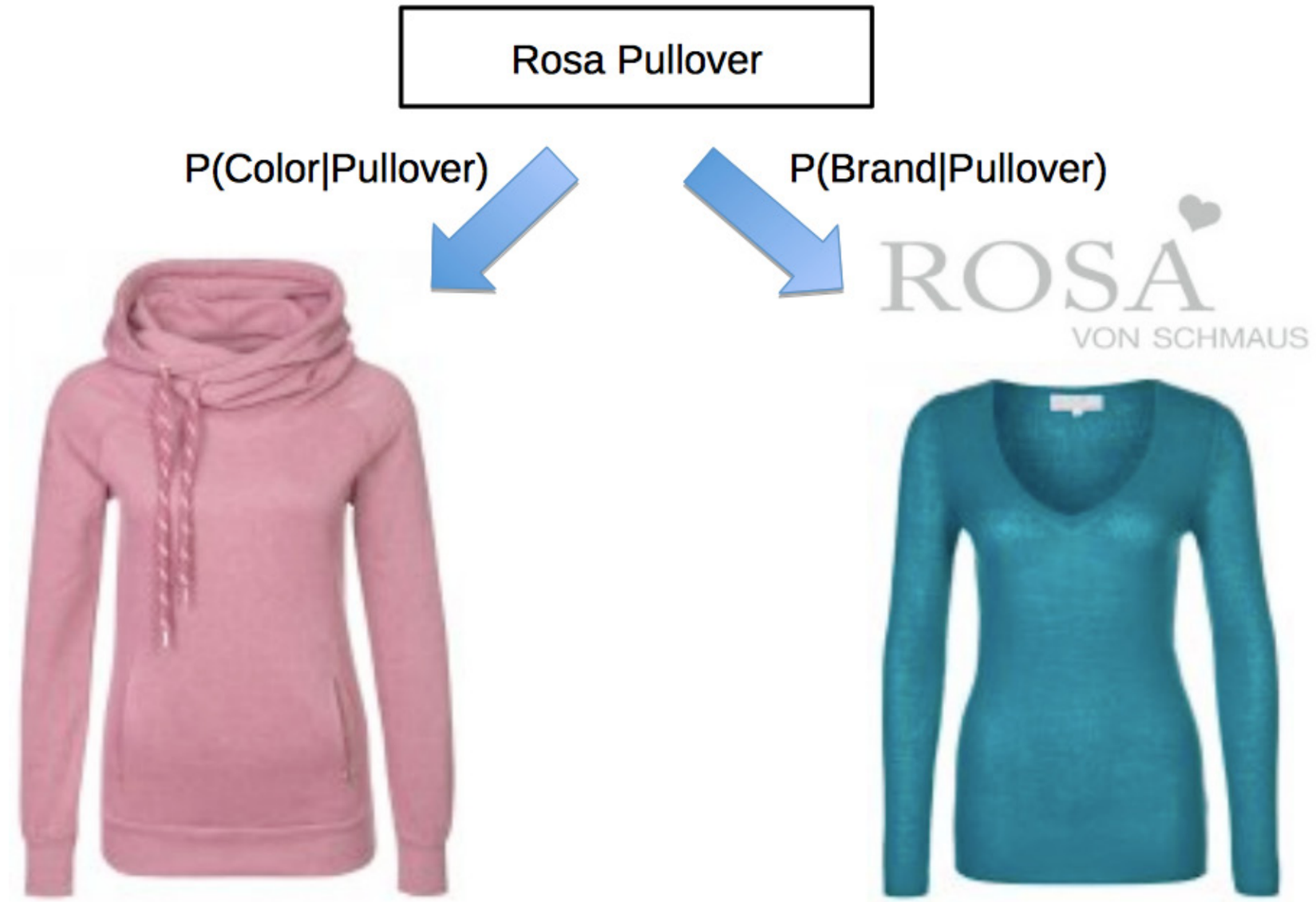
Language Model basierte Suchterm Klassifizierung

Marcus Fabarius
Kamila Kedzior
Philipp Liepert
Rim Sahnoun

Enterprise Data Management
SoSe 2014

Problem

Die Intention des Nutzers ist nicht immer eindeutig



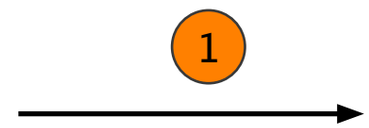
Mit freundlicher Genehmigung von Zalando

N-gram Language Model

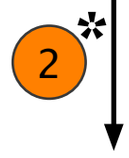
Ansatz

Article.json

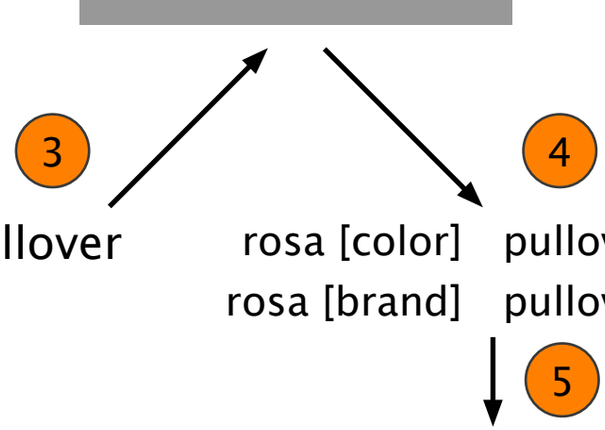

```
{
  "responseHeader":{
    "status":0,
    "QTime":11},
  "response":{"numFound":258751,"start":0,"docs":[
      {
        "sku":"HU722I005-701",
        "name_de":"SINADALIO - Strickpullover - medium brown",
        "brand":"HUGO",
        "brand_code":"HU7",
        "color_key":"color.702",
        "color":"rosa",
        "color_family_key":"color_family.700",
        "color_family":"Pink",
        "brandfamily":"BOSS",
        "brandfamily_code":"BOSS",
        "category":["pullover","herren",
         "premium-herrenbekleidung-strickpullover-sale",
          "alle","premium-herrenbekleidung-sale","outlet-herren",
          "premium-herrenbekleidung-strick-sweat-sale",
          "herrenbekleidung-strickpullover-sale",
          "katalog","outlet-herrenbekleidung",
          "outlet-bekleidung-herrenbekleidung-pullover-sweater",
          "outlet"],
        "tag_translated":[
          "Strickpullover"]},
      {
...
```

rosa [color] pullover [category]
pullover [category] rosa [color]
nike [brand] schuhe [category]
schuhe [category] nike [brand]
⋮
rosa [brand] pullover [category]



Lucene
FreeTextSuggester



Klassifiziertes Ergebnis:
color: rosa
category: pullover
} 80%
brand: rosa
category: pullover
} 20%

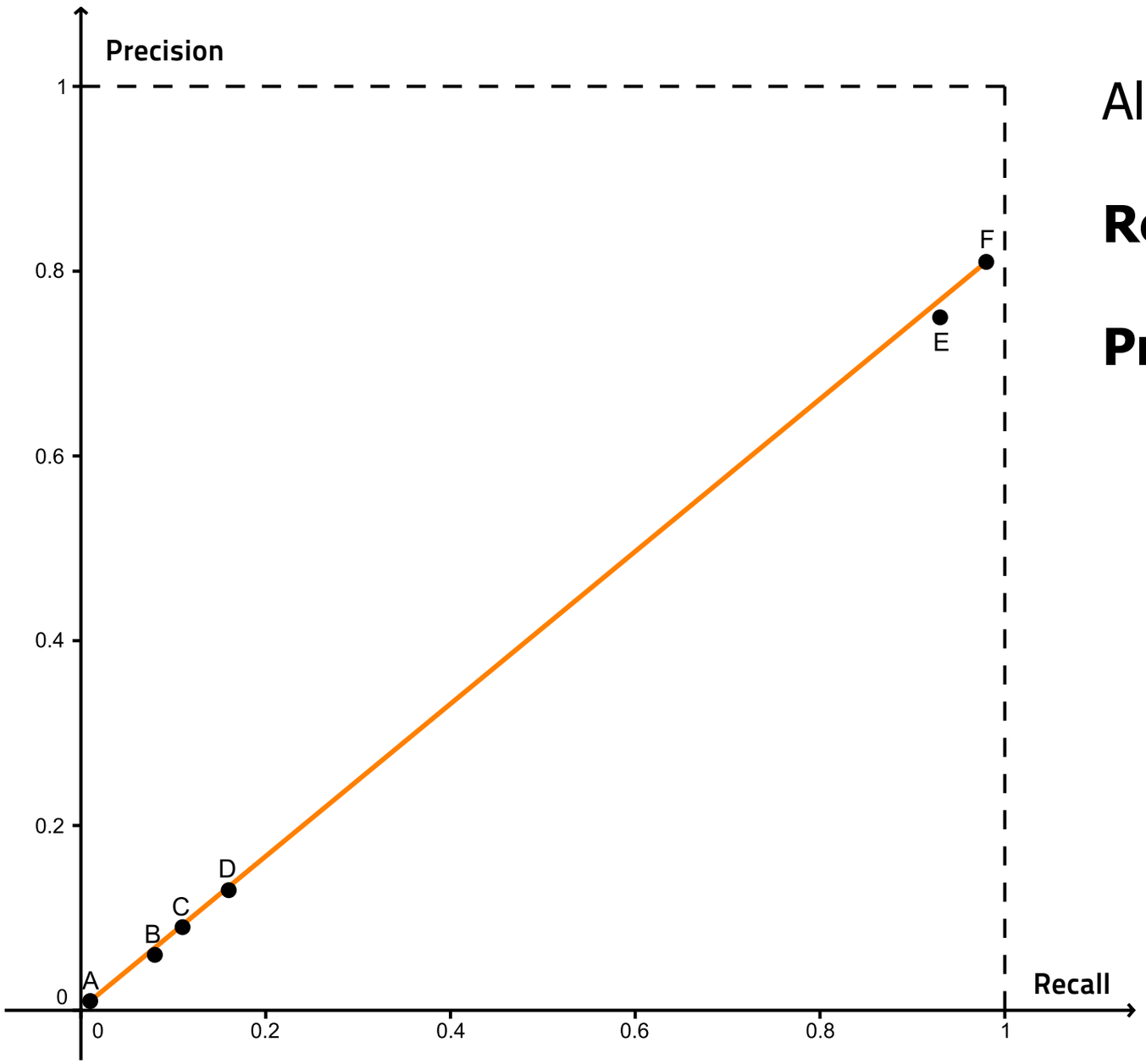
1 und 2 **Indexierungs-Zeit**

3 bis 5 **Query-Zeit**

* Fokus lag auf der Aufbereitung der Eingangsdaten. Dies macht ca. 90% der Programmlogik aus.

Ergebnisse: Mehr Features besseres Ergebnis

Precision / Recall



Alle Optimierungen zusammen:

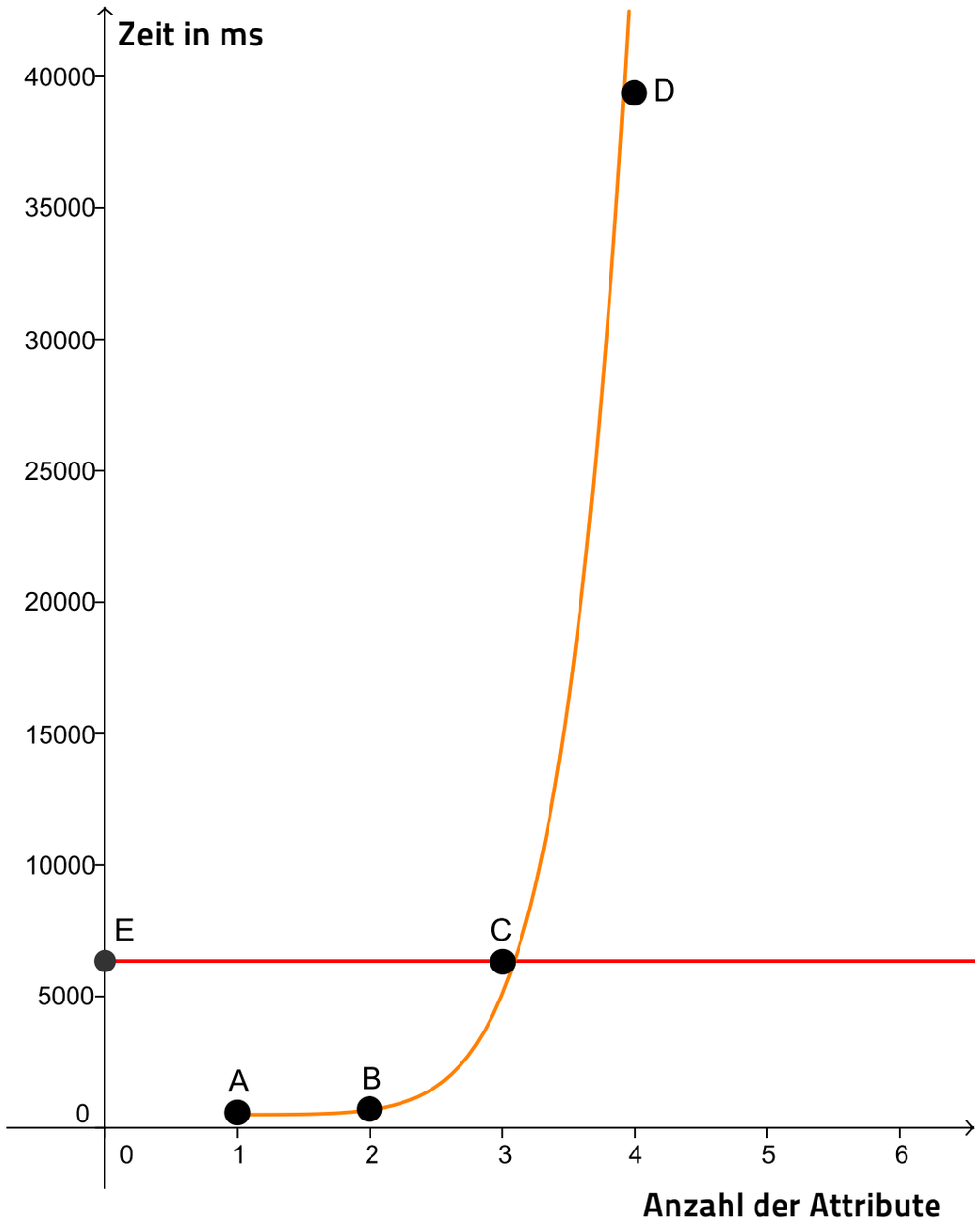
Recall: 59/60 ≙ 98%

Precision: 59/74 ≙ 80%

A: ohne Optimierungen
B: mit Sonderzeichen
C: zusätzlich Multi Words (Marken)
D: zusätzlich Synonyme (Farbe)
E: zusätzlich Permutation
F: optimierte Permutation

Ergebnisse: Mehr Features mehr Aufwand

Zeitaufwand zur Erzeugung des Suggesters



A bis D: Ein bis Vier Attribute.
E: Optimierte Attributenliste (mehrere Attributenlisten in dreier und zweier Kombinationen)

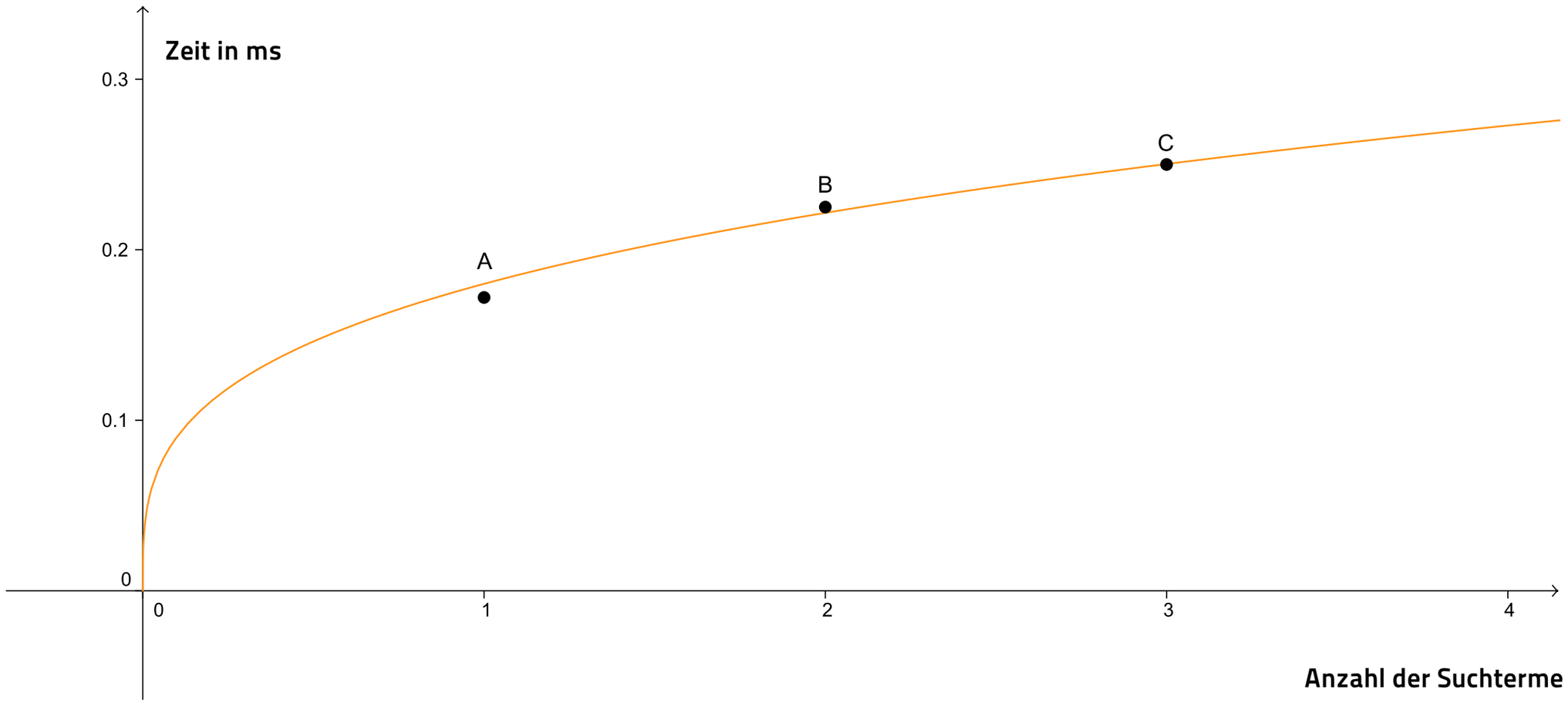
Testumgebung:

Java VM: 64 bit
RAM: 4 GB DDR3

CPU: Intel Core i7-4500U 1.8GHz,
Disk: Samsung SSD 840 EVO 250GB

Ergebnisse: Kurze Antwortzeit

Antwortzeit des Suggesters



A bis C: Länge der Suchphrase

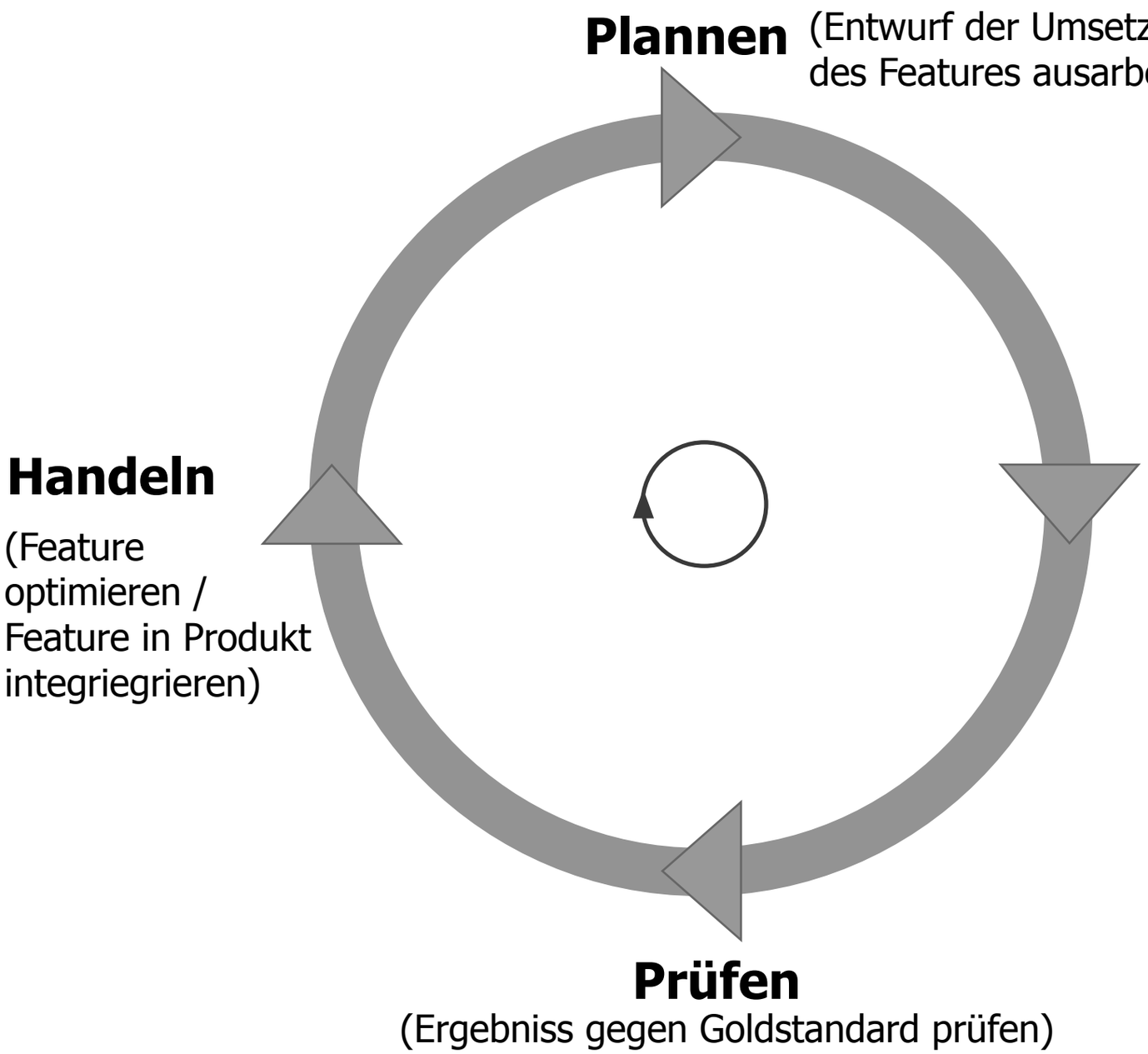
Testumgebung:

Java VM: 64 bit
RAM: 4 GB DDR3

CPU: Intel Core i7-4500U 1.8GHz,
Disk: Samsung SSD 840 EVO 250GB

Iterationsschritte

PDCA - plan-do-check-act



1. Permutationen
2. Synonyme
3. Multiterm Wörter
4. Format der Eingabedaten

Schwächen

Auf nur einen Attributtyp angewandte Features:
- Synonyme nur für Farben
- Multiterm Wörter nur für Marken
- Sonderzeichenbehandlung nur für Kategorien

Nicht implementierte Features:
- Keine Stoppwort Analyse
- Kein Stemming

Datenbestand:
- Geringe Anzahl unterstützter Atribut-Typen
- Zu kleiner Goldstandard

Ansatz ist nicht Kontext-spezifisch

Wie geht es weiter?

1. Qualität des Klassifiziers verbessern
2. Klassifizierer als Teil eines mehrstufigen Verfahrens nutzen
   - Mehrere Klassifizierer arbeiten paralell
   - Regelwerk entscheidet welche Klassifizierung verwendet wird

Wie geht es weiter?

Mehrstufiges Verfahren:

Rosa Pullover	Rosa	Pullover
Language Model	Farbe / Marke	Kategorie
aktueller Klassifizierer	Farbe	Kategorie
Social Klassifizierer	Person	Kleiderstück
...		

Tipps für das nächste Semester

1. Frühzeitig mit sinvollen Projektmanagement Tool beginen.
   - Klare Aufgaben verteilen
   - Teilprobleme rechtzeitig erkennen

2. Falls Anforderungen unklar, frühzeitig klären.

3. Kritik des Dozenten konstruktiver nutzen.

Fragen