# Word2Vec und das Skip-Gram-Modell von Grund auf skripten: Eigene Word Embeddings erstellen

Aleksandr Schamberger

Humboldt-Universität zu Berlin
Institut für Slawistik und Hungarologie
Sprachenübergreifend: Computerlinguistik II – Digitale Sprachmodelle und ihre Anwendung
Sommersemester 2024

2. Juli 2024

# Gliederung

# Gliederung

# Wörter (sinnvoll) als Zahlen darstellen

- Bank
- Stuhl
- Sitzgelegenheit
- dick
- dünn
- schlank

# ASCII-Tabelle

| Scan-code | ASCII hex | ASCII dez | Zeichen |
|---|---|---|---|
| | 00 | 0 | NUL ^@ |
| | 01 | 1 | SOH ^A |
| | 02 | 2 | STX ^B |
| | 03 | 3 | ETX ^C |
| | 04 | 4 | EOT ^D |
| | 05 | 5 | ENQ ^E |
| | 06 | 6 | ACK ^F |
| | 07 | 7 | BEL ^G |
| 0E | 08 | 8 | BS ^H |
| 0F | 09 | 9 | TAB ^I |
| | 0A | 10 | LF ^J |
| | 0B | 11 | VT ^K |
| | 0C | 12 | FF ^L |
| 1C | 0D | 13 | CR ^M |
| | 0E | 14 | SO ^N |
| | 0F | 15 | SI ^O |
| | 10 | 16 | DLE ^P |
| | 11 | 17 | DC1 ^Q |
| | 12 | 18 | DC2 ^R |
| | 13 | 19 | DC3 ^S |
| | 14 | 20 | DC4 ^T |
| | 15 | 21 | NAK ^U |
| | 16 | 22 | SYN ^V |
| | 17 | 23 | ETB ^W |
| | 18 | 24 | CAN ^X |
| | 19 | 25 | EM ^Y |
| | 1A | 26 | SUB ^Z |
| 01 | 1B | 27 | Esc ^[ |
| | 1C | 28 | FS ^\ |
| | 1D | 29 | GS ^] |
| | 1E | 30 | RS ^^ |
| | 1F | 31 | US ^_ |

| Scan-code | ASCII hex | ASCII dez | Zch. |
|---|---|---|---|
| | 20 | 32 | SP |
| 02 | 21 | 33 | ! |
| 03 | 22 | 34 | " |
| 29 | 23 | 35 | # |
| 05 | 24 | 36 | $ |
| 06 | 25 | 37 | % |
| 07 | 26 | 38 | & |
| 0D | 27 | 39 | ' |
| 09 | 28 | 40 | ( |
| 0A | 29 | 41 | ) |
| 1B | 2A | 42 | * |
| 1B | 2B | 43 | + |
| 33 | 2C | 44 | , |
| 35 | 2D | 45 | - |
| 34 | 2E | 46 | . |
| 08 | 2F | 47 | / |
| 0B | 30 | 48 | 0 |
| 02 | 31 | 49 | 1 |
| 03 | 32 | 50 | 2 |
| 04 | 33 | 51 | 3 |
| 05 | 34 | 52 | 4 |
| 06 | 35 | 53 | 5 |
| 07 | 36 | 54 | 6 |
| 08 | 37 | 55 | 7 |
| 09 | 38 | 56 | 8 |
| 0A | 39 | 57 | 9 |
| 34 | 3A | 58 | : |
| 33 | 3B | 59 | ; |
| 2B | 3C | 60 | < |
| 0B | 3D | 61 | = |
| 2B | 3E | 62 | > |
| 0C | 3F | 63 | ? |

| Scan-code | ASCII hex | ASCII dez | Zch. |
|---|---|---|---|
| | 40 | 64 | @ |
| 1E | 41 | 65 | A |
| 30 | 42 | 66 | B |
| 2E | 43 | 67 | C |
| 20 | 44 | 68 | D |
| 12 | 45 | 69 | E |
| 21 | 46 | 70 | F |
| 22 | 47 | 71 | G |
| 23 | 48 | 72 | H |
| 17 | 49 | 73 | I |
| 24 | 4A | 74 | J |
| 25 | 4B | 75 | K |
| 26 | 4C | 76 | L |
| 32 | 4D | 77 | M |
| 31 | 4E | 78 | N |
| 18 | 4F | 79 | O |
| 19 | 50 | 80 | P |
| 10 | 51 | 81 | Q |
| 13 | 52 | 82 | R |
| 1F | 53 | 83 | S |
| 14 | 54 | 84 | T |
| 16 | 55 | 85 | U |
| 2F | 56 | 86 | V |
| 11 | 57 | 87 | W |
| 2D | 58 | 88 | X |
| 2C | 59 | 89 | Y |
| 15 | 5A | 90 | Z |
| | 5B | 91 | [ |
| | 5C | 92 | \ |
| | 5D | 93 | ] |
| 29 | 5E | 94 | ^ |
| 35 | 5F | 95 | _ |

| Scan-code | ASCII hex | ASCII dez | Zch. |
|---|---|---|---|
| 0D | 60 | 96 | ` |
| 1E | 61 | 97 | a |
| 30 | 62 | 98 | b |
| 2E | 63 | 99 | c |
| 20 | 64 | 100 | d |
| 12 | 65 | 101 | e |
| 21 | 66 | 102 | f |
| 22 | 67 | 103 | g |
| 23 | 68 | 104 | h |
| 17 | 69 | 105 | i |
| 24 | 6A | 106 | j |
| 25 | 6B | 107 | k |
| 26 | 6C | 108 | l |
| 32 | 6D | 109 | m |
| 31 | 6E | 110 | n |
| 18 | 6F | 111 | o |
| 19 | 70 | 112 | p |
| 10 | 71 | 113 | q |
| 13 | 72 | 114 | r |
| 1F | 73 | 115 | s |
| 14 | 74 | 116 | t |
| 16 | 75 | 117 | u |
| 2F | 76 | 118 | v |
| 11 | 77 | 119 | w |
| 2D | 78 | 120 | x |
| 2C | 79 | 121 | y |
| 15 | 7A | 122 | z |
| | 7B | 123 | { |
| | 7C | 124 | \| |
| | 7D | 125 | } |
| | 7E | 126 | ~ |
| 53 | 7F | 127 | DEL |

# Wörter als deren ASCII-Zeichenkodierung

- Bank
  - 66 97 110 107 10
- Stuhl
  - 83 116 117 104 108
- Sitzgelegenheit
  - 83 105 116 122 103 101 108 101 103 101 110 104 101 105 116
- dick
  - 100 105 99 107
- dünn
  - 100 195 188 110 110
- schlank
  - 115 99 104 108 97 110 107

# Semantische Relationen (und Eigenschaften)

- **Hyponymie/Hyperonymie**: (Bank,Stuhl) $\rightarrow$ Sitzgelegenheit
- **Ko-Hyponymie**: Bank - Stuhl
- **Antonymie**: dünn - dick
- **Synonymie**: dünn - schlank
- **Ambiguität**: Bank - Bank

# Semantische Relationen (und Eigenschaften)

- **Hyponymie/Hyperonymie**: (Bank,Stuhl) $\rightarrow$ Sitzgelegenheit
- **Ko-Hyponymie**: Bank - Stuhl
- **Antonymie**: dünn - dick
- **Synonymie**: dünn - schlank
- **Ambiguität**: Bank - Bank

Und viele weitere, sprachliche Eigenschaften ...

# Zwischenlösung: One-Hot-Vektoren

- Bank
  - 1 0 0 0 0 0
- Stuhl
  - 0 1 0 0 0 0
- Sitzgelegenheit
  - 0 0 1 0 0 0
- dick
  - 0 0 0 1 0 0
- dünn
  - 0 0 0 0 1 0
- schlank
  - 0 0 0 0 0 1

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

- Aktiver Wortschatz: 12.000 - 16.000

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

- Aktiver Wortschatz: 12.000 - 16.000
- Passiver Wortschatz: 50.000

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

- Aktiver Wortschatz: 12.000 - 16.000
- Passiver Wortschatz: 50.000
- Im Allgemeinen angenommen: 300.000 - 500.000

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

- Aktiver Wortschatz: 12.000 - 16.000
- Passiver Wortschatz: 50.000
- Im Allgemeinen angenommen: 300.000 - 500.000
- Dudenkorpus: $>$ 18 Millionen

# Problem: Der Umfang des deutschen Wortschatzes

Umfang deutscher Wörter (Grundformen)

- Aktiver Wortschatz: 12.000 - 16.000
- Passiver Wortschatz: 50.000
- Im Allgemeinen angenommen: 300.000 - 500.000
- Dudenkorpus: $>$ 18 Millionen

10.000 Wörter: Bank $\rightarrow 1_1 0_2 0_3 0_4 0_5 0_6 0_7 0_8 0_9 \ldots 0_{10.000}$

# Gliederung

# Von weiten Vektoren zu dichten Vektoren

10.000 Wörter: Bank $\rightarrow 1_1 0_2 0_3 0_4 0_5 0_6 0_7 0_8 0_9 \ldots 0_{10.000}$

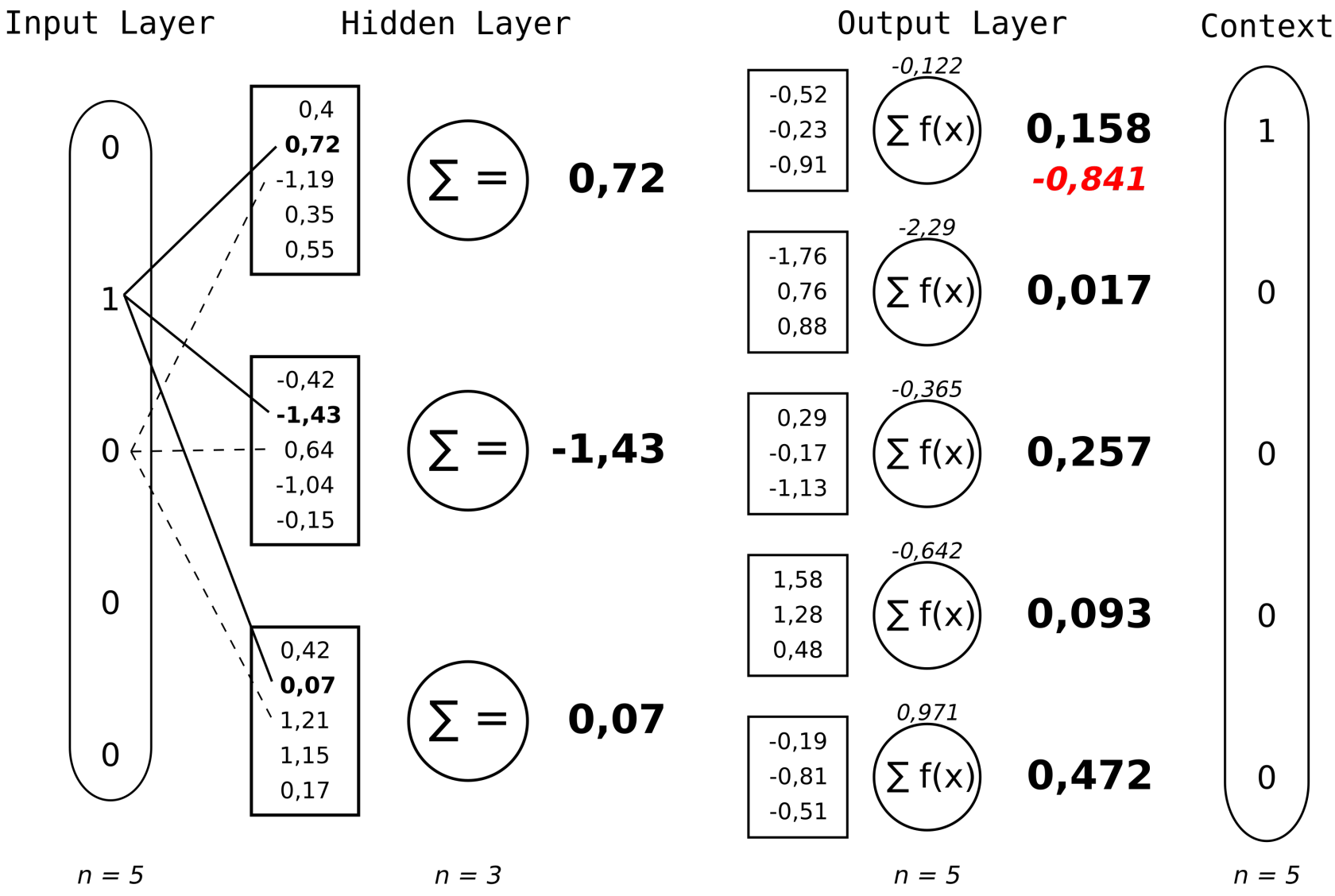
# Von weiten Vektoren zu dichten Vektoren

10.000 Wörter: Bank $\rightarrow 1_1 0_2 0_3 0_4 0_5 0_6 0_7 0_8 0_9 \ldots 0_{10.000}$

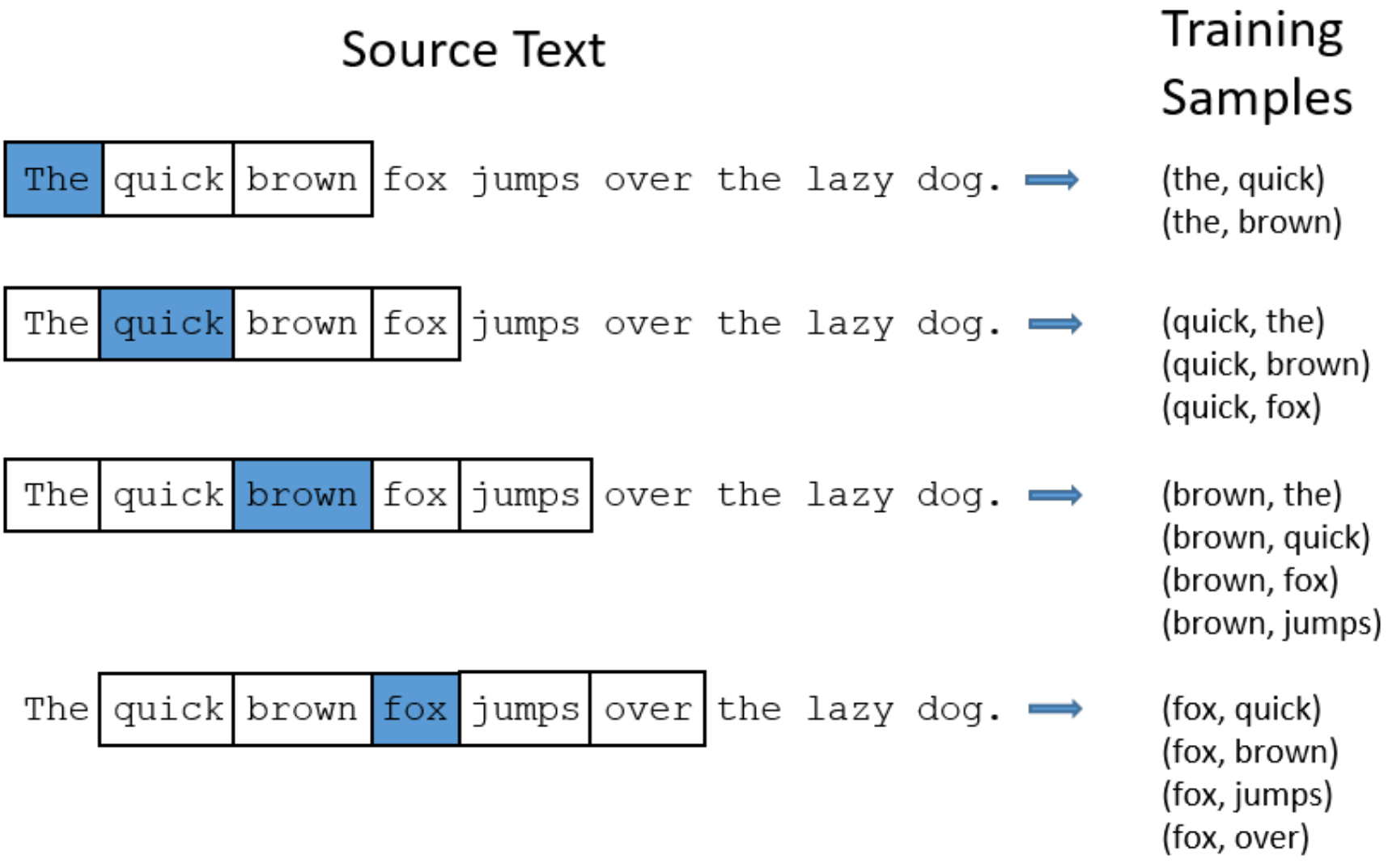
$\leftrightarrow$

10.000 Wörter: Bank $\rightarrow$
$-1, 23_1 3, 4_2 0, 09_3 0_4 1, 1_5 0, 98_6 - 2, 34_7 0, 11_8 - 1, 3_9 \ldots 0_{300}$

# Skip-Gram-Modell: Gesamtmodell mit Beispielrechnung



# Skip-Gram-Modell: Kontext (Syntagma)



# Quellen


- Tae, Jake (13.07.2020): Word2vec from Scratch <https://jaketae.github.io/study/word2vec/> (abgerufen am 02.07.2024)
- McCormick, Chris (19.04.2016): Word2Vec Tutorial - The Skip-Gram Model <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/> (abgerufen am 02.07.2024)
- Tae, Jake (05.02.2020): Building Neural Network From Scratch <https://jaketae.github.io/study/neural-net/> (abgerufen am 02.07.2024)
- Tae, Jake (21.12.2019): Demystifying Entropy (And More) <https://jaketae.github.io/study/information-entropy/> (abgerufen am 02.07.2024)

# Skip-Gram-Modell: Beispiele

**Beispiel(e) im Editor!**