DEUTSCHE NATIONAL BIBLIOTHEK

Matthias Nagelschmidt, Christoph Poley

Workshop: Vom Text zum Inhalt

DEUTSCHE NATIONAL BIBLIOTHEK

Fahrplan / Ziele

1. Grundlagen der natürlichen Sprachverarbeitung
   - Natürliche Sprache hat (k)eine Struktur
   - Natürliche Sprachverarbeitung – Welche Methoden stecken dahinter?
2. Hands on Session
   - Aufgabe / Herausforderung
   - Lösungsidee
   - Python NLTK
   - Diskussion

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprachverarbeitung – was ist das?

Natürliche Sprachverarbeitung bedeutet…

…automatische Analyse von textuell vorliegender, natürlicher (d.h. nicht formaler) Sprache.

Natürliche Sprachverarbeitung ist eine Vorverarbeitung für…

…Informationsextraktion (z.B. zur Abbildung auf ein normiertes Vokabular) und

…Informationsgenerierung (z.B. durch automatisches Abstracting etc.).

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur

Natürliche Sprache erhält ihre Struktur durch die grammatische Syntax. *Aber*: Textuelle Repräsentationen natürlicher Sprache gehören zu den sog. „unstrukturierten Daten"…

…Unstrukturierte Daten sind „formlos" und können nicht ohne Weiteres maschinell verarbeitet werden.

…Bevor die maschinelle Verarbeitung natürlicher Sprache stattfinden kann, muss der zugrundeliegende Text strukturiert werden.

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur

Strukturierung durch Mustererkennung…

…Text wird eingelesen und „tokenisiert", d.h. in einzelne Wörter und Satzzeichen (sog. Tokens) zerlegt [1, S. 109-111],

…Tokens werden mit Wortklassen-Informationen ausgestattet (sog. Part-of-Speech (PoS)-Tags) [1, S. 179-189],

…anhand er PoS-Tags werden bestimmte PoS-Sequenzen (sog. „Chunks") extrahiert [1, S. 264-270].

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur – Tokenisierung

Tokenizer…

…entfernen Leerzeichen (Whitespace-Tokenizing),

…basieren auf generischen Mustern (Regexp-Tokenizing),

…basieren auf intellektuell tokenisierten Referenzkorpora.

```
Noam Chomsky developed the transformational generative grammar.
```

```
['Noam', 'Chomsky', 'developed', 'the', 'transformational', 'generative', 'grammar', '.']
```

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur – Part-of-Speech-Tagging

PoS-Tagger…

…sind Verfahren der Mustererkennung,

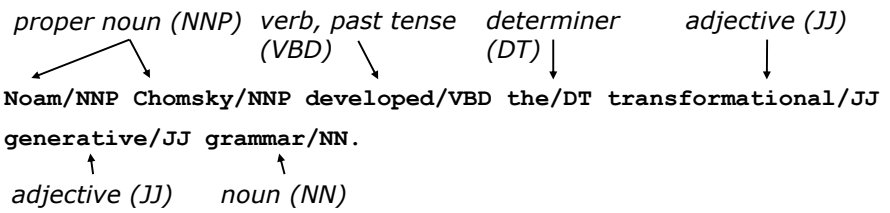
…übertragen bekannte Muster aus korrekt getaggten Korpora auf neue Texte,

…können auch Morphologie, insb. Suffix-Muster berücksichtigen, z.B. .*ed (VBD, past tense), .*ing (VBG, present participle), .*\'s (POS, possessive ending) etc.

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur – Part-of-Speech-Tagging

```
['Noam', 'Chomsky', 'developed', 'the', 'transformational',
'generative', 'grammar', '.']
```



DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprache hat (k)eine Struktur – Chunking

Chunking…

…basiert auf einer„Chunk Grammar".

```
<DT>?<JJ>*<NN>
```

```
Noam/NNP/O Chomsky/NNP/O developed/VBD/O the/DT/B-NP
transformational/JJ/I-NP generative/JJ/I-NP grammar/NN/I-NP
```

PoS-Tags werden um IOB-Tags ergänzt (B-NP: beginning noun phrase, I-NP: inside noun phrase, O: outside noun phrase). Durch IOB-Tags lassen sich Chunks auf einfache Weise in unstrukturierten textuellen Daten abbilden.

DEUTSCHE NATIONAL BIBLIOTHEK

Natürliche Sprachverarbeitung – Welche Methoden stecken dahinter?

Klassische Unterscheidung in drei Methoden-Typen [2, S. 245-246]…

…linguistisch-lexikonbasiert,

…linguistisch-regelbasiert,

…statistisch.

Linguistische Methoden: Natürliche Sprache analysieren, normalisieren, Unabhängigkeit von den vielfältigen Ausdruckformen schaffen. *Statistische Methoden*: Inhaltliche Relevanz einzelner Wörter bzw. Phrasen ermitteln.

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-lexikonbasierte Methoden

Bereitstellung von sprachbezogenem Wissen durch Auflistung des lexikalischen Inventars natürlicher Sprachen. Etablierte Lexika-Typen sind [2, S. 263]…

…Vollformen-Lexika (enthalten flektierte Wortformen),

…Grundformen-Lexika (enthalten Wortarten in ihrer jeweiligen Grundform, z.B. Nominativ, Infinitiv etc.),

…Stammformen-Lexika (enthalten Wortstämme, auf denen Wortbildung und Flexion aufbauen).

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-lexikonbasierte Methoden

```
#Vollformen-Lexikon
Wirtschaft
Wirtschaften
wirtschaften
Wirtschaftliche
Wirtschaftlichkeiten
…
```

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-lexikonbasierte Methoden

```
#Grundformen-Lexikon
{Wirtschaft, Wirtschaften}        Wirtschaft #NN
{wirtschaften}                    wirtschaften #VVFIN
{wirtschaftliche}                 wirtschaftlich #ADJD
{Wirtschaftlichkeiten}            Wirtschaftlichkeit #NN
```

…

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-lexikonbasierte Methoden

```
#Stammformen-Lexikon
{Wirtschaft, Wirtschaften,        wirtschaft
wirtschaften, wirtschaftliche,
Wirtschaftlichkeiten}
```

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-regelbasierte Methoden

Bereitstellung von sprachbezogenem Wissen durch ein generisches Set von Regeln, die aus einer flektierten Wortform eine Grund- bzw. eine Stammform erzeugen können. Bekannte Regelsets sind…

…der Porter-Stemmer [4] (ursprünglich für das Englische entwickeltes Regelset, das anhand von Verkürzungsregeln Stammformen erzeugt),

…der Kuhlen-Algorithmus [3] (für das Englische entwickeltes Regelset, das Grund- und Stammformen erzeugt).

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-regelbasierte Methoden

#Porter-Stemmer

[C](VC){m}[V] **Abstrakte Beschreibung eines beliebigen Terms, wobei…**

…[C] ein oder mehrere optionale Konsonanten,

…[V] ein oder mehrere optionale Vokale,

…(VC){m} Anzahl der Vokal-Konsonant-Folgen zwischen [C] und [V] bedeutet. In Abhängigkeit der Anzahl *m* greifen verschiedene Suffix-Stripping-Regeln.

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistisch-regelbasierte Methoden

```
#Porter-Stemmer
hopeful       m=3    FUL ->              hopeful -> hope
feudalism     m=3    ALISM -> AL         feudalism -> feudal
replacement   m=4    EMENT ->            replacement -> replac
```


DEUTSCHE NATIONAL BIBLIOTHEK

Statistische Methoden

Ermittlung einfacher Auftretenshäufigkeiten von Termen, um Rückschlüsse auf deren Relevanz für den Dokumentinhalt zu ziehen. Basale statistische Maße sind…

…TF (absolute Auftretenshäufigkeit eines Terms),

…relative TF (am Gesamtumfang der Kollektion relativierte Auftretenshäufigkeit eines Terms)

…TF*IDF (dokumentbezogene Auftretenshäufigkeit eines Terms in Relation zu dessen kollektionsbezogener Auftretenshäufigkeit).

DEUTSCHE NATIONAL BIBLIOTHEK

Linguistik oder Statistik? Oder beides?

Linguistische Methoden normalisieren natürliche Sprache und führen Terme in flektierten Wortformen auf Grund- bzw. Stammformen zurück.

Statistische Methoden führen einfache Zähloperationen auf Termen durch und schließen daraus auf deren inhaltliche Relevanz.

Daraus folgt: Linguistische Methoden „bereinigen" die Datenbasis für statistische Methoden.

DEUTSCHE NATIONAL BIBLIOTHEK

Sonderfall Eigennamen – Named Entity Recognition

Natürliche Sprache umfasst Eigennamen (von Personen, Organisationen, Produkten, Ländern, Städten, etc.)…

…Noam Chomsky,

…Deutsche Nationalbibliothek,

…Microsoft Powerpoint,

…Königreich Bhutan,

…Quadrath-Ichendorf,

…Essen.

DEUTSCHE NATIONAL BIBLIOTHEK

Sonderfall Eigennamen – Named Entity Recognition

Eigennamen sind oft mehrdeutig und nicht vorhersehbar, da durch den Sprachgebrauch ständig neue Eigennamen entstehen.

Das Erfassen von Eigennamen in einem Wörterbuch macht nur Sinn, wenn bewusst nur ganz bestimmte, wenige Eigennamen erkannt werden sollen.

Für eine umfassende Erkennung von Eigennamen (Named Entity Recognition bzw. NER) muss auf komplexere, algorithmische Lösungen zurückgegriffen werden.

DEUTSCHE NATIONAL BIBLIOTHEK

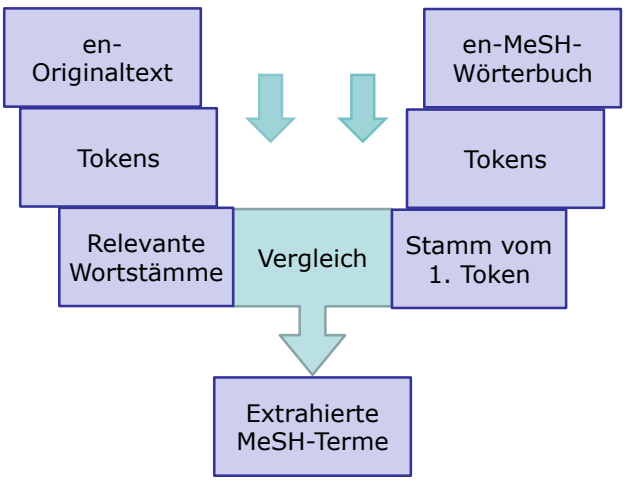
Hands on!

- Gegeben sei ein englischer medizinischer Text, z.B.: „An Atypical Presentation of Alagille Syndrome. Alagille syndrome is a multisystem disorder classically involving the liver, heart, vertebrae, facial features, and the eyes. …"
- Es sollen alle darin enthaltenen MeSH Terms (Medical Subject Headings) mit Konfidenzwerte ermittelt werden.
- Nutze linguistische und statistische Methoden
- Werkzeugkasten: Python3, NLTK, …

Python NLTK (Natural Language Toolkit)

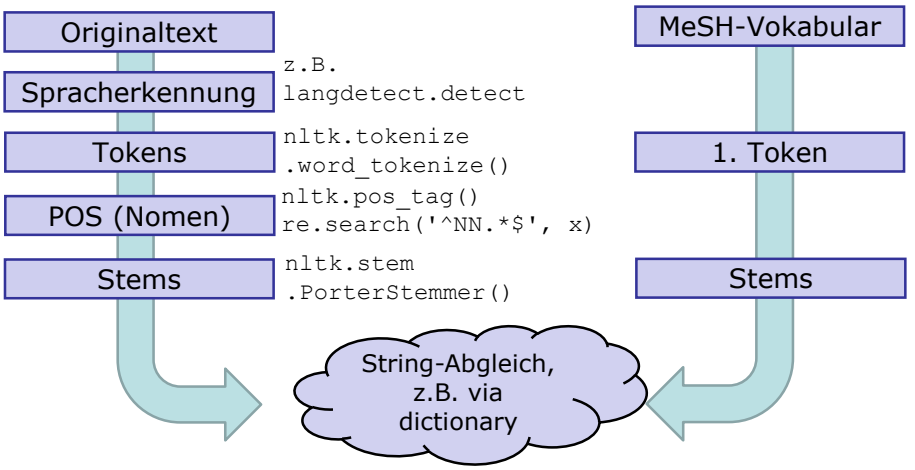
- Bibliothek für Anwendungen in der Computerlinguistik
- Apache Lizenz
- Ursprünge 2001 University of Pennsylvania unter Edward Loper und Steven Bird entwickelt
- Mehr als 50 Korpora (Tokenizer, Stemmer, POS-Tagger, City Database, …)
- Ist im Lehrbereich weit verbreitet

Lösungsidee



DEUTSCHE NATIONAL BIBLIOTHEK

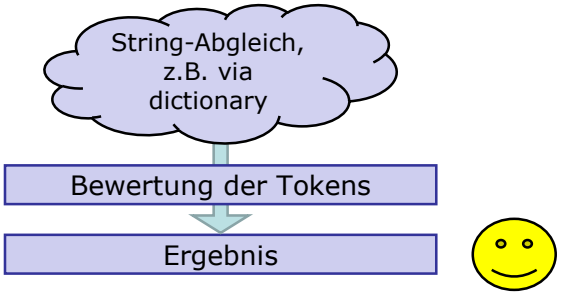
Details für möglichen Lösungsansatz



DEUTSCHE NATIONAL BIBLIOTHEK

Konfidenzwerte

- Statistische Methoden:
  - Auftretenshäufigkeit eines Tokens in einem Datensatz (TF)
  - Position des Tokens im Text / Ergebnismenge
  - …



DEUTSCHE NATIONAL BIBLIOTHEK

Literatur

[1] Bird, S./Klein, E./Loper, E. (2009): Natural Language Processing with Python. O'Reilly Media, Sebastopol.

[2] Gödert, W./Lepsky, K./Nagelschmidt, M. (2012): Informationserschließung und automatisches Indexieren. Ein Lehr- und Arbeitsbuch. Springer, Berlin et al., S. 245-246.

[3] Kuhlen, R. (1977): Experimentelle Morphologie in der Informationswissenschaft. Verl. Dokumentation, München. (DGD-Schriftenreihe ; Bd. 5). Zugl. Diss., Univ. Regensburg 1977 unter dem Titel: Flexive und Derivative in der maschinellen Verarbeitung englischer Sprache.

[4] Porter, M. (1980): An Algorithm for Suffix Stripping. In: Program, 14. Jg., H. 3, S. 130-137.

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank!

- Matthias Nagelschmidt: m.nagelschmidt@dnb.de
- Christoph Poley: c.poley@dnb.de