



# Introduction to Intelligent User Interfaces

Text and Natural Language Processing



# Overview - NLP

- **Applications**
- Definitions
- Terms, basic concepts and algorithms
- Milestones in the history of NLP
- Text Analytics

# Artificial languages and Natural Languages

```
42 <body><?php body_class='>
43 <div id="fb-root"></div>
44 <script>(function(d, s, id) {
45   var js, fjs = d.getElementsByTagName(s)[0];
46   if (d.getElementById(id)) return;
47   js = d.createElement(s); js.id = id;
48   js.src = "//connect.facebook.net/en_US/sdk.js#xfbml=1&version=v2.6&appId=209354264513715";
49   fjs.parentNode.insertBefore(js, fjs);
50 } (document, 'script', 'facebook-jssdk'));</script>
51 <div id="page" class="site">
52   <a class="skip-link screen-reader-text" href="#content"><?php esc_html_e( 'Skip to content' ) ?>
53   <header id="masthead" class="site-header" role="banner">
54     <div class="site-branding">
55       <div class="navbtn pull-left">
56         <?php if(is_home()) && $xpanel['homepage-style'] == 1 { ?>
57           <?php if(!is_page('')) { ?>
58             <a href="#" id="openMenu"><i class="fa fa-bars fa-3x"></i></a>
59           <?php } else { ?>
60             <a href="#" id="openMenu2"><i class="fa fa-bars fa-3x"></i></a>
61           <?php } ?>
62       </div>
63       <div class="logo pull-left">
64         <a href=<?php echo esc_url( home_url() ) ?>">
65           <img src=<?php echo $xpanel['logo']['url'] ?>">
66         </a>
67       </div>
68       <div class="search-box hidden-xs hidden-sm pull-left ml-10">
69         <?php get_search_form(); ?>
70       </div>
71       <div class="submit-btn hidden-xs hidden-sm pull-left ml-10">
72         <a href=<?php echo get_page_link($xpanel['submit-link']) ?>" class="header->
73       </div>
74       <div class="user-info pull-right mr-10">
75         <?php
76           if ( is_user_logged_in() ) {
```



CC-by-SA 3.0

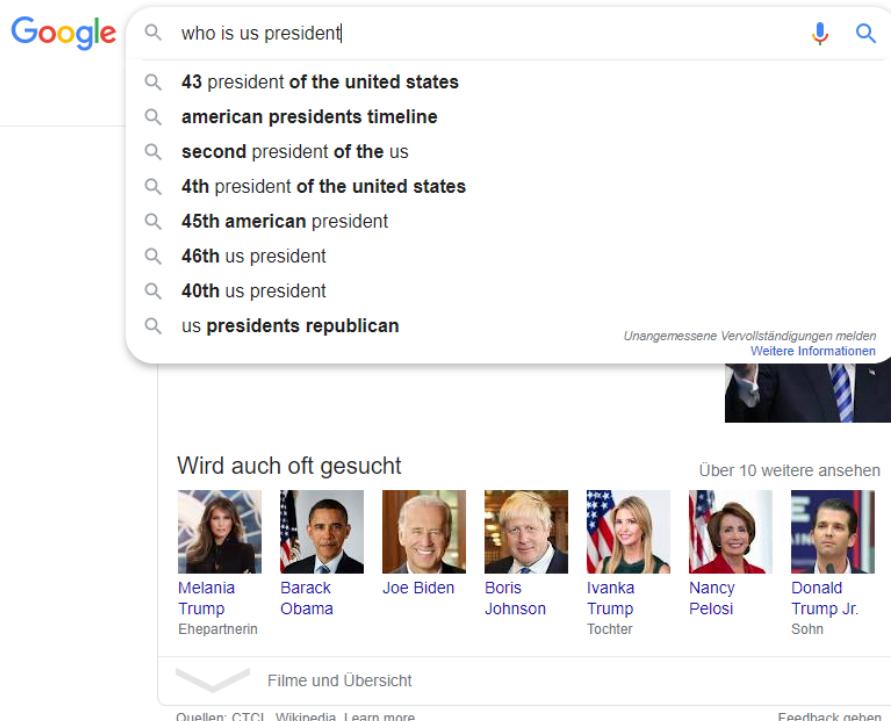
[https://commons.wikimedia.org/wiki/File:Mainz\\_Gutenberg\\_denkmal\\_Relief\\_2.JPG](https://commons.wikimedia.org/wiki/File:Mainz_Gutenberg_denkmal_Relief_2.JPG)

# Application of NLP?

- Where do you use NLP on a daily basis?
- What are typical tasks?

# Application of NLP?

- Where do you use NLP on a daily basis?
- What are typical tasks?



Google search results for "who is us president":

- Who is us president!
- 43 president of the united states
- american presidents timeline
- second president of the us
- 4th president of the united states
- 45th american president
- 46th us president
- 40th us president
- us presidents republican

Wird auch oft gesucht:

- Melania Trump Ehepartnerin
- Barack Obama
- Joe Biden
- Boris Johnson
- Ivanka Trump Tochter
- Nancy Pelosi
- Donald Trump Jr. Sohn

Über 10 weitere ansehen

Unangemessene Vervollständigungen melden Weitere Informationen

Feedback geben

Quellen: CTCL, Wikipedia. Learn more



Nutzer fragen auch:

- Who is the 52 president?
- Who is the richest president?
- Is the President of the United States a federal employee?
- Is a former president still called President?

Feedback geben

List of presidents of the United States - Wikipedia 

[https://en.wikipedia.org/wiki/List\\_of\\_presidents\\_of\\_the\\_United\\_States](https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States)

Diese Seite übersetzen

The president of the United States is the head of state and head of government of the United States; the 45th and current president is Donald Trump. ... Of those who have served as the nation's president, four died in office of...  
List of presidents of the United States · List of vice presidents · Disambiguation

President of the United States - Wikipedia 

[https://en.wikipedia.org/wiki/President\\_of\\_the\\_United\\_States](https://en.wikipedia.org/wiki/President_of_the_United_States)

Diese Seite übersetzen

Donald Trump is the 45th and current president of the United States.

Member of: Cabinet; Domestic Policy Council; ... Term length: Four years, renewable once  
Salary: \$400,000 annually Appointer: Electoral College

# Application of NLP?

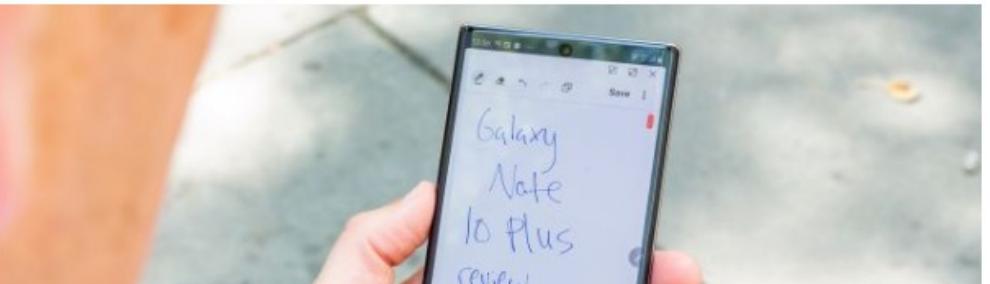
- Where do you use NLP on a daily basis?
- What are typical tasks?

**How to Transcribe and Export Written Notes on the Galaxy Note 10**

By Adam Ismail September 01, 2019 Phones 

Convert your handwriting to text instantly

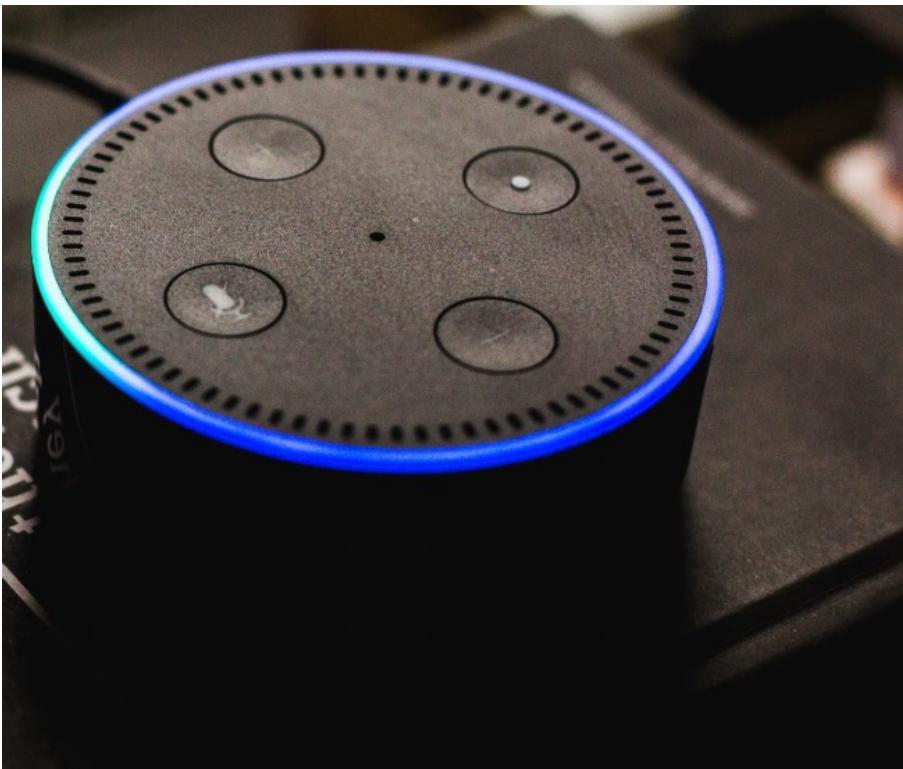




<https://www.tomsguide.com/how-to/how-to-transcribe-and-export-written-notes-on-the-galaxy-note-10>

# Application of NLP?

- Where do you use NLP on a daily basis?
- What are typical tasks?



# Overview - NLP

- Applications
- **Definitions**
- Terms, basic concepts and algorithms
- Milestones in the history of NLP
- Text Analytics

# Artificial languages and Natural Languages

- Vocabulary as set of words
- Text a sequence of words
- Language all “valid” texts
- Syntax
  - The rules and structure that govern how words and symbols are arranged
- Semantic
  - Meaning conveyed by words, phrases, and sentences
- Pragmatics
  - Involvement of the Context

# Natural Language Processing (NLP)

## A Definition

“Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.”

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

# Natural Language Processing (NLP)

## A Definition

“Natural Language Processing is a theoretically motivated range of **computational techniques** for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.”

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

# Natural Language Processing (NLP)

## A Definition

“Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of **achieving human-like language processing** for a range of tasks or applications.”

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

# Natural Language Processing (NLP)

## A Definition

“Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a **range of tasks or applications.**”

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.



IBM's Watson on Jeopardy!

<https://www.youtube.com/watch?v=Sp4q60BsHoY>

# Challenges in NLP Examples

- Paraphrase an input text
- Translate text from one language into another
- Answer questions about the contents of the text (or corpus)
- Draw inferences and conclusions from the text (or corpus)

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

# Typical algorithms for

- High-level Tasks
  - Responding to phrases (e.g., chatbot)
  - Discover topics and concepts a text describes
  - Summarizing texts (to different lengths)
  - Extracting relevant keywords from texts
  - Identify the sentiment of a text or phrase
  - ...
- Low-level Tasks
  - Break up texts in tokens
  - Reduce words to the root/stem
  - ...

duckduckgo.com/?q=lmu+münchen&t=h\_&ia=web

lmu münchen

Web Bilder Videos Nachrichten Karten Einstellungen

Deutschland Sichere Suche: Moderat Irgendwann

Ludwig Maximilian University - Hotels. Best price guarantee. WERBUNG  
[www.booking.com/Munich/Hotels](http://www.booking.com/Munich/Hotels) | Werbung melden  
Hotels near Ludwig Maximilian University. Save up to 50% on your reservation.

**Luxury Hotels**  
Easy and Secure Online Booking.  
Read Real Reviews and Book Now!

**Budget Hotels**  
New deals listed every day! Easy and Secure Booking

**Book for Tomorrow**  
Easy, fast and secure booking! New deals listed every day

**Book for Tonight**  
Your booking instantly confirmed!  
Around-the-clock customer service

**LMU München**  
<https://www.uni-muenchen.de>  
Musikalisches Engagement an der LMU. Freude, schöner Götterfunken Ob Chöre, Ensembles, Orchester - für LMU-Studierende gibt es unzählige Möglichkeiten, sich während der Studienzeit musikalisch zu engagieren. Neben dem Spaß am Musizieren ergeben sich neue Freundschaften oder Konzertreisen rund um den Globus.

**Studien- und Lehrangebot - LMU München**  
<https://www.uni-muenchen.de/studium/studienangebot/index.html>  
Über die LMU Einrichtungen Studium Studien- und Lehrangebot Studienangebote ...

**Biomedizinisches Centrum München - LMU München**  
<https://www.bmc.med.uni-muenchen.de/index.html>  
Das neue Biomedizinische Centrum München (BMC) ist einer der deutschlandweit größten Forschungsbauten der letzten Jahre - mit Laboren für derzeit etwa 60 Forschergruppen und insgesamt ca. 450 Mitarbeiter. In der Strategie der LMU, Wissenschaft und Klinik eng zu verzahnen, nimmt das BMC einen zentralen Platz ein. Mit seinem Profil und ...

**Zentrum Seniorenstudium - LMU München**  
<https://www.seniorenstudium.uni-muenchen.de/index.html>  
Die Ludwig-Maximilians-Universität bietet akademisch Interessierten ein umfangreiches Studienangebot aus allen Fakultäten. Es kommt den Wünschen nach ...

Institut für Musikwissenschaft - LMU München

<https://www.musikwissenschaft.uni-muenchen.de/index.html>  
Von Montag, 28.10.2019, bis Donnerstag, 31.10.2019, können interessierte Schülerinnen und Schüler in das Fach Musikwissenschaft an der Uni hineinschnuppern.

## Herzlich Willkommen am Brustzentrum der LMU München!

[www.klinikum.uni-muenchen.de/Brustzentrum/de/index.html](http://www.klinikum.uni-muenchen.de/Brustzentrum/de/index.html)  
Brustzentrum der LMU München Brustkrebs zählt zu den häufigsten bösartigen Erkrankungen der Frau. Nur durch eine sichere Diagnosestellung gefolgt von einer von Anfang an unter Berücksichtigung der Tumobiologie interdisziplinär geplanten und qualitativ hochwertigen Therapie können heutzutage die besten Überlebenschancen erreicht werden.

## Klassische Archäologie - LMU München

<https://www.klass-archaeologie.uni-muenchen.de/index.html>  
11.11.2019 China und Rom. Archäologisches zu einer antiken Fernhandelsbeziehung  
Vortrag von Prof. Dr. Lorenz E. Baumer ...

## Ethikkommission - Medizinische Fakultät - LMU München

<https://www.med.uni-muenchen.de/ethik/index.html>  
Vor der Durchführung eines biomedizinischen Forschungsvorhabens oder einer klinischen Prüfung am Menschen hat sich jeder Forscher, der Mitglied der LMU ist, durch die Ethikkommission bei der Med. Fakultät der LMU beraten zu lassen und ein zustimmendes Votum einzuholen.

## Alphabetische Liste aller Personen - Juristische Fakultät - LMU Mün...

<https://www.jura.uni-muenchen.de/personen/index.html>  
Hinweise zur Datenübertragung bei der Google™ Suche. Links und Funktionen.  
www.lmu.de; LMU-Portal; Personen ... Sprachumschaltung. English

Ungefähr 5.880.000 Ergebnisse (0,47 Sekunden)

## LMU München

<https://www.uni-muenchen.de> ▾

Die LMU ist eine der renommiertesten und traditionsreichsten Universitäten Europas. Sie verbindet hervorragende Forschung mit einem anspruchsvollen ...

Ergebnisse von uni-muenchen.de



### Studien- und Lehrangebot

Studiengänge und -  
Studienangebote - Sprachkurse

### Studienangebote

Bitte beachten Sie: Die auf diesen  
Seiten eingestellten ...

### Fakultäten

Medizinische Fakultät - Fakultät für  
Sprach - Juristische Fakultät - ...

### Studiengänge und -

Studiengänge von A bis Z. Studiengänge ...

### Studium

Studieninteressierte - Studium AZ -  
Hochschulzugang - Studierende

### Einrichtungen

Fakultäten - Medizinische  
Fakultät - Bibliotheken - ...

## Ludwig-Maximilians-Universität München – Wikipedia

[https://de.wikipedia.org/wiki/Ludwig-Maximilians-Universität\\_München](https://de.wikipedia.org/wiki/Ludwig-Maximilians-Universität_München) ▾

Die Ludwig-Maximilians-Universität München (kurz Universität München oder LMU) ist eine Universität in der bayerischen Landeshauptstadt München.

**Gründung:** 1472 in Ingolstadt, seit 1826 in Mü... **Präsident:** Bernd Huber

**Land:** Deutschland

Davon Professoren: 762 (2017)

## Schlagzeilen



Oliver Welke an der LMU - "Du darfst nichts Halbgares servieren, auch in der Satire nicht"

Süddeutsche · vor 1 Tag



SZ-Veranstaltung - Was kann Satire? Ein Gespräch mit Oliver Welke

[→ Mehr zu LMU münchen](#)

## Ludwig-Maximilians-Universität München - Home | Facebook

[https://www.facebook.com/Places/Munich\\_Germany](https://www.facebook.com/Places/Munich_Germany)

★★★★★ Bewertung: 4,4 - 589 Abstimmungsergebnisse

Sie begleiten "Herr der Ringe" live im Gasteig und standen schon unter der Leitung von Ennio Morricone in der Olympiahalle auf der Bühne - der Münchner ...

## Department Psychologie - Fakultät für Psychologie und ...

[https://www.fak11.lmu.de/dep\\_psychologie](https://www.fak11.lmu.de/dep_psychologie)

... der Psychologie, der Pädagogik/Bildungswissenschaft und des Lehramts bei und ist an übergreifenden Programmen wie der "LMU excellent Graduate School" ...

## Zentrale Lernplattform • LMU München

<https://moodle.lmu.de>

Zentrale Lernplattform der Ludwig-Maximilians-Universität München.

## Ludwig-Maximilians-Universität München - Das offizielle ...

<https://www.muenchen.de/sehenswuerdigkeiten/orte>

★★★★★ Bewertung: 5 - 6 Abstimmungsergebnisse

München hat nicht nur die zweitgrößte Universität Deutschlands, sondern auch eine der schönsten: Das beeindruckende Hauptgebäude der ...

## Ähnliche Suchanfragen zu LMU münchen

[lmu münchen klinikum](#)[lmu münchen adresse](#)[ludwig maximilians universität münchen namhafte absolventen](#)[lmu portal](#)[lsf lmu](#)[lmu münchen online portal](#)[lmu münchen jura](#)[lmu münchen stellenangebote](#)

bing.com/search?q=lmu+münchen&qs=n&form=QBLH&sp=-1&pq=lmu+mü&sc=8-6&sk=&cvid

Alle Bilder Videos Karten News Shopping | Meine gespeicherten I

1.370.000 Ergebnisse Datum Sprache Region

**LMU München**  
<https://www.uni-muenchen.de> ▾  
Musikalisches Engagement an der LMU. Freude, schöner Götterfunken Ob Chöre, Ensembles, Orchester – für LMU-Studierende gibt es unzählige Möglichkeiten, sich während der Studienzeit musikalisch zu engagieren. Neben dem Spaß am Musizieren ergeben sich ...

**Studien- und Lehrangebot**  
Ringvorlesung LMU Studium Generale  
Seniorenstudium Frauenstudien und Gender ...

**Studium**  
Von Ägyptologie bis Zahnmedizin bietet die LMU München fast 200 Studiengänge mit ...

**LSF**  
Hier sollte eine Beschreibung angezeigt werden, diese Seite lässt dies jedoch nicht zu.

**Einrichtungen**  
Organisation der LMU Hochschulleitung  
Organigramm der LMU Gremien Beauftragte ...

**Stellenangebote**  
Über die LMU Einrichtungen Studium  
Forschung Kooperationen Weiterbildung ...

Ergebnisse von uni-muenchen.de suchen

## Lokale Ergebnisse für lmü münchen

Bing Lokale Suche

Geschwister-Scholl-Platz 1 · 80539 München · 089 21800  
Routenplaner · Details ·  67 TripAdvisor Bewertungen

In Partnerschaft mit **Das Örtliche**

**en.uni-muenchen.de - LMU Munich** Diese Seite übersetzen  
<https://www.en.uni-muenchen.de/index.html> ▾

LMU Research Fellowship. Getting to Grips with Clouds Postdoc Linda Forster is an Outgoing LMU Research Fellow and is now engaged on a project on cloud formation at the California Institute of Technology. The results should help to improve forecasts of future climate change.

**Biomedizinisches Centrum München - LMU München**  
<https://www.bmc.med.uni-muenchen.de/index.html> ▾  
Das neue Biomedizinische Centrum München (BMC) ist einer der deutschlandweit größten Forschungsbauten der letzten Jahre – mit Laboren für derzeit etwa 60 Forschergruppen und insgesamt ca. 450 Mitarbeiter. In der Strategie der LMU, Wissenschaft und Klinik eng zu verzahnen, nimmt das BMC einen zentralen Platz ein. Mit seinem Profil und ...

**Qualitative Sozialforschung - LMU München**  
<https://www.qualitative-sozialforschung.soziologie.uni-muenchen.de/index.html> ▾  
Lehr- und Forschungsbereich für Qualitative Methoden der empirischen Sozialforschung, Prof. Dr. Hella von Unger - Institut für Soziologie - LMU München

**Zentrum Seniorenstudium - LMU München**  
<https://www.seniorenstudium.uni-muenchen.de/index.html> ▾  
Die Ludwig-Maximilians-Universität bietet akademisch Interessierten ein umfangreiches Studienangebot aus allen Fakultäten. Es kommt den Wünschen nach wissenschaftlicher Information, geistiger Orientierung und Zusammenführung der Generationen entgegen und möchte einen Beitrag zur sinnvollen Gestaltung des Lebens nach der Zeit aktiver ...

## News über Lmu München

bing.com/news



München: Oliver Welke spricht an der LMU über Satire

Süddeutsche Zeit... · 22 Std.



Nach Klage gegen LMU Geld aus Schließfach verschwunden: Bayern ...

Abendzeitung · 4 T.



Hassanrufe? - "Ja, von meiner Frau!"

Mittelbayerische · 21 Std.

Weitere Nachrichten zu lmü münchen anzeigen

## Herzlich Willkommen am Brustzentrum der LMU München!

[www.klinikum.uni-muenchen.de/Brustzentrum](http://www.klinikum.uni-muenchen.de/Brustzentrum) ▾

Als zertifiziertes universitäres Brustzentrum in einem Tumorzentrum der Spitzenklasse (Comprehensive Cancer Center CCC München) bieten wir unter Leitung von Frau Prof. Harbeck an zwei Standorten in München (Frauenkliniken Maistrasse-Innenstadt und Großhadern) alle Bestandteile einer modernen Brustkrebstherapie unter einem Dach

# Discussion: Naive implementation of search

- Given
  - 10000 text documents with an average length of 1000 words
  - A search term of up to 4 words
- Aim: the 10 text documents that best match the query
- Approach?
- Where can Machine Learning and AI help?

## Scenario:

You have stored all articles of the New York Times from 2010 to 2020

Someone asks for the article about “Merkel visiting the US”

Which articles do you give back?

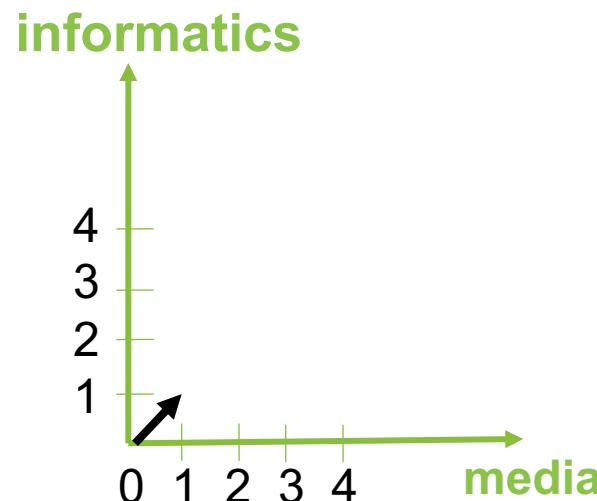
# Python Example

## Counting word occurrences

Query = “media informatics”

- Transferring the query into a coordinate
  - One cell / dimension for each word
  - Count occurrences

media    informatics



Example based on: <https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55>

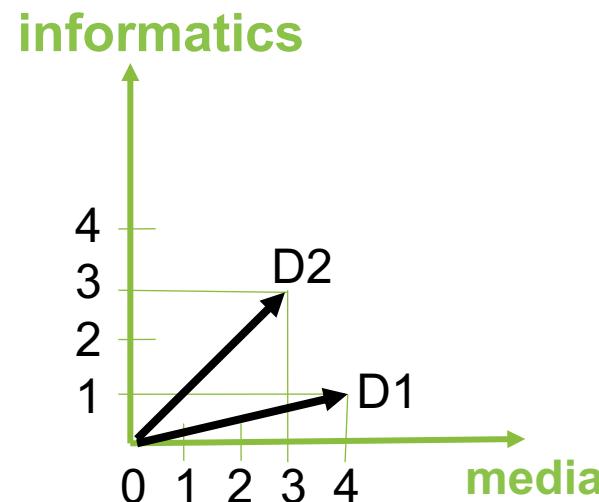
# Python Example

## Counting word occurrences

Query = “media informatics”

- Transferring documents into coordinates
  - D1: media informatics media media media
  - D2: media informatics media informatics media informatics
  - Count occurrences

	media	informatics
D1	4	1
D2	3	3



Example based on: <https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55>

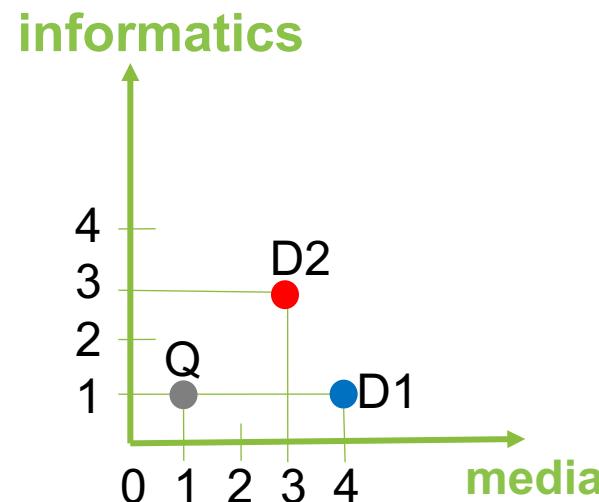
# Python Example

## Counting word occurrences

Query = “media informatics”

- Transferring documents into coordinates
  - D1: media informatics media media media
  - D2: media informatics media informatics media informatics
  - Count occurrences

	media	informatics
D1	4	1
D2	3	3



Example based on: <https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55>

# Python Example

## Counting word occurrences

```
d1 = "media ... informatics media ... media ... media"  
d2 = "media informatics ... media ... informatics ... media informatics"  
...  
vector = [0, 0]  
for word in d1.split(" "):  
    if word=="media":  
        vector[0] = vector[0] + 1  
    if word=="informatics":  
        vector[1] = vector[1] + 1  
print (vector)
```

Example based on: <https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55>

# Overview - NLP

- Applications
- Definitions
- **Terms, basic concepts and algorithms**
- Milestones in the history of NLP
- Text Analytics

# Tokenization

- Separating a text into individual words
- Words are called tokens
- Removing punctuation, (multiple) spaces, separators
- Approach:
  - Search along the text and extract tokens separated by space and punctuation
  - Store all tokens in a list
- Any difficulties?

# Tokenization

- Separating a text into individual words
- Words are called tokens
- Removing punctuation, (multiple) spaces, separators
- Approach:
  - Search along the text and extract tokens separated by space and punctuation
  - Store all tokens in a list
- Any difficulties?  
*Dr. Max von Mayer-Hauser is today in New York and we will meet him-hopefully.*

# Stop Words Removal

- Remove „small“ words in the text, such as articles, pronouns, and prepositions
- Examples
  - English: the, and, a, to, ...
  - German: der, die, das, und, es, ...
- Approaches include stop word lists or stop word learning (based on frequency)
- Discussion: Approach based on ML/AI?

# Text Normalization

- **Aim:** match the “same” words
- Syntactic matching
- What is the problem? How to do this?

# Text Normalization

- Aim: match the “same” words
- Upper case / lower case letters (e.g., all lower case)
  - Tricky if upper case is required to detect names, grammar
- Acronyms (U.K. → UK)
- Expanding contractions ("don't" → "do not").
- Umlauts (für → fuer or fuer → für)
- Dealing with numbers and symbols in text (“three” → “3”)
- Correcting misspelling
- Normalizing word forms (stemming, lemmatization)

# Stemming

- Reducing the word to its stem
- Extract the morphological root
- Removing affixes and suffixes
- Heuristic process (works most of the time)
  
- Approaches and algorithms
  - Set of rules (language dependent)
  - E.g., as automaton
- Typical algorithms: Porter Stemmer, Snowball Stemmer
- Example:
  - player, playing, playful, plays, played → play
  - newer, newest → new ... what happens to news

# Stemming

## Example: Porter Stemmer

- Set of rules for removing/changing suffixes
- Rules are grouped
- Rules for their application of the rules

```
from nltk.stem.porter import PorterStemmer
```

In a set of rules written beneath each other, only one is obeyed, and this will be the one with the longest matching S1 for the given word. For example, with:

SSES → SS  
IES → I  
SS → SS  
S →

(here the conditions are all null) CARESSES maps to CARESS since SSES is the longest match for S1. Equally CARESS maps to CARESS (S1 = 'SS') and CARES to CARE (S1 = 'S').

In the rules below, examples of their application, successful or otherwise, are given on the right in lower case. The algorithm now follows:

Step 1a

SSES → SS	caresses	→ caress
IES → I	ponies	→ poni
SS → SS	ties	→ ti
S →	caress	→ caress
	cats	→ cat

Step 1b

(m > 0) EED → EE	feed	→ feed
(*v*) ED →	agreed	→ agree
(*v*) ING →	plastered	→ plaster
	bled	→ bled
	motoring	→ motor
	sing	→ sing

If the second or third of the rules in Step 1b is successful, the following is done:

AT → ATE	conflat(ed)	→ conflate
BL → BLE	troubl(ing)	→ trouble
IZ → IZE	siz(ed)	→ size
(*d and not (*L or *S or *Z)) → single letter	hopp(ing)	→ hop
	tann(ed)	→ tan
	fall(ing)	→ fall
	hiss(ing)	→ hiss
	fizz(ed)	→ fizz
(m = 1 and *o) → E	fail(ing)	→ fail
	fil(ing)	→ file

Porter, Martin F (1980) "An algorithm for suffix stripping." *Program*, 14(3)

Reprinted 2006: <https://cl.lingfil.uu.se/~marie/undervisning/textanalys16/porter.pdf>

# Lemmatization

- Return the base (dictionary) form of a word
- Uses linguistic knowledge (vocabulary, grammar, morphological analysis)
- “*Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called **Lemma**. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.*”  
<https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>
- Examples: was→be, running→run, has→have, swims→swim, caring→care, ...

```
from nltk.stem import WordNetLemmatizer
```

# Stemming/Lemmatization based on ML?

- How to implement this not rule-based? Aka, with ML?

# Part-Of-Speech Tagging

- Determining the type of a word in the context of a sentence
- Identifying words with examples according to Penn Treebank POS tag set [1]:
  - NN - Noun, singular
  - VB - Verb, base form
  - JJ - Adjective
  - DT - Determiner

```
from nltk import pos_tag, word_tokenize
text = "The quick brown fox jumps over the lazy dog."
print(pos_tag(word_tokenize(text)))
[('The', 'DT'), ('quick', 'JJ'), ('brown', 'JJ'), ('fox', 'NN'),
 ('jumps', 'VBZ'), ('over', 'IN'), ('the', 'DT'),
 ('lazy', 'JJ'), ('dog', 'NN')]
```

[1] [https://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

# Named-Entity Disambiguation and Entity linking

- Determining the meaning of a word, if word have more meanings
- Using context and knowledge
- Example

*I did not use an apple to make these slides.*

vs.

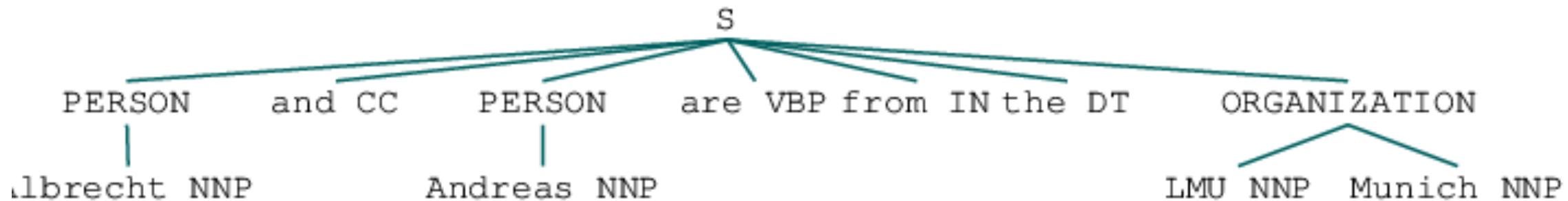
*I really like to bake apple crumble.*

# Named-Entity Recognition

- Gets entities from unstructured texts
  - Assigning entities/words to categories
  - Examples of named entities: people, places, companies, organizations, industries, products, product categories, time, location, brands, etc.
  - Application and domain specific, e.g. abbreviations for trading stocks, medical conditions, addresses
- 
- Library: <https://spacy.io/api/annotation#section-named-entities>
  - Example in Python: <https://nlpforhackers.io/named-entity-extraction/>
  - Software: Stanford Named Entity Recognizer (NER)  
<https://nlp.stanford.edu/software/CRF-NER.shtml>

# Named-Entity Disambiguation and Entity linking

```
text = "Albrecht and Andreas are from the LMU Munich."  
pos_tags = pos_tag(word_tokenize((text))  
named_entities = ne_chunk(pos_tags)  
IPython.core.display.display(named_entitled)
```



# Bag of Words

- Texts are considered a set of words
- Simplified representation
- Ignores grammar (and generally word order)
- Typical calculation:
  - Frequency of occurrence of words
  - Frequency of n-grams (preserves some word order)
- “My cat likes to sleep. I sleep a lot. Do you have a cat?”  
cat 3x, sleep 2x, a 2x, ...

# Corpus

## Definition

“A corpus is a large body of natural language text used for accumulating statistics on natural language text. The plural is corpora. Corpora often include extra information such as a tag for each word indicating its part-of-speech, and perhaps the parse tree for each sentence.”

NLP Dictionary <http://www.cse.unsw.edu.au/~billw/nlpdict.html>

# Corpus, Corpora

- Monolingual corpora
  - data from one single language
- Parallel corpora
  - original texts in one language
  - translations in other languages
- Examples:
  - Gutenberg, archive of free electronic books, <https://www.gutenberg.org/> or <https://www.projekt-gutenberg.org/>
  - <https://www.collinsdictionary.com/api/collins-english-dictionary,61,HCA.html>
  - Swiss SMS Corpus, <https://sms.linguistik.uzh.ch/>
  - The National University of Singapore SMS Corpus <https://www.kaggle.com/rtatman/the-national-university-of-singapore-sms-corpus>

# What's next?

Transformers



Attention based



Bi-directional LSTM



LSTM based models



RNN based models



Words Embedding's



Bag of words

This file is licensed under the Creative Commons Attribution-Share Alike 4.0 (CC BY-SA) license:

<https://creativecommons.org/licenses/by-sa/4.0>

Attribution: Sven Mayer and Albrecht Schmidt

For more content see: <https://iui-lecture.org>

