Recap

- Types and tokens
- Zipf distribution
- Type-Token-Ratio
- Encoding
- Unicode
- Concordances

Collocations

Sprachverarbeitung (VL + Ü)

Nils Reiter

April 20, 2023

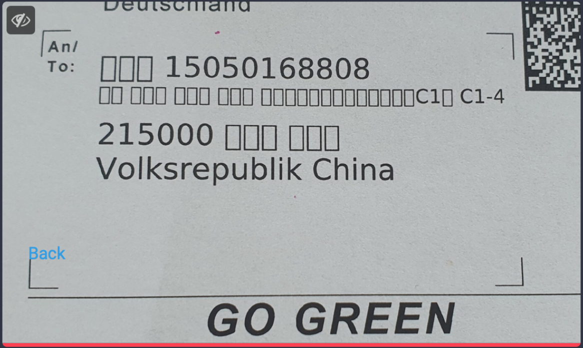
INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Unicode

Not a solved problem

Marina Weisband
@afelia@chaos.social

Much wow, DHL.
Versandmarken bei einem weltweiten Logistiker kaufen, der seine IT voll unter Kontrolle hat.

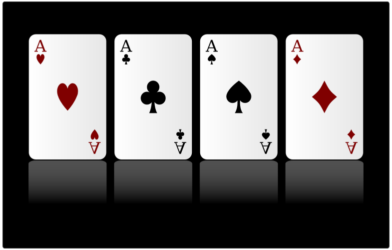


17 Apr 2023 at 13:11 · 54 · 176

Section 1

Basic Probability Theory

Example: Cards



- 32 cards Ω (sample space)
- 4 ›colors‹: $C = \{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$
- 8 values: $V = \{7, 8, 9, 10, J, Q, K, A\}$
- Individual cards (›outcomes‹) are denoted with value and color: $8\heartsuit$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond«

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- »We draw a queen«

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- »We draw a queen« – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- »We draw a queen« – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- »We draw a heart eight or diamond 10«

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- »We draw a queen« – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- »We draw a heart eight or diamond 10« – $E = \{8\heartsuit, 10\diamondsuit\}$
- »We draw any card«

Basics

Events

- Generally, we draw cards from a (well shuffled) deck
- We define what events we are interested in
- An event can be any subset of the sample space Ω
- Events will be denoted with E

Examples

- »We draw a heart eight« – $E = \{8\heartsuit\}$
- »We draw card with a diamond« – $E = \{7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}$
- »We draw a queen« – $E = \{Q\clubsuit, Q\spadesuit, Q\diamondsuit, Q\heartsuit\}$
- »We draw a heart eight or diamond 10« – $E = \{8\heartsuit, 10\diamondsuit\}$
- »We draw any card« – $E = \Omega$

Basics

Probabilities

- Probability $p(E)$: Ratio of size of E to size of Ω (Laplace)
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Basics

Probabilities

- Probability $p(E)$: Ratio of size of E to size of Ω (Laplace)
  - $0 \leq p \leq 1$
  - $p(E) = 0$: Impossible event $\qquad p(E) = 1$: Certain event
  - $p(E) = 0.000001$: Very unlikely event

Example

- If all outcomes are equally likely: $p(E) = \frac{|E|}{|\Omega|}$
- $p(\{8\heartsuit\}) = \frac{1}{32}$
- $p(\{9\clubsuit, 9\spadesuit, 9\diamondsuit, 9\heartsuit\}) = \frac{4}{32}$
- $p(\Omega) = 1$ (must happen, certain event)

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
    - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{\spadesuit\}) = \frac{8}{10}$

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
  - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{\spadesuit\}) = \frac{8}{10}$

- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$

Basics

Probability and Relative Frequency

- Probability p: Theoretical concept, idealization, expectation
- Relative Frequency f: Concrete measure
  - Normalised number of *observed* events

Example

After 10 cards (with returning and shuffling), the event ♠ took place 8 times: $f(\{♠\}) = \frac{8}{10}$

- For large numbers of drawings, relative frequency approximates the probability
  - $\lim_{\infty} f = p$
- In practice, we will often use determine probabilities by counting relative frequencies
  - Assumption: Frequency is measured on representative and large data set

Independent Events

Joint Probability

- We are often interested in multiple events (and their relation)
- E: We draw $8\heartsuit$ two times in a row (putting the first card back)
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$

Independent Events

Joint Probability

- We are often interested in multiple events (and their relation)
- E: We draw $8\heartsuit$ two times in a row (putting the first card back)
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$
- E: We draw $\heartsuit$ two times in a row (putting the first card back)
  - E_1: First card is $X\heartsuit$
  - E_2: Second card is $X\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$

Independent Events

Joint Probability

- We are often interested in multiple events (and their relation)
- E: We draw $8\heartsuit$ two times in a row (putting the first card back)
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{32} * \frac{1}{32} = 0.0156$
- E: We draw $\heartsuit$ two times in a row (putting the first card back)
  - E_1: First card is $X\heartsuit$
  - E_2: Second card is $X\heartsuit$
  - $p(E) = p(E_1, E_2) = p(E_1) * p(E_2) = \frac{1}{4} * \frac{1}{4} = 0.0625$
- These events are independent
  - because we return and re-shuffle the cards all the time
  - Drawing $8\heartsuit$ the first time has no influence on the second drawing
  - Default case with dice

Dependent Events

Conditional Probability

- We no longer return the card
- E: We draw $8\heartsuit$ two times in a row
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $\cancel{p(E_1, E_2) = p(E_1) * p(E_2)}$
  - This no longer works, because the events are not independent
    - Obvious: Only one $8\heartsuit$ in the game, and $p(E_2)$ has to express that it might be gone

Dependent Events

Conditional Probability

- We no longer return the card
- E: We draw $8\heartsuit$ two times in a row
  - E_1: First card is $8\heartsuit$
  - E_2: Second card is $8\heartsuit$
  - $\cancel{p(E_1, E_2) = p(E_1) * p(E_2)}$
  - This no longer works, because the events are not independent
    - Obvious: Only one $8\heartsuit$ in the game, and $p(E_2)$ has to express that it might be gone
  - This is done with the notion of conditional probability
  - $p(E_1, E_2) = p(E_1) * p(E_2|E_1)$
    - $p(E_2|E_1) = 0$, therefore $p(E) = 0$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$\begin{aligned} p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\ &= \end{aligned}$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

$$\begin{aligned} p(E_{\heartsuit}, E_{\heartsuit}) &= p(E_{\heartsuit}) * p(E_{\heartsuit} | E_{\heartsuit}) \\ &= \frac{8}{32} * \end{aligned}$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$
\begin{aligned}
p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\
&= \frac{8}{32} * \frac{7}{31} = 0.056
\end{aligned}
$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_{\heartsuit}$: Card is $X\heartsuit$
- $E_{\diamondsuit}$: Card is $X\diamondsuit$

$$
\begin{aligned}
p(E_{\heartsuit}, E_{\heartsuit}) &= p(E_{\heartsuit}) * p(E_{\heartsuit} | E_{\heartsuit}) \\
&= \frac{8}{32} * \frac{7}{31} = 0.056 \\
p(E_{\diamondsuit}, E_{\heartsuit}) &= p(E_{\diamondsuit}) * p(E_{\heartsuit} | E_{\diamondsuit}) \\
&=
\end{aligned}
$$

Dependent Events

Conditional Probability

A less obvious example:

- We draw two cards in a row
- $E_\heartsuit$: Card is $X\heartsuit$
- $E_\diamondsuit$: Card is $X\diamondsuit$

$$
\begin{aligned}
p(E_\heartsuit, E_\heartsuit) &= p(E_\heartsuit) * p(E_\heartsuit | E_\heartsuit) \\
&= \frac{8}{32} * \frac{7}{31} = 0.056 \\
p(E_\diamondsuit, E_\heartsuit) &= p(E_\diamondsuit) * p(E_\heartsuit | E_\diamondsuit) \\
&= \frac{8}{32} * \frac{8}{31} = 0.064
\end{aligned}
$$

Conditional and Joint Probabilities

Another Example

- Setup: We make a survey in a street in Cologne
- We count four types of events in two random variables:
  - Person has brown hair ($H = B$)
  - Person has red hair ($H = R$)
  - Person likes to wake up late ($W = L$)
  - Person likes to wake up early ($W = E$)

Conditional and Joint Probabilities

Another Example

- Setup: We make a survey in a street in Cologne
- We count four types of events in two random variables:
  - Person has brown hair ($H = B$)
  - Person has red hair ($H = R$)
  - Person likes to wake up late ($W = L$)
  - Person likes to wake up early ($W = E$)
- Assumption: B / R and L / E are mutually exclusive
  - I.e., a single person cannot have red *and* brown hair
- A single person can be encoded with two symbols (e.g., »BL«)
  - ⚠ But this combination is not unique – in contrast to the cards example
- All following numbers are made up

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W \ / \ H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W \ / \ H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

If we pick a random person, what's the probability that this person has brown hair?

$$p(H = \text{brown}) =$$

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W \ / \ H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

If we pick a random person, what's the probability that this person has brown hair?

$$p(H = \text{brown}) = \frac{50}{65}$$

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W \ / \ H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

$$\left.\begin{array}{ll} p(H = \text{brown}) = \frac{50}{65} & p(H = \text{red}) = \frac{15}{65} \\ p(W = \text{early}) = \frac{30}{65} & p(W = \text{late}) = \frac{35}{65} \end{array}\right\} \text{sums per row or column}$$

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W \;/\; H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

- Joint probability: $p(W = \text{late}, H = \text{brown}) = \frac{30}{65}$
  - Probability that someone has brown hair *and* prefers to wake up late
  - Denominator: Number of all items

Conditional and Joint Probabilities

Example

Relation between **hair color** H and preferred **wake-up time** W

$\downarrow W$ / $H \rightarrow$	brown	red	sum
early	20	10	30
late	30	5	35
sum	50	15	65

Table: Survey Results, Ω: Group of questioned people

- Joint probability: $p(W = \text{late}, H = \text{brown}) = \frac{30}{65}$
  - Probability that someone has brown hair *and* prefers to wake up late
  - Denominator: Number of all items
- Conditional probability: $p(W = \text{late} | H = \text{brown}) = \frac{30}{50}$
  - Probability that one of the brown-haired participants prefers to wake up late
  - Denominator: Number of remaining items (after conditioned event has happened)

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$p(A|B) \quad = \quad \frac{p(A, B)}{p(B)} \quad \text{definition of conditional probabilities}$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$p(A|B) = \frac{p(A, B)}{p(B)} \quad \text{definition of conditional probabilities}$$

$$p(W = \text{late}|H = \text{brown}) = \frac{30}{50} = 0.6 \quad \text{intuition from previous slide}$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W = e, H = b) = 0.31$	$p(W = e, H = r) = 0.15$	$p(W = e) = 0.46$
late	$p(W = l, H = b) = 0.46$	$p(W = l, H = r) = 0.08$	$p(W = l) = 0.54$
margin	$p(H = b) = 0.77$	$p(H = r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$
\begin{aligned}
p(A|B) &= \frac{p(A, B)}{p(B)} && \text{definition of conditional probabilities} \\
p(W = \text{late}|H = \text{brown}) &= \frac{30}{50} = 0.6 && \text{intuition from previous slide} \\
&= \frac{p(W = \text{late}, H = \text{brown})}{p(H = \text{brown})} && \text{by applying definition}
\end{aligned}
$$

Conditional and Joint Probabilities

Example

	brown	red	margin
early	$p(W=e, H=b) = 0.31$	$p(W=e, H=r) = 0.15$	$p(W=e) = 0.46$
late	$p(W=l, H=b) = 0.46$	$p(W=l, H=r) = 0.08$	$p(W=l) = 0.54$
margin	$p(H=b) = 0.77$	$p(H=r) = 0.23$	$p(\Omega) = 1$

Table: (Joint) Probabilities, derived by dividing everything by $|\Omega|$

$$
\begin{aligned}
p(A|B) &= \frac{p(A,B)}{p(B)} && \text{definition of conditional probabilities} \\
p(W = \text{late}|H = \text{brown}) &= \frac{30}{50} = 0.6 && \text{intuition from previous slide} \\
&= \frac{p(W = \text{late}, H = \text{brown})}{p(H = \text{brown})} && \text{by applying definition} \\
&= \frac{0.46}{0.77} = 0.6
\end{aligned}
$$

Section 2

Collocations

Introduction

A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. *(Manning/Schütze, 1999, 151)*

Introduction

A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. (Manning/Schütze, 1999, 151)

Examples

- »Das ist mein zweites Frühstück« (adjective noun)
- »Da müssen wir Abhilfe schaffen« (noun verb)
- »Es regnet in Strömen« (verb preposition noun)

Limited Compositionality

- Compositionality: The meaning of linguistic expressions can be understood from understanding their parts
- Collocations: Not entirely true
  - I.e., they are learned by heart and stored in lexicon

Limited Compositionality

- Compositionality: The meaning of linguistic expressions can be understood from understanding their parts
- Collocations: Not entirely true
  - I.e., they are learned by heart and stored in lexicon
- Related concepts
  - Idiomatic expressions, metaphors, figure of speech ...

Why are Collocations Interesting?

- Generation: Produce natural sounding expressions
  E.g., »Da müssen wir Abhilfe schaffen« instead of »Da müssen wir Abhilfe erzeugen«
- Parsing: Collocations are more likely to also be syntactic phrases
- Lexicography: Collocations should be included in dictionaries
- Social justice: Collocations may be important in reinforcing cultural stereotypes

How to Detect Collocations Quantitatively?

Multiple methods

- Frequency
- (Pointwise) Mutual Information

Subsection 1

Frequency

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?
- Zipf's law: Two words that are highly frequent have much higher chance to co-occur with high frequency

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

Content Words

- Content words: Nouns, verbs, adjectives, adverb
- My operationalization here: Remove everything that doesn't contain one upper-case letter
  - Because verb-verb combinations are rare (as bigrams)
  - But we're missing verb-adverb combinations

Counting Bigrams

Content Words

- Content words: Nouns, verbs, adjectives, adverb
- My operationalization here: Remove everything that doesn't contain one upper-case letter
  - Because verb-verb combinations are rare (as bigrams)
  - But we're missing verb-adverb combinations

Bigram	Frequency
Jahre alt	56
Bevölkerung waren	47
Prozent waren	46
Jahre später	45
of Fame	44
Hall of	43
New York	41
als Nachfolger	41
Olympischen Spielen	35
Professor für	32
ersten Mal	32
er Mitglied	29
Fame aufgenommen	28
selben Jahr	28
Zweiten Weltkrieg	26
zum Mitglied	25
zum Professor	24
Jahr später	23
zwei Jahre	22
University of	21
Professor an	20
nach Deutschland	20
Betrieb genommen	18
Bevölkerung war	18
Los Angeles	18
drei Jahre	18
als Professor	17
Im Jahr	16
Lehrstuhl für	16

Focus Words

- Look at bigrams that contain a specific word
- In this case: »Gründen«

Focus Words

- Look at bigrams that contain a specific word
- In this case: »Gründen«

Bigram	Frequency
gesundheitlichen Gründen	7
Gründen von	3
finanziellen Gründen	2
Gründen abgeben	1
Gründen als	1
Gründen auf	1
Gründen aus	1
Gründen den	1
Gründen die	1
Gründen gab	1
Gründen ihre	1
Gründen interessierte	1
Gründen nach	1
Gründen um	1
Gründen zurück	1
disziplinarischen Gründen	1
gesundheitlichen Problemen	1
nationalpolitischen Gründen	1
paläographischen Gründen	1
persönlichen Gründen	1
politischen Gründen	1
strategischen Gründen	1

Subsection 2

Point-wise Mutual Information

Introduction

Example

»1910 wurde Gerland _______________ ________ in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Introduction

Example

»1910 wurde Gerland außerordentlicher ________ in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Introduction

Example

»1910 wurde Gerland <u>außerordentlicher</u> <u>Professor</u> in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Intuition

- We are interested in the (potential) collocation »außerordentlicher Professor«

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	$19\,811\,129$	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret relative frequencies as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(W = \text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	$19\,811\,129$	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret relative frequencies as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(W = \text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher« and »Professor«?

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	$19\,811\,129$	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret relative frequencies as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(W = \text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher« and »Professor«? $p(W = \text{außerordentlicher}) \times p(W = \text{Professor}) = 5.5 \times 10^{-10}$

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	19 811 129	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret relative frequencies as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(W = \text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher« and »Professor«? $p(W = \text{außerordentlicher}) \times p(W = \text{Professor}) = 5.5 \times 10^{-10}$
- This is the probability that these two words appear together – if they are distributed randomly / independent events

Pointwise Mutual Information

$$\mathrm{pmi}(w_1, w_2) =$$

Pointwise Mutual Information

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{}{p(W = w_1)p(W = w_2)}$$

- Denominator: Probability that the words appear together, if they are distributed randomly

Pointwise Mutual Information

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{p(B = \langle w_1, w_2 \rangle)}{p(W = w_1)p(W = w_2)}$$

- Denominator: Probability that the words appear together, if they are distributed randomly
- Numerator: Probability that they *actually* appear together

Pointwise Mutual Information

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{}{p(W = w_1)p(W = w_2)}$$

- Denominator: Probability that the words appear together, if they are distributed randomly
- Numerator: Probability that they *actually* appear together
- $\log_2$: Scales

Interpretations

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability

Interpretations

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 1 = 0$

Interpretations

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 1 = 0$
  - Co-occurrence is 8 times more probable than by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.008, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 \frac{0.0008}{0.0001} = 3$

Interpretations

$$\mathrm{pmi}(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 1 = 0$
  - Co-occurrence is 8 times more probable than by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.008, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 \frac{0.0008}{0.0001} = 3$
  - Co-occurrence is 8 times less probable than by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0000125, \Rightarrow \mathrm{pmi}(w_1, w_2) = \log_2 \frac{0.0000125}{0.0001} = -3$

Section 3

Summary

Summary

- Probability theory
  - Probability: Ratio of events of interest to all possible events (within event space)
  - Joint probability: Two events take place simultaneously
  - Conditional probability: One event takes place under the assumption that another event took place
  - Dependent and independent events
- Collocations
  - Multiple words that have a meaning beyond their parts (non-compositionality)
  - Counting n-grams: Function word combinations are most frequent
  - Pointwise Mutual information: Metric how much information one word yields about another

References I

Manning, Christopher D./Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.