

Sebastian Marr

Kookkurrenzbasierte Link Discovery am Beispiel von Produkttags

Masterkolloquium

Zielsetzung der Arbeit

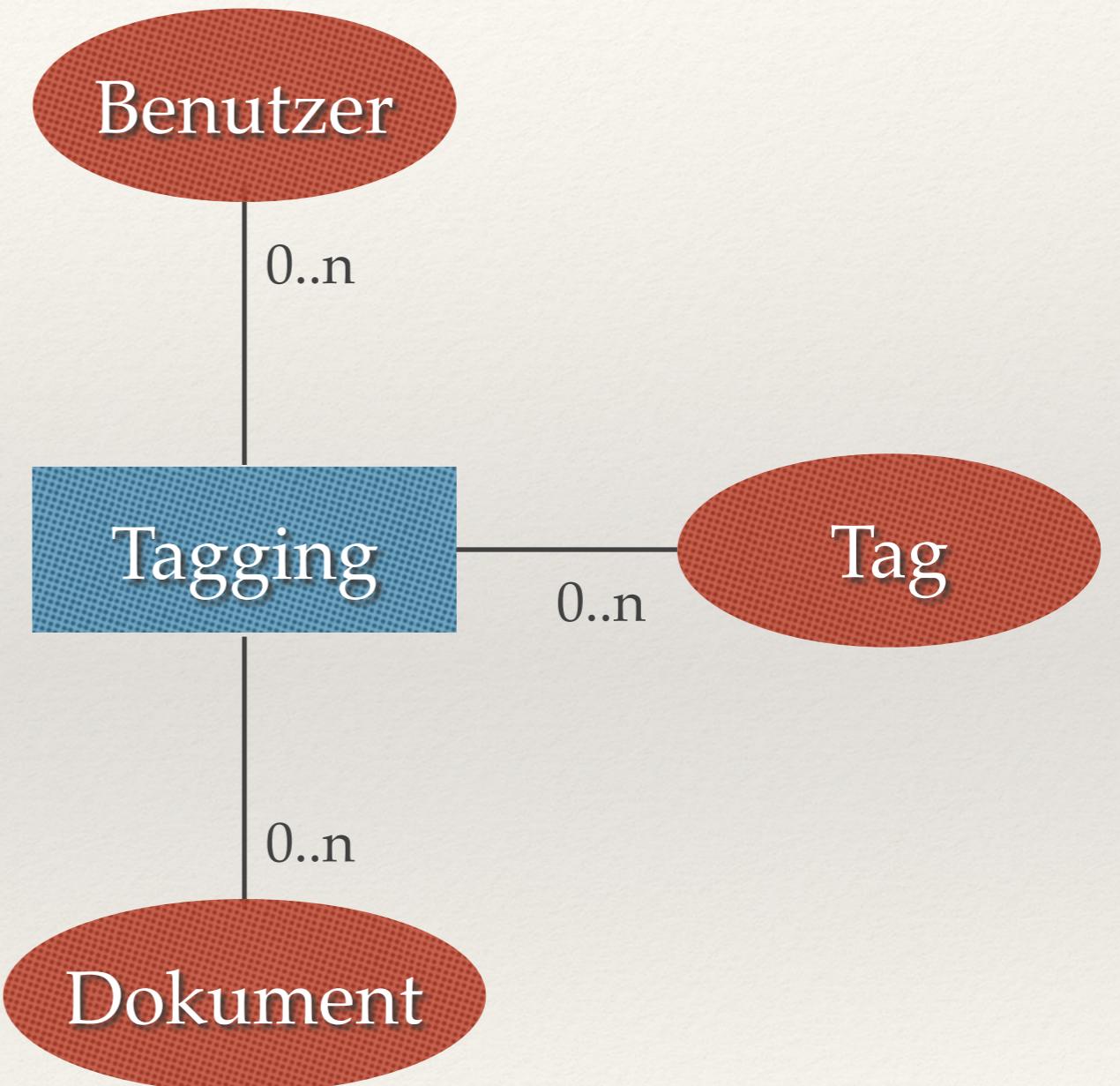
- ❖ *Link Discovery*: Methoden des Data Minings, die zum Ziel haben, Verbindungen zwischen Objekten herzustellen
- ❖ Beispiel: benutzergenerierte Begriffswelt eines Tagging-Systems
- ❖ Anwendungsgebiete:
 - ❖ dynamische Navigationskonzepte
 - ❖ Erweiterung von Suchräumen

Aufgaben

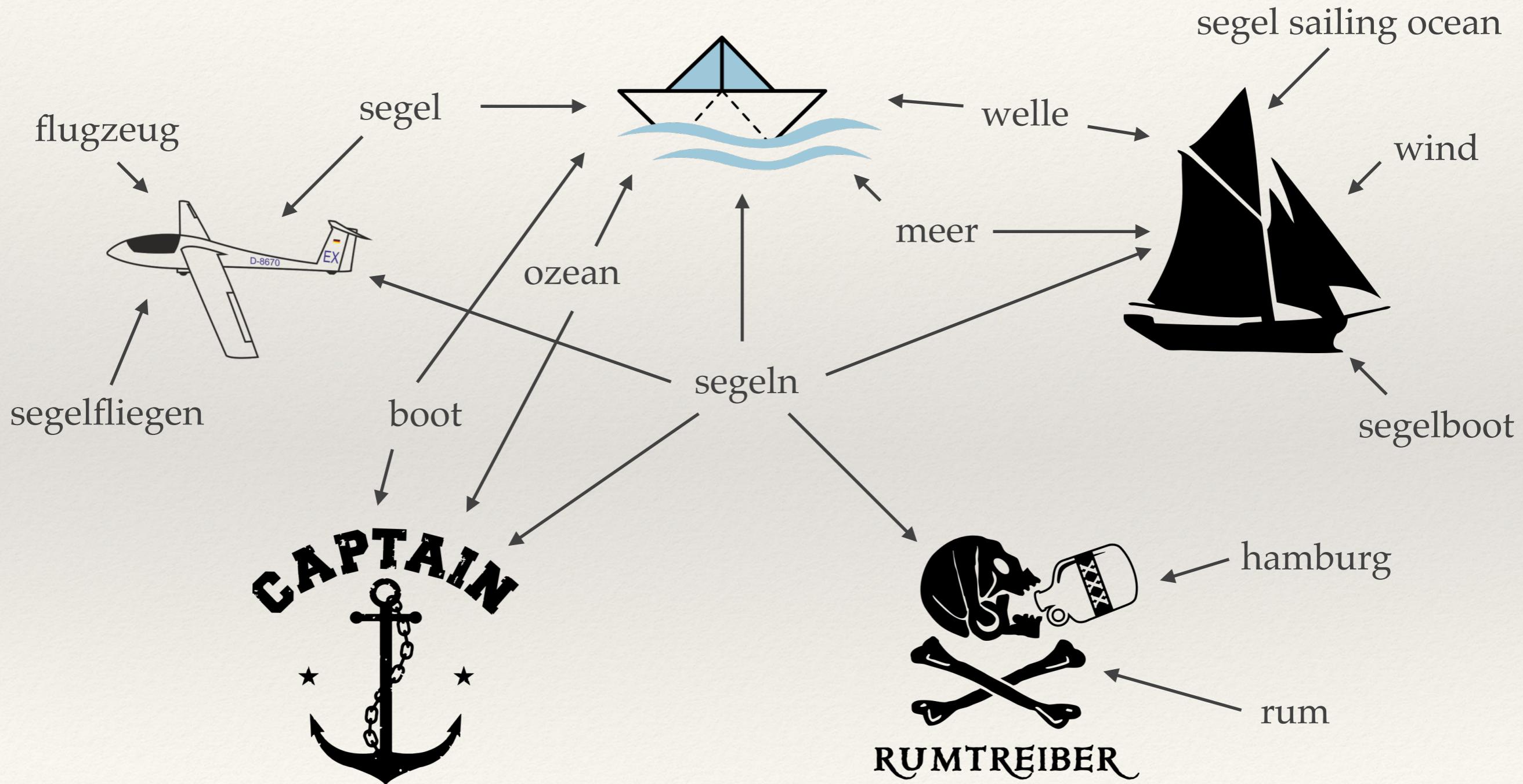
- ❖ Modellierung des betrachteten Weltausschnittes und Überführung in eine geeignete Datenstruktur
- ❖ Entwicklung und Durchführung eines Prozesses zur Extraktion von Beziehungen
- ❖ Anreicherung der Beziehungen durch weitere Datenquellen
- ❖ Priorisierung der Beziehungen

Ausgangssituation

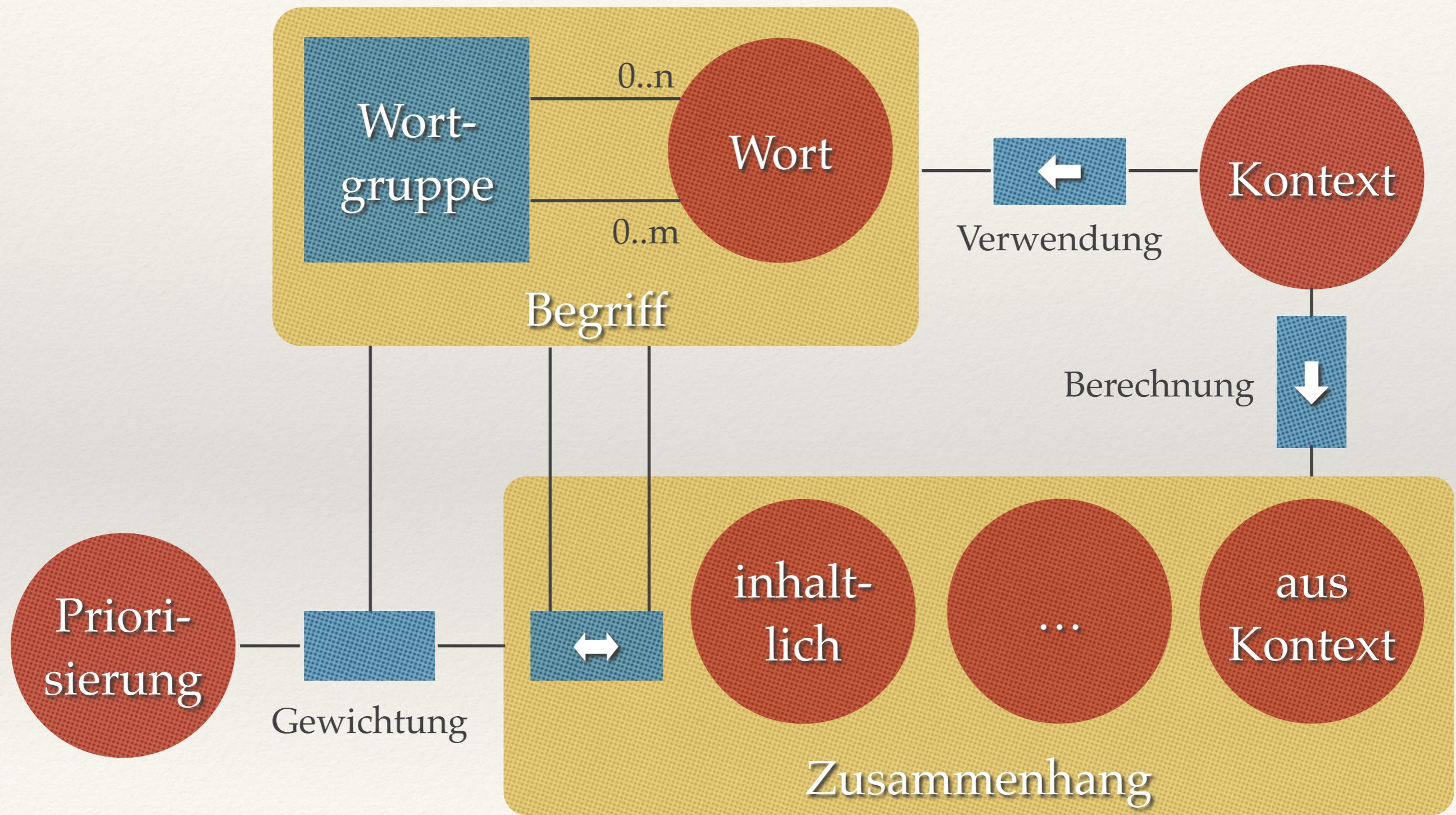
- ❖ 2 Mio. Tags
- ❖ 6,5 Mio. Benutzer
- ❖ 26 Mio. Dokumente
- ❖ 72 Mio. Taggings



Ausgangssituation



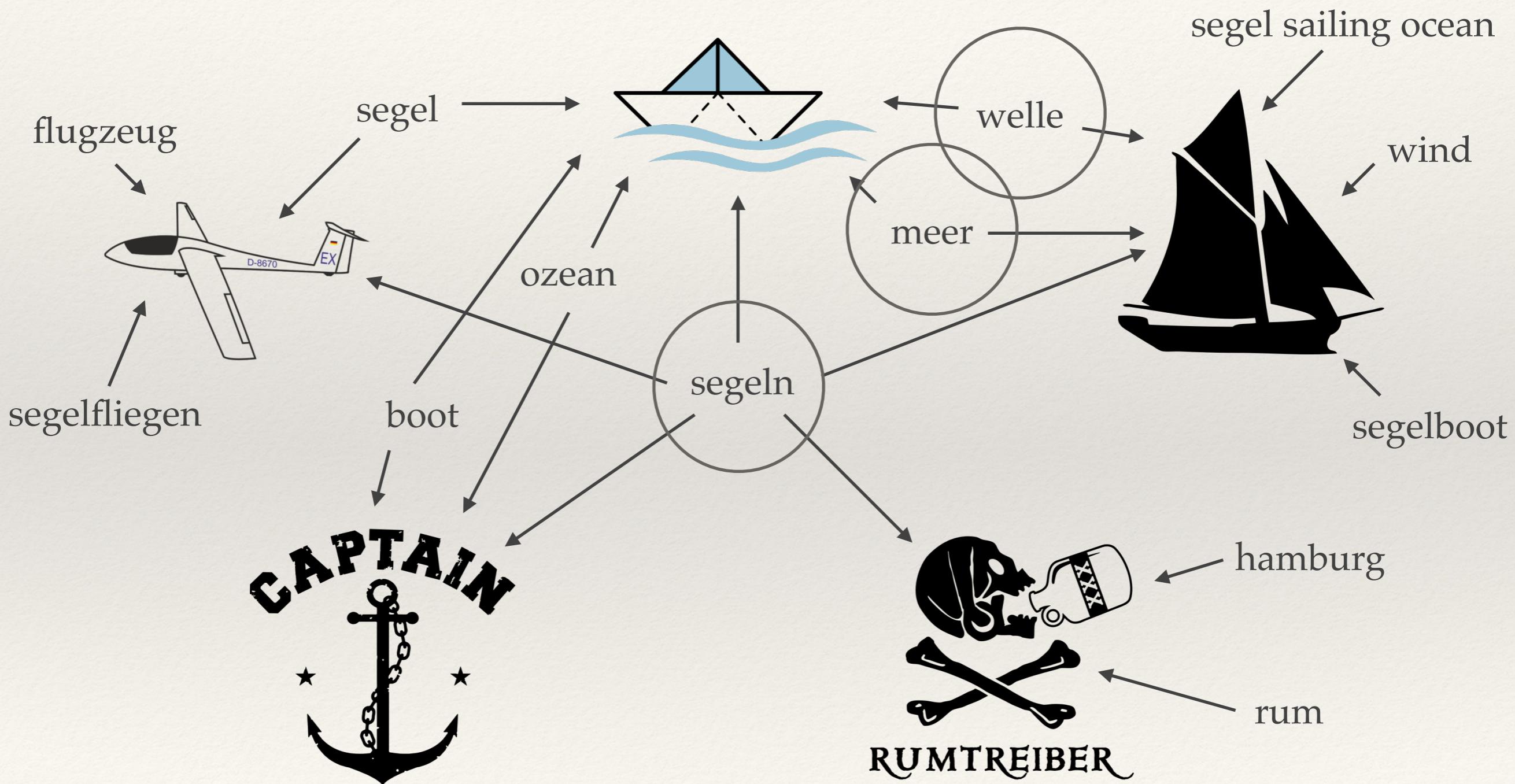
Modell - Weltausschnitt



Prozess

- ❖ Initiale Erstellung des Weltausschnittes aus Tagging-Daten
- ❖ Anreicherungsschritte
 - ❖ Integration von Clicktracking-Daten
 - ❖ Zerlegung von Wortgruppen
 - ❖ Integration des Wortschatzes der Universität Leipzig
- ❖ Priorisierung mittels interaktiver evolutionärer Algorithmen

Kookkurrenz



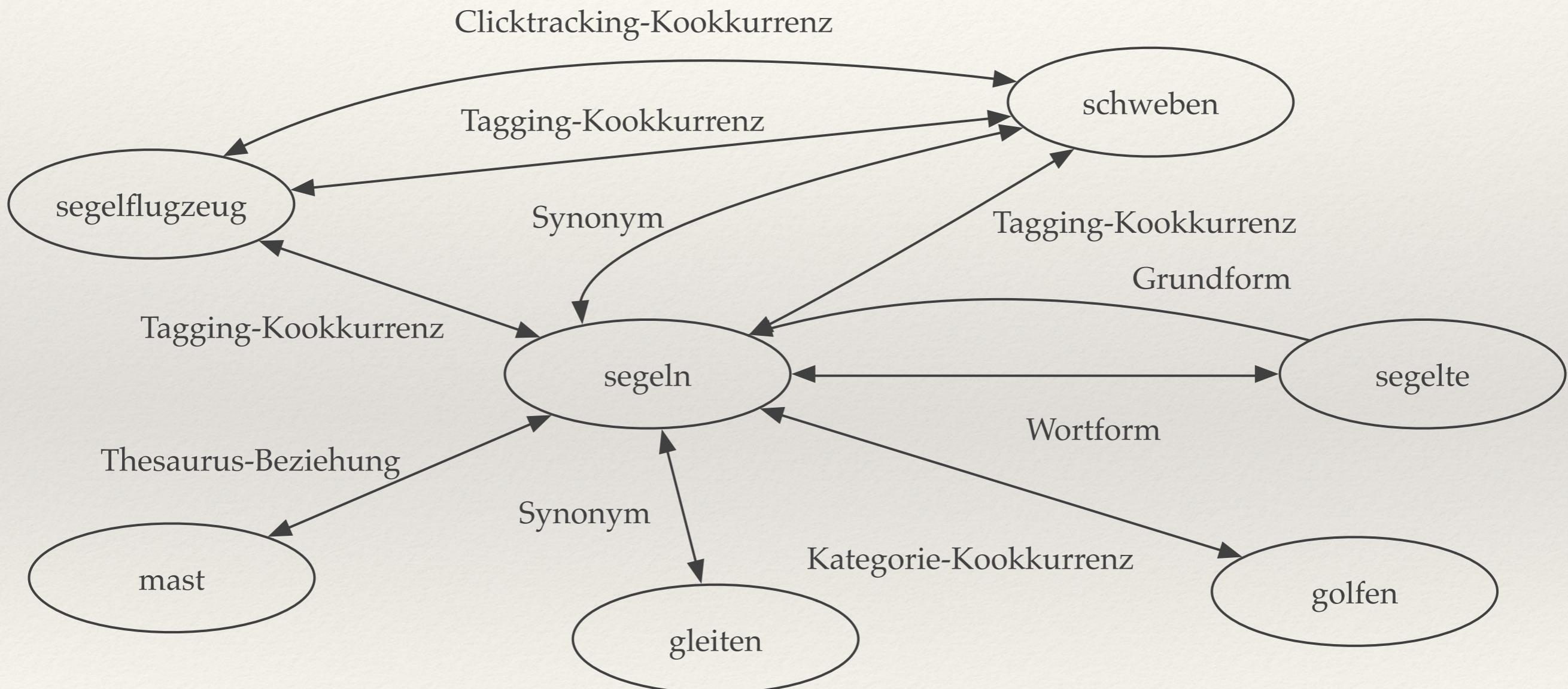
Kookkurrenzmaße

$$\delta_{Dice}(a, b) = \frac{2|A \cap B|}{|A| + |B|}$$

$$\delta_{Jaccard}(a, b) = \frac{|A \cap B|}{|A \cup B|}$$

$$\delta_{Cosine}(a, b) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

Graphenrepräsentation

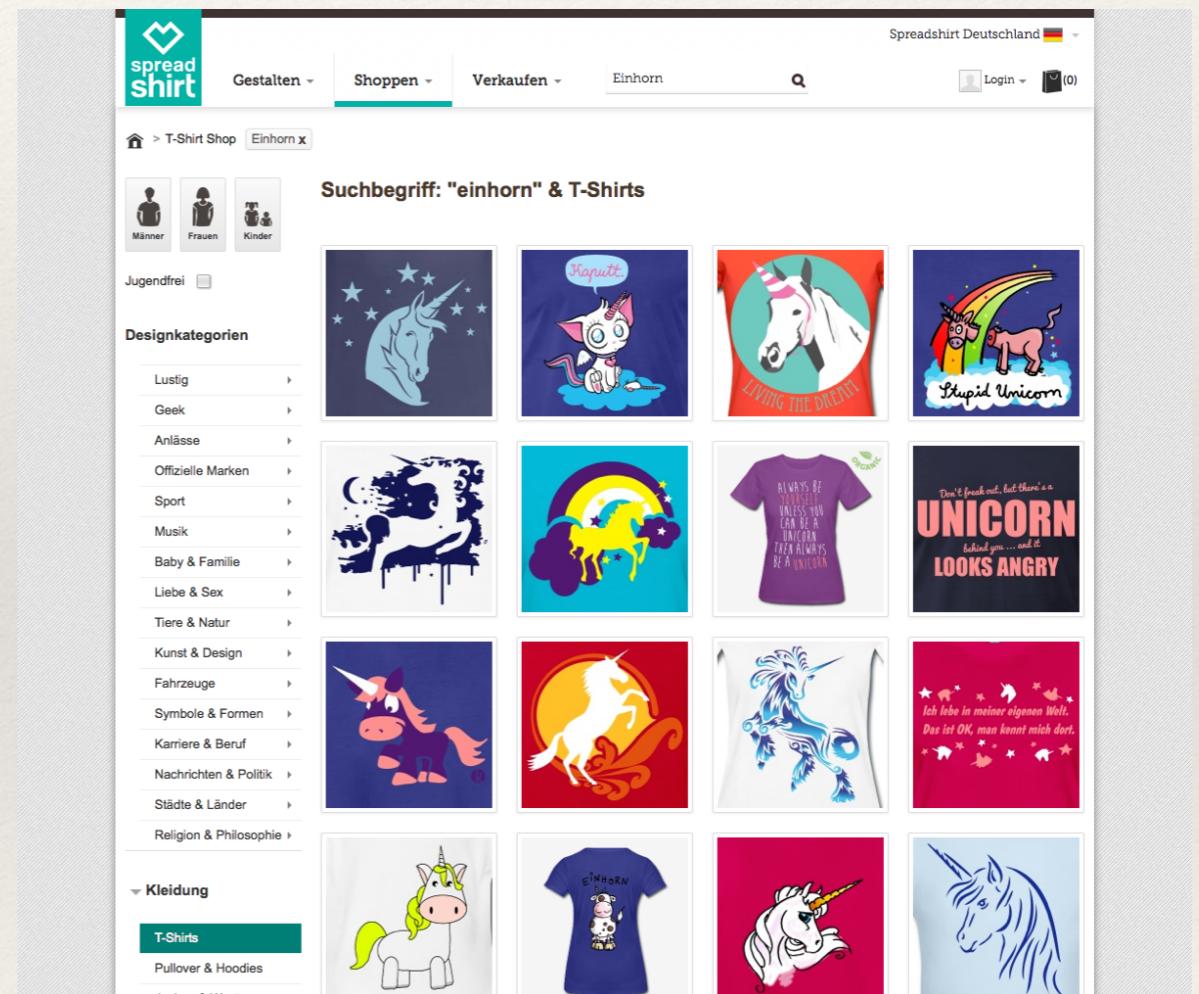


Tagging-Daten

- ❖ Tags \Rightarrow Knoten
 - ❖ Beschränkung auf deutsche Tags
 - ❖ Entfernung von Groß-/Kleinschreibung und Bereinigung: Reduktion um 68%
- ❖ Kanten vom Typ Tagging-Kookkurrenz
- ❖ Ergebnis: 314 351 Knoten und 21 834 868 Kanten

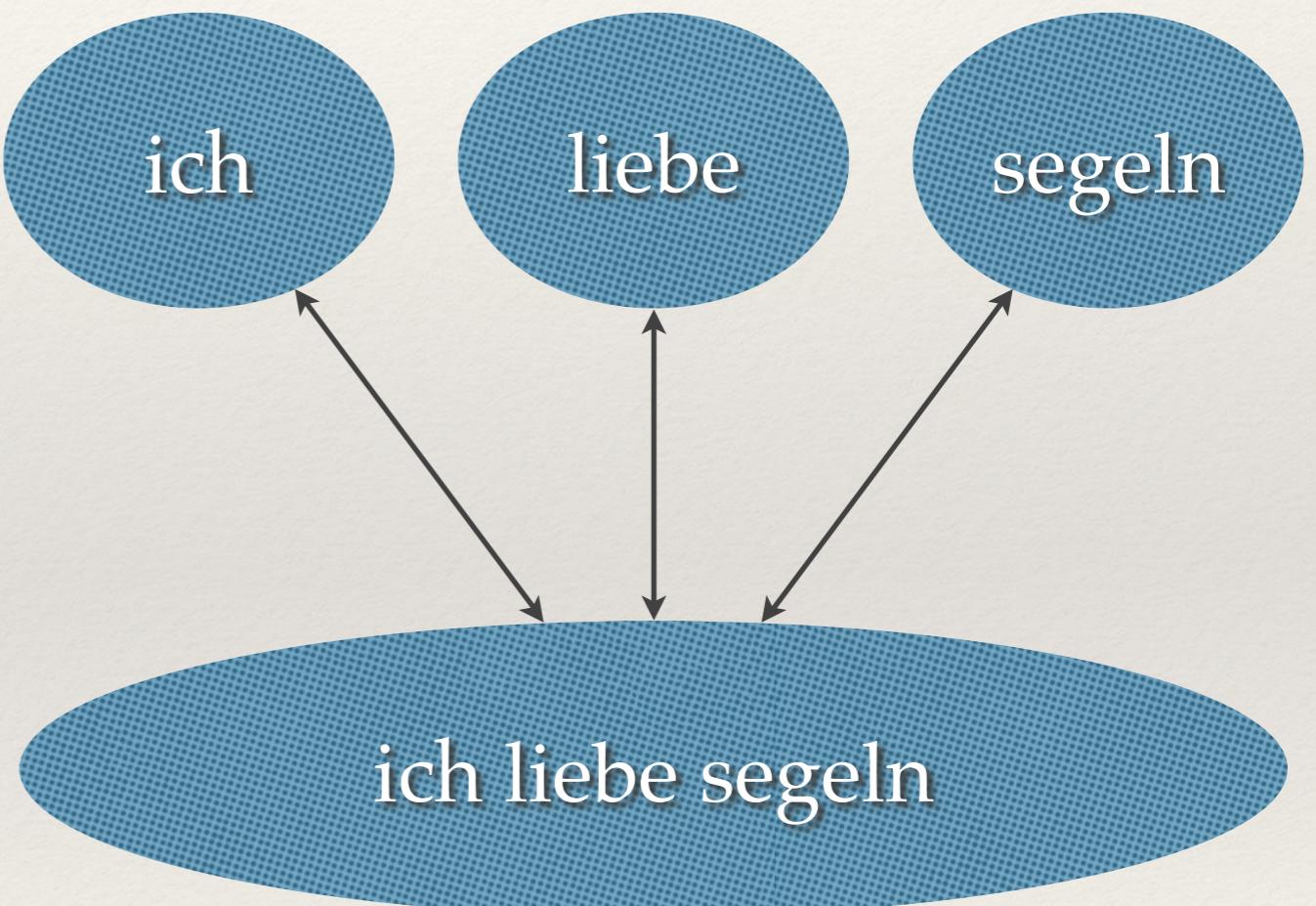
Clicktracking

- ❖ Klicks auf Suchergebnisseiten
- ❖ Gleiche Artikel zu verschiedenen Suchbegriffen: Kookkurrenz
- ❖ Ergebnis:
 - ❖ 78 237 neue Knoten
 - ❖ 310 860 Kanten



Zerlegung von Wortgruppen

- ❖ 47% aller Begriffe bestanden aus mehreren Wörtern
- ❖ Zerlegung:
 - ❖ 38 349 neue Knoten
 - ❖ 1 238 900 neue Kanten der Typen Zerlegung und Zusammensetzung



Wortschatz der Universität Leipzig

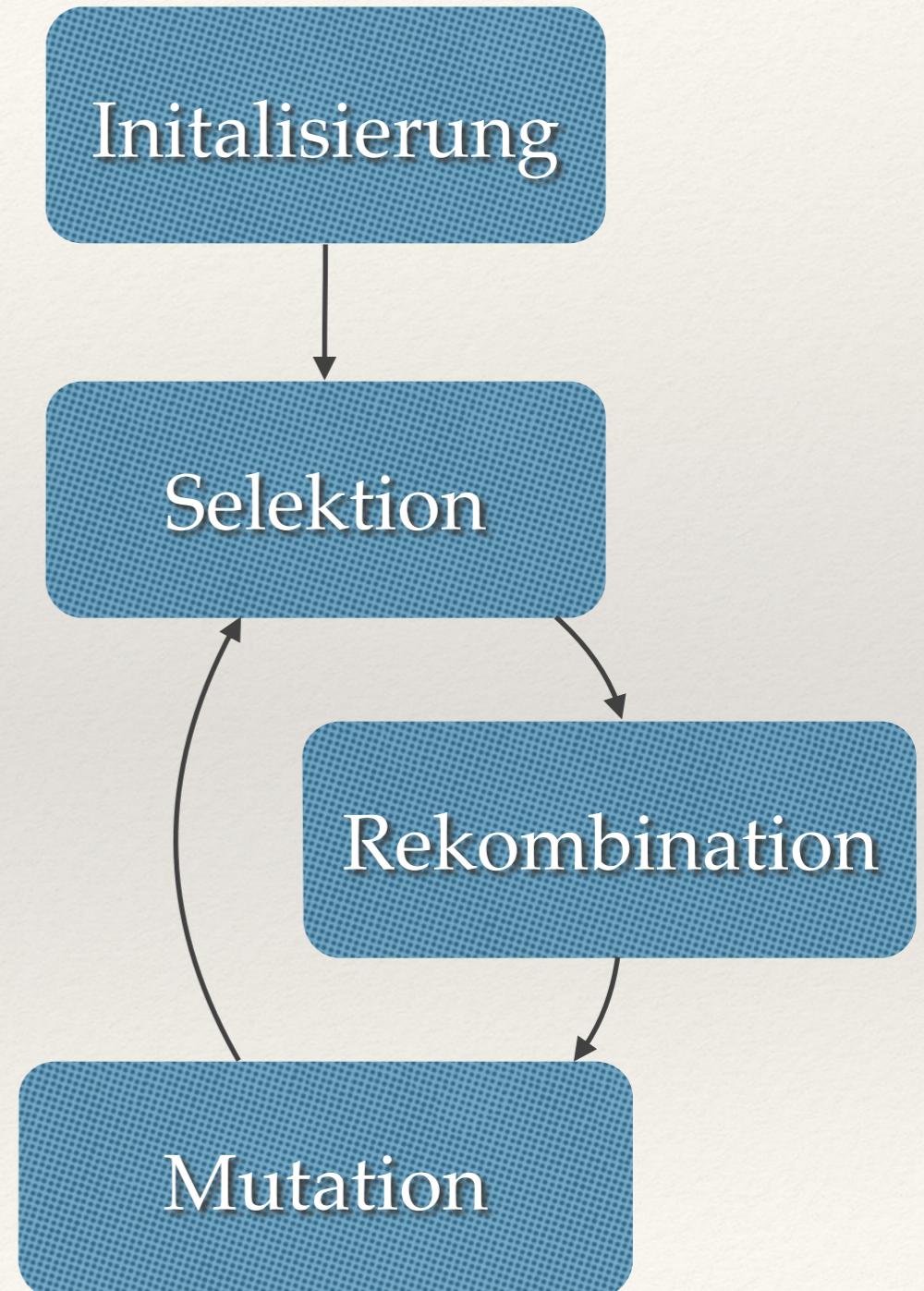
- ❖ Informationen: *Grundform, Wortformen, Synonyme, Thesaurus-Beziehungen und Kategorien*
- ❖ *Kategorien*: Kookkurrenz
- ❖ Restliche Beziehungen direkt als Kantentypen integriert
- ❖ Ergebnis: 145 023 neue Knoten, 50 227 965 Kanten
 - ❖ davon ca. 48 Mio. *Kategorie-Kookkurrenz*

575 960 Knoten

73 612 593 Kanten

Priorisierung

- ❖ 9 Kantentypen, 3 Kookkurrenzmaße
- ❖ 15 Stichproben
- ❖ interaktiver evolutionärer Algorithmus
- ❖ Selektion: Auswahl eines Gewinners durch vergleich der höchstgewichteten Nachbarn
- ❖ 13 Generationen
- ❖ 975 Selektionen

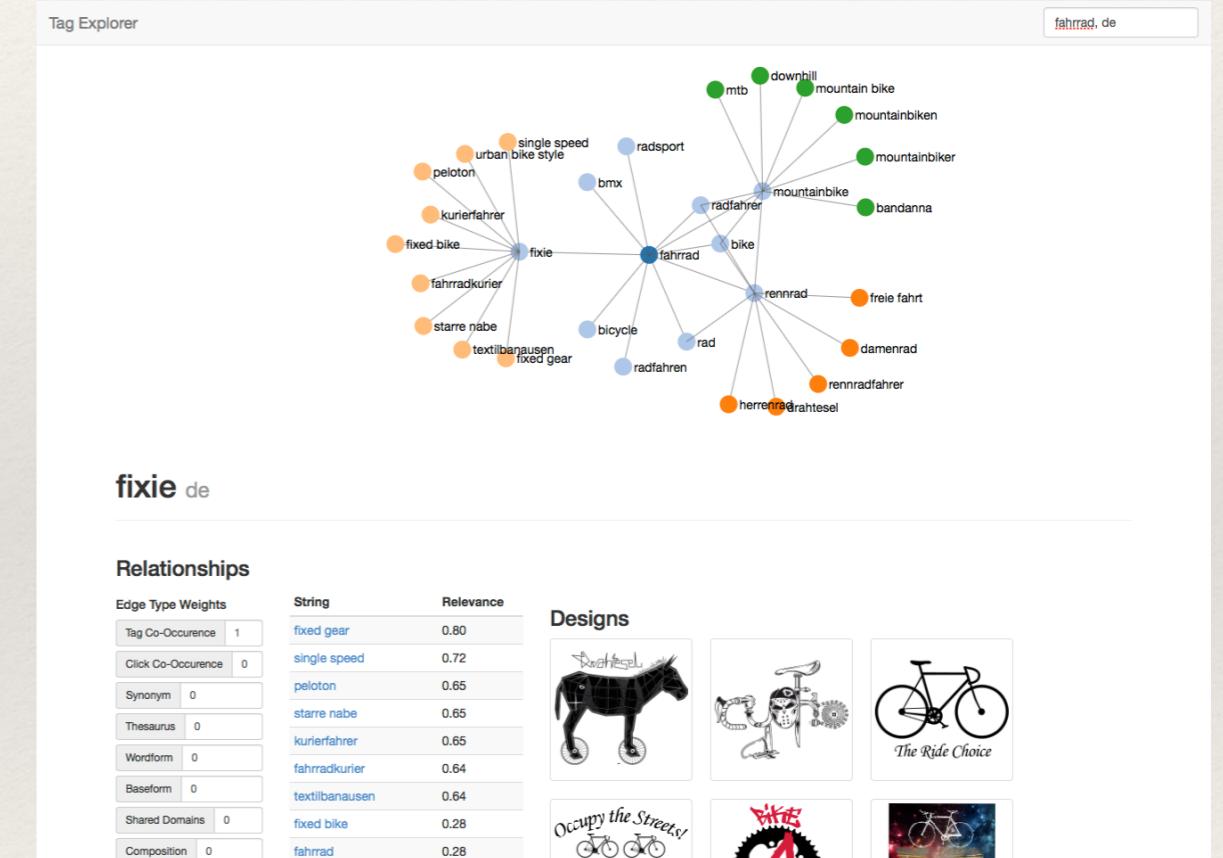


Priorisierung - Ergebnisse

- ❖ Nach der mit der Periodisierung ermittelten Gewichtung nächste Nachbarn:
 - ❖ **kind:** kleinkind, säugling, junge, nachwuchs, dreikäsehoch
 - ❖ **leipzig:** i heart leipzig, tshirts leipzig, t-shirts leipzig, leipzig stadt, deutschland leipzig
 - ❖ **mountainbike:** mountainbikes, fahrrad, rennrad, rad, gangschaltung

Technische Aspekte

- ❖ Datenbanksystem: MongoDB
- ❖ JavaScript zur Implementierung der Algorithmen mittels MapReduce
- ❖ JavaScript und node.js zur serverseitigen Priorisierung, API und *Tag Explorer*



Ausblick

- ❖ Qualitative Untersuchung der erzeugten Zusammenhänge
- ❖ Integration weiterer Datenquellen
- ❖ Clusteranalyse
- ❖ Interaktiver Trainingsschritt zum manuellen Entfernen und Hinzufügen von Zusammenhängen

Zusammenfassung

- ❖ *Link Discovery*: Herstellung von Zusammenhängen
- ❖ Modellierung: Begriffe, Zusammenhänge, Kontexte
- ❖ Kookkurrenz als primäres Mittel der Beziehungserzeugung
- ❖ Tagging-Daten als Ausgangspunkt
- ❖ Integration weiterer Datenquellen
- ❖ Priorisierung mittels interaktiver evolutionärer Algorithmen