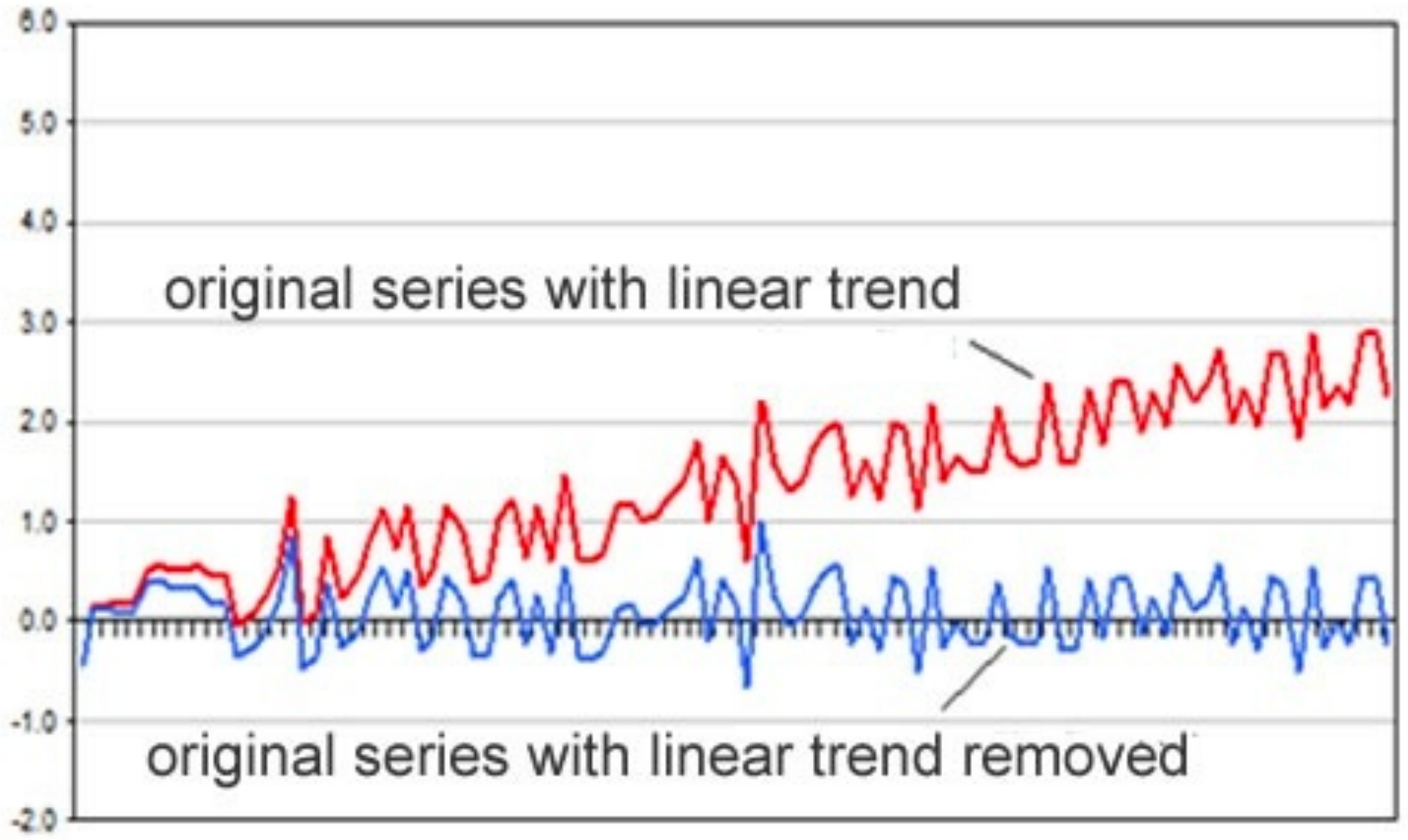
GPT-2 Language model

Donald Trump told the Times he is preparing a "major speech" on his economic plans, but did not provide details on what it will entail.

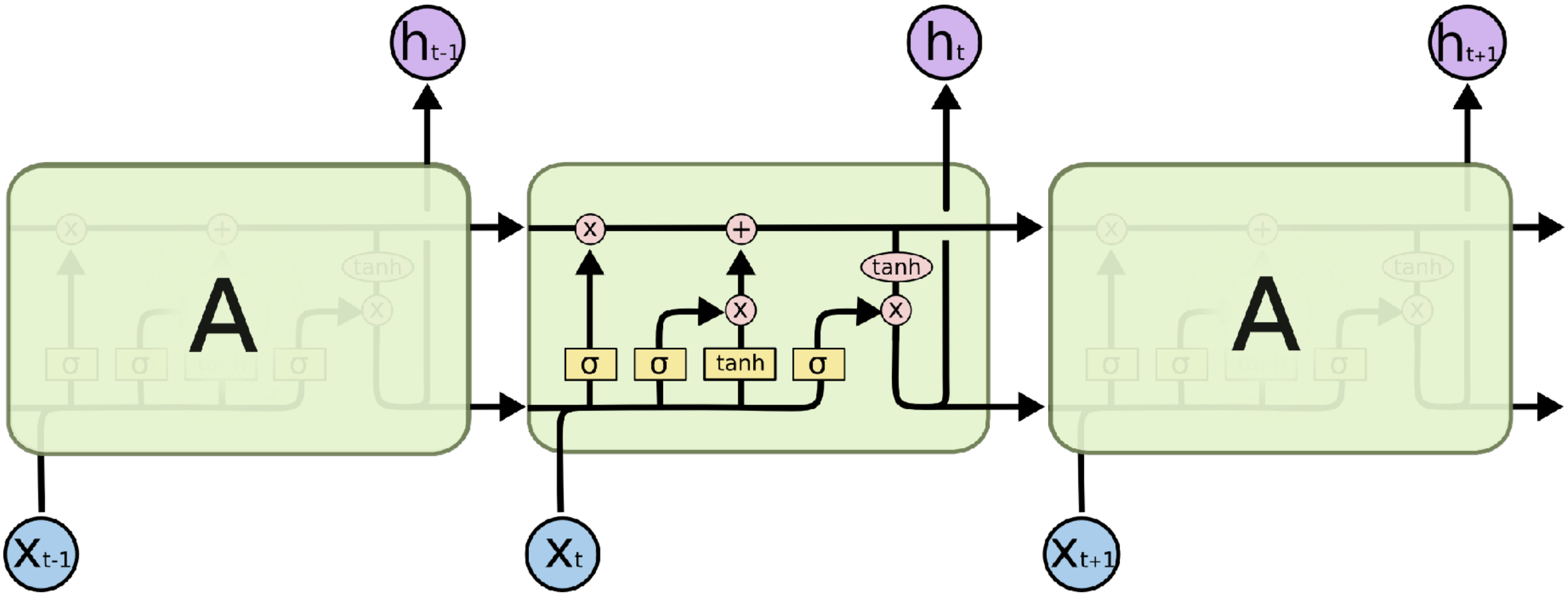
“I’m getting ready for the speech. And I will have a major speech on Tuesday.” Trump said during an interview in the White House residence.

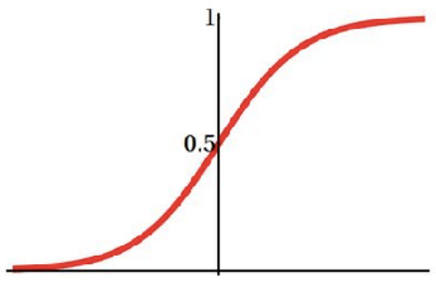
GPT-3: **https://openai.com/blog/openai-api/**

Time series specifics - stationarity

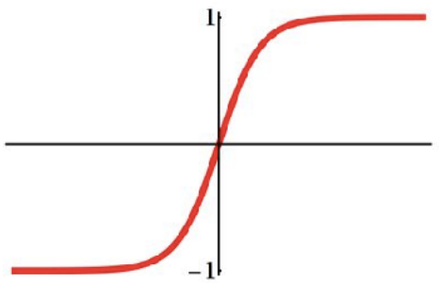


Long Short-Term Memory





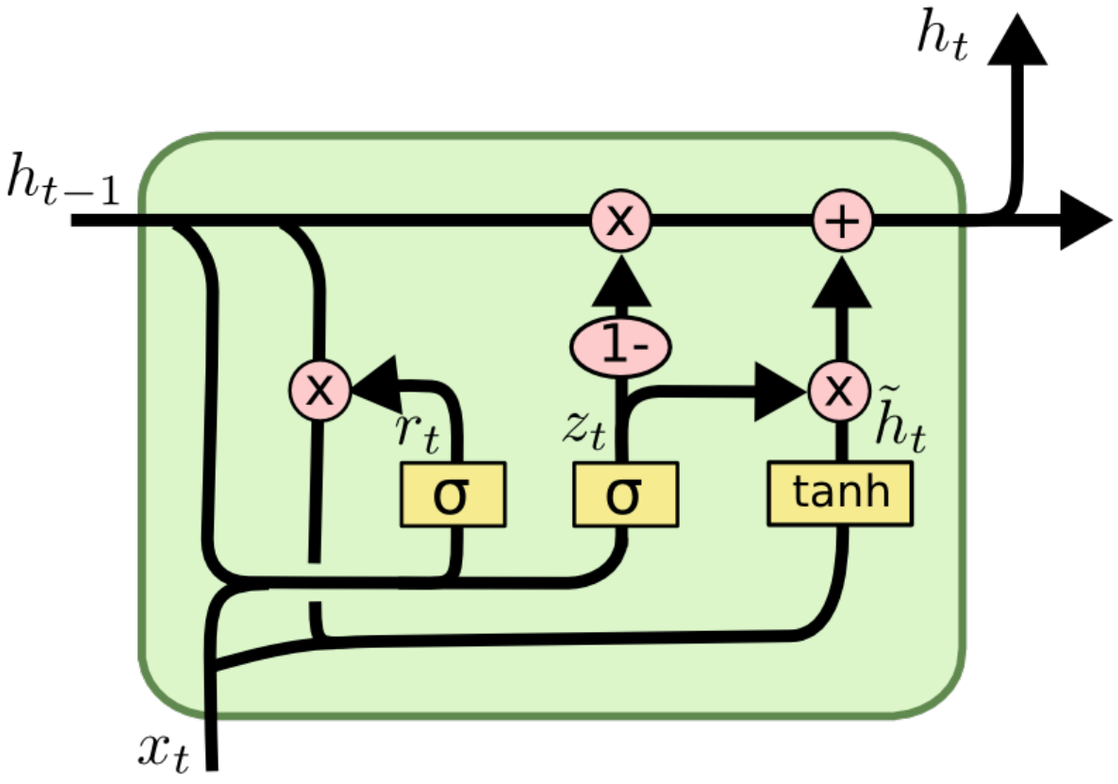
sigmoid



tanh

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

Gated Recurrent Unit



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

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

Implementation of the regression task using Recurrent neural network

06-Regression-rnn.ipynb

Thank you for your attention

e-mail: jiri@mlcollege.com

Web: www.mlcollege.com

Twitter: @JiriMaterna

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

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