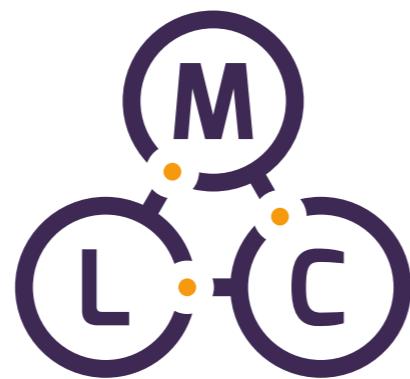


Natural Language Processing for Phonexia

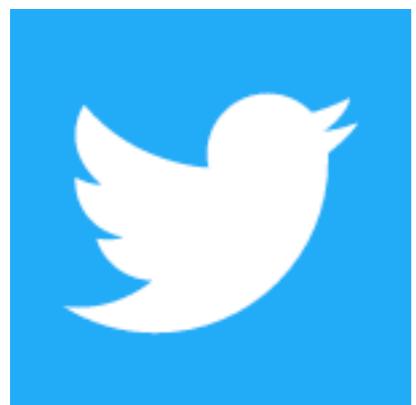
Jiří Materna



Machine
Learning
College



@mlcollegecom



@mlcollegecom



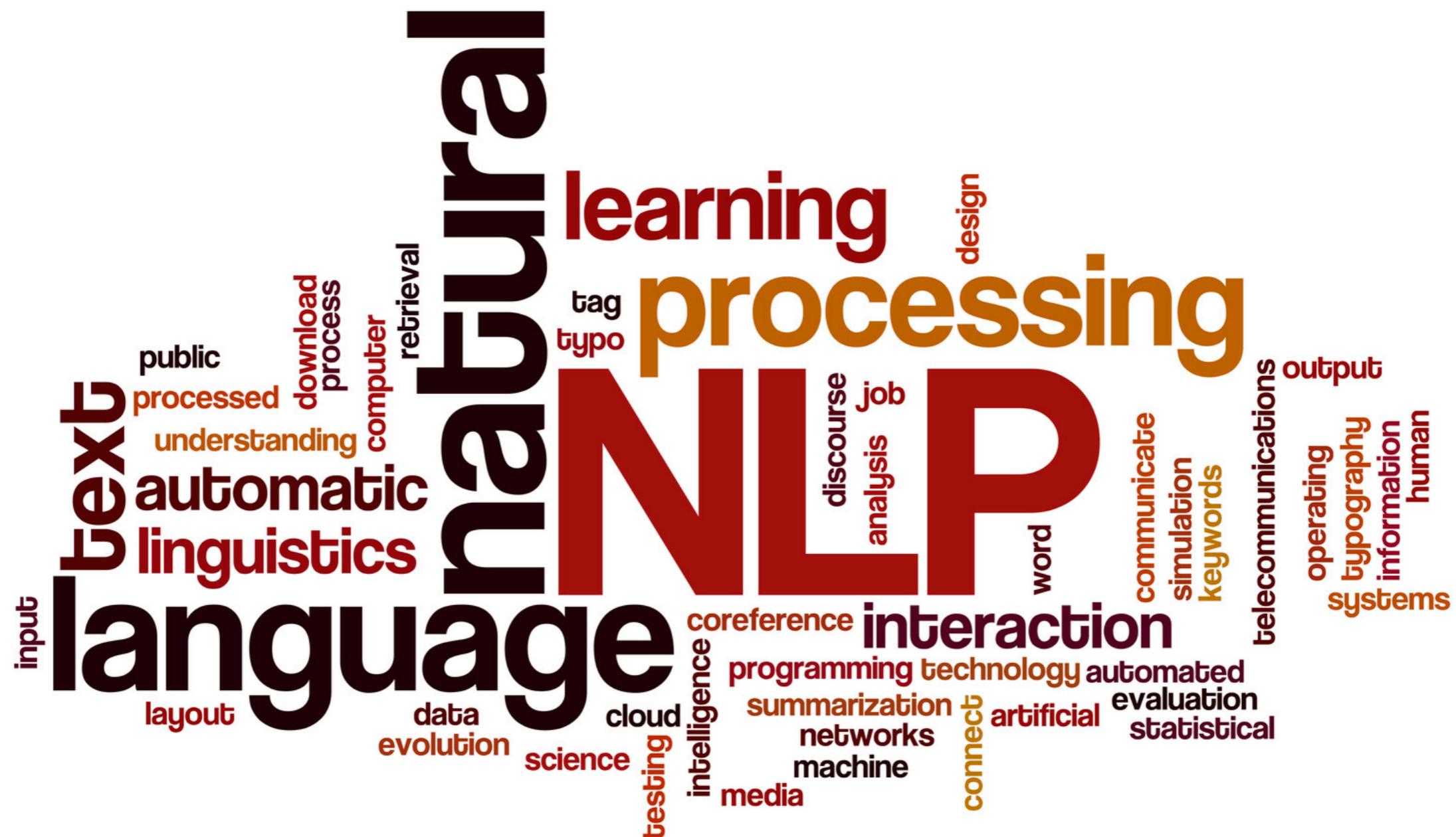
About me

- 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
- Mentor at StartupYard
- ML Freelancer and consultant

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
- Transformers

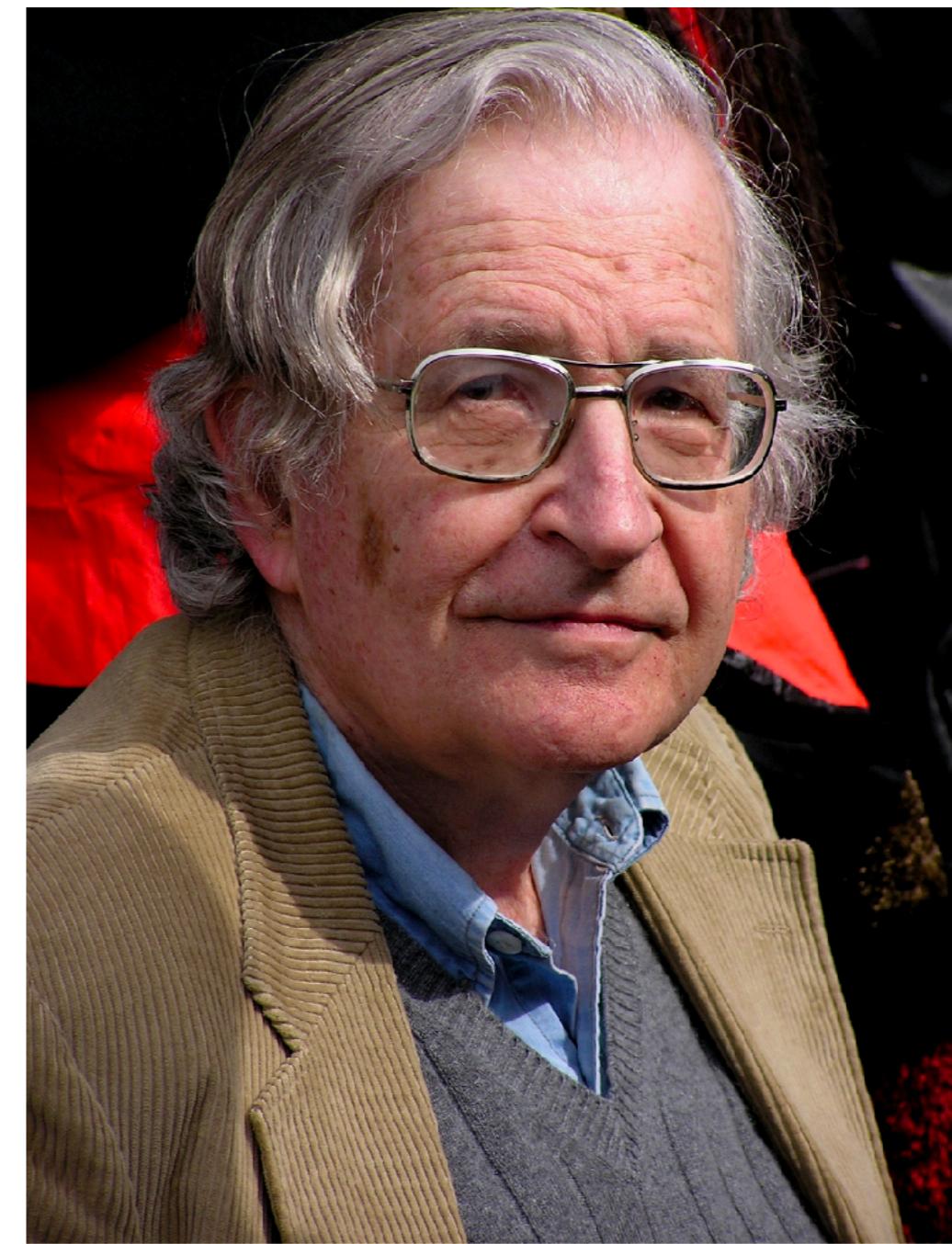
What is Natural Language Processing?



Norvig vs. Chomsky



source: <https://www.commarts.com>



source: <https://citaty.net>

Text corpus



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 of 13 |

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 . </p><p> 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. </p><p> 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**. </p><p> 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 of 13 |

Token & tokenization

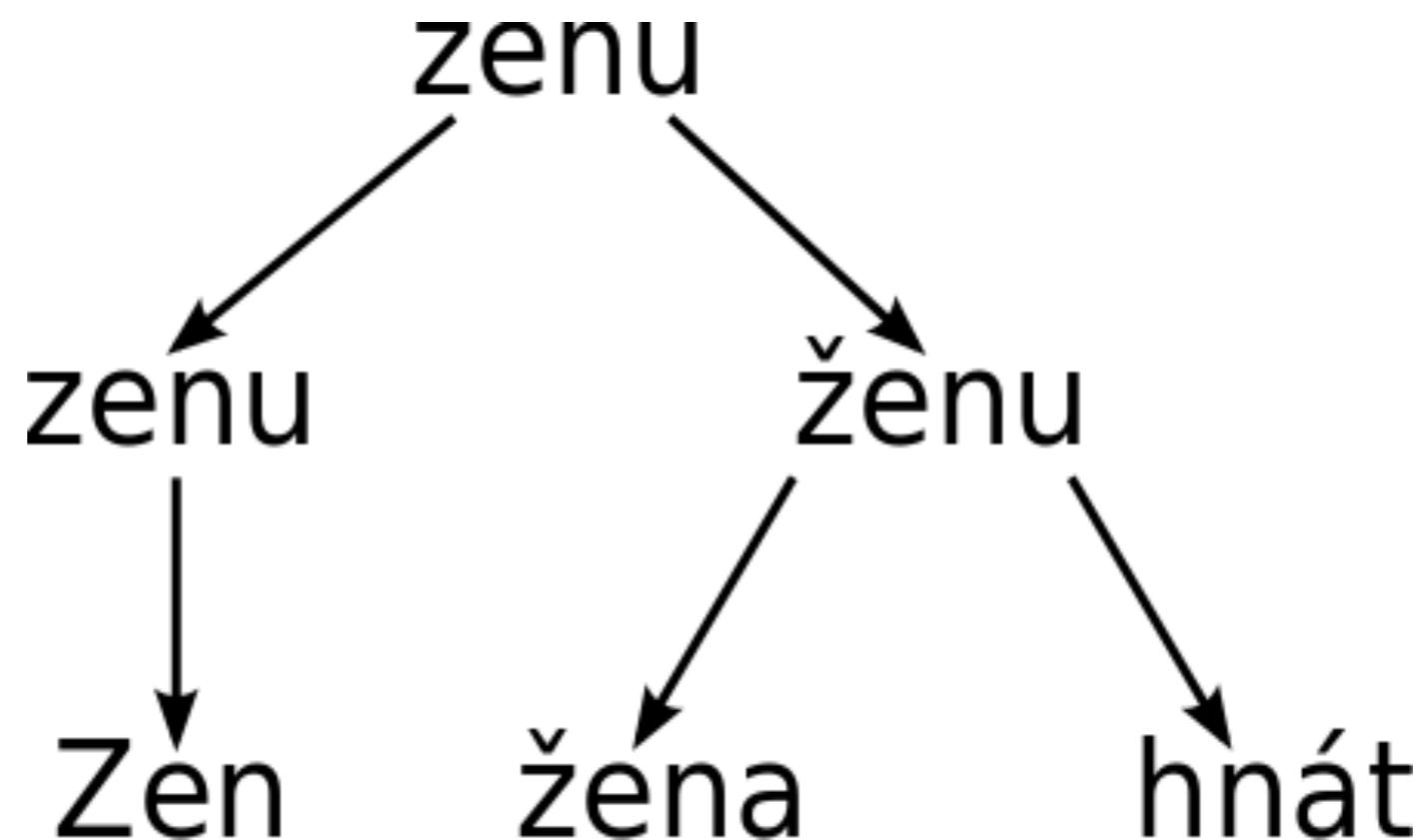
This is a non-trivial English sentence: Ludolph's number is approx. 3.14.

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

Stemming & lemmatization

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

Ambiguity in lemmatization



Stemming & lemmatization

English:

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

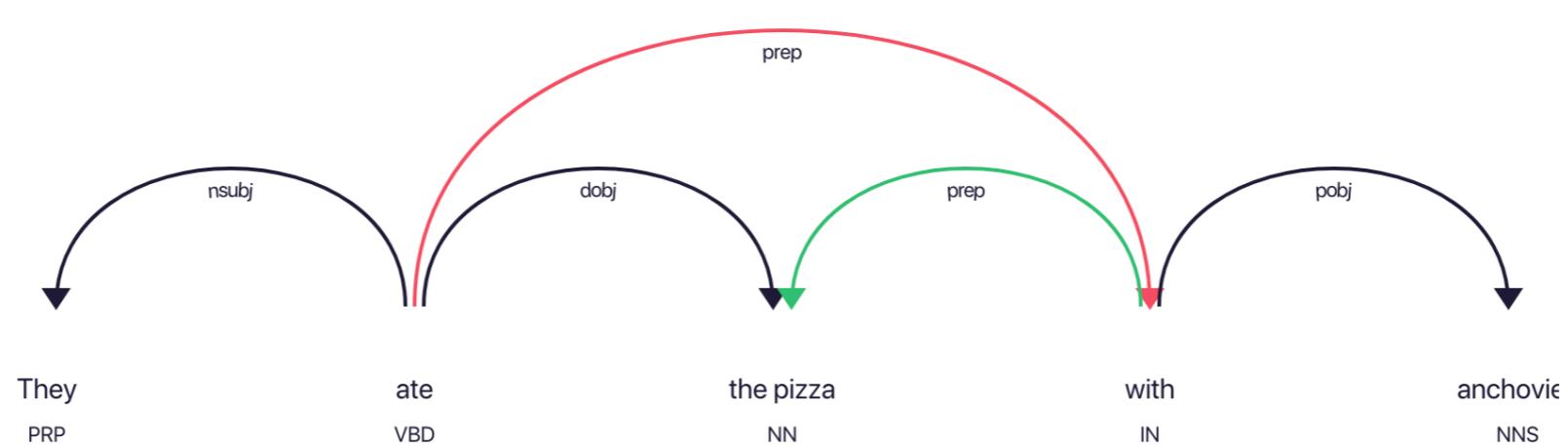
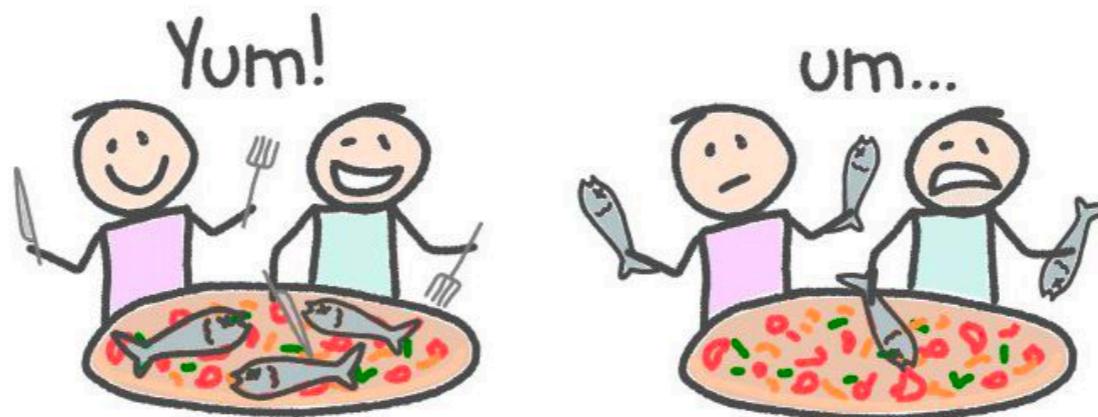
<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Czech:

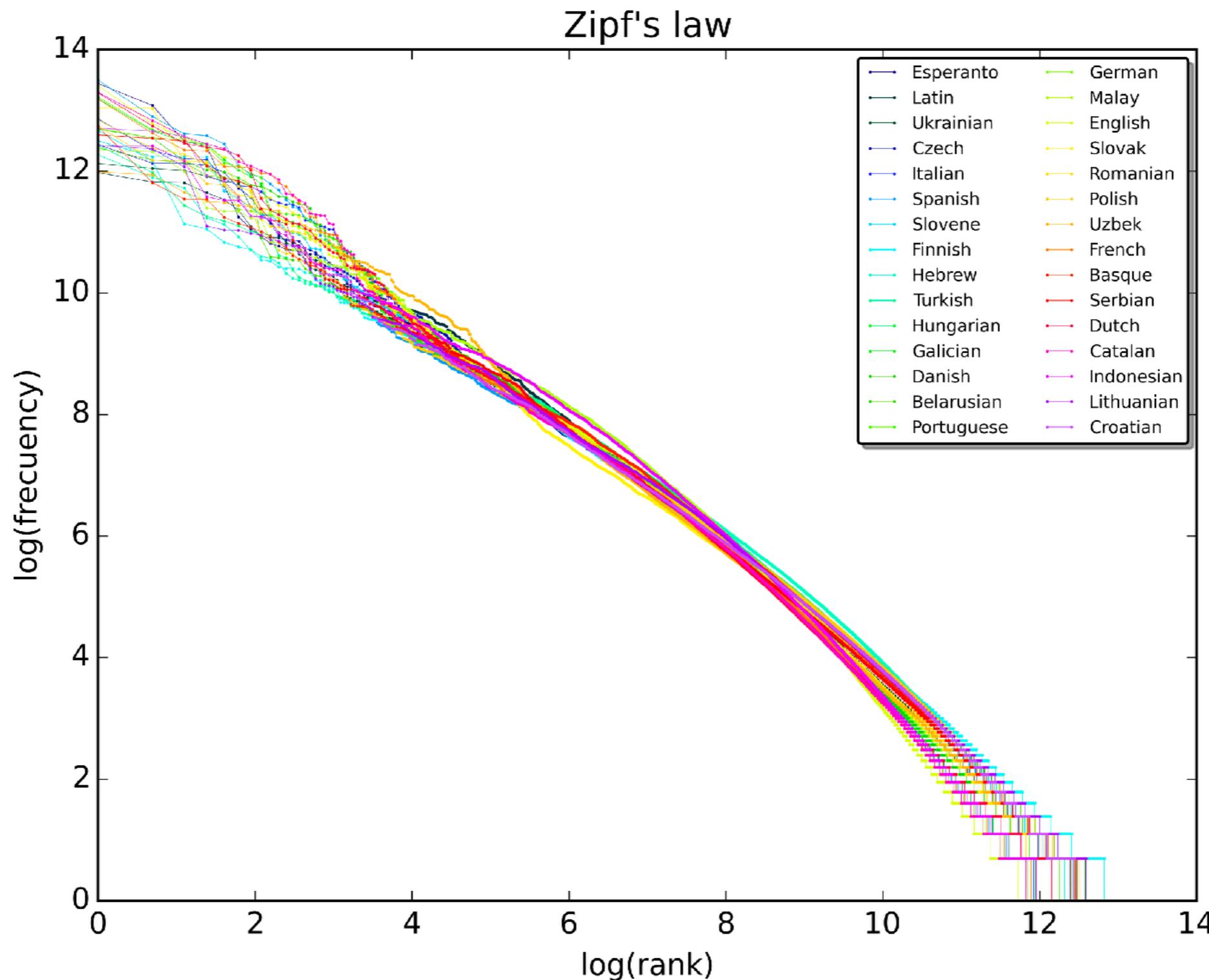
<http://ufal.mff.cuni.cz/morphodita>

Parsing

They ate the pizza with anchovies



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

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

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

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

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

Language model smoothing

- Laplace smoothing (plus one)

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

- interpolation
- Good-Turing
- Witten-Bell
- ...

Perplexity

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

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

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

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

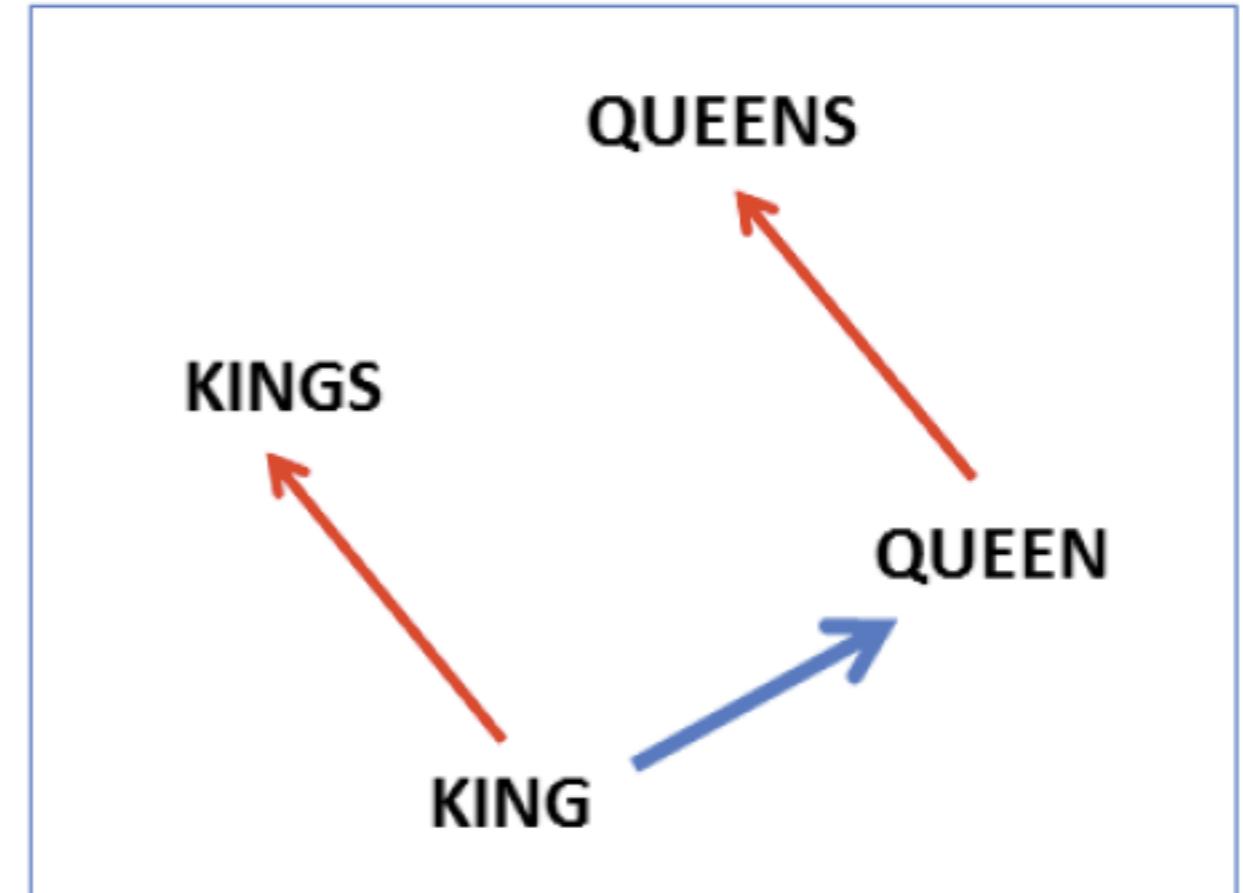
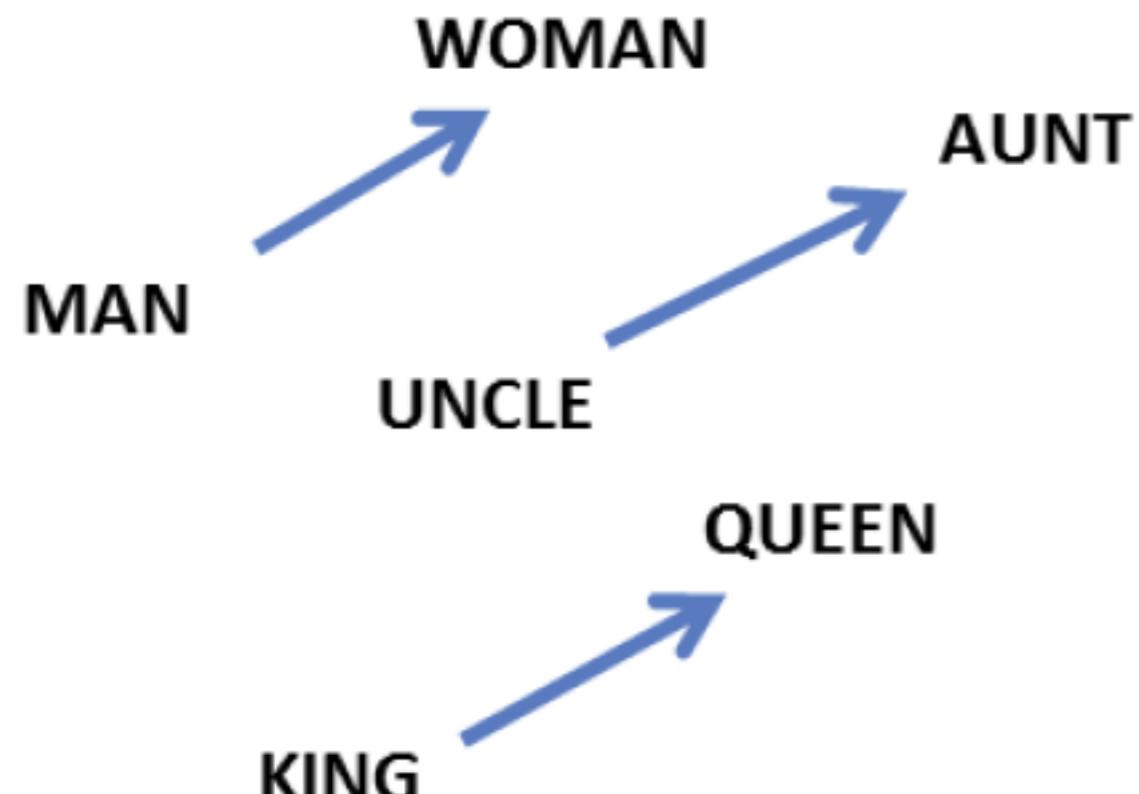
Language detection using language models

02-Language-detection-assignment.ipynb

Travel agency review classification

03-Review-classification-assignment.ipynb

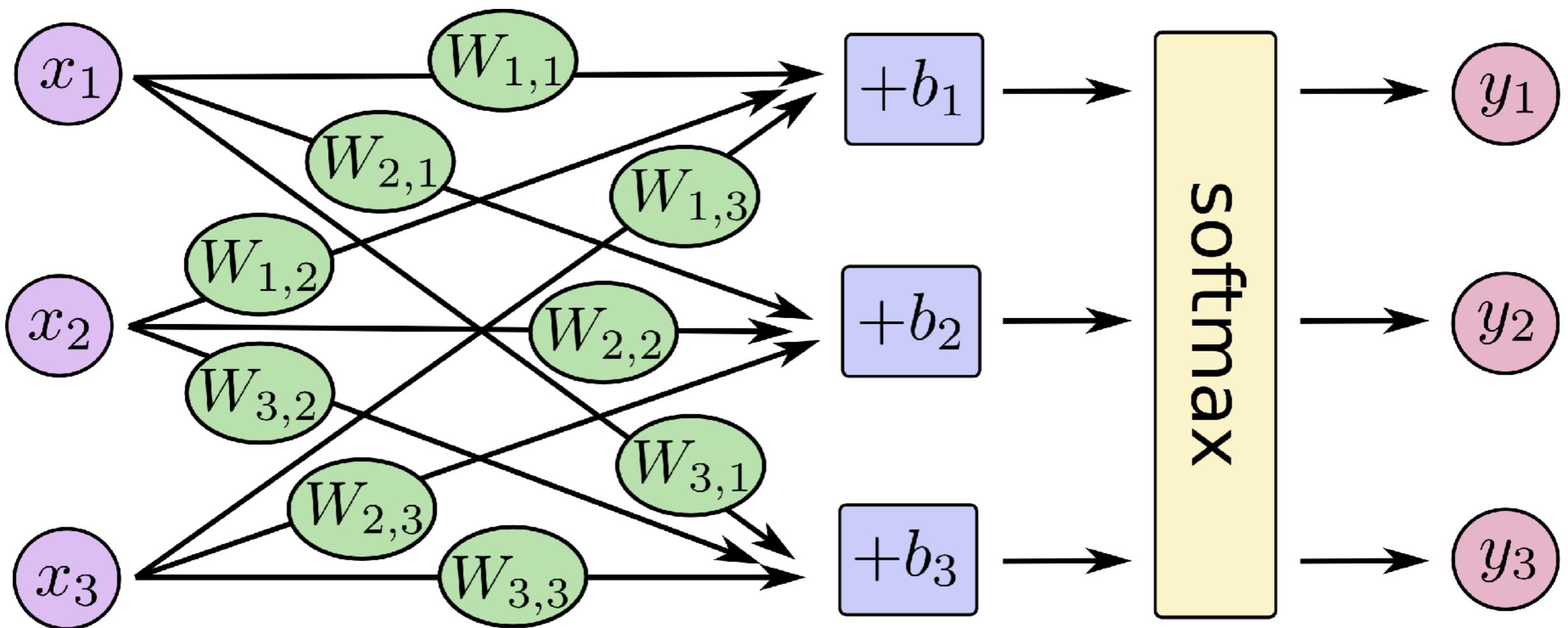
word2vec



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

$$v(\text{kings}) - v(\text{king}) = v(\text{queens}) - v(\text{queen})$$

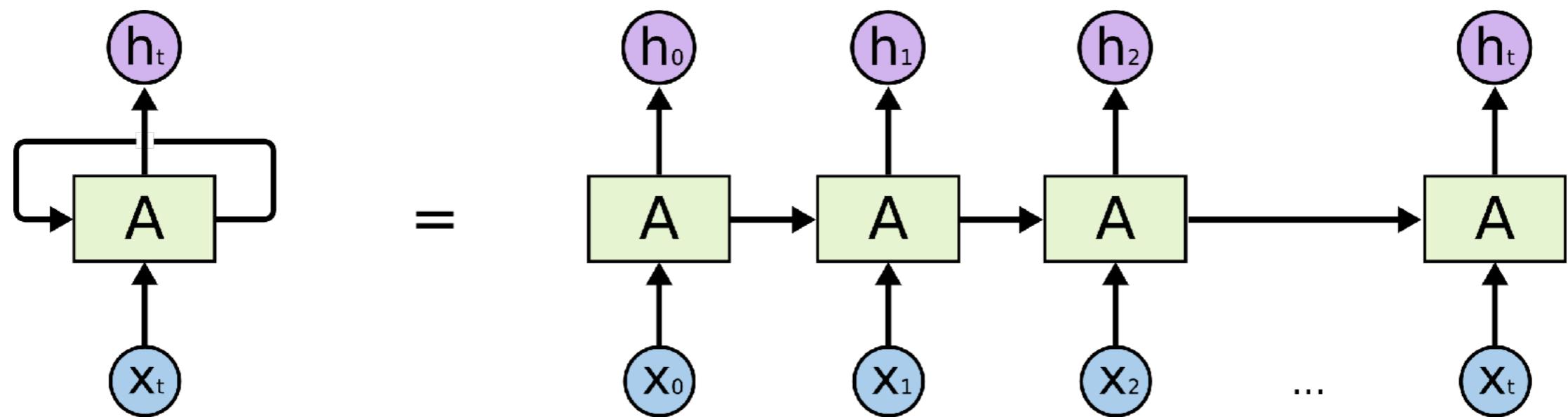
Feed-Forward Neural Network



source: <https://www.tensorflow.org>

Recurrent Neural networks

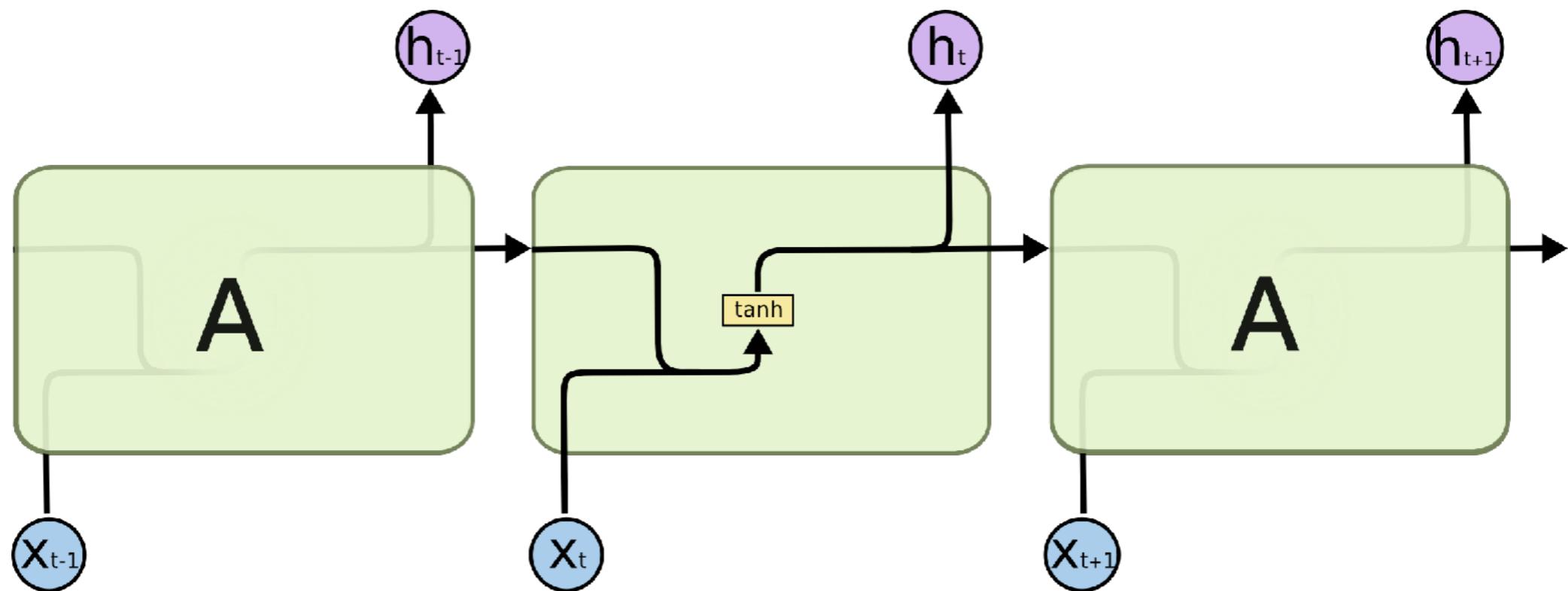
1/2



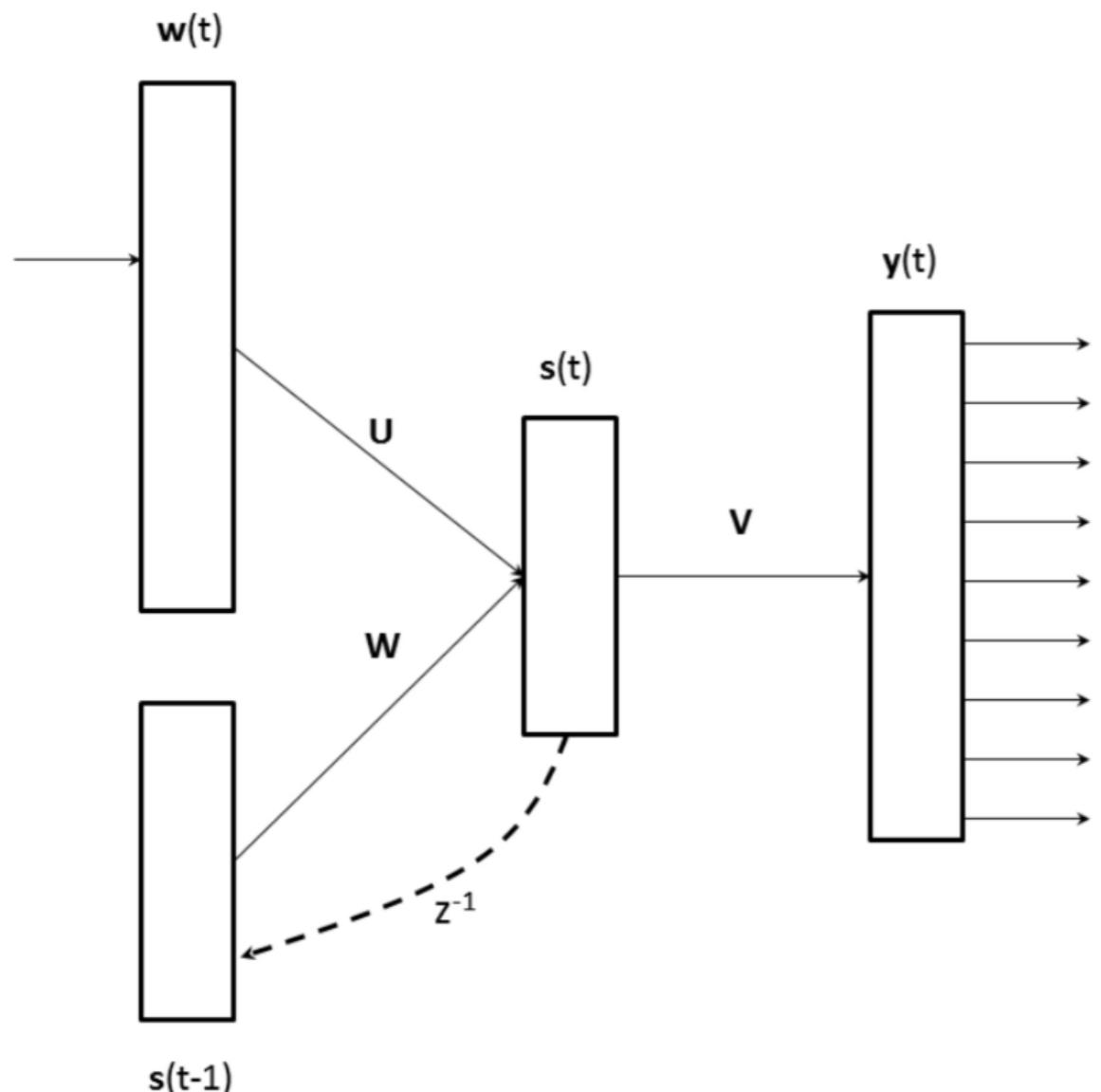
source: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Recurrent Neural Networks

2/2



Recurrent Neural Network Language Modeling Toolkit

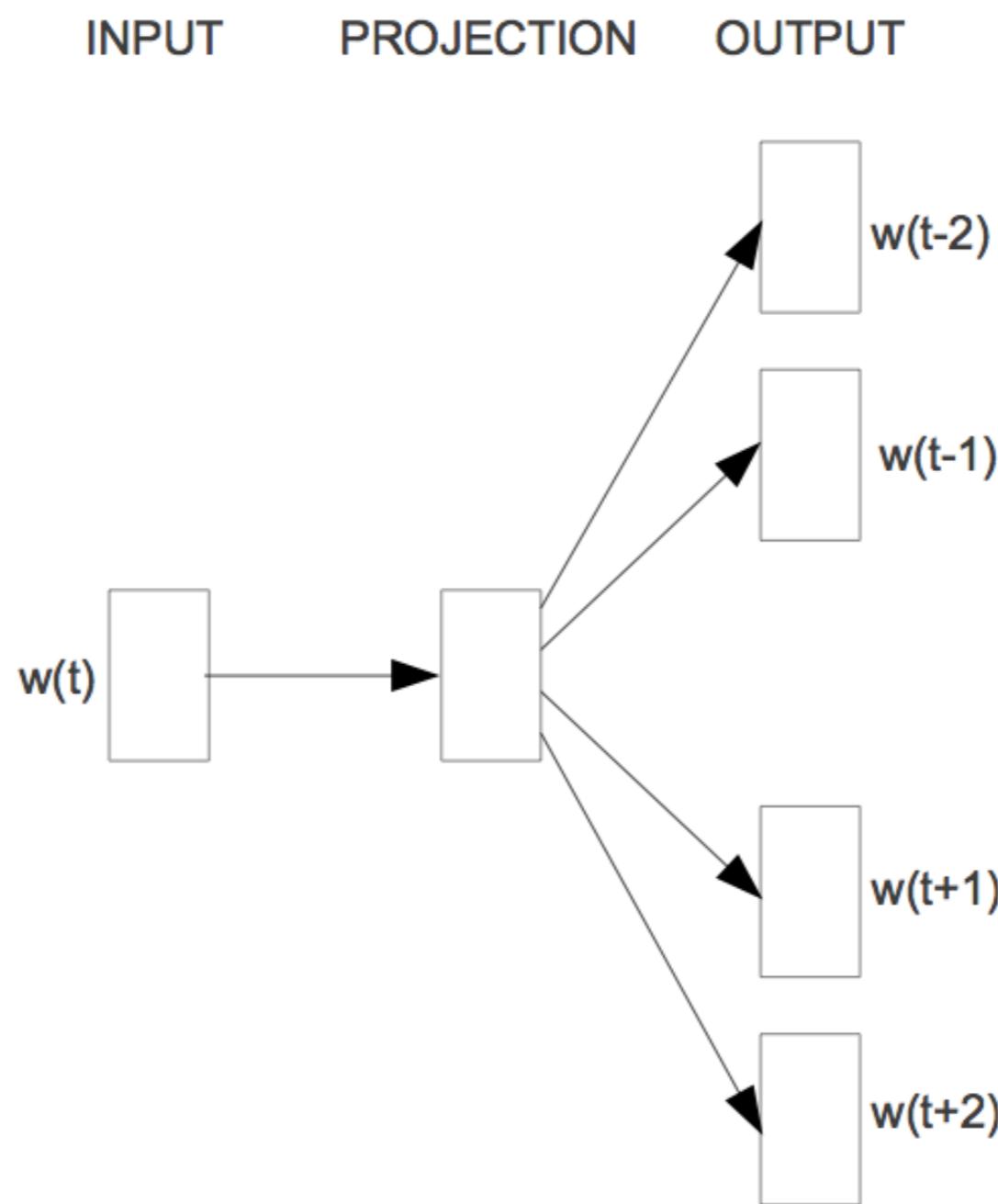


$$\mathbf{s}(t) = f(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1))$$

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

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

The skip-gram model



Skip-gram improvements

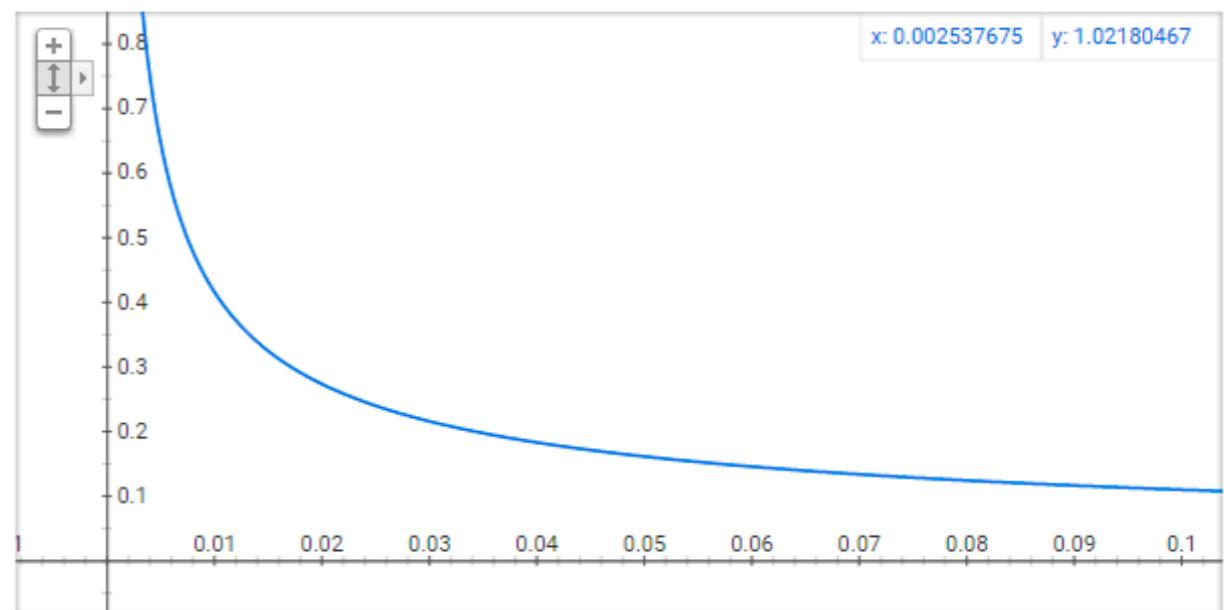
Subsampling frequent inputs

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

$z(w)$ Relative frequency of word w

$P(w)$ Probability of keeping word w

Graph for $(\sqrt{x/0.001}+1)*0.001/x$



Negative sampling

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

Experiments with word2vec

04-Word2vec-in-gensim.ipynb

05-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



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 pomocí umělých neuronových sítí. 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?
cítim tvou slabost
a whisky

NOVEMBER

I am falling asleep, crying, dying, thinking
what do you feel?
I feel your weakness
and whisky

SPRAVEDLNOST

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

JUSTICE

for your decadent soul
in the morning, in the evening
the god has prepared a crossbow

Metaphores

...tělo plné červánků...

...body full of blush of dawn...

...tak vzácný jako listí...

...as rare as leaves of trees...

Language models for text generating

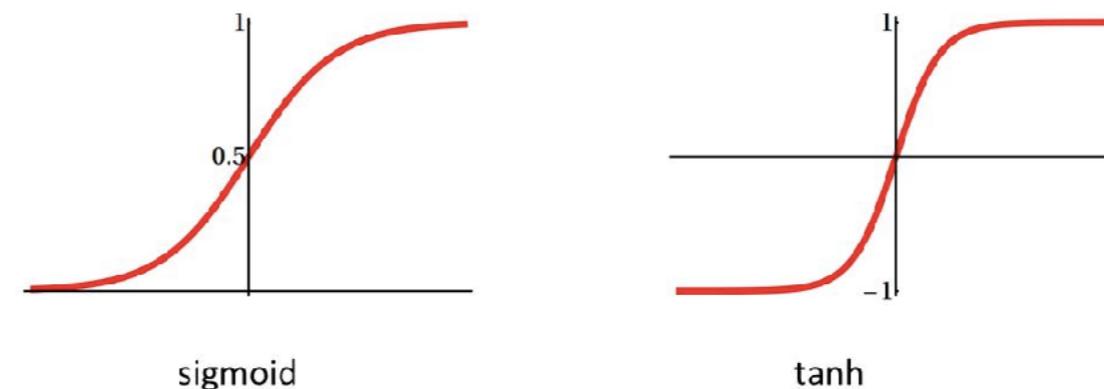
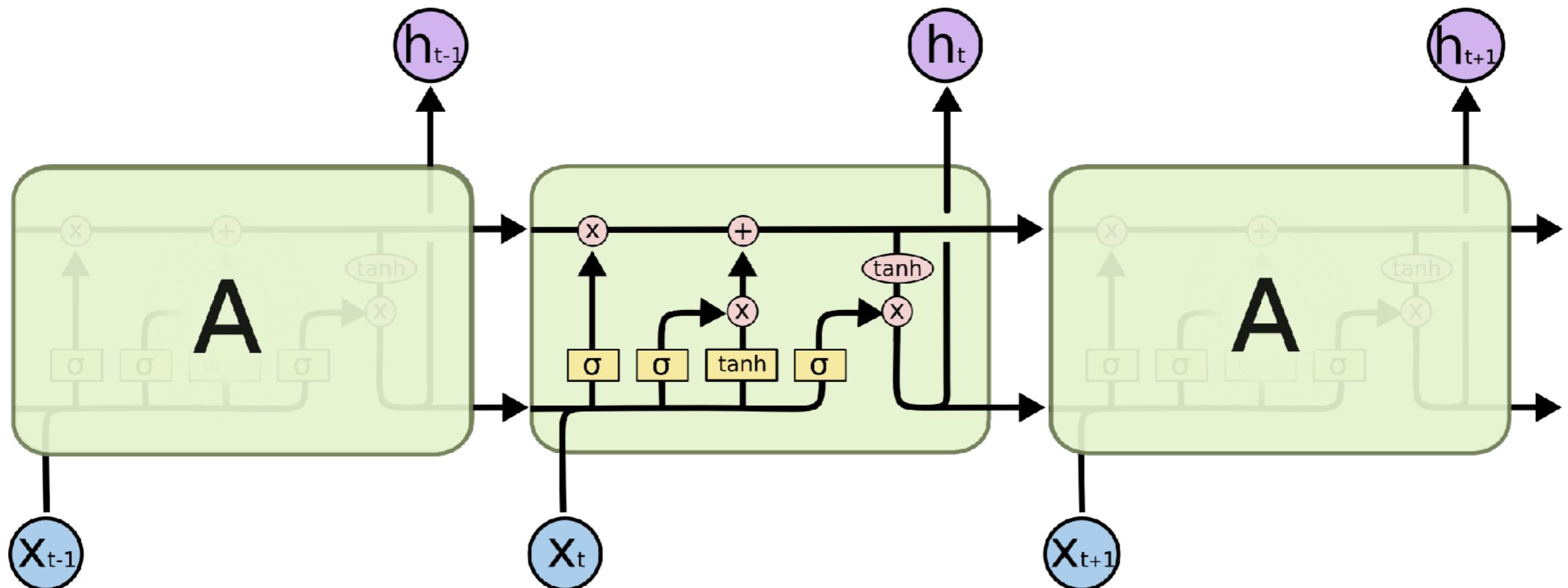
$$P(\text{maso} | \text{máma}, \text{mele}) = 0.5$$

$$P(\text{Emu} | \text{máma}, \text{mele}) = 0.3$$

$$P(\text{tátu} | \text{máma}, \text{mele}) = 0.2$$

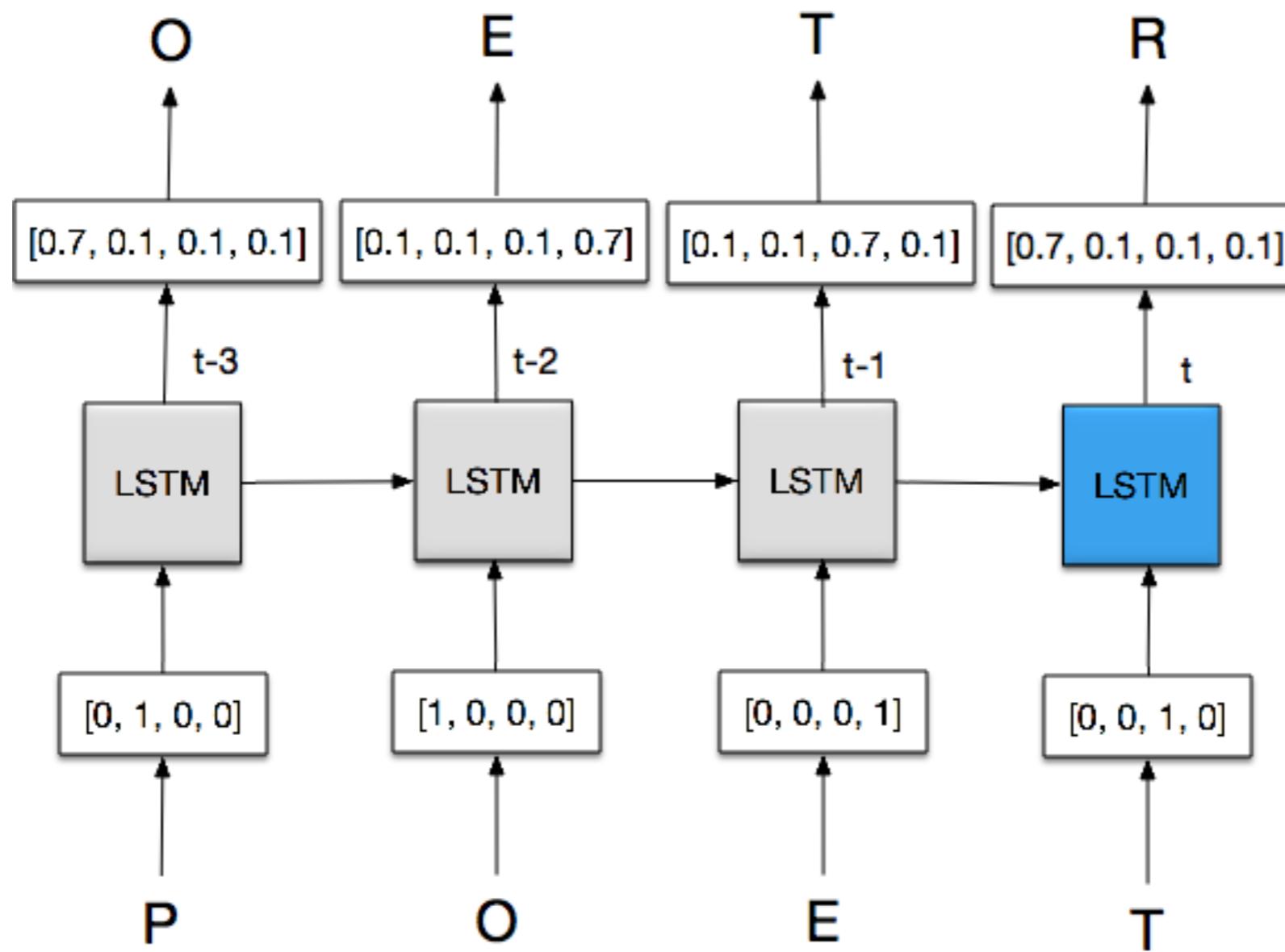
```
t ~ Uniform(0, 1)
s = 0
for v in Vocabulary:
    s += v.prob
    if t < s:
        return v.word
```

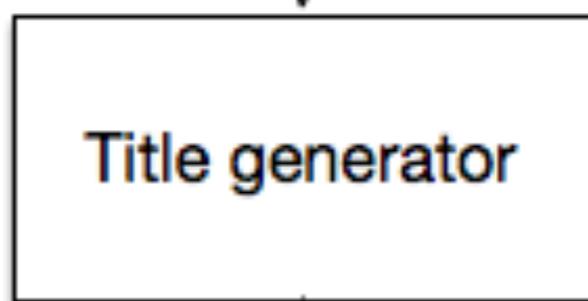
Long Short-Term Memory



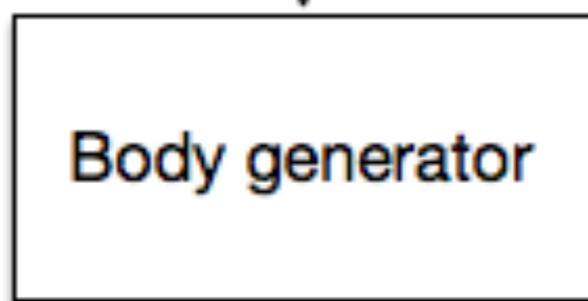
Zdroj: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

LSTM language model





AUTUMN SONG



why don't you kill yourself?
a phone call isn't hope
this planet is still your home
your time is still going on

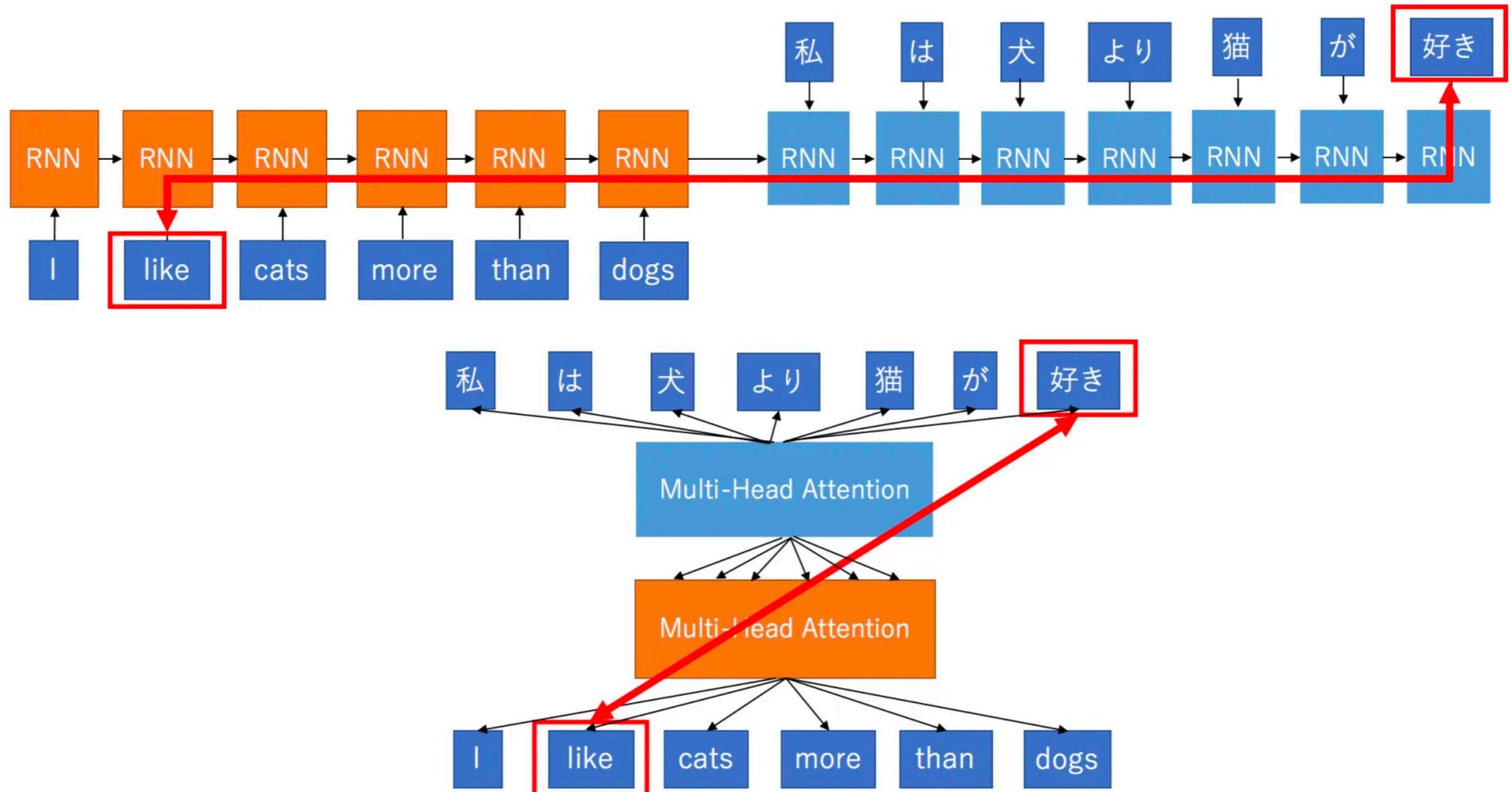
na na na...

I know, I'm revealing a book of dreams
I'll find out I'm nothing more than this

LSTM review generator

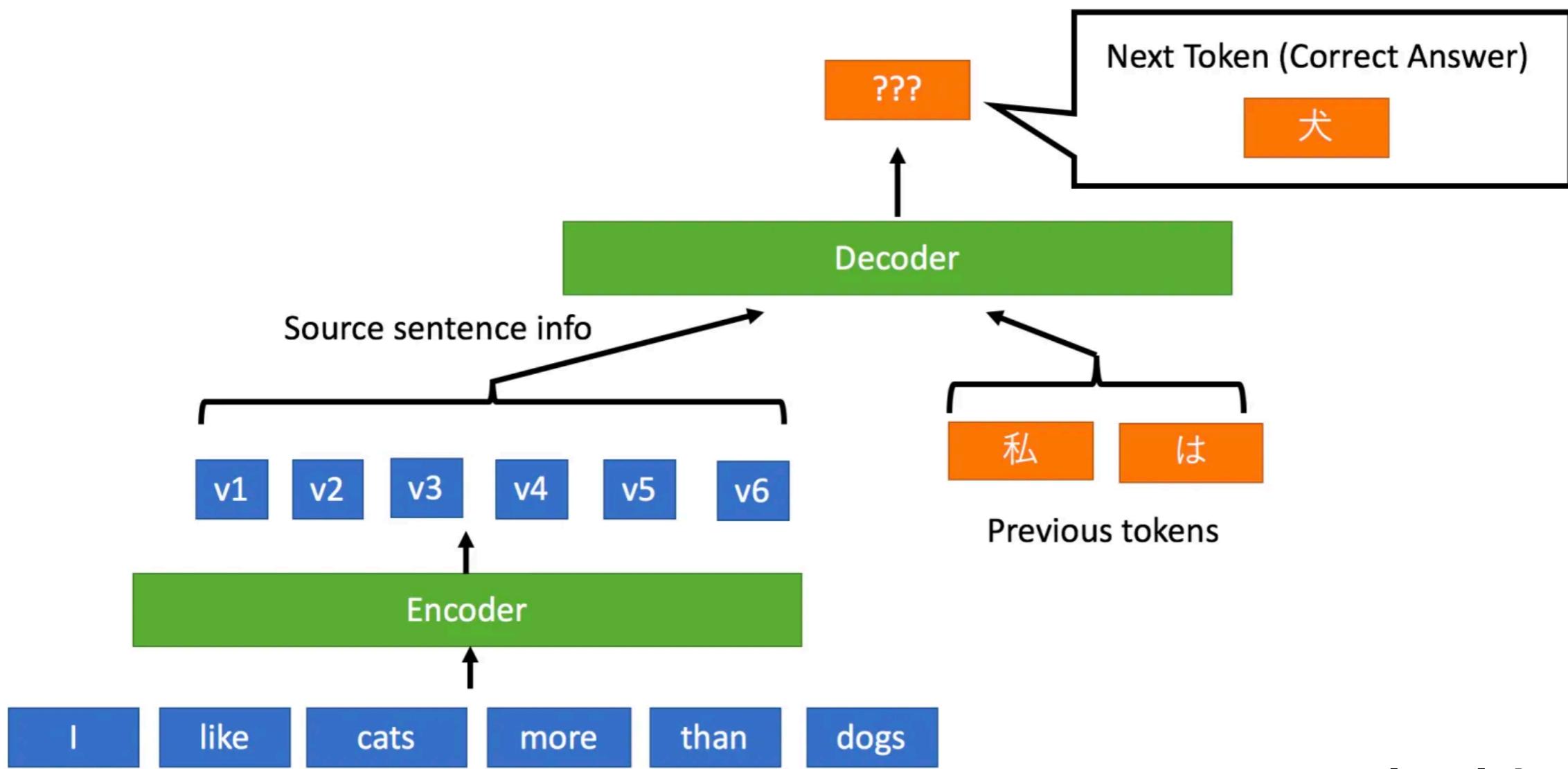
06-Review-generator.ipynb

Transformer



source: www.mlexplained.com

Translation with Transformers



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.

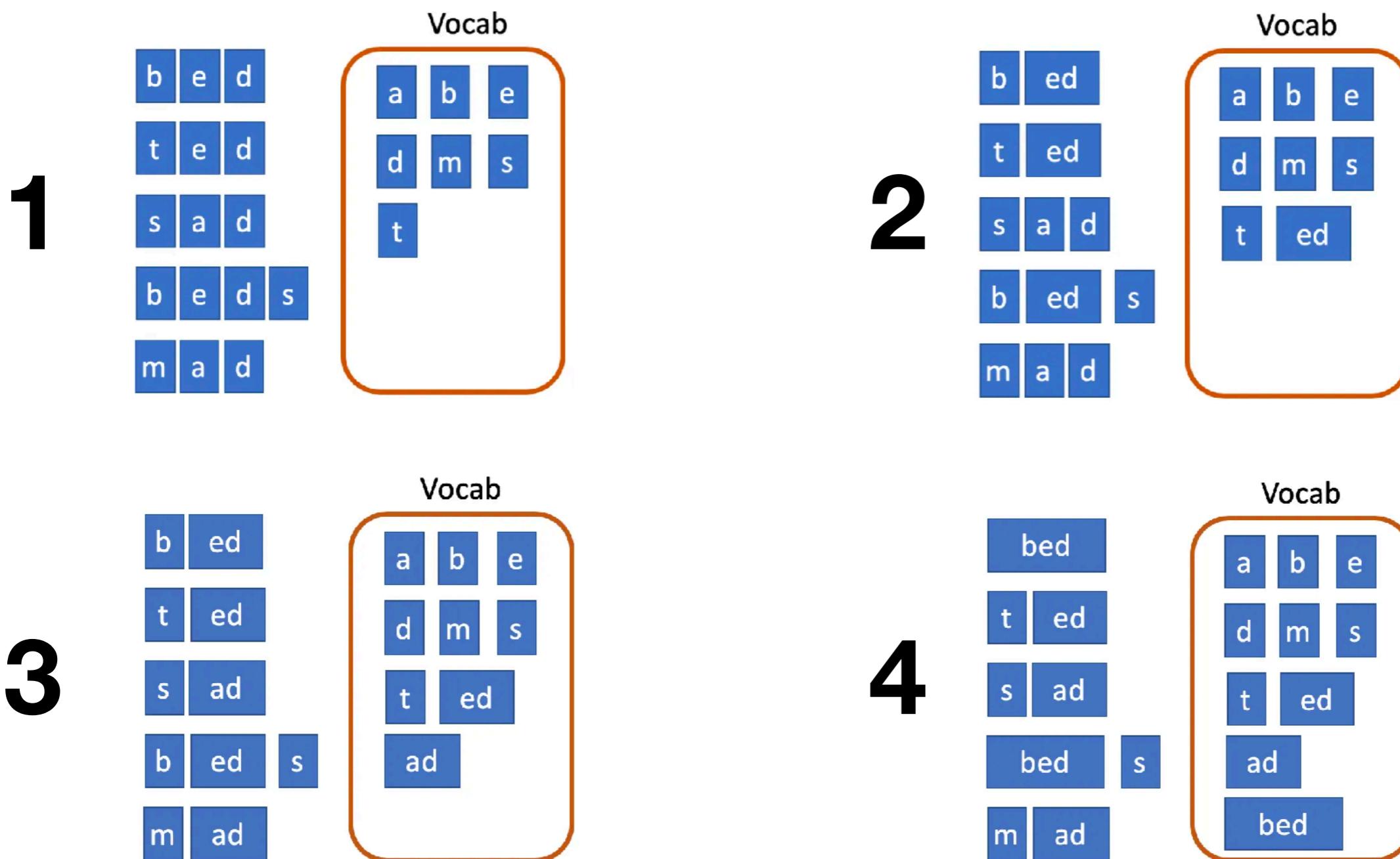
"I'm getting ready for the speech. And I will have a major speech on Tuesday." Trump said during an interview in the White House residence.

GPT-3: <https://openai.com/blog/openai-api/>

BERT

Subword tokenization

Byte-pair encoding



Subword tokenization

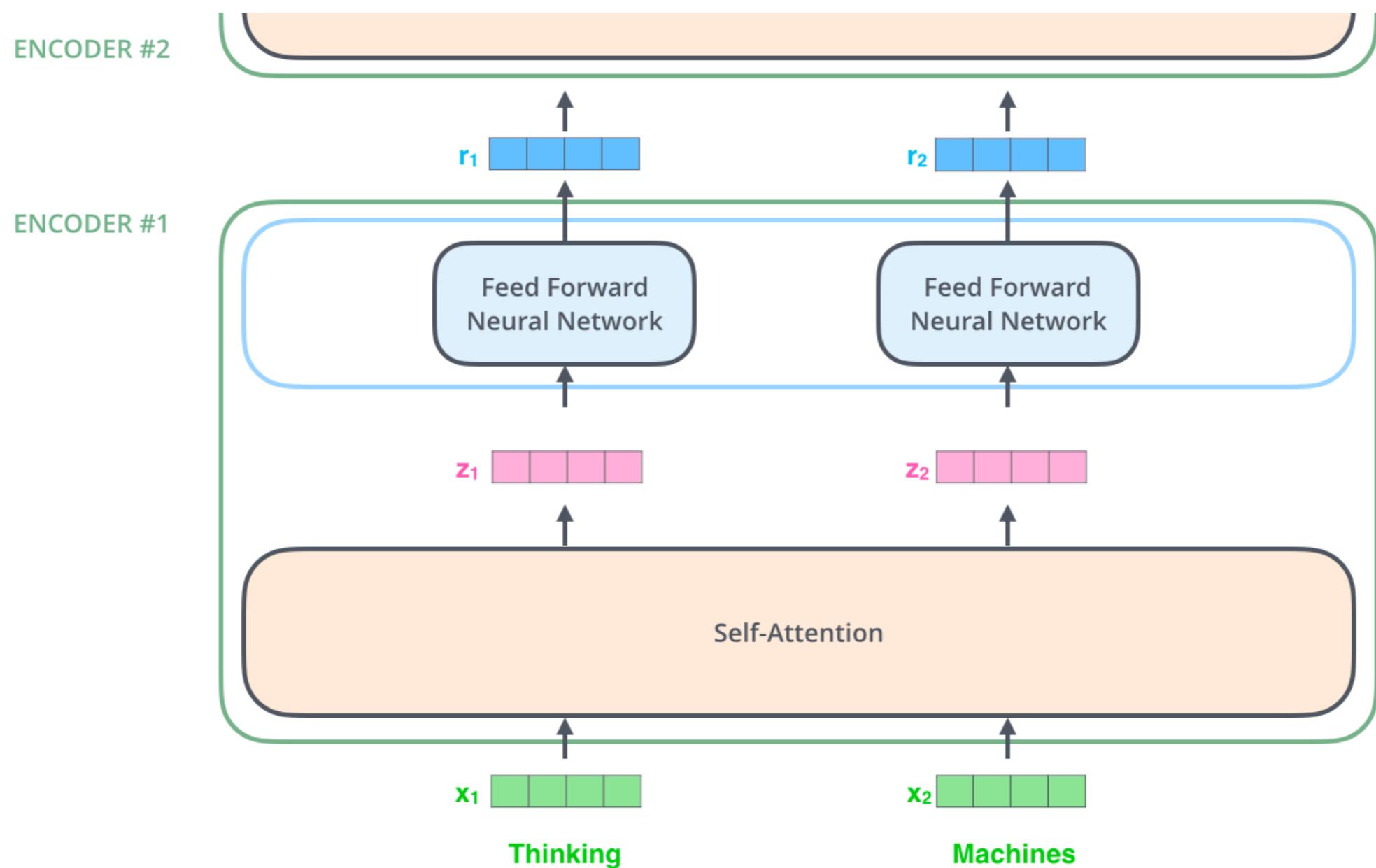
Wordpiece and sentencepiece tokenization

Merges bigrams with maximum mutual information instead of maximum frequency.

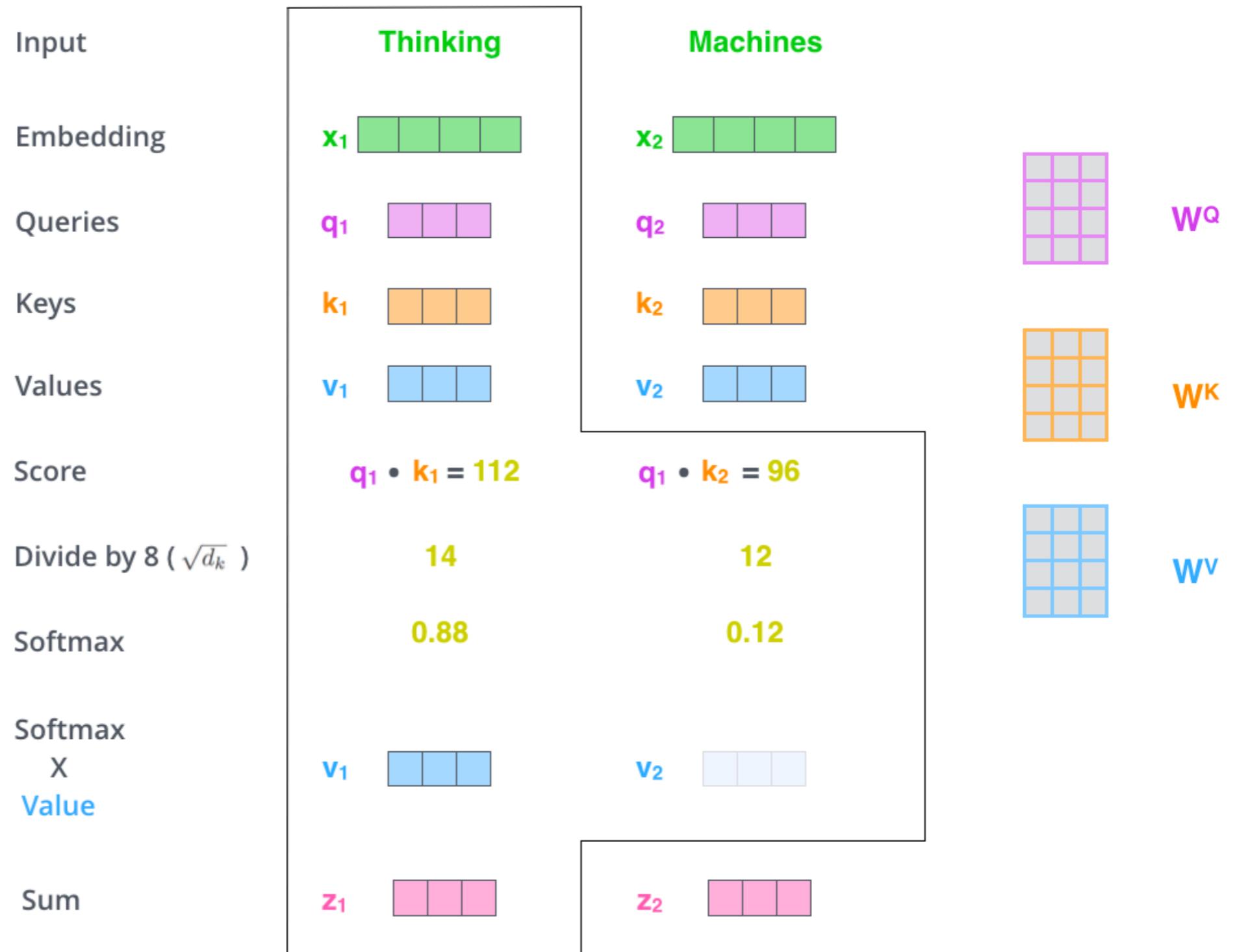
$$I(x, y) = \log \left(\frac{p(x, y)}{p(x) p(y)} \right)$$

playing -> play, ##ing

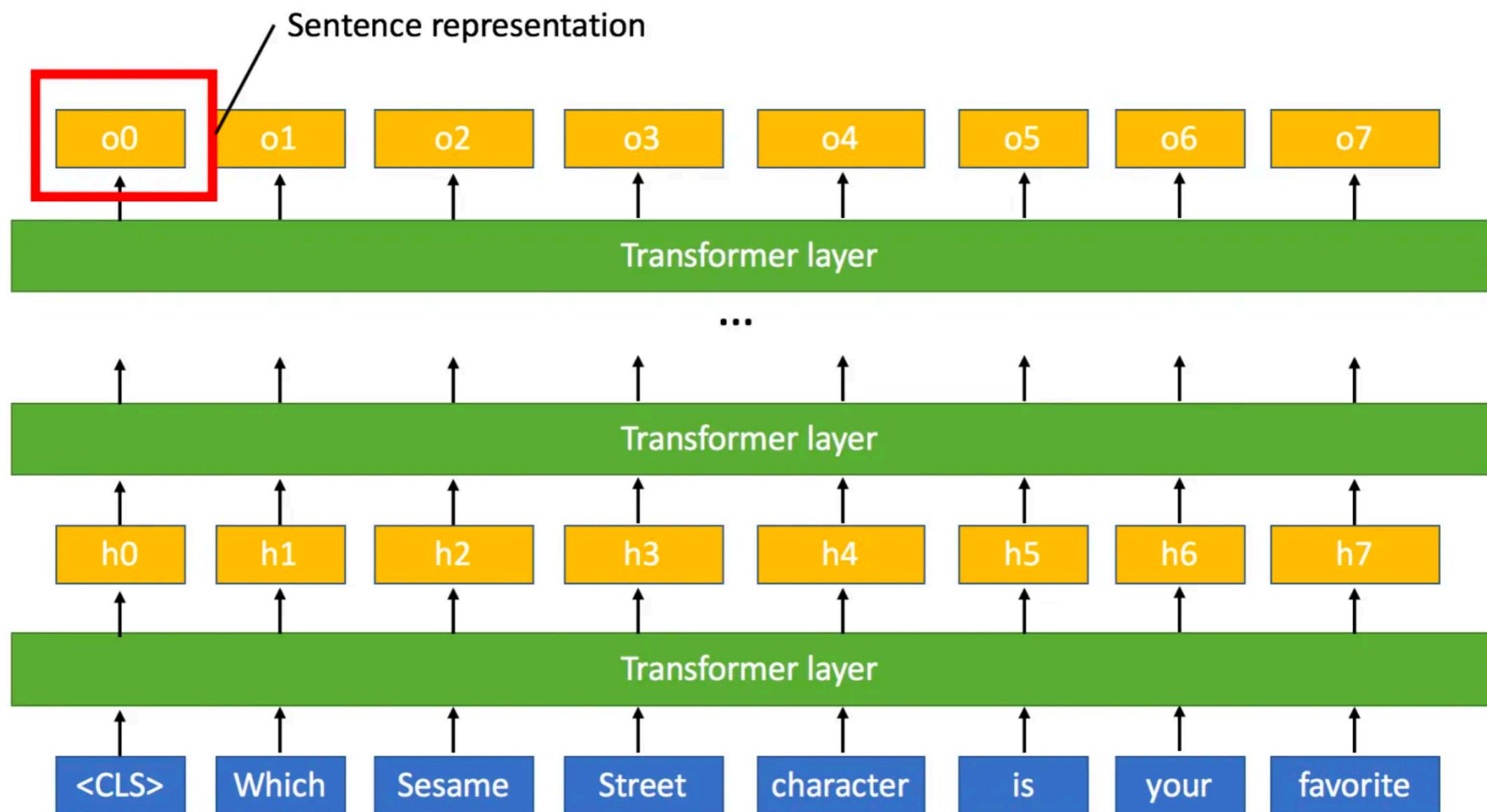
Attention is all you need



Self-attention



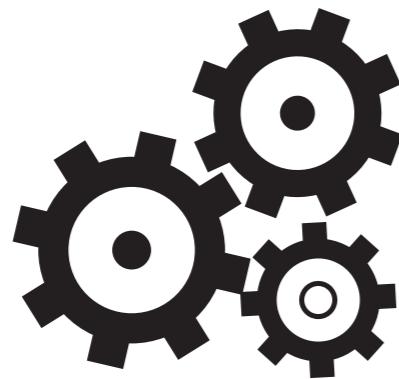
BERT (classification)



Text classification using BERT

08-Review-classification-BERT.ipynb

What next?



Machine Learning Prague

ML MACHINE LEARNING
MU meetups

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/>