

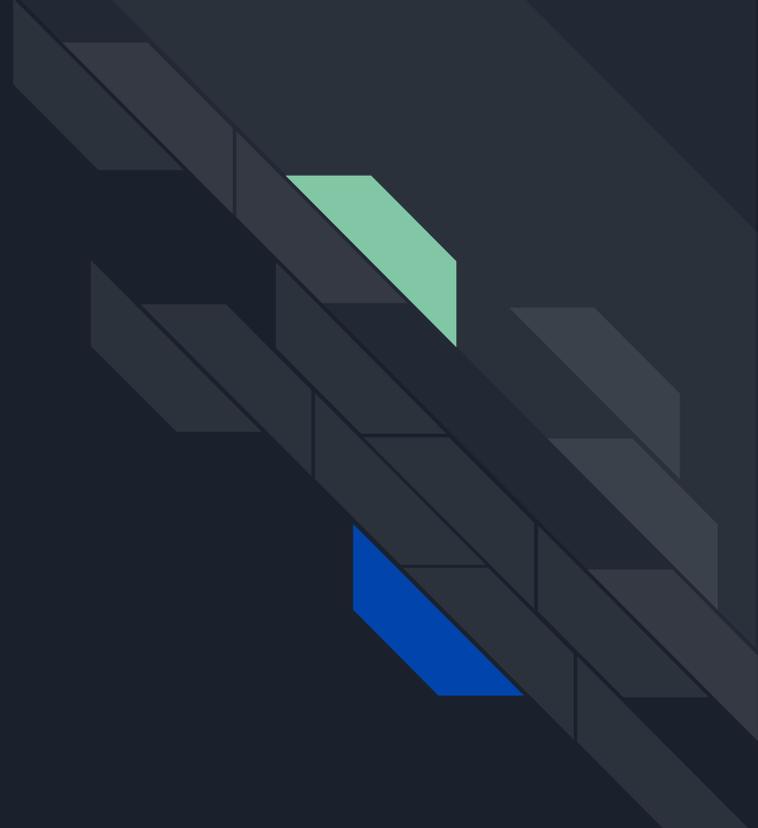


Information Extraction

von Simon Herding und Sophie Jorns

Inhalt

- Definition
- Anwendungsmöglichkeiten
- Funktionsweise
- Unterscheidung verschiedener Herangehensweisen
- Abwägung von Vor- und Nachteilen
- Anwendungsbeispiel
- Quellen





Definition

- Prozess unstrukturierte Daten zu analysieren und wesentliche Informationen in strukturierte Datenformate zu extrahieren
- Ziel: relevante Informationen zu identifizieren, diese zu extrahieren und in Datenbanken abzulegen
- Weiterverarbeitung zum Text-Mining, Meinungsforschung, Textzusammenfassung etc.



Definition

- Abgrenzung von Nachbargebieten:
 - *Text-Extraction*: umfassende Zusammenfassung des Inhaltes eines Textes
 - *Textclustering*: selbständiges Gruppieren von Texten
 - *Textklassifikation*: das Einordnen von Texten in vorgegebene Gruppen
 - *Information Retrieval*: Dokumentensuche innerhalb einer Datenbank
 - *Data-Mining*: Prozess Muster in Daten zu erkennen



Anwendungsmöglichkeiten

- Menschliche Nutzung -> unstrukturierte
Ergebnisdarstellung
- Maschinelle Weiterverarbeitung -> strukturierte
Ergebnisdarstellung



Grundlegende Funktionsweise

- Knowledge Engineering-Ansatz
 - Manuelle Eingabe von Regeln
 - IE sucht nach vordefinierten Mustern im Text
- Maschinelles Lernverfahren
 - IE lernt anhand von Beispieldokumenten die gesuchten Informationen und Textmerkmale



Grundlegende Funktionsweise

- Einfügen eines/ mehrerer Fließtexte
- Daten verarbeitbar machen
- Festlegen einer Domäne
- Filtern nach Informationen, Entitäten, Zusammenhängen, Ereignissen
- Zusammenfassung der Informationen
- Zwei Perspektiven
 - IE als Erkennen und Sammeln von Informationen
 - IE als Löschen unwesentlicher Informationen



Unterscheidung verschiedener Herangehensweisen

- Tokenisierung: Text wird erst in Sätzen und dann in Tokens aufgeteilt
- Part of Speech Tagging: Tokens werden mit Wörterbucheinträgen verglichen (Wortart, Attribute)
- Word Sense Tagging: Eigennamen werden identifiziert und klassifiziert
- morphologische Analyse: Wörter werden in ihre Bedeutungen zerlegt



Unterscheidung verschiedener Herangehensweisen

- lexikalische Analyse: Wörter werden kategorisiert (Part of Speech Tagging) und deren Bedeutung festgelegt (Word Sense Tagging)
- syntaktische Analyse: Extraktion von Fakten und Events, Beziehungen der Nominalphrasen eines Textes
- Koreferenzanalyse: Anaphorische Abhängigkeiten werden aufgelöst, zum Beispiel Pronomen
- Domänenanalyse: Informationen aus dem jeweiligen Fachgebiet werden analysiert



Vorzüge der Information Extraction

- Große Arbeitersparnis
- Lösung für ein Teilproblem der Sprachverarbeitung
- Wichtig als Datengrundlage für andere Bereiche

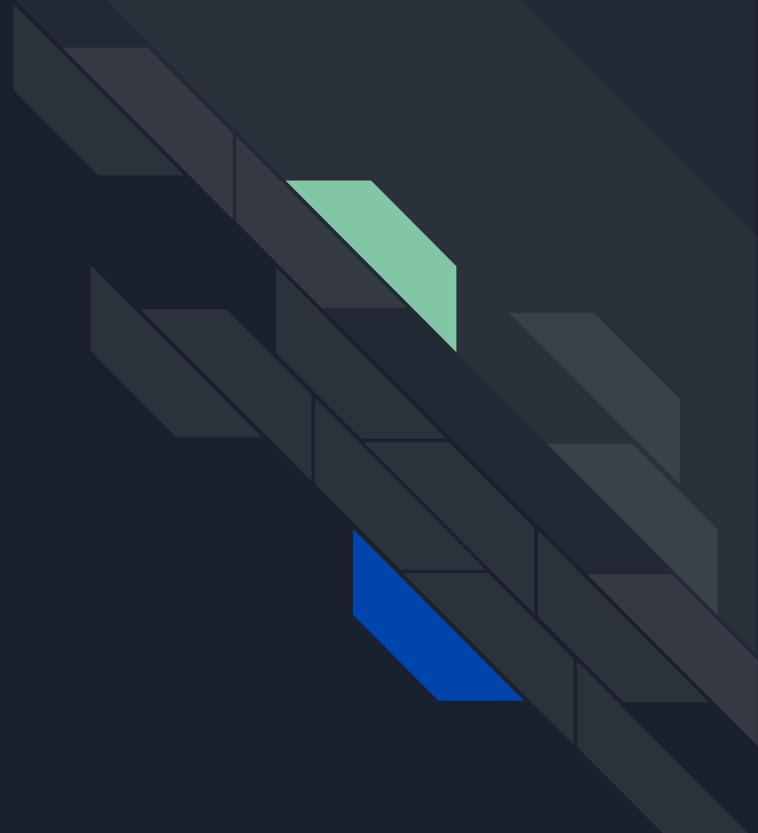


Nachteile der Information Extraction

- System sprachabhängig
- System domänenabhängig

Anwendungsbeispiel

<https://demo.natif.ai>



Quellen

- <https://gi.de/informatiklexikon/informationsextraktion>
- https://reposit.haw-hamburg.de/bitstream/20.500.12738/5938/1/BA_Probst.pdf
- <https://de.wikipedia.org/wiki/Informationsextraktion>
- <https://nanonets.com/blog/information-extraction/>
- <https://esb-dev.github.io/mat/swt-domain-l.pdf>

