FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Johannes Franke

IR: Query Understanding WiSe 24/25

Query Segmentation

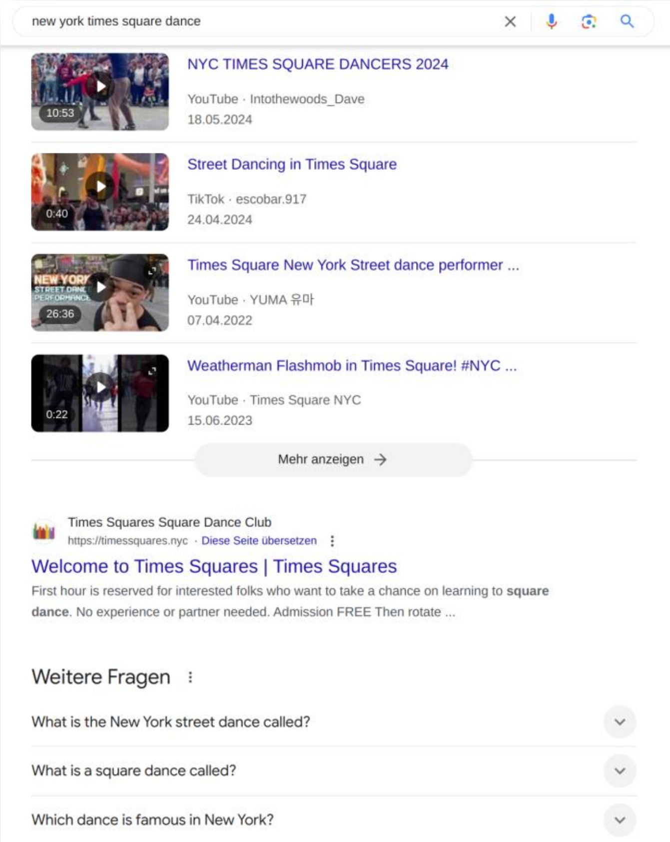
10.12.2024

Beispiel

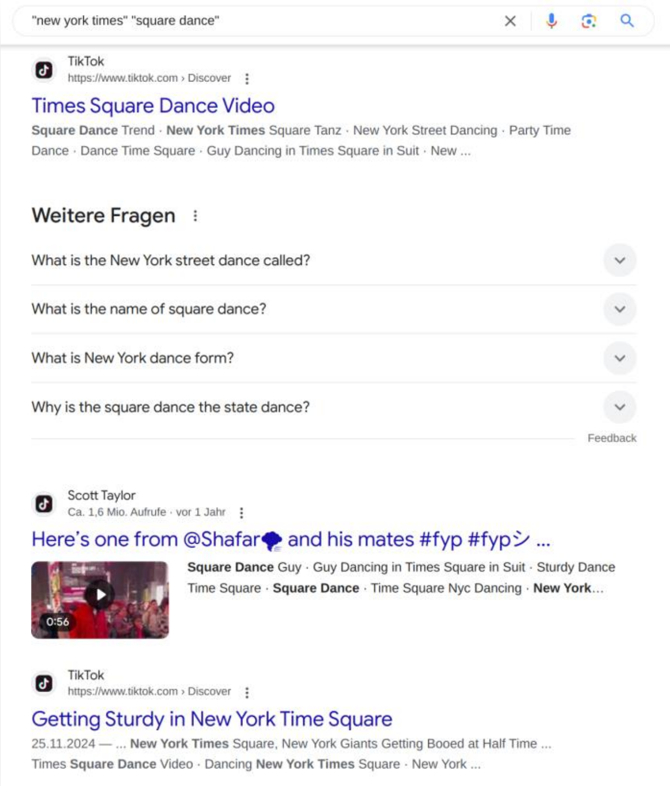
"new york" "times square" dance

"new york times" "square dance"

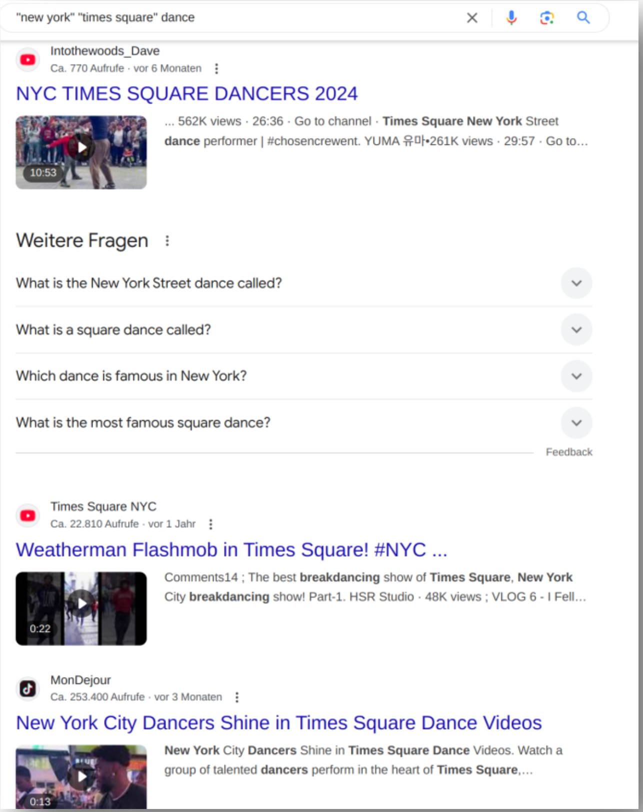
Beispiel


new york times square dance
NYC TIMES SQUARE DANCERS 2024
YouTube · Intothewoods_Dave
18.05.2024
10:53
Street Dancing in Times Square
TikTok · escobar.917
24.04.2024
0:40
Times Square New York Street dance performer ...
YouTube · YUMA 유마
07.04.2022
26:36
Weatherman Flashmob in Times Square! #NYC ...
YouTube · Times Square NYC
15.06.2023
0:22
Mehr anzeigen
Times Squares Square Dance Club
https://timessquares.nyc · Diese Seite übersetzen
Welcome to Times Squares | Times Squares
First hour is reserved for interested folks who want to take a chance on learning to **square dance**. No experience or partner needed. Admission FREE Then rotate ...
Weitere Fragen
What is the New York street dance called?
What is a square dance called?
Which dance is famous in New York?

Beispiel

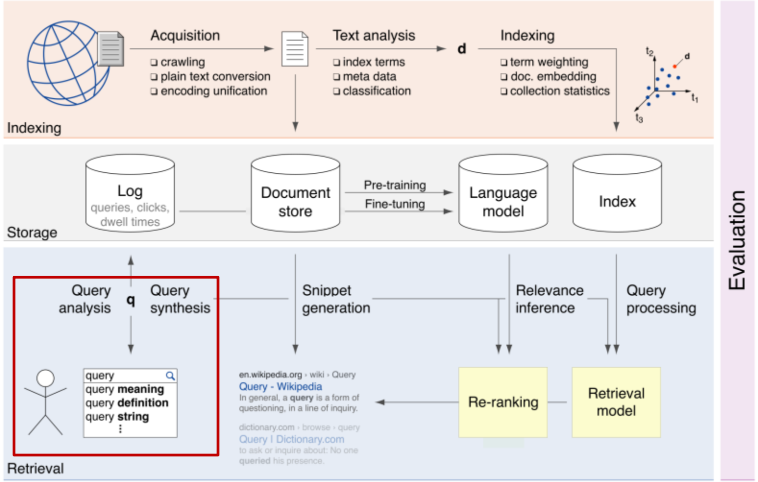

"new york times" "square dance"
TikTok
https://www.tiktok.com › Discover
Times Square Dance Video
Square Dance Trend · New York Times Square Tanz · New York Street Dancing · Party Time Dance · Dance Time Square · Guy Dancing in Times Square in Suit · New ...
Weitere Fragen
What is the New York street dance called?
What is the name of square dance?
What is New York dance form?
Why is the square dance the state dance?
Feedback
Scott Taylor
Ca. 1,6 Mio. Aufrufe · vor 1 Jahr
Here's one from @Shafar and his mates #fyp #fypシ ...
0:56
Square Dance Guy · Guy Dancing in Times Square in Suit · Sturdy Dance Time Square · Square Dance · Time Square Nyc Dancing · New York...
TikTok
https://www.tiktok.com › Discover
Getting Sturdy in New York Time Square
25.11.2024 — ... New York Times Square, New York Giants Getting Booed at Half Time ...
Times Square Dance Video · Dancing New York Times Square · New York ...

Beispiel


"new york" "times square" dance
Intothewoods_Dave
Ca. 770 Aufrufe · vor 6 Monaten
NYC TIMES SQUARE DANCERS 2024
10:53
... 562K views · 26:36 · Go to channel · **Times Square New York** Street **dance** performer | #chosencrewent. YUMA 유마•261K views · 29:57 · Go to...
Weitere Fragen
What is the New York Street dance called?
What is a square dance called?
Which dance is famous in New York?
What is the most famous square dance?
Feedback
Times Square NYC
Ca. 22.810 Aufrufe · vor 1 Jahr
Weatherman Flashmob in Times Square! #NYC ...
0:22
Comments14 ; The best **breakdancing** show of **Times Square**, **New York** City **breakdancing** show! Part-1. HSR Studio · 48K views ; VLOG 6 - I Fell...
MonDejour
Ca. 253.400 Aufrufe · vor 3 Monaten
New York City Dancers Shine in Times Square Dance Videos
0:13
New York City **Dancers** Shine in **Times Square Dance** Videos. Watch a group of talented **dancers** perform in the heart of **Times Square**,...

Grundlagen

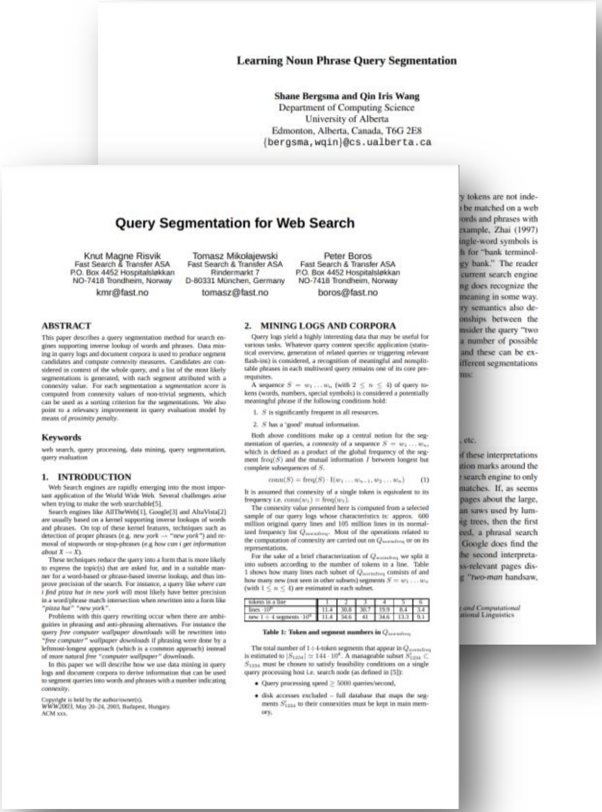
- Aufteilung in Token-Sequenzen
- Query mit ***n*** Token hat ***n* – *1*** Breakpoints
- Segmentierungen haben Einfluss auf Bedeutung & Retrieval



Paper

- ***Query Segmentation for Web Search***
  - Risvik et al., 2003
  - Publisher: The Web Conference

- ***Learning Noun Phrase Query Segmentation***
  - Bergsma et al., 2007
  - Publisher: EMNLP-CoNLL
    (Empirical Methods in Natural Language Processing and Computational Natural Language Learning)


Learning Noun Phrase Query Segmentation

Shane Bergsma and Qin Iris Wang
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2E8
{bergsma,wqin}@cs.ualberta.ca

Query Segmentation for Web Search

Knut Magne Risvik, Fast Search & Transfer ASA, P.O. Box 4452 Hospitalsløkkan, NO-7418 Trondheim, Norway, kmr@fast.no
Tomasz Mikolajewski, Fast Search & Transfer ASA, Rindermarkt 7, D-80331 München, Germany, tomasz@fast.no
Peter Boros, Fast Search & Transfer ASA, P.O. Box 4452 Hospitalsløkkan, NO-7418 Trondheim, Norway, boros@fast.no

ABSTRACT

Keywords

1. INTRODUCTION

2. MINING LOGS AND CORPORA

Table 1: Token and segment numbers in Q_{sampling}

Query Segmentation for Web Search

- Segmente = Bedeutungsvolle Phrasen mit
  - signifikanter Frequenz im Corpus
  - hoher "Mutual Information"

$$\mathbf{conn(S) = freq(S) \cdot I(w_1 \ldots w_{n-1}, w_2 \ldots w_n)}$$

- Berechnung des Connexity-Scores für alle 2^{n-1} Segmentierungen
- Sortierung nach max. kumulativen Scores

Query Segmentation for Web Search

$$\mathrm{conn}(S) = \mathrm{freq}(S) \cdot I(w_1 \dots w_{n-1}, w_2 \dots w_n)$$

S_1 = "new york" "times square" dance	S_2 = "new york times" "square dance"
• freq("new york") = 7.500 • freq("times square") = 100 • I("new", "york") = 0.04 • I("times", "square") = 0.02	• freq("new york times") = 3.400 • freq("square dance") = 200 • I("new york", "york times") = 0.03 • I("square", "dance") = 0.01
Score(S_1) = conn("new york") + conn("times square") = 300 + 2 = 302	Score(S_2) = conn("new york times") + conn("square dance") = 102 + 2 = 104

Learning Noun Phrase Query Segmentation

- Segmentation als Classification Task mittels
  - supervised Machine Learning
  - Support Vector Model (SVM)
- Trennung wird an jedem Breakpoint einer Query entschieden
- Einbezug des Kontext in Form eines Token-Fensters
  - 3 Token links/rechts von Breakpoint (falls vorhanden)
  - genannt "Decision Boundary"

$$\{\ldots, w_{L2}, w_{L1}, w_{L0}, w_{R0}, w_{R1}, w_{R2}, \ldots\}$$

Learning Noun Phrase Query Segmentation

Q_1 = sugar free tea
Q_2 = …
[…]

Features hinzufügen

Ground Truth annotieren

sugar free tea

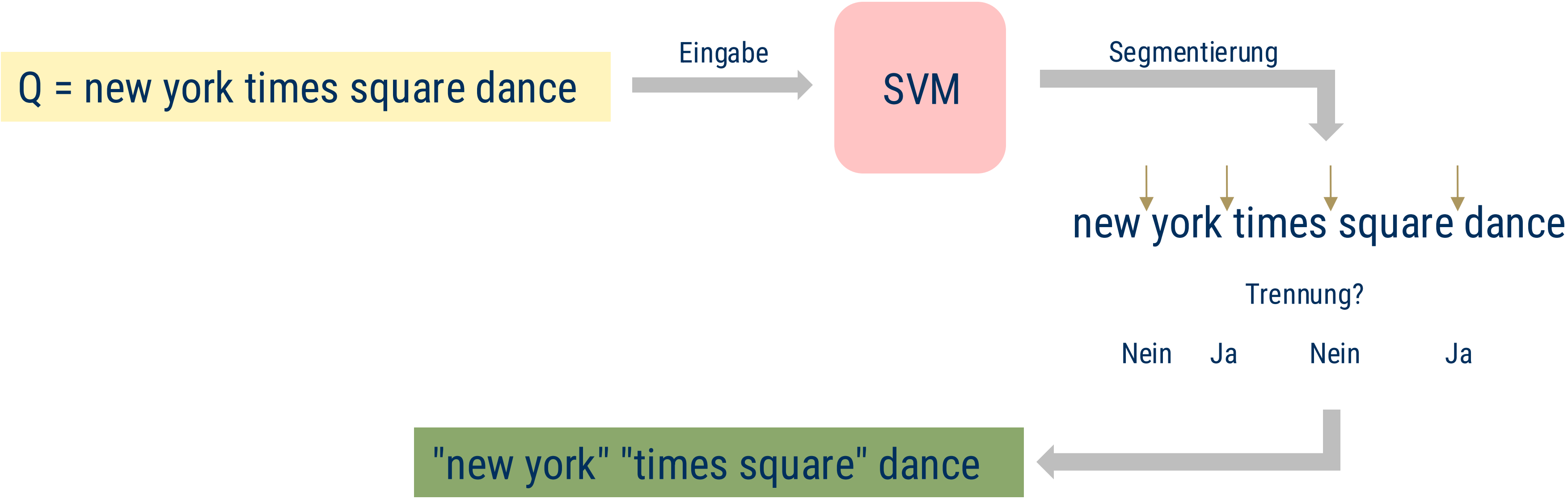
Feature-List 1	Feature-List 2	Feature-List 3
• is-free = false	• is-free = true	• is-free = false
• fwd-pos = 0	• fwd-pos = 1	• fwd-pos = 2
[…]	[…]	[…]

"sugar free" "tea"

Training

SVM

Learning Noun Phrase Query Segmentation



Learning Noun Phrase Query Segmentation

Annotation

- Aufteilung in Training, Validation und Test (je 500 Queries)
- "Ground Truth" von 3 Annotatoren manuell erstellt
- Agreement $\kappa = 0.69$

Naive Baselines

Ansatz	Seg.-Acc.	Qry.-Acc.
"always split"	0.44	0.04
"never split"	0.55	0.04

Learning Noun Phrase Query Segmentation

Decision Boundary

- Features für x_{L0}, x_{R0}
- Aufgeteilt in
  - Indikator F.
  - Statistische F.

Context

- Wie Decision Boundary
- Features für ganzes Fenster x_{L2} bis x_{R2}

Dependency

- Nimmt Rücksicht auf Einfluss zwischen entfernten Token
- Count von
  - x_{L0} und x_{R1}
  - x_{L1} und x_{R0}

Learning Noun Phrase Query Segmentation

Indikator	Beschreibung
is-free	Token x = "free"
fwd-pos	Position von Anfang
rev-pos	Position von Ende
[…]	

Statistik	Beschreibung
web-count	Häuf. von *x* im Web
Qcount-1	Häuf. von *x* in Query-Log
[…]	

Learning Noun Phrase Query Segmentation

Feature Type	Feature Span	Seg.-Acc.	Qry.-Acc.
MI	Decision Boundary	0.68	0.26
Basic	Decision Boundary	0.71	0.29
Basic	Decision Boundary, Context	0.80	0.52
Basic	Decision Boundary, Context, Dependency	0.81	0.53
All	Decision Boundary	0.84	0.57
All	Decision Boundary, Context	0.86	0.63
All	Decision Boundary, Context, Dependency	0.85	0.61

Takeaways & Fragen

- Query Segmentation: Grundlagen & Motivation
- Connexity-Score: Bedeutung & Berechnung
- Segmentation als Classification-Task
- Feature-Engineering und Training eines SVMs

Dankeschön!

Mutual Information

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{(X,Y)}(x,y) \log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)\,P_Y(y)}\right)$$

$$\Leftrightarrow \log C(x_{L0}x_{R0}) + \log K - \log C(x_{L0}) - \log C(x_{R0})$$

- Ein Maß für die gegenseitige Abhängigkeit zweier Variablen
- Auch bekannt als "Information Gain"
- Benötigte Größen im IR Kontext:
  - rel. Wahrscheinlichkeiten von Token "x", "y" separat &
  - rel. Wahrscheinlichkeit von Token "x y" zusammen

Support Vector Machine (SVM)

- Überwachter, feature-basierter Klassifikationsalgorithmus
- Gilt für viele Probleme als guter Default-Ansatz
- Maximiert den Abstand zwischen den Trainingsinstanzen
  - Margin:
    max. Abstand zwischen Support Vectoren
  - Support Vectoren:
    Trainingsinstanzen am nächsten zur Margin