Korpuskonstruktion und Entwicklung einer Pipeline für Clickbait-Spoiling

Bachelorthesis

Bagrat Ter-Akopyan

Bauhaus-Universität Weimar, 07.12.2017

Gutachter : Jun.-Prof. Dr. Matthias Hagen, Prof. Dr.-Ing. Volker Rodehorst

Betreuer : Jun.-Prof. Dr. Matthias Hagen, Dr. Martin Potthast, Tim Gollub

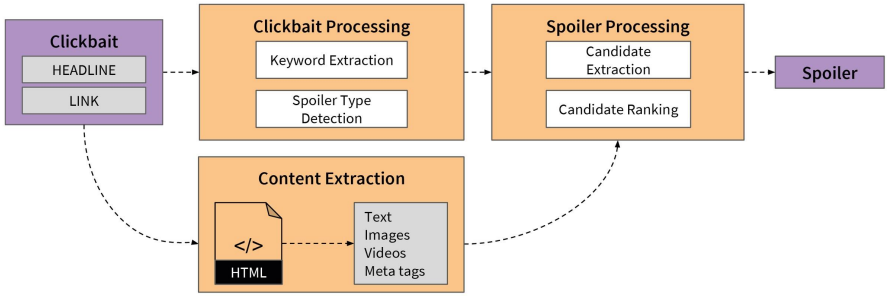
Begriffsdefinition:

Was ist Clickbait-Spoiling?

Begriffsdefinition:

Was ist Clickbait-Spoiling?

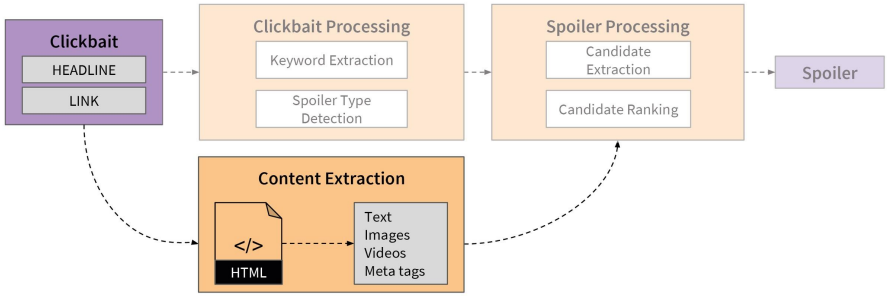
Clickbait Spoiling



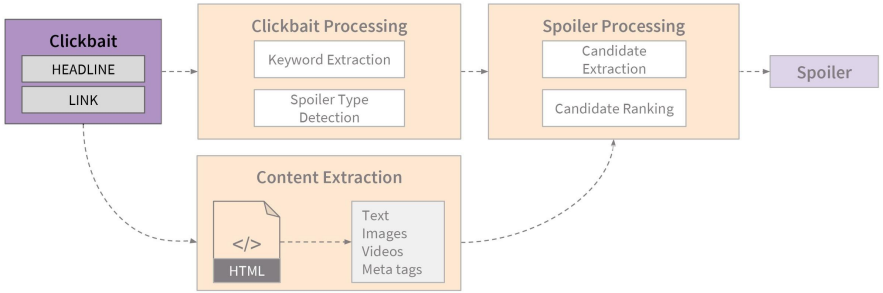
Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft). Januar, 2017. https://web.stanford.edu/~jurafsky/slp3/

Clickbait Spoiling

Datenkorpus



Korpuskonstruktion



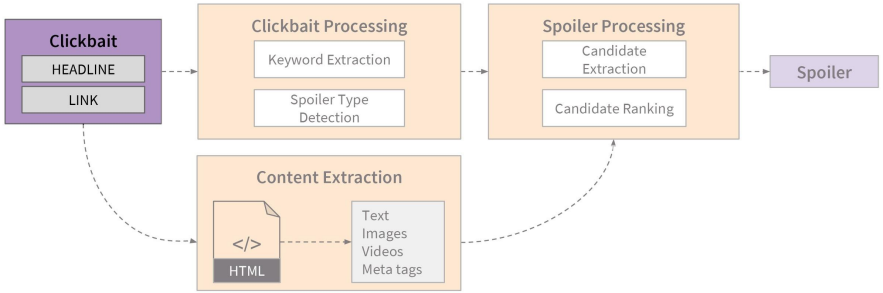
Rohdaten

Rohdaten:

- Reddit (/r/savedyouaclick) → 3024 Clickbaits
- Facebook (@Stop Clickbait) → 1271 Clickbaits
- Twitter (→ 2934 Clickbaits
  @SavedYouAClick,
  @UmworthySpoiler,
  @HuffPoSpoilers)

- Gesamt → 7229 Clickbaits

Korpus



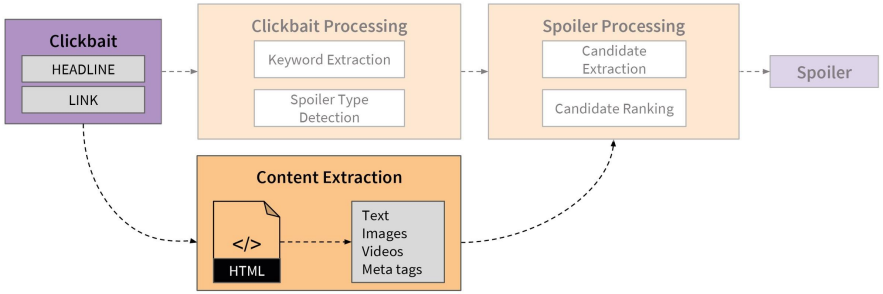
{

post_id: “567763744483467265”
cb_headline: “@HuffPostPol: First California Republican wad..”
cb_spoiler: “CA Assemblyman Rocky Chávez”
spoiler_publisher: “HuffPoSpoilers”
social_media_platform: “Twitter”
target_url: “http://huff.to/1ySOrZo”

article_title: “First Republican Lawmaker Wades Into 2016 California...”
article_text: “California Assemblyman Rocky Chávez announced Tuesday...”
article_imgs: [“http://i.huffpost.com/...ROCKY-CHAVEZ-CALIF-facebook.jpg”]
article_description: “California Assemblyman Rocky Chávez”

}

Korpus



{

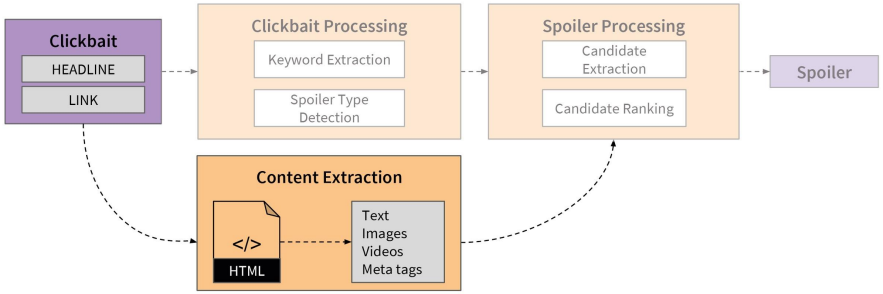
post_id: “567763744483467265”
cb_headline: “@HuffPostPol: First California Republican wad...”
cb_spoiler: “CA Assemblyman Rocky Chávez”
spoiler_publisher: “HuffPoSpoilers”
social_media_platform: “Twitter”
target_url: “http://huff.to/1ySOrZo”

article_title: “First Republican Lawmaker Wades Into 2016 California...”
article_text: “California Assemblyman Rocky Chávez announced Tuesday...”
article_imgs: [“http://i.huffpost.com/...ROCKY-CHAVEZ-CALIF-facebook.jpg”]
article_description: “California Assemblyman Rocky Chávez”

}

Korpusanalyse

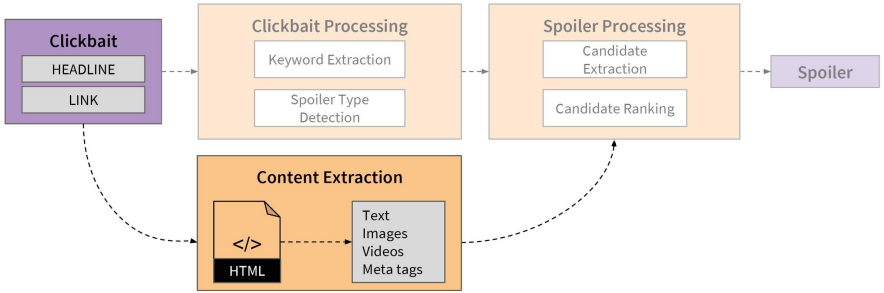
Welche Arten von Clickbaits gibt es?



- Vorwärtsverweis als Aussagesatz → 4190 Clickbaits
  - Clickbait: **First California Republican wades into 2016 Senate race**
  - Spoiler: **Rocky Chávez**
- Vorwärtsverweis als Fragesatz → 934 Clickbaits
  - Clickbait: **Which City Is the Worst for Spring Allergies?**
  - Spoiler: **Mississippi**
- Abgeschlossene Sätze → 663 Clickbaits
  - Clickbait: **Netflix will pay you to watch Netflix all day**
  - Spoiler: **If you live in the UK or Ireland**

Exkurs

Formen der Koreferenz

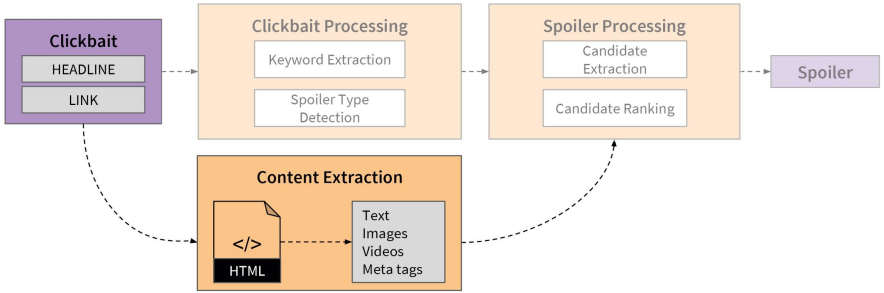


- Anapher - Rückwärtsverweis:
  - **Hans** geht heute essen. **Er** mag besonders Pizza.

- Katapher - Vorwärtsverweis:
  - **Er** mag besonders Pizza. **Hans** geht heute essen.

Korpusanalyse

Vorwärtsverweise in Clickbaits

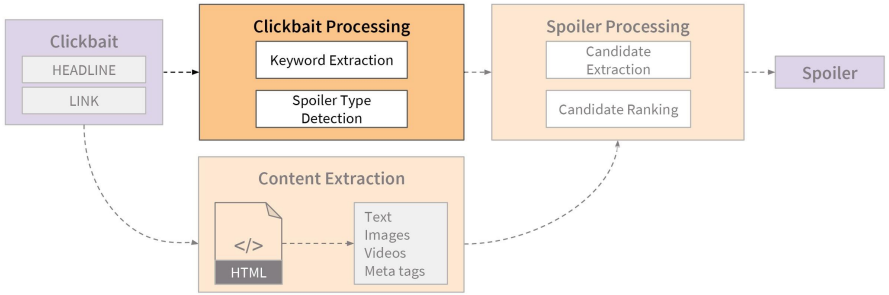


- Clickbaits bedienen sich aller Mittel der Vorwärtsreferenz:[1]
  1. **This** state could be next to legalize pot... → Demonstrativpronomen
  2. **She** planted **these** tea bags in her garden → Personalpronomen
  3. **Here** Is One Thing You Can Do That... → Adverbien des Ortes
  4. **The** world's busiest airport → Bestimmter Artikel
  5. **See** How Much Money Floyd Mayweather... → Imperative
  6. **What** Trump is Hiding About Canada Will... → Interrogativpronomen
  7. Girls **star** opens up about her eating disorder... → Allgemeine Nomen

[1]Blom, Jonas Nygaard, and Kenneth Reinecke Hansen. "Click bait: Forward-reference as lure in online news headlines." Journal of Pragmatics 76 (2015): 87-100.

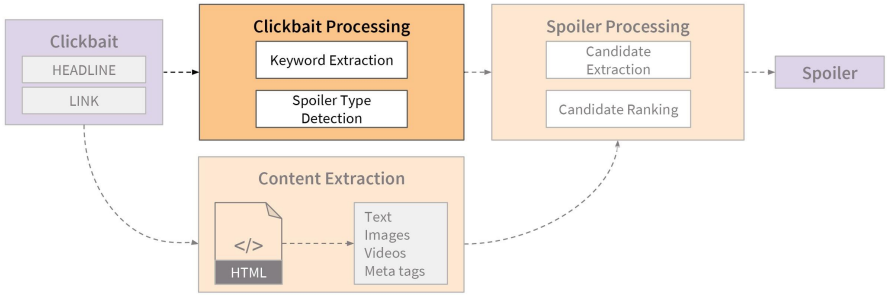
Clickbait Spoiling

Clickbait Processing



Clickbait Processing

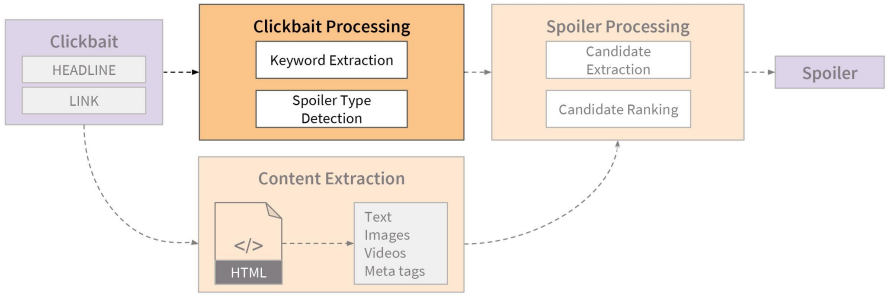
Metainformation aus dem Korpus



- Extraktion der Schlüsselwörter (Dan Moldovan et al. 1999)
- Sentimentanalyse (StanfordNLP)
- Auflösung der Koreferenzen (StanfordNLP)
- Manuelle Zuweisung des Spoilertyps

Clickbait Processing

Auflösung der Koreferenzen
(StanfordNLP)

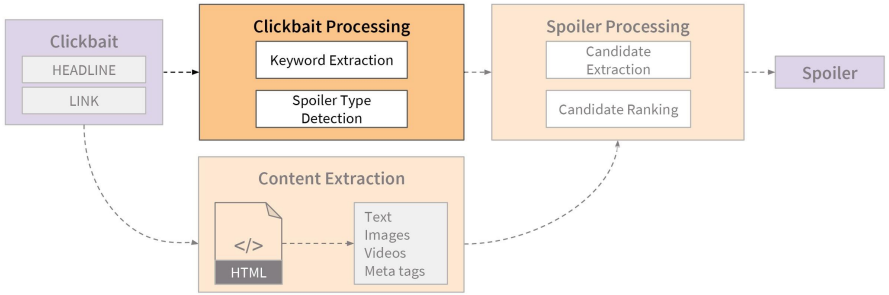


Clickbait: @HuffPostPol: **First California Republican** wades into 2016 Senate race

Text: "California Assemblyman **Rocky Chávez** announced Tuesday that he is exploring a bid for the Golden State's open U.S. Senate seat in 2016, becoming the first sitting Republican lawmaker to take a formal step toward running for retiring Democratic Sen. Barbara Boxer's seat. **Chávez**, who represents part of San Diego County, has opened an exploratory ... growth for almost two decades," **Chávez** said in a statement.

Clickbait Processing

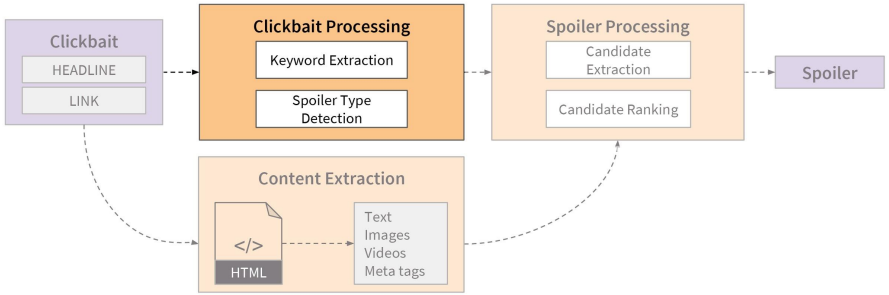
Spoiler Type Detection



- Faktenbasierte Clickbaits
  - A cup of coffee will cost you 8$ in this **city**
- Komplexe (narrative) Clickbaits
  - What Does Your Eyebrow Shape Say About Your Personality?

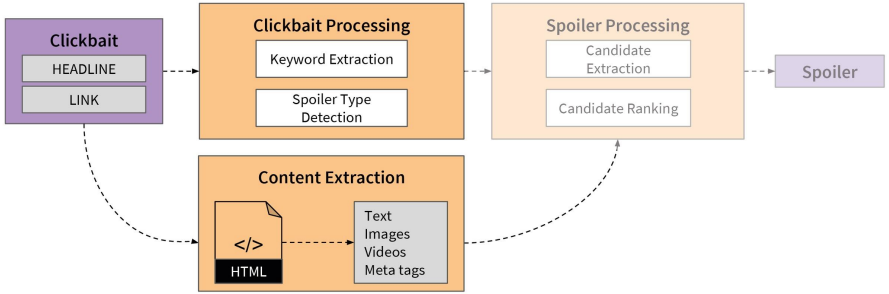
Clickbait Processing

Spoiler Type Detection



Spoilertyp	Anzahl
PERSON	**429**
LOCATION	**219**
DATE	47
ORGANIZATION	50
Sonstige	5042

Korpusreduktion



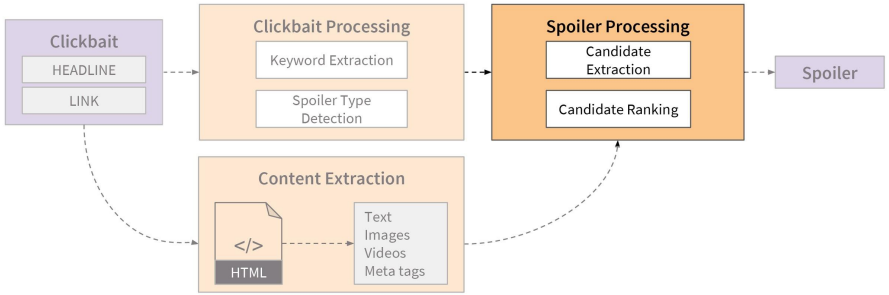
Personen + Orte

648 Clickbaits

- Clickbaits, die genau einen Spoiler erwarten — 370 Pers. + 189 Orte → 559 Clickbaits
- Der korrekte Spoiler ist im verlinkten Dokument — 345 Pers. + 167 Orte → 512 Clickbaits

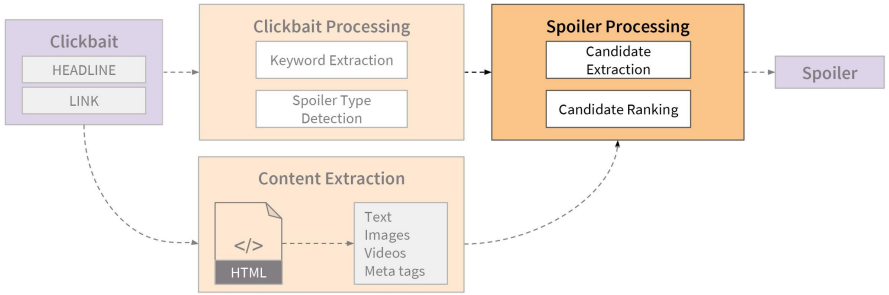
Clickbait Spoiling

Spoiler Processing



Spoiler Processing

Candidate Extraction

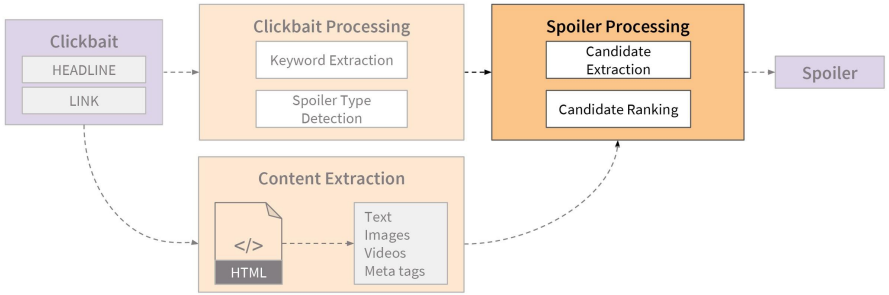


Clickbait: @HuffPostPol: First California Republican wades into 2016 Senate race
Spoilertyp: PERSON

"California Assemblyman **Rocky Chávez** announced Tuesday that he is exploring a bid for the Golden State's open U.S. Senate seat in 2016, becoming the first sitting Republican lawmaker to take a formal step toward running for retiring Democratic Sen. **Barbara Boxer**'s seat. **Chávez**, who represents part of San Diego County, has opened an exploratory committee, allowing him to fundraise for the statewide race. Well-known Republicans like Rep. **Darrell Issa, San Diego Mayor Kevin Faulconer** and 2014 gubernatorial candidate **Neel Kashkari**..."

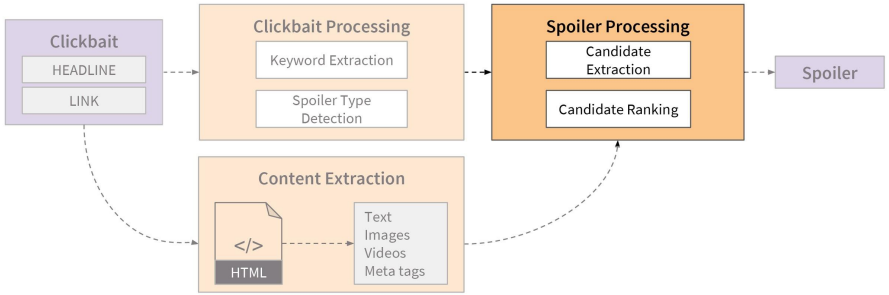
Spoiler Processing

Candidate Ranking



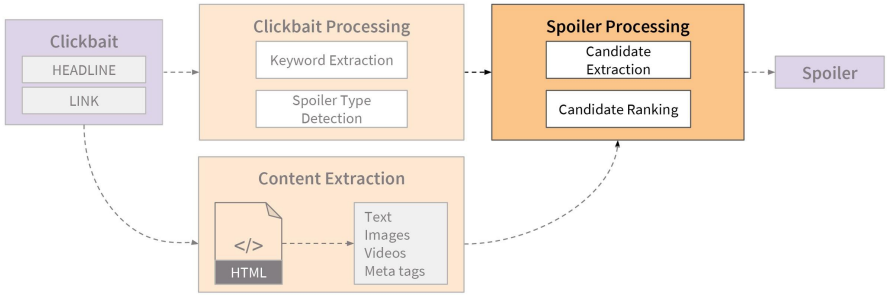
- Nach der Position im verlinkten Dokument
- Nach der Häufigkeit im verlinkten Dokument
- Regelbasierter Ansatz
- Machine Learning Ansatz

Regelbasierter Ansatz



1. Sätze mit mindestens einem Kandidaten
2. Kosinusähnlichkeit zu der Clickbait-Überschrift
3. *N* Sätze mit höchster Ähnlichkeit
4. Häufigster Kandidat aus den *N* selektierten Sätzen
5. Falls mehrere: häufigster Kandidat aus dem gesamten Dokument
6. Falls mehrere: erst genannter Kandidat im gesamten Dokument

Machine Learning Ansatz



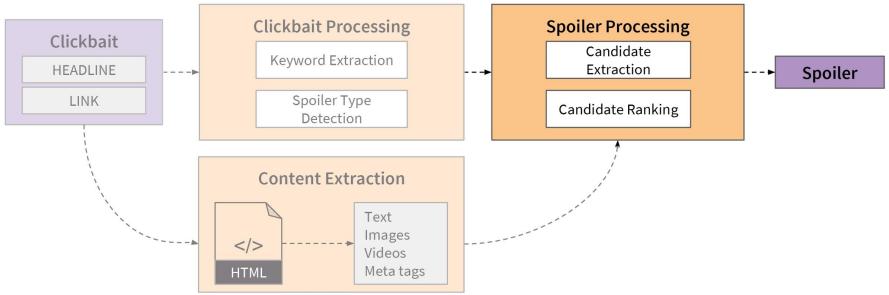
Features:

(Kandidatensätze - kommt mind. 1 Kandidat vor)

1. Häufigkeit des Kandidaten im Text
2. Maximale Kosinusähnlichkeit zwischen Kandidatensätzen und der Clickbait-Überschrift
3. Anzahl aller erkannten Entitäten in den Kandidatensätzen
4. Anzahl der Schlüsselwörter aus der Clickbait-Überschrift in den Kandidatensätzen
5. Erster Treffer im Text (Satzindex)
6. Kandidat in der Liste mit Koreferenzen
7. Durchschnittlicher Stimmungswert über alle Kandidatensätze

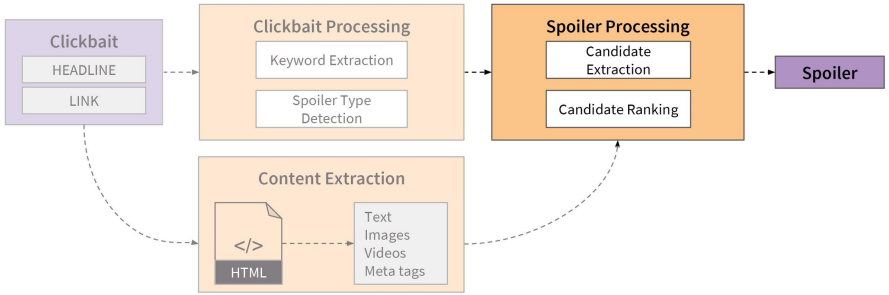
Clickbait Spoiling

Evaluation



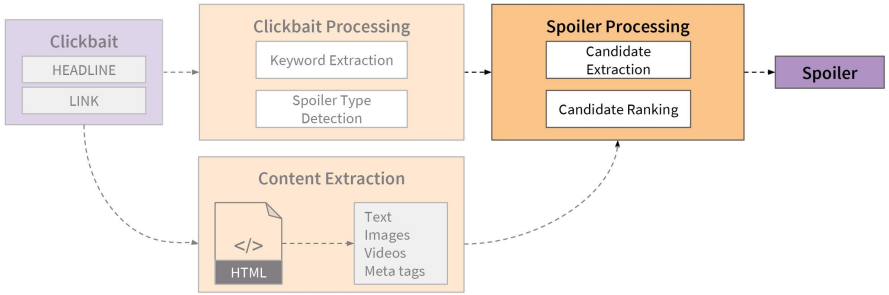
Evaluation

Trainings- und Testdaten



- In jedem Dokument gibt es mindestens einen richtigen Kandidaten
- Insgesamt: 512 Dokumente mit 888 richtigen und 3380 falschen Kandidaten
- 80% Trainings-, 20% Testdaten
- Trainingsdaten:
  - 409 Dokumente: 706 richtige und 2634 falsche Kandidaten
- Testdaten:
  - 103 Dokumente: 182 richtige und 746 falsche Kandidaten

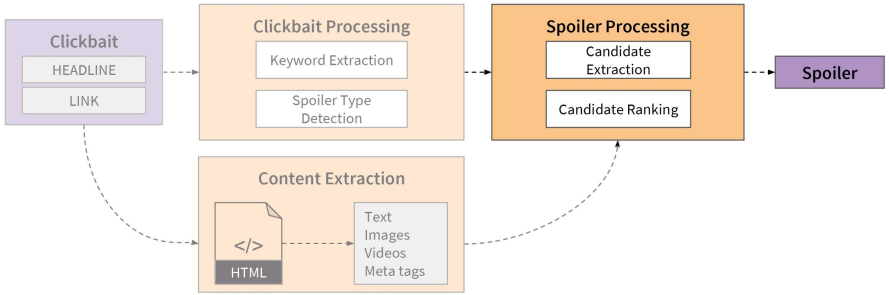
Evaluationsmaße



1. *Accuracy*
   - Entspricht der höchst gerankte Kandidat dem Goldstandard?
2. *Mean Reciprocal Rank*
   - Für jedes Dokument eine Liste mit gewichteten Kandidaten
   - Bewertung: 1/Rang des ersten korrekten Kandidaten
     - Wenn der erste Kandidat korrekt ist: 1
     - Wenn der zweite Kandidat korrekt ist: ½ usw.
     - Wenn keiner korrekt ist: 0
   - Bildung des Mittelwertes über alle Dokumente

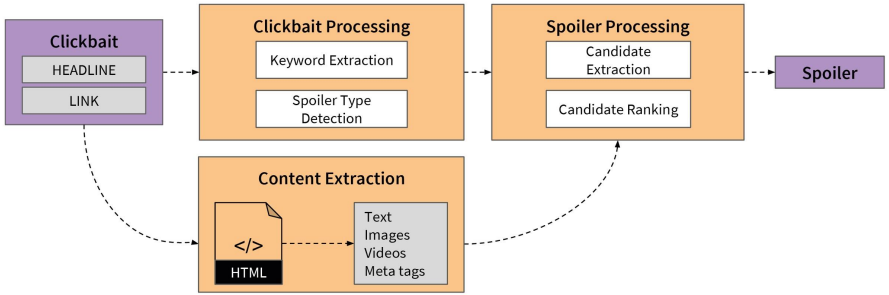
$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_1}}{N}$$

Evaluation



Ansatz	Accuracy	MRR@3
Erstgenannter Kandidat	0.2524	0.4498
Häufigster Kandidat	0.6893	0.8009
Regelbasiert (N=5)	0.7087	-
Machine Learning (DT, 1+5+6)	0.7378	0.8171

Ausblick



- Erweiterung des bestehenden Systems:
  - Signifikantere Features für den ML-Ansatz
  - Andere Entitäten
  - Clickbaits z.B. mit Bildern spoilern

- Andere Verfahren:
  - Text Summarization (für narrative Clickbaits)

Danke für Ihre Aufmerksamkeit!