Sequence Learning

Markov Chains and n-grams

Korbinian Riedhammer

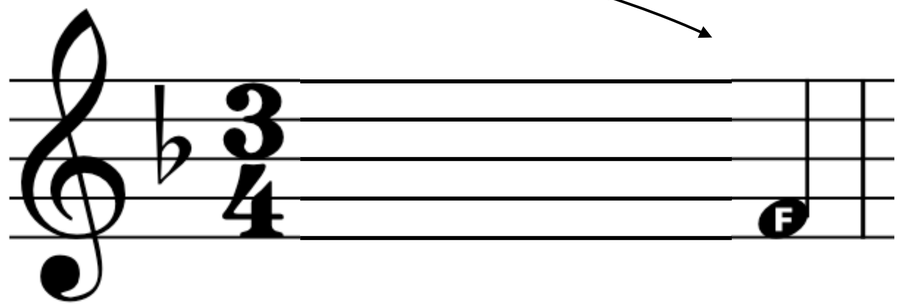
TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Markov Assumption

Next prediction depends only on the recent (rather than entire) history

1st order Markov assumption

great _____



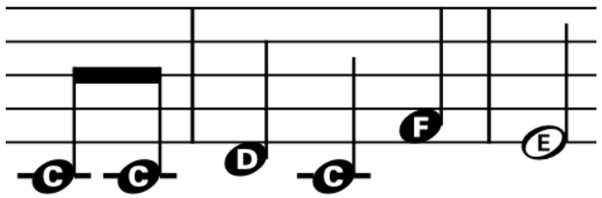
?

Some Terminology

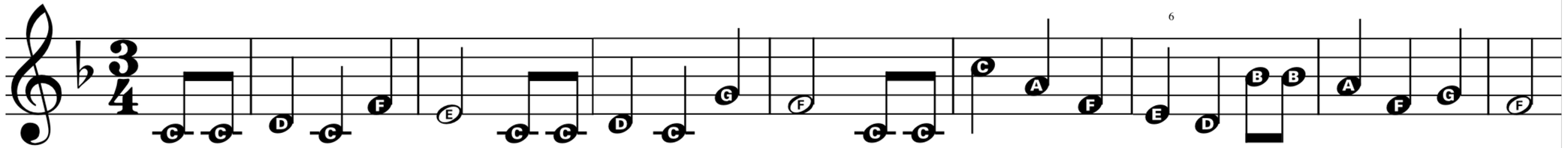
- *Type*: unique symbol or word (programming: class)
- *Token*: instance of a type or word occurrence (programming: object)

Make America great again!

5 types and tokens



5 tyes and 8 tokens



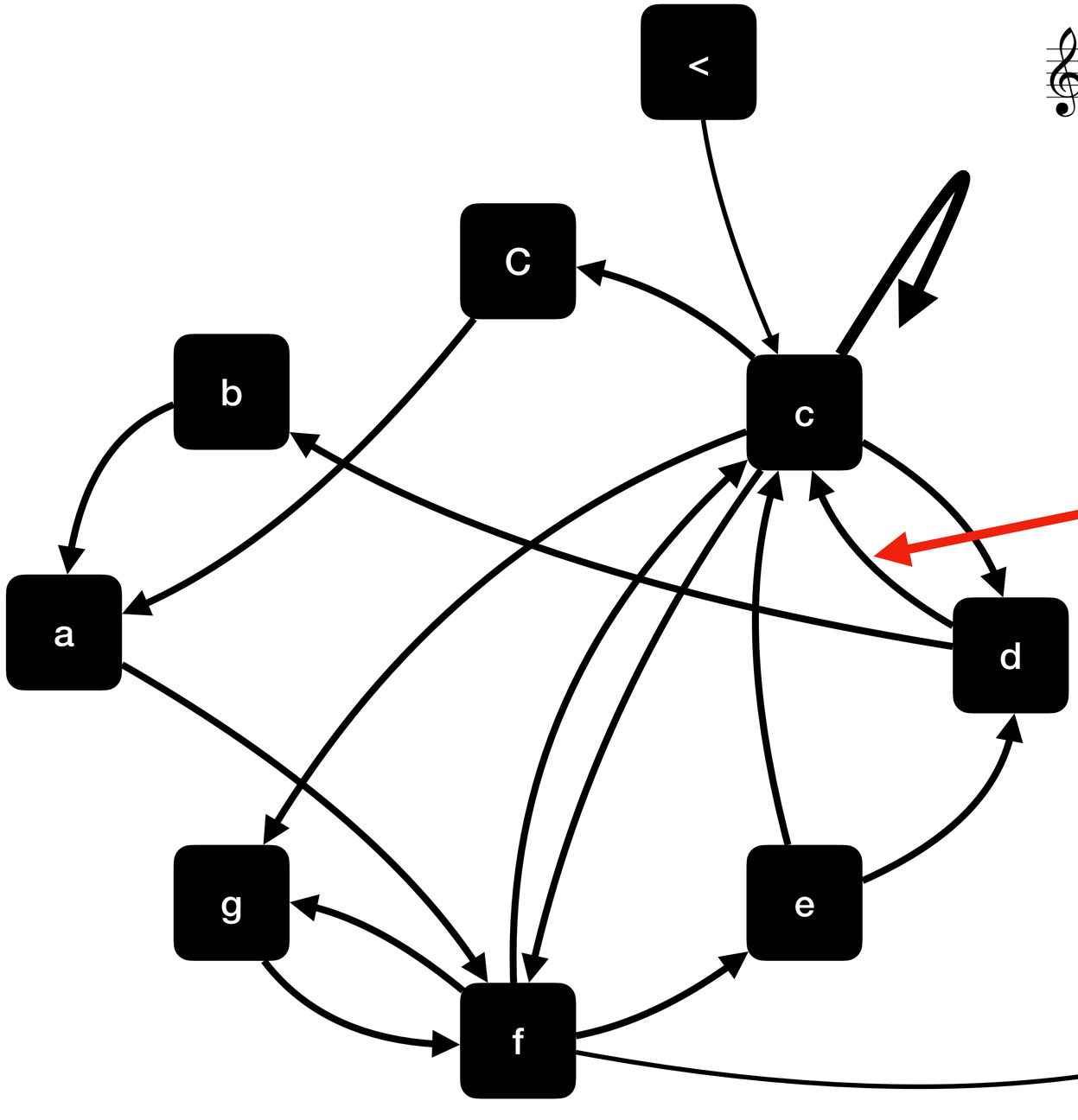
`ccdcfe ccdcgf ccCafed bbafgf`

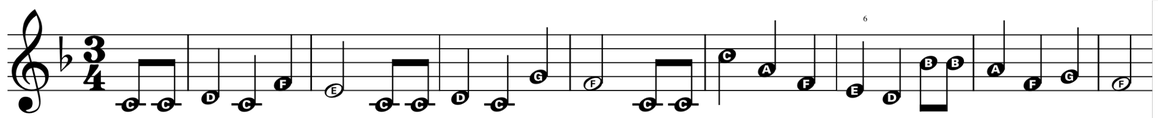
- 8 types: `Cabcdefg`
- 25 tokens
- Probability of each type?
- Probability of type given current token?

$$p(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_i)}{C(w_{n-1})}$$

C: 1
a: 2
b: 2
c: 8
d: 3
e: 2
f: 5
g: 2

$$p(w_i) = \frac{C(w_i)}{\sum_j C(w_j)}$$





<ccdcfeccdcgfccCafedbbafgf/>

$$p(c \mid d) = \frac{C(dc)}{\sum_x C(dx)} = \frac{C(dc)}{C(d)}$$

1st order Markov Chain ~ bi-gram

Probability of Sequences

- Sequence or “sentence” $\boldsymbol{w} = w_1 w_2 \cdots w_n$
- Expand via chain rule
  $p(\boldsymbol{w}) = p(w_1, w_2, \ldots, w_n) = p(w_1) \cdot p(w_2 \mid w_1) \cdots p(w_n \mid w_1 \ldots w_{n-1})$
- Probability of w_i depends on all prior words (=history)
- $p(w_i \mid w_1 w_2 \cdots w_{n-1})$ cannot be estimated for large i
- Limit context by defining equivalence classes for preceding words

Probability of Sequences

…of limited context

- Map longer histories to same (shortened) history

$$p(\boldsymbol{w}) = p(w_1) \cdot \prod_{i=2}^{N} p(w_i \,|\, \phi(w_1 \cdots w_{i-1}))$$

- Bi-gram *language model*

$$p(\boldsymbol{w}) = p(w_1)] \cdot \prod_{i=2}^{N} p(w_i \,|\, w_{i-1})$$

- Tri-gram language model

$$p(\boldsymbol{w}) = p(w_1) \cdot p(w_2 \,|\, w_1) \cdot \prod_{i=3}^{N} p(w_i \,|\, w_{i-2} w_{i-1})$$

Compute in log-domain to avoid underflows!

N-gram models

…or (n-1)-order Markov chains

- Uni-gram: $p(w_i) = \frac{C(w_i)}{\sum_j C(w_j)}$
- Bi-gram: $p(w_n \mid w_{n-1}) = \frac{C(w_{n-1} w_n)}{C(w_{n-1})}$
- N-gram: $p(w_n \mid w_{n-N+1}, \ldots, x_{n-1}) = \frac{C(w_{n-N+1} \cdots w_n)}{C(w_{n-N+1} \cdots w_{n-1})}$

ML Estimates of N-Gram Language Models

- Compute relative frequencies for each 1…n-gram (formulas on prev. slide)
- Relative frequencies ~ ML estimate (for exhaustive context size)

Trumps Language

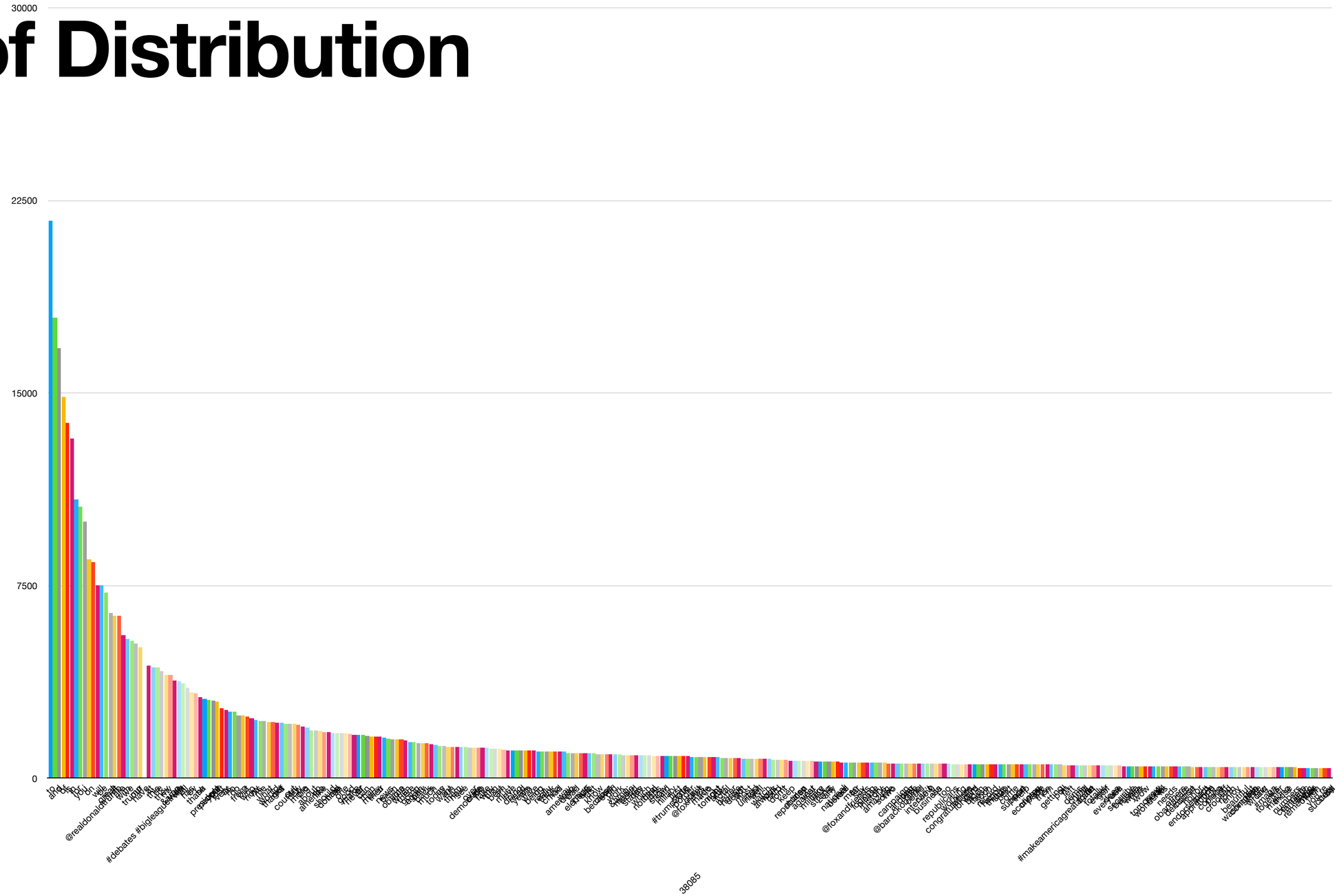
https://www.thetrumparchive.com

- ca. 45k tweets (excluding retweets)
- account permanently suspended Jan 8, 2021
- ca. 44k types, including
  - ca. 15k mentions
  - ca. 2k hashtags
  - ca. 9k that appear more than 5 times

Word counts for Trump's tweets

1	<s>	45417
2	</s>	45417
3	the	38085
4	to	21714
5	and	17937
6	a	16763
7	of	14835
8	is	13838
9	in	13240
10	for	10860
11	you	10555
12	i	9995
13	on	8529
14	@realdonaldtrump	8417
15	will	7522
16	be	7516
17	great	7222

Zipf Distribution



Natural Language “Generation”

…by randomly sampling from n-grams

- if only donald trump appearing today on crony green energy projects
- @srwmichellej @realdonaldtrump and @terrellowens sounds like it thugs
- at 9 00 with top representatives from the inside
- scotland thank you danny boy trump is you the ability to lead normal lives and money the special prosecutor to
- sleepy joe biden is the best businessman in the u s senate from the best evidence is the leadership of
- apprentice tonight at 10 30am est enjoy #trump2016

Natural Language “Generation”

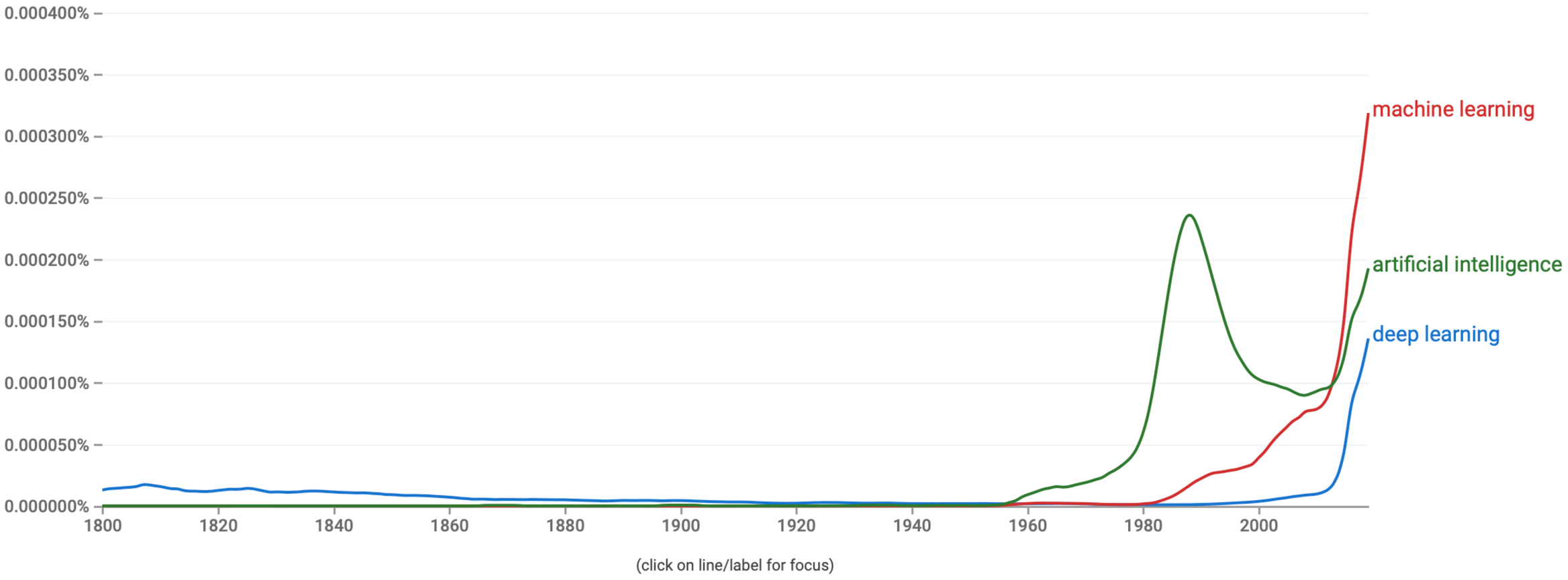
…using seed words

- crooked hillary her rallies are held in the first class funeral
- make america great again people that got rich selling out
- sleepy as he traveled with a coffee in his career tax
- sleepy joe he is the media would come in and they knew

Google Books Ngram Viewer

deep learning,machine learning,artificial intelligence

1800 - 2019 | English (2019) | Case-Insensitive | Smoothing



(click on line/label for focus)

Problems with N-gram Models

- Problem: many n-grams (and words) do not occur in the training corpus $\rightarrow p(\boldsymbol{w}) = 0$
  - Example: size of vocabulary $|V| = 20{,}000 \rightarrow 8 \cdot 10^{12}$ possible trigrams
- Use
  - Special symbol for unknown word (“unk”)
  - smoothing of the estimates
  - Categories/bins of words

Smoothing of N-grams

Lapcace (“add-one”) Smoothing

- To prevent zero counts, add 1 to any count.
- $\hat{C}(w_{i-n+1} \cdots w_i) = 1 + C(w_{i-n+1}$
- $p(w_i \mid w_{i-n+1} \cdots w_{i-1}) = \frac{\hat{C}(w_{i-n+1} \cdots w_i)}{\hat{C}(w_{i-n+1} \cdots w_{i-1})} = \frac{1 + C(w_{i-n+1} \cdots w_i)}{|V| + \sum_{w_i \in V} C(w_{i-n+1} \cdots w_i)}$
- Drawback: rather high probability for non-existing (=missing) n-grams

Smoothing of N-grams

Good-Turing Smoothing (1953)

- Idea: n-grams with the same count r in the training corpus should be assigned the same probability
- Good-Turing (GT) estimate $r^\star = (r+1)\frac{\eta_{r+1}}{\eta_r}$
- η_r is the number of n-grams which occurred exactly r times
- $\hat{p}_k = r^\star / N$ with $r = N_k$
- GT requires that $\eta_r \neq 0$ for all n-grams (→ other smoothing may be necessary)
- Frequencies not arbitrarily changed but **redistributed**

Smoothing of N-grams

Back-Off

- Idea: missing n-grams are mapped onto a lower order n-gram
- $$p_{\text{bo}}(w \mid \boldsymbol{v}) = \begin{cases} p(w \mid \boldsymbol{v}) & \text{if } C(\boldsymbol{v}w) > 0 \\ \beta(\boldsymbol{v}) \cdot p_{\text{bo}}(w \mid \boldsymbol{v}') & \text{if } C(\boldsymbol{v}w) = 0 \end{cases}$$
- $\boldsymbol{v}'$ is the reduced/shortened history
- Weight $\beta(\boldsymbol{v})$ (“back-off weight”) guarantees stochasticity of p_{bo}

Smoothing of N-grams

Interpolation

- Observation: For small r, the ML estimate is very imprecise
- Mitigation: include statistics of lower order even if $r > 0$
- Linear interpolation

  $$p_I(w_n \mid w_1 \cdots w_{n-1}) = \rho_0 \frac{1}{|V|} + \rho_1 p(w_n) + \rho_2 p(w_n \mid w_{n-1}) + \ldots$$

  with $\sum_i \rho_i = 1$
- Learn weights using cross-validation and expectation maximization (EM)

Smoothing of N-grams

Other Methods

- Kneser-Ney interpolation: non-linear interpolation of statistics
- Interpolation using max-ent
- Log-linear interpolation

Using Categories

- Instead of actual types, use **categories**
- Group words with similar functional an statistical properties $\mathscr{C} = \{\mathscr{C}_1, \ldots, \mathscr{C}_n\}$ with $\bigcup_k \mathscr{C}_k = V$
  - Numerals, names, cities, persons, months, etc.
  - Should be dissect; what about “Essen” (city and food)
- For a categorical bi-gram: $p(w_1 \cdots w_n) = \prod_i p(w_i \,|\, \mathscr{C}(w_i)) \cdot p(\mathscr{C}(w_i) \,|\, \mathscr{C}(w_{i-1}))$

Cache Models

- Context may shift: dynamically adapt statistics to the current circumstances
- Interpolate a static model $p^{(s)}$ linearly with a cache model $p^{(c)}$
  $\bar{p}(w_i \mid w_{i-n+1} \cdots w_{i-1}) = \rho_c p^{(s)}(w_i \mid w_{i-n+1} \cdots w_{i-1}) + (1 - \rho_c) p^{(c)}(w_i \mid w_{i-1})$
- Cache model is continuously updated, typically low order
- Interpolation weight ρ_c dependent on cache size

Measuring Fit

How well does a model fit an observation?

- Compute the (log-)probability (“production probability”) → directly impacted by sentence length
- Perplexity $\mathrm{PP}(\boldsymbol{w}) = \sqrt[n]{\frac{1}{p(w_1 \cdots w_n)}}$, normalized by sentence length
- Measures the weighted average branching factor in predicting the next word
- Lower is better

Using N-gram Models for Classification

- Train an n-gram model for each class (1) and (2)
- For binary classification: compute log ratio

$$s(\boldsymbol{w}) = \log \frac{p^{(1)}(\boldsymbol{w})}{p^{(2)}(\boldsymbol{w})} = \begin{cases} > 0 & \text{if (1) is more likely} \\ \leq 0 & \text{if (2) is more likely} \end{cases}$$

- Other class could be “background class”
- Example usages:
  - Discriminate speech of healthy and people with Alzheimer’s
  - Author attribution

Limits of N-grams

:-)

Recognizer output (actually impressive!)

im Bundesrat klar als Bedingung. Das heißt also Absenkung des na des des des des des eh na… ist es Alters das ist das Alter der Kinder wenn sie des Nachzugsalters. Dann kommt der fünfte Punkt und das sechste Punkt kommen sicherlich die Fragen gleich gleichgeschl gleich die gleiche schließen dann ob ich aufhöre Asylgründe schaffte außerhalb der politischen und der rassistischen Verfolgung also auch Gründe wenn aus wenn wenn andere Gründe sozusagen also aus dem Geschlecht oder ähnlichen stattfinden das er seine Frauen die irgendwie wegen ihres Frau sei es irgendwo verfolgt werden. Ob ich denen jetzt ein Asyl Asylgrund gebe…

Edmund Stoiber bei Christiansen, Jan 20, 2002