Natural Language Processing and Bots

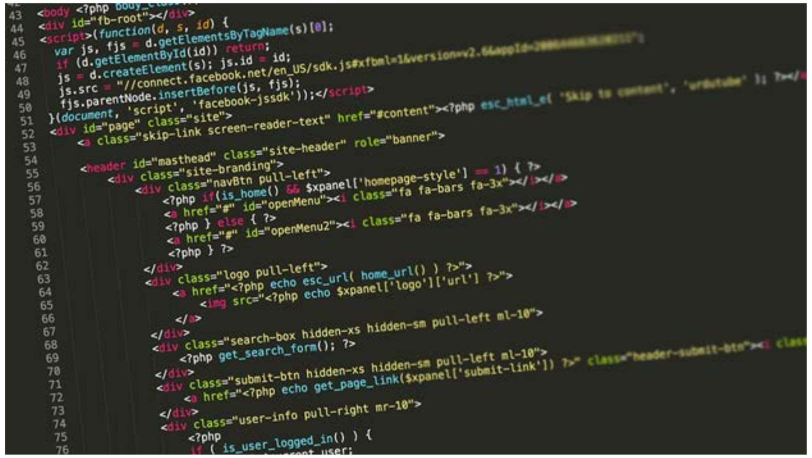
Natural Language Processing

Overview - NLP

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

Artificial languages and Natural Languages

- Syntax and semantic
- Vocabulary as set of words
- Text a sequence of words
- Language all “valid” texts



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

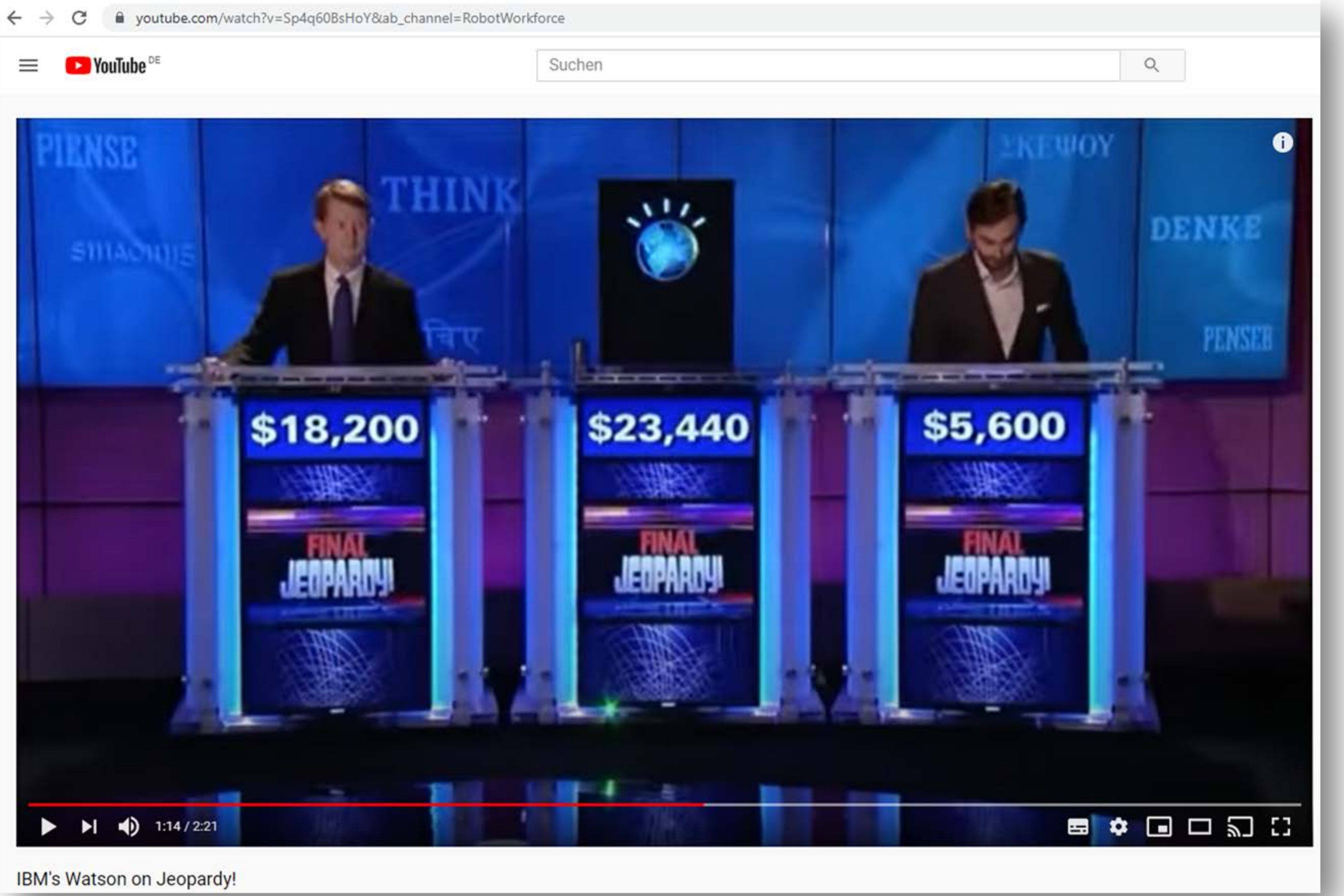
"Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of **achieving human-like language processing** for a range of tasks or applications."

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.



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

- Responding to phrases (e.g. chat bot)
- Discover topics and concepts a text describes
- Summarizing texts (to different length)
- Extracting relevant keywords from texts
- Identify the sentiment of a text or phrase
- …
- …
- Break up texts in tokens
- Reduce words to the root / stem

https://algorithmia.com/blog/introduction-natural-language-processing-nlp duckduckgo.com/?q=lmu+münchen&t=h_&ia=web

lmu münchen

Web Bilder Videos Nachrichten Karten Einstellungen

Deutschland Sichere Suche: Moderat Irgendwann

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
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 Tumorbiologie 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

Google

LMU münchen

Alle | Maps | Bilder | News | Videos | Mehr | Einstellungen | Tools

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
Studienfächer und - Studienangebote - Sprachkurse

Studienfächer und
Studienfächer und Studiengänge von A bis Z. Studiengänge ...

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

Studium
Studieninteressierte - Studium AZ - Hochschulzugang - Studierende

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

Einrichtungen
Fakultäten - Medizinische Fakultät - Bibliotheken - ...

Ludwig-Maximilians-Universität München – Wikipedia
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
★★★★★ 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
... 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

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 ...

LMU Intern
Die internen Webangebote der LMU sind vielfältig – Portale, Lern-Management ...

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

Online-Katalog
Funktionen | Katalog der UB | Aufsätze & mehr | Web Opac App und mobile Oberfläche. ...

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

Immatrikulation
Sobald die Zahlung der Semesterbeiträge auf dem Konto der LMU eingegangen ist (im ...

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

Aktuelles
Preis für LMU-Forscher Mehr Nachhaltigkeit im Unterricht! 30.10.2019. Weitere News ...

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

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

Ergebnisse von uni-muenchen.de suchen   Suche

Lokale Ergebnisse für lmu 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 lmu münchen anzeigen

Herzlich Willkommen am Brustzentrum der LMU München!
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

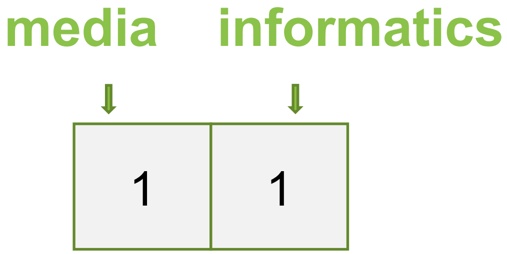
Someone asks for the article about “Merkel visiting the US”

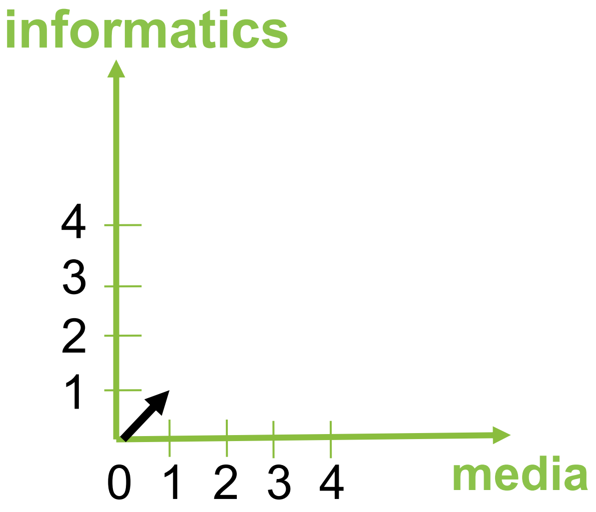
Which articles do you give back?

Python Example
Counting word occurrences

Query = “media informatics”

- Transfering the query into a coordinate
  - One cell / dimension for each word
  - Count occurances





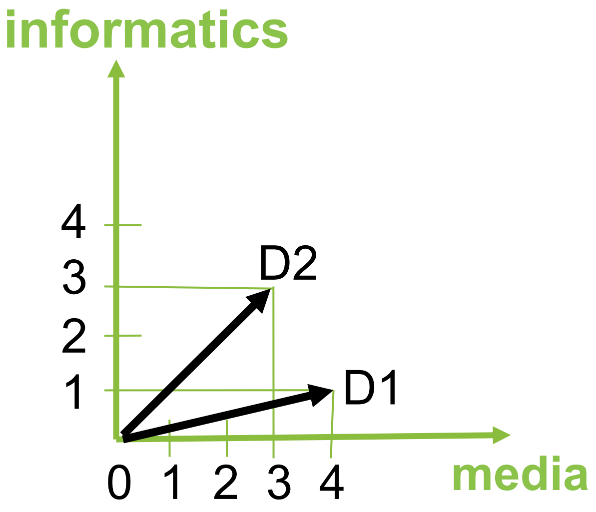
Example based on: https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55

Python Example
Counting word occurrences

Query = “media informatics”

- Transfering documents into coordinates
  - D1: media informatics media media media
  - D2: media informatics media informatics media informatics
  - Count occurances

	media	informatics
D1	4	1
D2	3	3



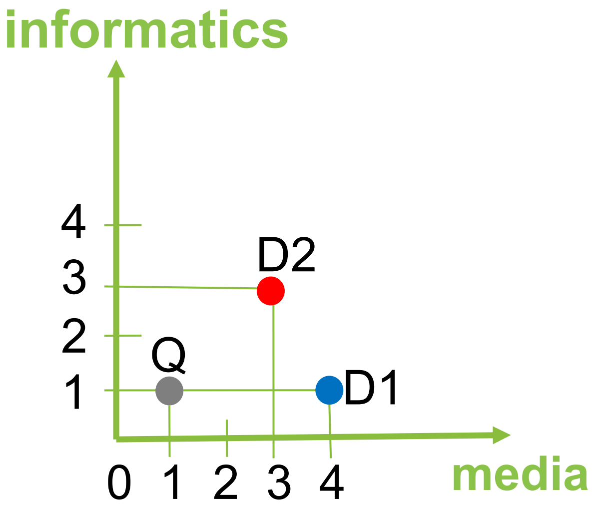
Example based on: https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/55

Python Example
Counting word occurrences

Query = “media informatics”

- Transfering documents into coordinates
  - D1: media informatics media media media
  - D2: media informatics media informatics media informatics
  - Count occurances

	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

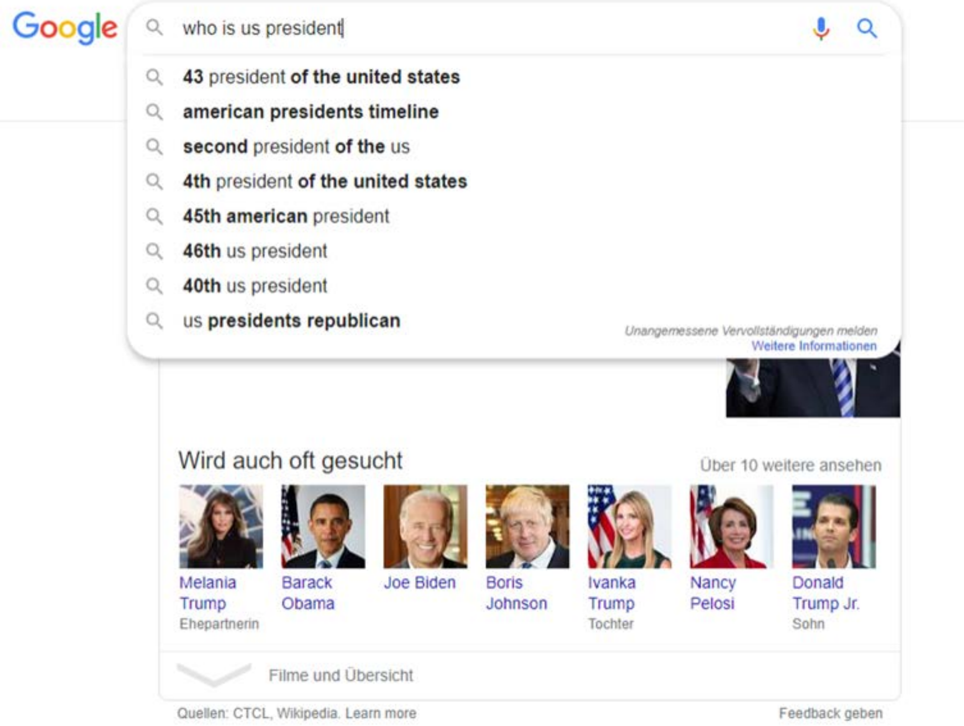
- Definition and Motivation
- **Applications**
- Terms, basic concepts and algorithms
- Milestones in the history of NLP
- Text Analytics

Application of NLP?

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

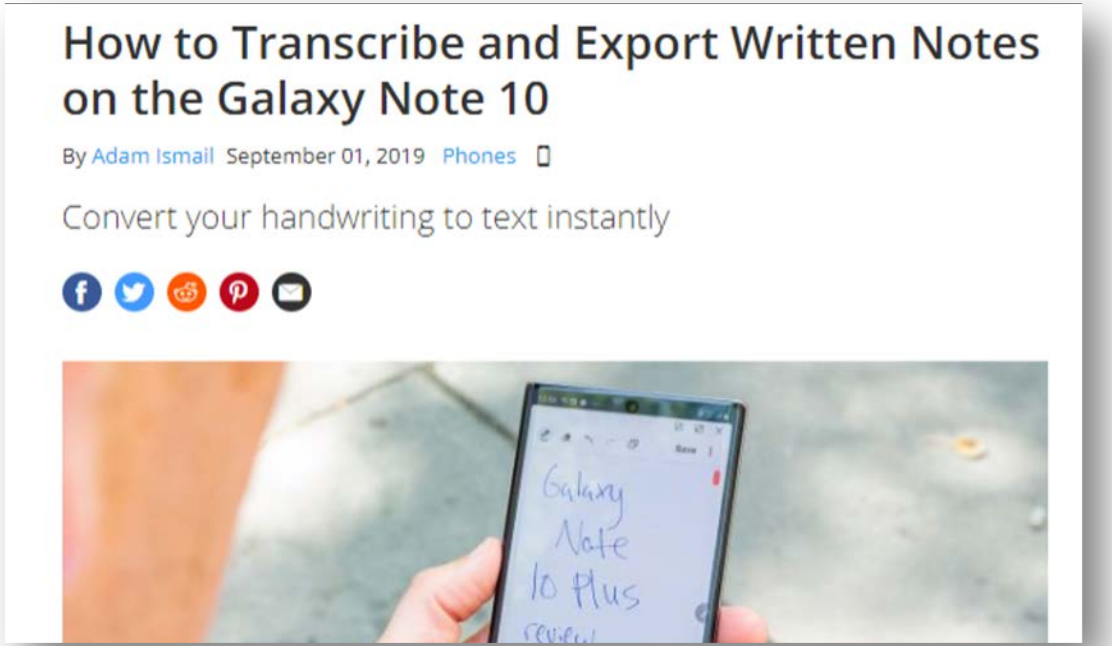
Application of NLP?

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



Application of NLP?

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



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

- Definition and Motivation
- Applications
- **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
- Normalizing word forms (stemming, lemmatization)
- Acronyms (U.K. □ UK)
- Umlauts (für □ fuer or fuer □ für)
- Dealing with numbers and symbols in text
- Correcting misspelling

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

Example
Porter Stemmer

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

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
	ties	→ ti
SS → SS	caress	→ caress
S →	cats	→ cat

Step 1b

(m > 0) EED → EE	feed	→ feed
	agreed	→ agree
(*v*) ED →	plastered	→ plaster
	bled	→ bled
(*v*) ING →	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, …

Stemming/Lemmatization based on ML?
How would you do this?

Part-Of-Speech Tagging

- Determining the type of a word in the context of a sentence
- Identifying words as nouns, verbs, adjectives, adverbs, …

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 crumple.

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

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, 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

Corpus

- "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
- Further Examples:
  - 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

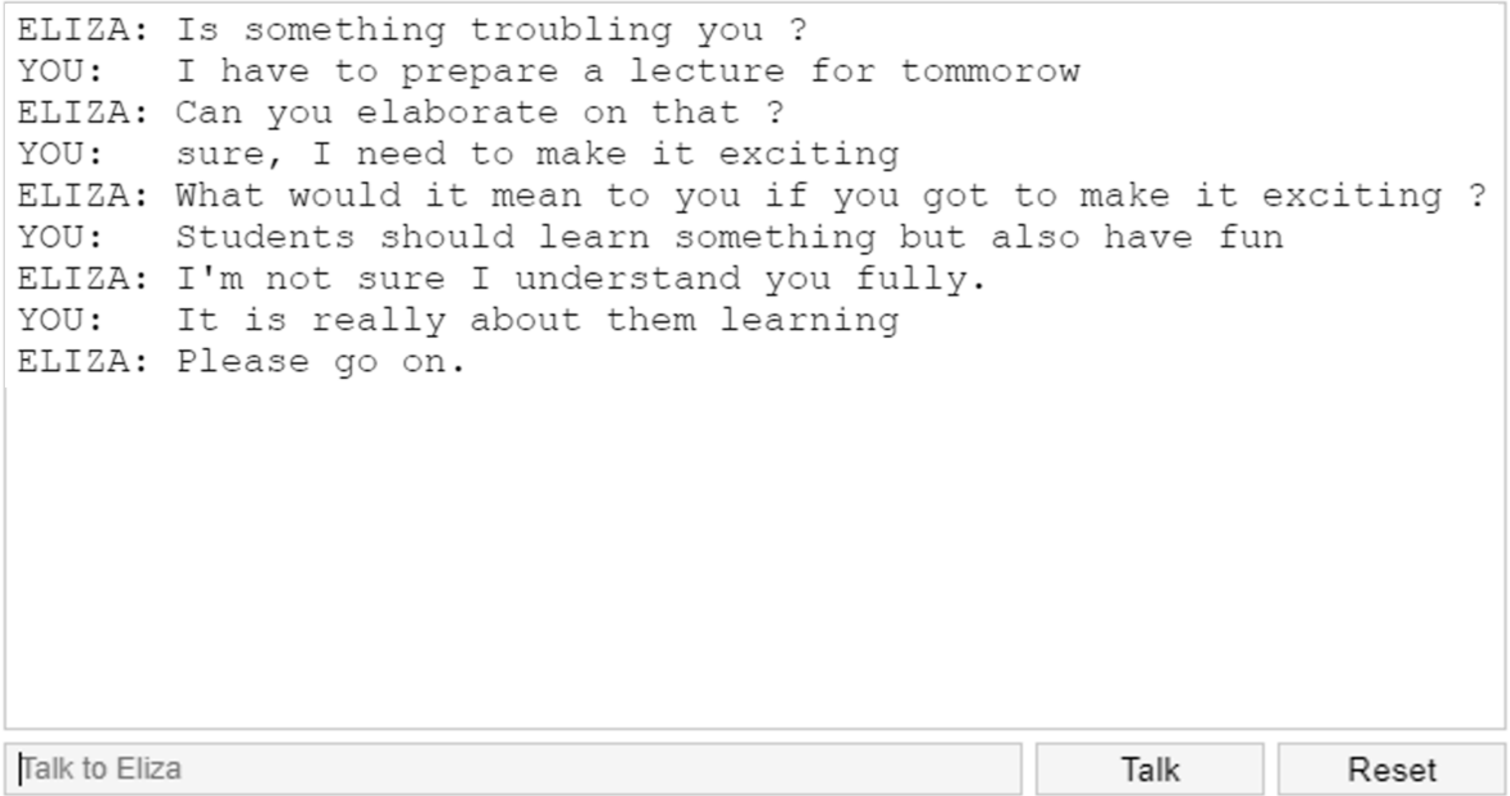
Live-Coding

Overview - NLP

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

ELIZA by Joseph Weizenbaum

https://katzlberger.ai/2018/08/31/eliza-der-erste-chatbot/



Weizenbaum, Joseph. "ELIZA---a computer program for the study of natural language communication between man and machine." Communications of the ACM 9.1 (1966): 36-45.
https://cse.buffalo.edu/~rapaport/572/S02/weizenbaum.eliza.1966.pdf

ELIZA by Joseph Weizenbaum (1966)

- “ELIZA is a program […] which makes certain kinds of natural language conversation between man and computer possible. Input sentences are analyzed on the basis of decomposition rules which are triggered by key words appearing in the input text. Responses are generated by reassembly rules associated with selected decomposition rules. The fundamental technical problems with which ELIZA is concerned are:
- the identification of key words,
- the discovery of minimal context,
- the choice of appropriate transformations,
- generation of responses in the absence of keywords, and
- the provision of an ending capacity for ELIZA "scripts".”

Weizenbaum, Joseph. "ELIZA---a computer program for the study of natural language communication between man and machine." Communications of the ACM 9.1 (1966): 36-45. https://doi.org/10.1145/365153.365168

Eliza

- Try Eliza out at: http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm
- Source code in Python: https://github.com/wadetb/eliza

Can be run interactively:

```
$ python eliza.py
How do you do.  Please tell me your problem.
> I would like to have a chat bot.
You say you would like to have a chat bot ?
> bye
Goodbye.  Thank you for talking to me.
```

...or imported and used as a library:

```python
import eliza

eliza = eliza.Eliza()
eliza.load('doctor.txt')

print(eliza.initial())
while True:
    said = input('> ')
    response = eliza.respond(said)
    if response is None:
        break
    print(response)
print(eliza.final())
```

Task:
How to implement a chatbot for Alexa?

https://www.masswerk.at/eliza/

NLP timeline
three different types of approaches

- Since 1950s (early days of NLP):
  - **Rule-based Approaches**
- Since 1980s (statistical approaches):
  - **Machine Learning Approaches**
- Since 2010s (advances in neural networks):
  - **Deep Learning Approaches**

https://livebook.manning.com/book/essential-natural-language-processing/chapter-1/15

Overview - NLP

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

Text Analytics

- Text is a key media:
  - in personal communication (e.g. texting, email)
  - in communication media (e.g. news, web pages, social media)
  - for knowledge sharing and acquisition (e.g. books, reports)

- Most user interfaces include texts
- Text reception (reading, understanding, or skimming) is often a key factor that defines the require time for a (knowledge work) task
- Big individual differences in text reception (e.g. reading speed, understanding)

Definitions of text analytics

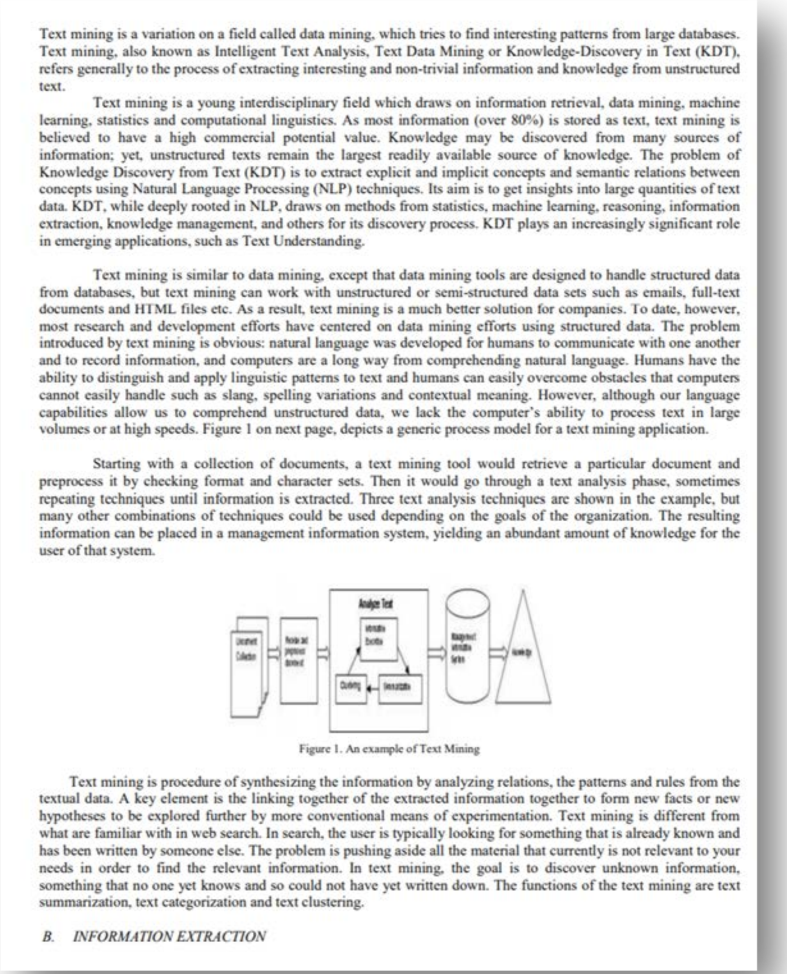
- Definition of “text data mining”: “as the application of algorithms and methods from the fields machine learning and statistics to texts with the goal of finding useful patterns” [1]
- *“Text mining* is the process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts.” [2]

[1] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." *Ldv Forum*. Vol. 20. No. 1. 2005.
[2] https://www.ibm.com/docs/en/wsd?topic=analytics-about-text-mining

Text analytics – Why and Where?

- Answering questions like
  - What is this text about?
  - What did the person communicate?
  - What is the key information in this document?
  - What feelings are communicated?
  - Who is saying something?
  - Is this different from what was said before?

- Application areas
  - Social media analytics, e.g. twitter
  - Communication and reading interfaces
  - Customer reviews and feedback
  - Chat bots
  - Forensics



Text mining is a variation on a field called data mining, which tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding.

Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Figure 1 on next page, depicts a generic process model for a text mining application.

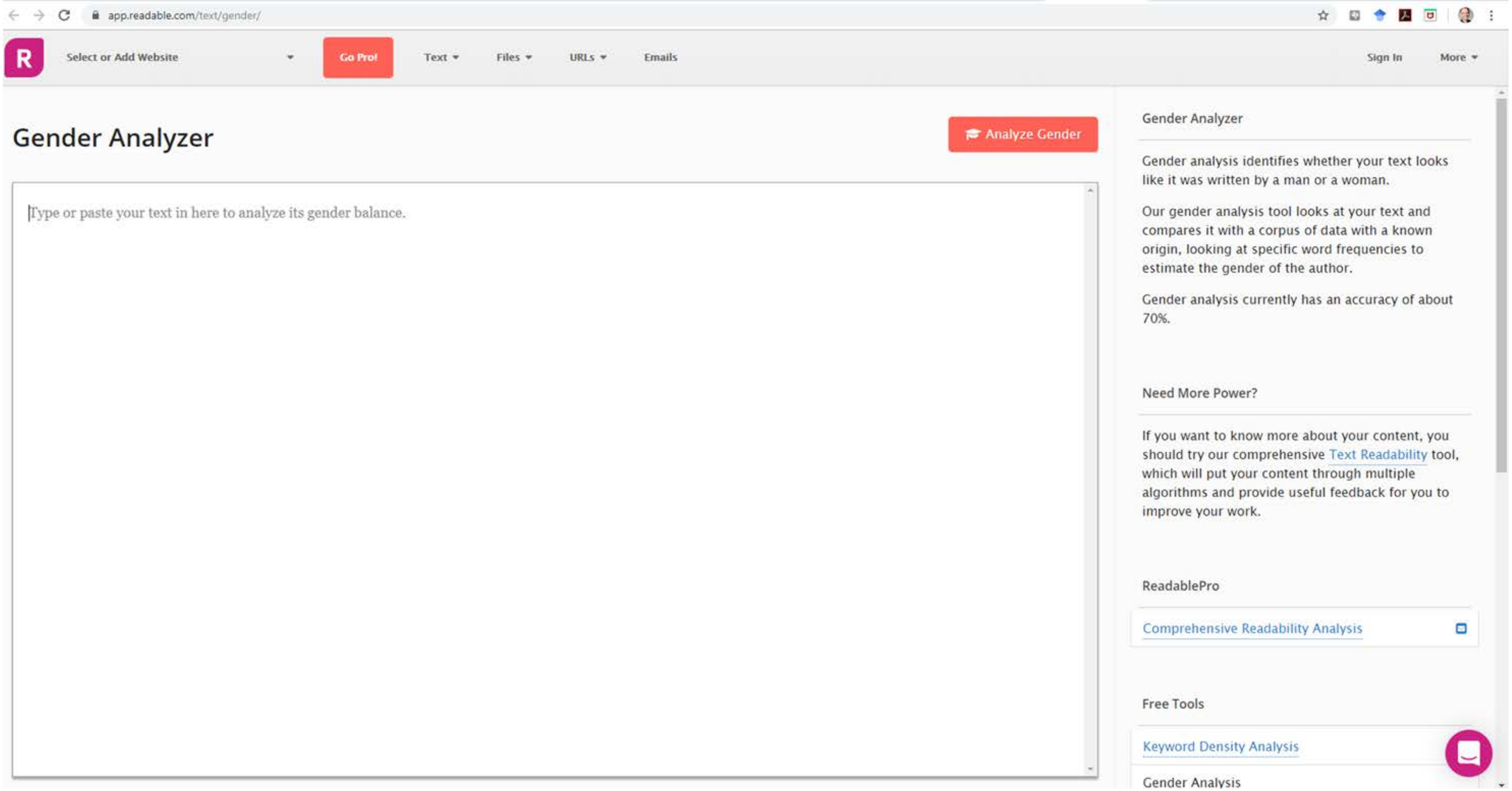
Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

Figure 1. An example of Text Mining

Text mining is procedure of synthesizing the information by analyzing relations, the patterns and rules from the textual data. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. The functions of the text mining are text summarization, text categorization and text clustering.

B. INFORMATION EXTRACTION

http://ijiet.com/wp-content/uploads/2015/04/17.pdf



https://app.readable.com/text/gender/

Gender Analyzer

nature

International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog | natu

comments on this story

Stories by subject

- Technology
- Brain and behaviour

Stories by keywords

- linguistics
- gender
- male
- female
- literature
- book
- computer
- program
- algorithm

This article elsewhere

Blogs linking to this article

Add to Digg
Add to Facebook
Add to Newsvine
Add to Del.icio.us
Add to Twitter

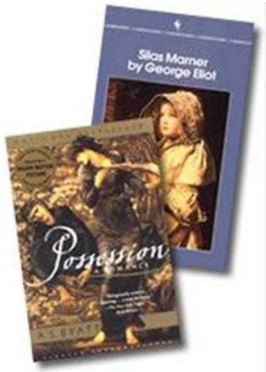
Published online 18 July 2003 | Nature | doi:10.1038/news030714-13

News

Computer program detects author gender

Simple algorithm suggests words and syntax bear sex and genre stamp.

Philip Ball



A.S Byatt confuses the computer; will it see through George Elliot?

A new computer program can tell whether a book was written by a man or a woman. The simple scan of key words and syntax is around 80% accurate on both fiction and non-fiction[1,2].

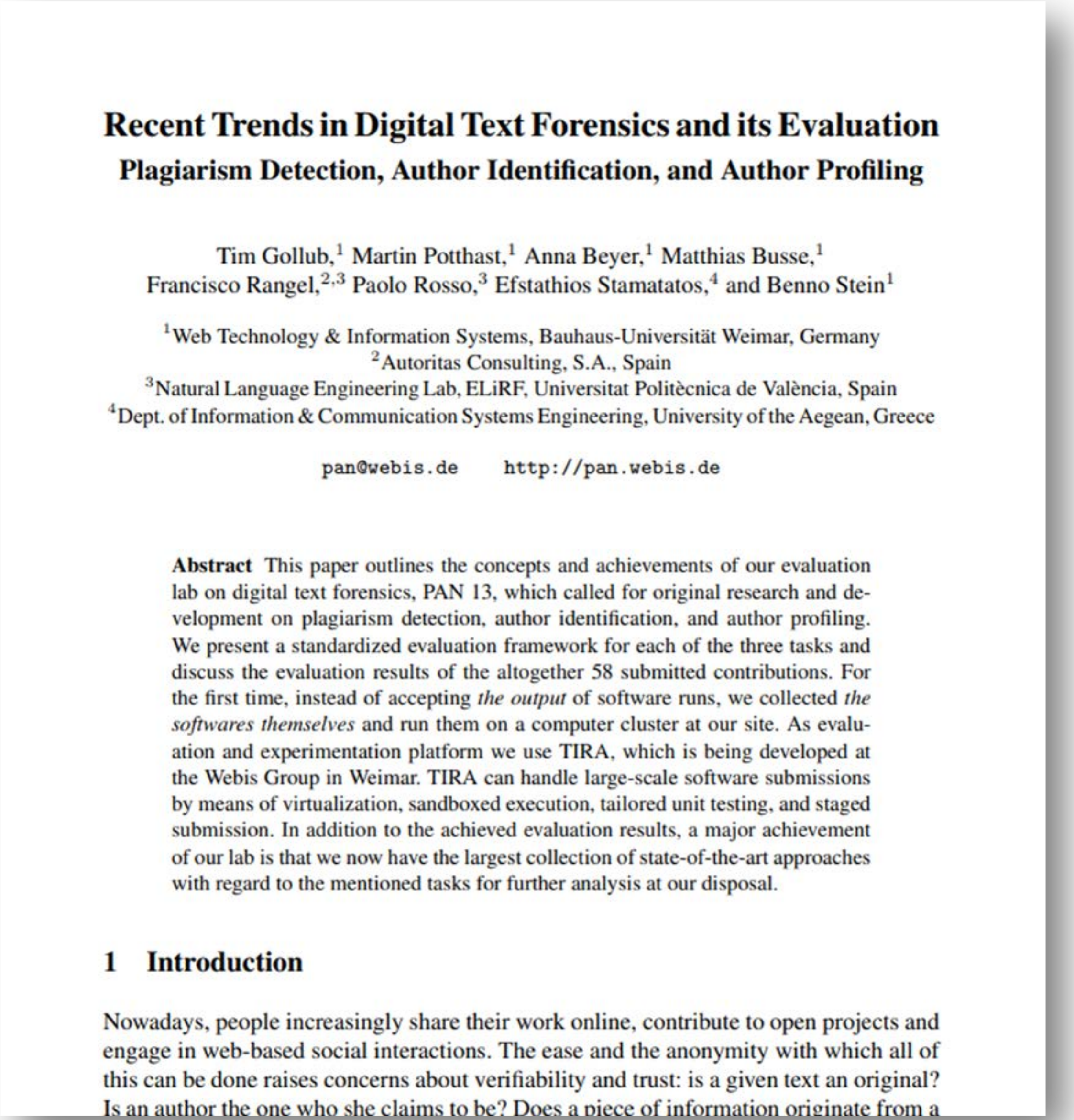
The program's success seems to confirm the stereotypical perception of differences in male and female language use. Crudely put, men talk more about objects, and women more about relationships.

Female writers use more pronouns (I, you, she, their, myself), say the program's developers, Moshe Koppel of Bar-Ilan University in Ramat Gan, Israel, and colleagues. Males prefer words that identify or determine nouns (a, the, that) and words that quantify them (one, two, more).

So this article would already, through sentences such as this, have probably betrayed its author as male: there is a prevalence of plural pronouns (they, them), indicating the male tendency to categorize rather than personalize.

https://www.nature.com/news/2003/030714/full/news030714-13.html

https://app.readable.com/text/gender/



Recent Trends in Digital Text Forensics and its Evaluation

Plagiarism Detection, Author Identification, and Author Profiling

Tim Gollub,[1] Martin Potthast,[1] Anna Beyer,[1] Matthias Busse,[1]
Francisco Rangel,[2,3] Paolo Rosso,[3] Efstathios Stamatatos,[4] and Benno Stein[1]

[1]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[2]Autoritas Consulting, S.A., Spain
[3]Natural Language Engineering Lab, ELiRF, Universitat Politècnica de València, Spain
[4]Dept. of Information & Communication Systems Engineering, University of the Aegean, Greece

pan@webis.de   http://pan.webis.de

Abstract This paper outlines the concepts and achievements of our evaluation lab on digital text forensics, PAN 13, which called for original research and development on plagiarism detection, author identification, and author profiling. We present a standardized evaluation framework for each of the three tasks and discuss the evaluation results of the altogether 58 submitted contributions. For the first time, instead of accepting *the output* of software runs, we collected *the softwares themselves* and run them on a computer cluster at our site. As evaluation and experimentation platform we use TIRA, which is being developed at the Webis Group in Weimar. TIRA can handle large-scale software submissions by means of virtualization, sandboxed execution, tailored unit testing, and staged submission. In addition to the achieved evaluation results, a major achievement of our lab is that we now have the largest collection of state-of-the-art approaches with regard to the mentioned tasks for further analysis at our disposal.

1 Introduction

Nowadays, people increasingly share their work online, contribute to open projects and engage in web-based social interactions. The ease and the anonymity with which all of this can be done raises concerns about verifiability and trust: is a given text an original? Is an author the one who she claims to be? Does a piece of information originate from a

https://webis.de/downloads/publications/papers/stein_2013g.pdf

Discussion

Gollub, Tim, et al. "Recent trends in digital text forensics and its evaluation." *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, Berlin, Heidelberg, 2013.

Text analytics – typical tasks?

- Language detection
- Named entity extraction
- Detecting themes, categories, topics
- Detecting intentions
- Sentiment analysis
- Document summarization
- Basis for translation

Text analytics – Identification of the Language

- Can tell what language the text is, e.g. English, German, Spanish,…
- Relevant for understanding and translation

- Example (Online) APIs:
  - https://console.bluemix.net/apidocs/language-translator
  - https://docs.microsoft.com/en-us/azure/cognitive-services/translator/
  - https://cloud.google.com/translate/docs/detecting-language
  - https://pypi.org/project/langdetect/
- Language identification using NLTK, examples
  - https://avital.ca/notes/language-identification-using-nltk
  - http://www.algorithm.co.il/blogs/programming/python/cheap-language-detection-nltk/

Text analytics – Sentiment analysis
Example – how to… classify reviews?

★★★★★ **Klare Kaufempfehlung**
29. November 2017
Format: Kindle Ausgabe | Verifizierter Kauf

Wer sich mit dem Thema künstliche Intelligenz und Neuronalen Netzen beschäftigen und auch selbst anhand von Python ein Neuronales Netz zur Schriftzeichen-Erkennung programmieren möchte, liegt hier Gold richtig.

Gute Erklärung der Grundlagen, ohne zu theoretisch zu werden. Schneller Einstieg in die Praxis mit gut verständlichen Code-Beispielen ohne unnötigen Ballast. Grundlegende Python Kenntnisse sind hilfreich aber nicht unbedingt erforderlich.

Kundenbilder

Alle Kundenbilder anzeigen

Lesen Sie Rezensionen, die folgende Stichworte enthalten

neuronalen netzen | neuronales netz | schritt schritt | neuronaler netze | tariq rashid | neuronalen netzes | zweiten teil | thema neuronale | künstliche intelligenz | grundlagen neuronaler | mathematischen grundlagen

67 Kundenrezensionen

Spitzenrezensionen

kkr

★★★★★ **Ausgesprochen hilfreiches Einsteigerbuch**
16. Januar 2018
Format: Taschenbuch | Verifizierter Kauf

Mir hat das Buch sehr dabei geholfen praktisch ohne Vorkenntnisse (ich konnte kein Python, nur etwas Pascal und C, noch aus den Neunzigern) und mit mäßigem mathematischem Hintergrund das Thema neuronale Netze zu erfassen und - und darum ging es mir in erster Linie - selbst ein neuronales Netz in Python zusammenzubasteln, um damit für ein Projekt Bilder zu erkennen. Das Ergebnis ist überraschend gut und ich kann dieses Buch dementsprechend begeistert weiter empfehlen.

Das neuronale Netz Herrn Rashids funktioniert ohne auf Frameworks wie Tensorflow oder Torch zurückzugreifen und ohne GPU-Unterstützung. Theoretisch kann man es auf einem Raspberry Pi berechnen (was allerdings quälend langsam ist), ein einfaches Notebook ohne Nvidia Graphikkarte reicht aber völlig aus.

Für den praktischen Einstieg ist das Buch perfekt und hebt sich deutlich von allem anderen ab, was der deutsch- und englischsprachige Markt so her gibt. Im Übrigen ist auch das Blog des Autors sehr zu empfehlen.

39 Personen fanden diese Informationen hilfreich

Nützlich | Kommentar | Missbrauch melden

★★☆☆☆ **Trifft keine Zielgruppe (nichts Halbes und nichts Ganzes)**
5. März 2019
Format: Taschenbuch | Verifizierter Kauf

Das Buch beginnt als ob es für Kinder geschrieben ist. Mit sehr vielen unnötigen Wiederholungen. Scheinbar merkt der Autor aber recht schnell, dass es sich um ein komplexes Thema handelt und pendelt dann zwischen sehr leicht oder sehr knapp und anspruchsvoll. Er setzt mindestens Mathe-Leistungskurs(Abitur) voraus.
Ich habe das Buch vor allem wegen der Vielzahl an Wiederholungen als sehr anstrengend empfunden. Zeitgleich denke ich nicht, dass Laien den mathematischen Ausführungen folgen können.

Text analytics – Sentiment analysis

- In the sentiment analysis the algorithm determines if text is positive, neutral, or negative
- Used to analyze reports, social media posts, customer reviews, forums, news items, communication, etc.
- Typically a text is broken up in parts (e.g. sentences or phrases) and for each part the sentiment is estimated. The score for the parts is then combined to get an overall score
- Sentiment library and rules
  - Sentiment library (collection of adjectives and phrases that are either positive or negative, e.g. good, brilliant, great, amazing)
  - Rules are used to assign a sentiment score based on the library and rules
- Typical problems
  - “Not good”, “the cake wasn’t bad”, …

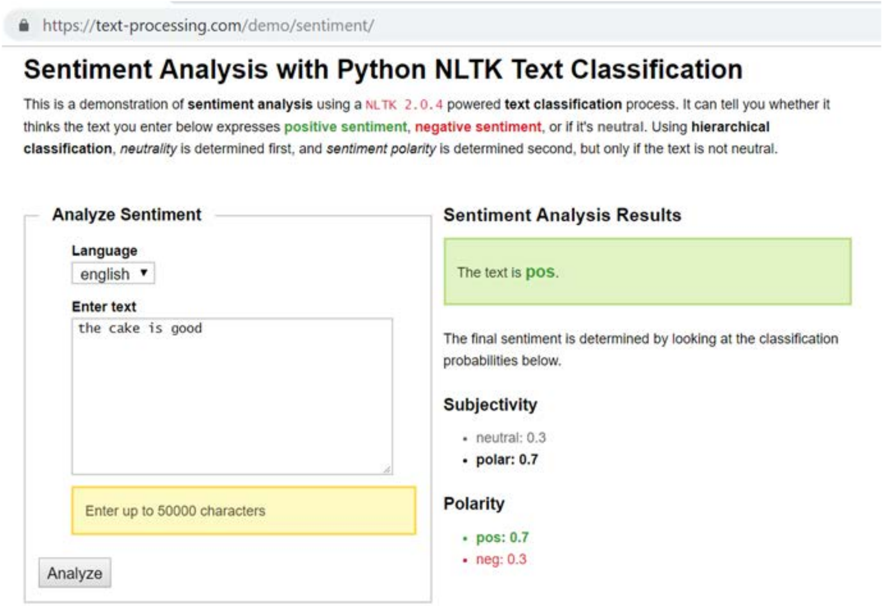
Example: https://text-processing.com/demo/sentiment/

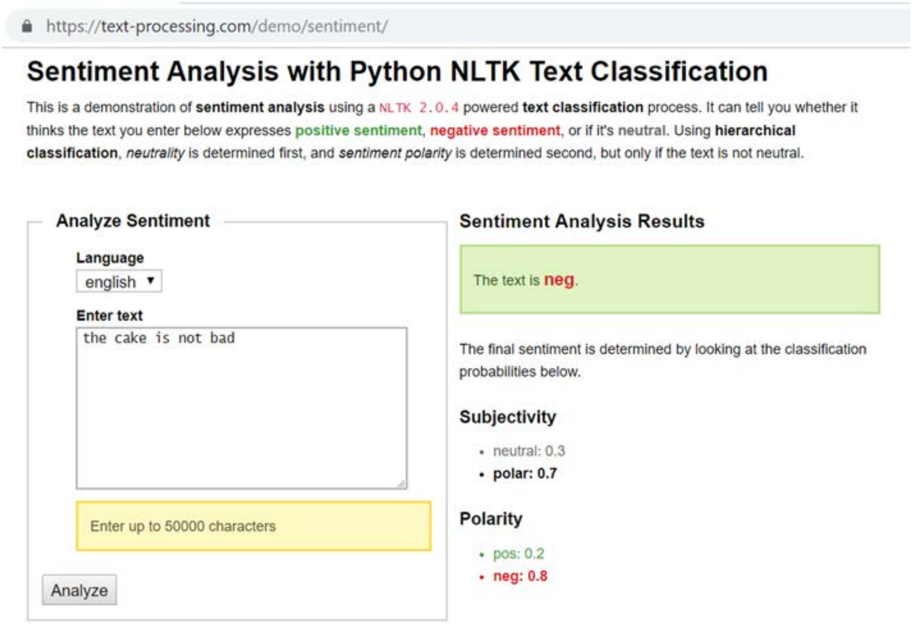
The cake is good

- pos: 0.7
- neg: 0.3

The cake is not bad

- pos: 0.2
- neg: 0.8





NLTK vader

https://www.nltk.org/_modules/nltk/sentiment/vader.html

god	1.1	1.51327	[0, 0, 0, 1, 0, 3, 0, 3, 0, 4]
goddam	-2.5	1.28452	[0, -3, -3, -4, -3, -1, -4, -1, -3, -3]
goddammed	-2.4	0.91652	[-2, -3, -1, -1, -2, -2, -4, -3, -3, -3]
goddamn	-2.1	1.75784	[-3, -3, -2, -4, -4, -3, -3, -1, 1, 1]
goddamned	-1.8	2.03961	[-3, -3, -3, -4, -1, 2, -2, -3, 2, -3]
goddamns	-2.1	1.51327	[-3, -2, -4, 2, -2, -3, -3, -2, -2, -2]
goddams	-1.9	1.92094	[-3, -3, -2, -4, -4, -2, -3, -1, 2, 1]
godsend	2.8	0.87178	[2, 3, 3, 2, 4, 3, 3, 1, 4, 3]
good	1.9	0.9434	[2, 1, 1, 3, 2, 4, 2, 2, 1, 1]
goodness	2.0	1.54919	[2, 2, 2, 3, 1, 2, -2, 4, 3, 3]
gorgeous	3.0	0.63246	[3, 3, 2, 3, 3, 3, 4, 4, 3, 2]
gorgeously	2.3	0.78102	[2, 2, 2, 3, 1, 2, 4, 3, 2, 2]
gorgeousness	2.9	0.9434	[3, 4, 3, 1, 4, 4, 2, 2, 3, 3]

Sentiment ratings from 10 independent human raters […]. Over 9,000 token features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". We kept every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of those ten independent raters. This left us with just over 7,500 lexical features […] For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon :(is –2.2, and "sucks" and it's slang derivative "sux" are both –1.5.

NLTK vader

https://www.nltk.org/_modules/nltk/sentiment/vader.html

```
# booster/dampener 'intensifiers' or 'degree adverbs'
# http://en.wiktionary.org/wiki/Category:English_degree_adverbs

BOOSTER_DICT = {
    "absolutely": B_INCR,
    "amazingly": B_INCR,
    "awfully": B_INCR,
    "completely": B_INCR,
    "considerably": B_INCR,
    "decidedly": B_INCR,
    "deeply": B_INCR,
    "effing": B_INCR,
    "enormously": B_INCR,
    "entirely": B_INCR,
    "especially": B_INCR,
    "exceptionally": B_INCR,
    "extremely": B_INCR,
    "fabulously": B_INCR,
    "flipping": B_INCR,
    "flippin": B_INCR,
    "fricking": B_INCR,
    "frickin": B_INCR,
    "frigging": B_INCR,
    "friggin": B_INCR,
    "fully": B_INCR,
    "fucking": B_INCR,
```

```
SPECIAL_CASE_IDIOMS = {
    "the shit": 3,
    "the bomb": 3,
    "bad ass": 1.5,
    "yeah right": -2,
    "cut the mustard": 2,
    "kiss of death": -1.5,
    "hand to mouth": -2,
}
```

```
def negated(input_words, include_nt=True):
    """
    Determine if input contains negation words
    """
    neg_words = NEGATE
    if any(word.lower() in neg_words for word in input_words):
        return True
    if include_nt:
        if any("n't" in word.lower() for word in input_words):
            return True
    for first, second in pairwise(input_words):
        if second.lower() == "least" and first.lower() != 'at':
            return True
    return False
```

```
NEGATE = {
    "aint",
    "arent",
    "cannot",
    "cant",
    "couldnt",
    "darent",
    "didnt",
    "doesnt",
    "ain't",
    "aren't",
    "can't",
    "couldn't",
    "daren't",
    "didn't",
    "doesn't",
    "dont",
    "hadnt",
    "hasnt",
    "havent",
    "isnt",
    "mightnt",
    "mustnt",
    "neither",
    "don't",
    "hadn't",
    "hasn't",
    "haven't",
    "isn't",
    "mightn't",
    "mustn't",
    "neednt",
    "needn't",
```

http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

```python
def negated(input_words, include_nt=True):
    """
    Determine if input contains negation words
    """
    neg_words = NEGATE
    if any(word.lower() in neg_words for word in input_words):
        return True
    if include_nt:
        if any("n't" in word.lower() for word in input_words):
            return True
    for first, second in pairwise(input_words):
        if second.lower() == "least" and first.lower() != 'at':
            return True
    return False
```

```python
NEGATE = {
    "aint",
    "arent",
    "cannot",
    "cant",
    "couldnt",
    "darent",
    "didnt",
    "doesnt",
    "ain't",
    "aren't",
    "can't",
    "couldn't",
    "daren't",
    "didn't",
    "doesn't",
    "dont",
    "hadnt",
    "hasnt",
    "havent",
    "isnt",
    "mightnt",
    "mustnt",
    "neither",
    "don't",
    "hadn't",
    "hasn't",
    "haven't",
    "isn't",
    "mightn't",
    "mustn't",
    "neednt",
    "needn't",
```

NLTK vader: https://www.nltk.org/_modules/nltk/sentiment/vader.html

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

Text analytics – Sentiment analysis

- Sentiment Analysis with Python NLTK Text Classification (online example)
  https://text-processing.com/demo/sentiment/
- Twitter Sentiment Analysis using Python
  https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
- nltk.sentiment.sentiment_analyzer module
  facilitate Sentiment Analysis tasks using NLTK features and classifiers
  https://www.nltk.org/api/nltk.sentiment.html
- https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis

Text analytics – Summarization

- For a given text a short version is created that keep a maximum of the content and should still relay the same message
- Important, especially if dealing with a lot of text (reports, social media, communication)
- Optimum: Reduce text in a way that only the relevant information remains
- Applications:
  - Reduce reading time for human reader
  - Improve indexing of documents
  - Simplify overview of larger texts and collections
- Manual text summarization is common
  - Headings in newspapers, synopses from a book, abstracts in papers, reviews of a film or book

- See: https://machinelearningmastery.com/gentle-introduction-text-summarization/

Text analytics – Summarization Example

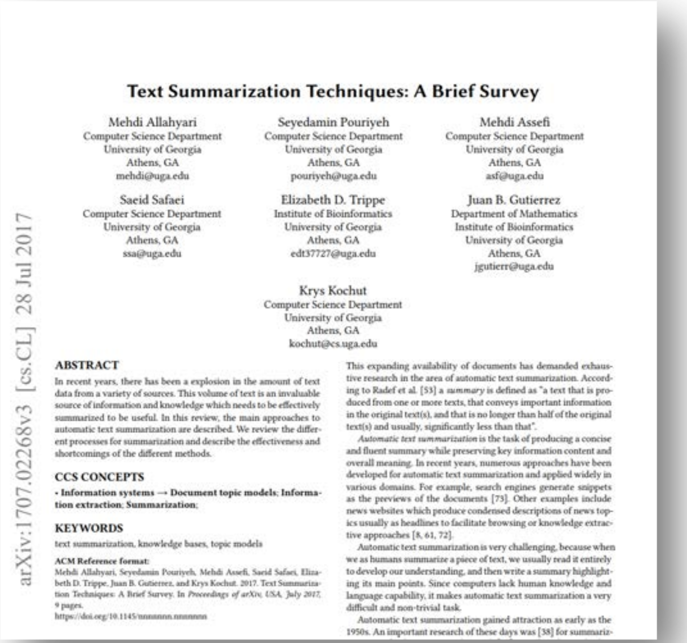
"So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work."

Task: Summarize the above paragraph in 1 sentence. The sentence must not be longer than 10 words.

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization

- Approaches:
  - **Extraction**: “identifying important sections of the text and generating them verbatim; thus, they depend only on extraction of sentences from the original text”.
  - **Abstraction**: “aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text.”
- Extraction is easier in typically better than abstraction as it requires less semantic understanding



arXiv:1707.02268v3 [cs.CL] 28 Jul 2017

Text Summarization Techniques: A Brief Survey

Mehdi Allahyari
Computer Science Department
University of Georgia
Athens, GA
mehdi@uga.edu

Seyedamin Pouriyeh
Computer Science Department
University of Georgia
Athens, GA
pouriyeh@uga.edu

Mehdi Assefi
Computer Science Department
University of Georgia
Athens, GA
asf@uga.edu

Saeid Safaei
Computer Science Department
University of Georgia
Athens, GA
ssa@uga.edu

Elizabeth D. Trippe
Institute of Bioinformatics
University of Georgia
Athens, GA
edt37727@uga.edu

Juan B. Gutierrez
Department of Mathematics
Institute of Bioinformatics
University of Georgia
Athens, GA
jgutierr@uga.edu

Krys Kochut
Computer Science Department
University of Georgia
Athens, GA
kochut@cs.uga.edu

ABSTRACT

In recent years, there has been a explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods.

CCS CONCEPTS

• **Information systems → Document topic models**; **Information extraction**; **Summarization**;

KEYWORDS

text summarization, knowledge bases, topic models

ACM Reference format:
Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. In *Proceedings of arXiv, USA, July 2017,* 9 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

This expanding availability of documents has demanded exhaustive research in the area of automatic text summarization. According to Radef et al. [53] a *summary* is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that".

Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. For example, search engines generate snippets as the previews of the documents [73]. Other examples include news websites which produce condensed descriptions of news topics usually as headlines to facilitate browsing or knowledge extractive approaches [8, 61, 72].

Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task.

Automatic text summarization gained attraction as early as the 1950s. An important research of these days was [38] for summariz-

Mehdi Allahyari et al. Text Summarization Techniques: A Brief Survey. 2017. arXiv:1707.02268 [cs.CL]

Text analytics – Summarization Example

> So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work.

Step 1: Convert Paragraphs to Sentences

1. So, keep working
2. Keep striving
3. Never give up
4. Fall down seven times, get up eight
5. Ease is a greater threat to progress than hardship
6. Ease is a greater threat to progress than hardship
7. So, keep moving, keep growing, keep learning
8. See you at work

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization Example

So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work.

Step 1: Convert Paragraphs to Sentences

1. So, keep working
2. Keep striving
3. Never give up
4. Fall down seven times, get up eight
5. Ease is a greater threat to progress than hardship
6. Ease is a greater threat to progress than hardship
7. So, keep moving, keep growing, keep learning
8. See you at work

Step 2: Text Preprocessing

1. keep working
2. keep striving
3. never give
4. fall seven time get eight
5. ease greater threat progress hardship
6. ease greater threat progress hardship
7. keep moving keep growing keep learning
8. see work

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization Example

So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work.

Step 1: Convert Paragraphs to Sentences

1. So, keep working
2. Keep striving
3. Never give up
4. Fall down seven times, get up eight
5. Ease is a greater threat to progress than hardship
6. Ease is a greater threat to progress than hardship
7. So, keep moving, keep growing, keep learning
8. See you at work

Step 2: Text Preprocessing

1. keep working
2. keep striving
3. never give
4. fall seven time get eight
5. ease greater threat progress hardship
6. ease greater threat progress hardship
7. keep moving keep growing keep learning
8. see work

Step 3. Tokens

```
['keep',
 'working',
 'keep',
 'striving',
 'never',
 'give',
 'fall',
 'seven',
 'time',
 'get',
 'eight',
 'ease',
 'greater',
 'threat',
 'progress',
 'hardship',
 'ease',
 'greater',
 'threat',
 'progress',
 'hardship',
 'keep',
 'moving',
 'keep',
 'growing',
 'keep',
 'learning',
 'see',
 'work']
```

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization Example

Step 3. Tokens

```
['keep',
 'working',
 'keep',
 'striving',
 'never',
 'give',
 'fall',
 'seven',
 'time',
 'get',
 'eight',
 'ease',
 'greater',
 'threat',
 'progress',
 'hardship',
 'ease',
 'greater',
 'threat',
 'progress',
 'hardship',
 'keep',
 'moving',
 'keep',
 'growing',
 'keep',
 'learning',
 'see',
 'work']
```

Step 4: Find weighted frequency of occurrence

Word	Frequency	Weighted Frequency
ease	2	0.40
eight	1	0.20
fall	1	0.20
get	1	0.20
give	1	0.20
greater	2	0.40
growing	1	0.20
hardship	2	0.40
keep	5	1.00
learning	1	0.20
moving	1	0.20
never	1	0.20
progress	2	0.40
see	1	0.20
seven	1	0.20
striving	1	0.20
threat	2	0.40
time	1	0.20
work	1	0.20
working	1	0.20

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization Example

Step 4: Find weighted frequency of occurrence

Word	Frequency	Weighted Frequency
ease	2	0.40
eight	1	0.20
fall	1	0.20
get	1	0.20
give	1	0.20
greater	2	0.40
growing	1	0.20
hardship	2	0.40
keep	5	1.00
learning	1	0.20
moving	1	0.20
never	1	0.20
progress	2	0.40
see	1	0.20
seven	1	0.20
striving	1	0.20
threat	2	0.40
time	1	0.20
work	1	0.20
working	1	0.20

> So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work.

5. Replace Words by Weighted Frequency in Original Sentences

Sentence	Sum of Weighted Frequencies
So, keep working	1 + 0.20 = 1.20
Keep striving	1 + 0.20 = 1.20
Never give up	0.20 + 0.20 = 0.40
Fall down seven times, get up eight	0.20 + 0.20 + 0.20 + 0.20 + 0.20 = 1.0
Ease is a greater threat to progress than hardship	0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0
Ease is a greater threat to progress than hardship	0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0
So, keep moving, keep growing, keep learning	1 + 0.20 + 1 + 0.20 + 1 + 0.20 = 3.60
See you at work	0.20 + 0.20 = 0.40

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Text analytics – Summarization Example

> So, keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, keep moving, keep growing, keep learning. See you at work.

5. Replace Words by Weighted Frequency in Original Sentences

Sentence	Sum of Weighted Frequencies
So, keep working	1 + 0.20 = 1.20
Keep striving	1 + 0.20 = 1.20
Never give up	0.20 + 0.20 = 0.40
Fall down seven times, get up eight	0.20 + 0.20 + 0.20 + 0.20 + 0.20 = 1.0
Ease is a greater threat to progress than hardship	0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0
Ease is a greater threat to progress than hardship	0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0
So, keep moving, keep growing, keep learning	1 + 0.20 + 1 + 0.20 + 1 + 0.20 = 3.60
See you at work	0.20 + 0.20 = 0.40

6. Results

> So, keep moving, keep growing, keep learning

> So, keep moving, keep growing, keep learning. Ease is a greater threat to progress than hardship.

Example from https://stackabuse.com/text-summarization-with-nltk-in-python/

Recap
… do you remember what we did last time?

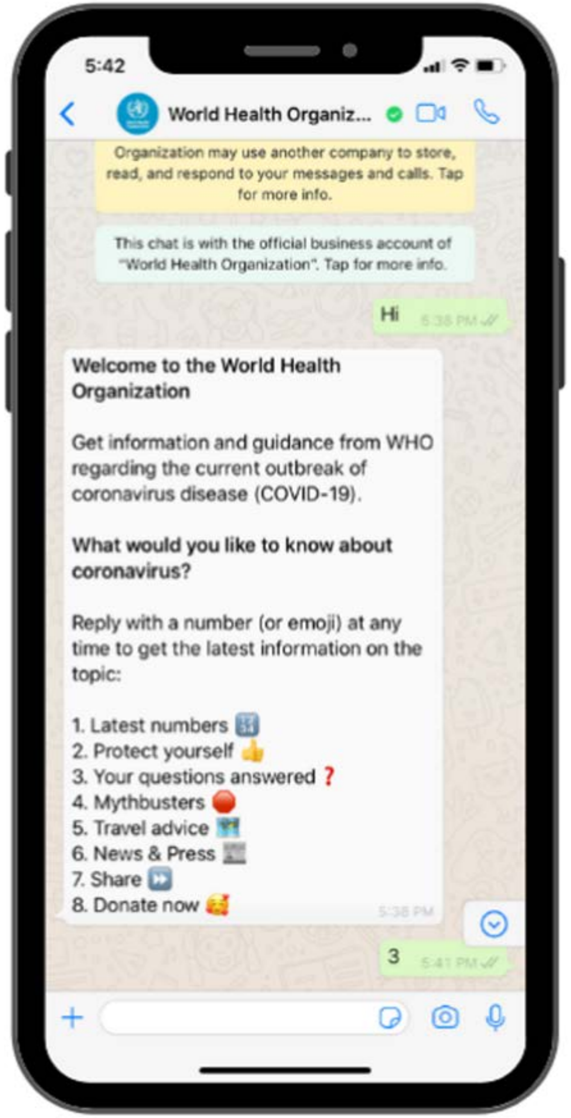
- Tokenization
- Stop Words Removal
- Text normalization
- Stemming / Porter Stemmer
- Lemmatization
- Part-Of-Speech Tagging
- Named Entity Disambiguation
- Named Entity Extraction
- Bag of Words
- Corpus, Corpora

Conversational UIs & Bots

Recent Example

Chatbot by the World Health Organization

- Gives covid19 infos via chat (WhatsApp)
- What are the benefits of this approach? Why not just a website?



https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp

https://www.whatsapp.com/coronavirus/who

Why use Conversation as a User Interface?

- Natural language
- Non-graphical UIs (voice)
- Hands-free (voice)
- Social contexts
- Personification / antropomorphism
- Personalisation
- Integration
- …

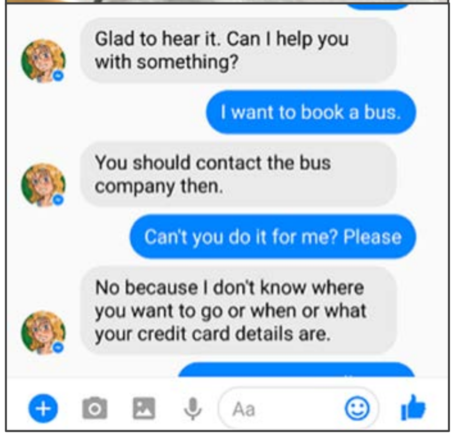
Involved Tasks & Technologies


Photo by Ryutaro Tsukata

Voice recognition / natural language understanding

Dialogue model (e.g. contextual references)

Voice synthesis / natural language generation



CUI Design

Overview & Examples: CUI Design Factors

What is there to *design* about a conversational UI?

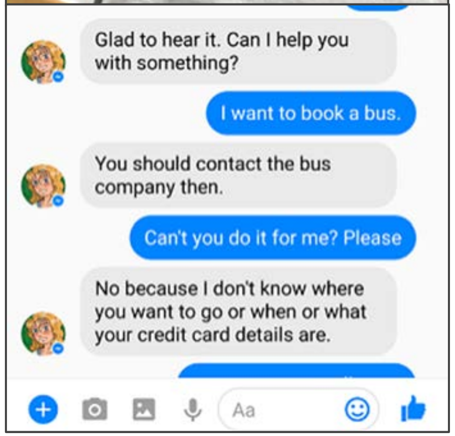

Photo by Ryutaro Tsukata

Input

Conversation

Output

Personification
of the System

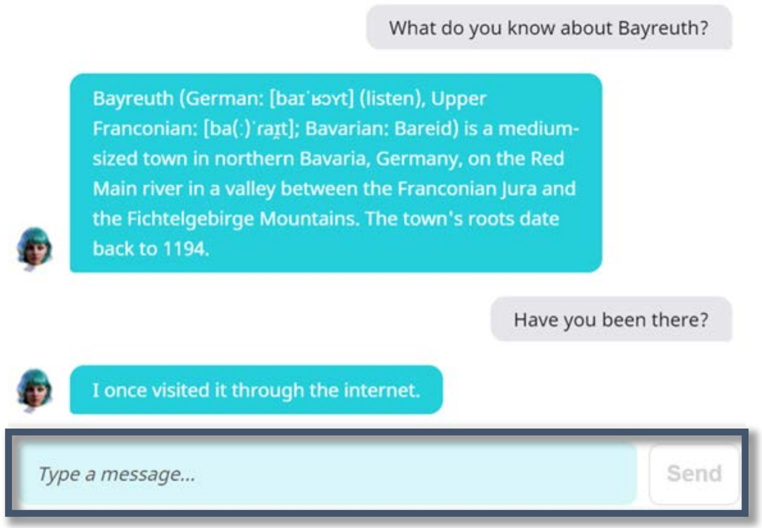


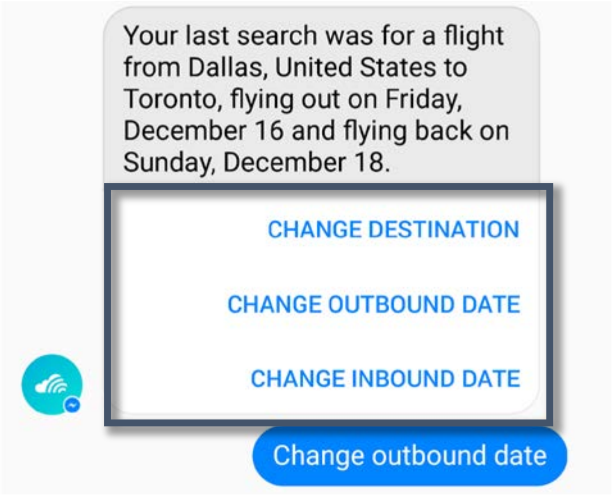
Integration

Input Design Factors

Free text vs fixed options

	Free text	Fixed options
+	Realistic impression, broad scope, free user expression	Full control (for designer), user guidance, clear capabilities, predictable, less typing
-	Many edge cases or undefined cases, potentially unwanted in/output, annoying if bot can‘t handle input	Limited scope, impersonal, artificial, not useful if user‘s intent not anticipiated





https://chatbotsmagazine.com/19-best-practices-for-building-chatbots-3c46274501b2

Output Design Factors

Conversational elements

- **Linguistic features:**
  e.g. word choice, sentence length, level of formality, slang, puns, irony, cultural proverbs, …

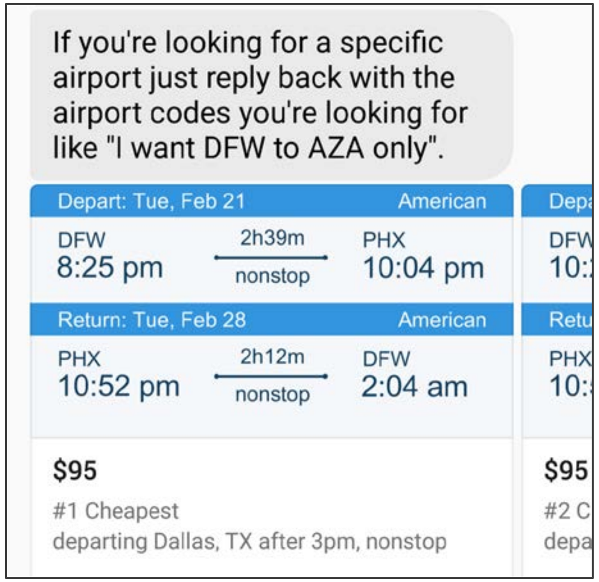
- **Paraverbal features:**
  e.g. prosody, tone of voice, dialect, intonation, gender, …

https://www.genderlessvoice.com/

Output Design Factors

Additional (visual) elements

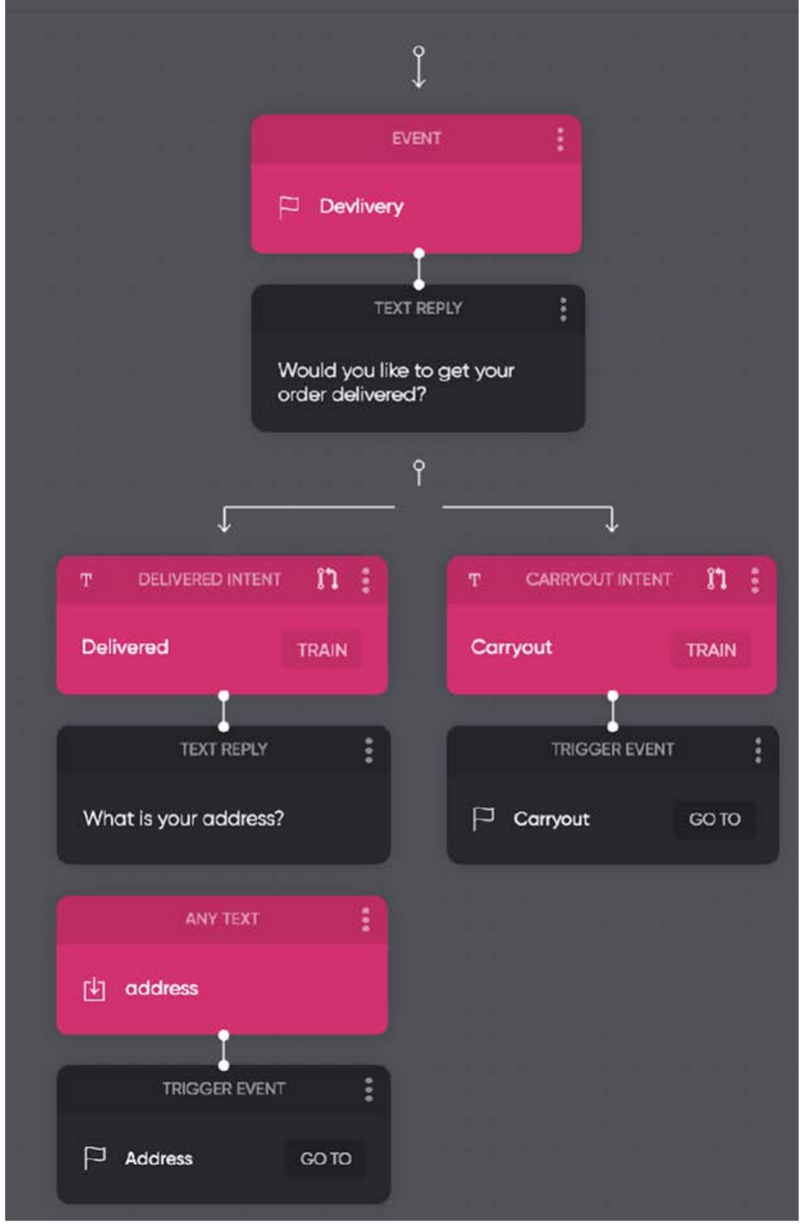
- **Non-conversational UI:** e.g. status, progress, shortcuts, tables, images
- **Surrounding presentation:** e.g. avatar, device design



Conversation Design

Conversation flow

Graphical representation of conversation, branches, start and end points, e.g.:



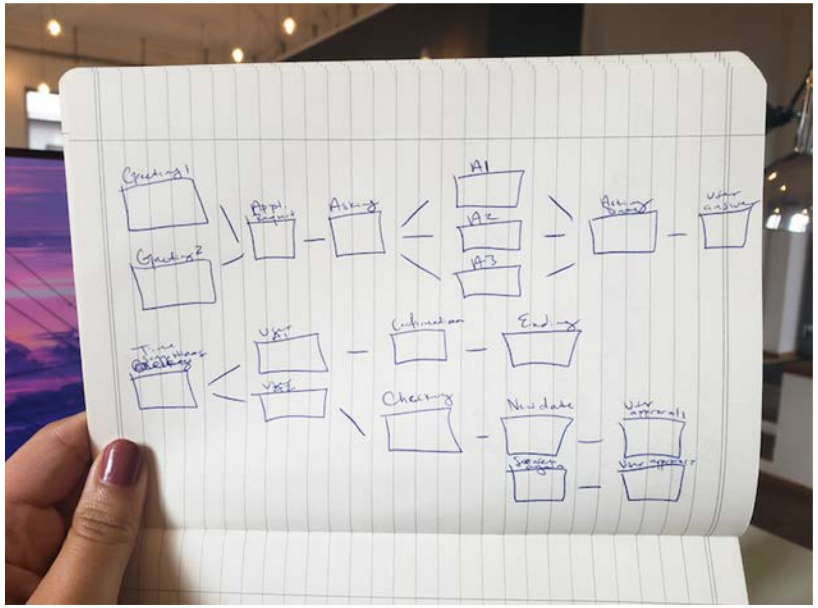
https://flow.ai

Breakout Activity

Conversation Design

Scenario: You develop a chatbot for scheduling appointments at a local business (e.g. hair stylist, bookshop, doctor).

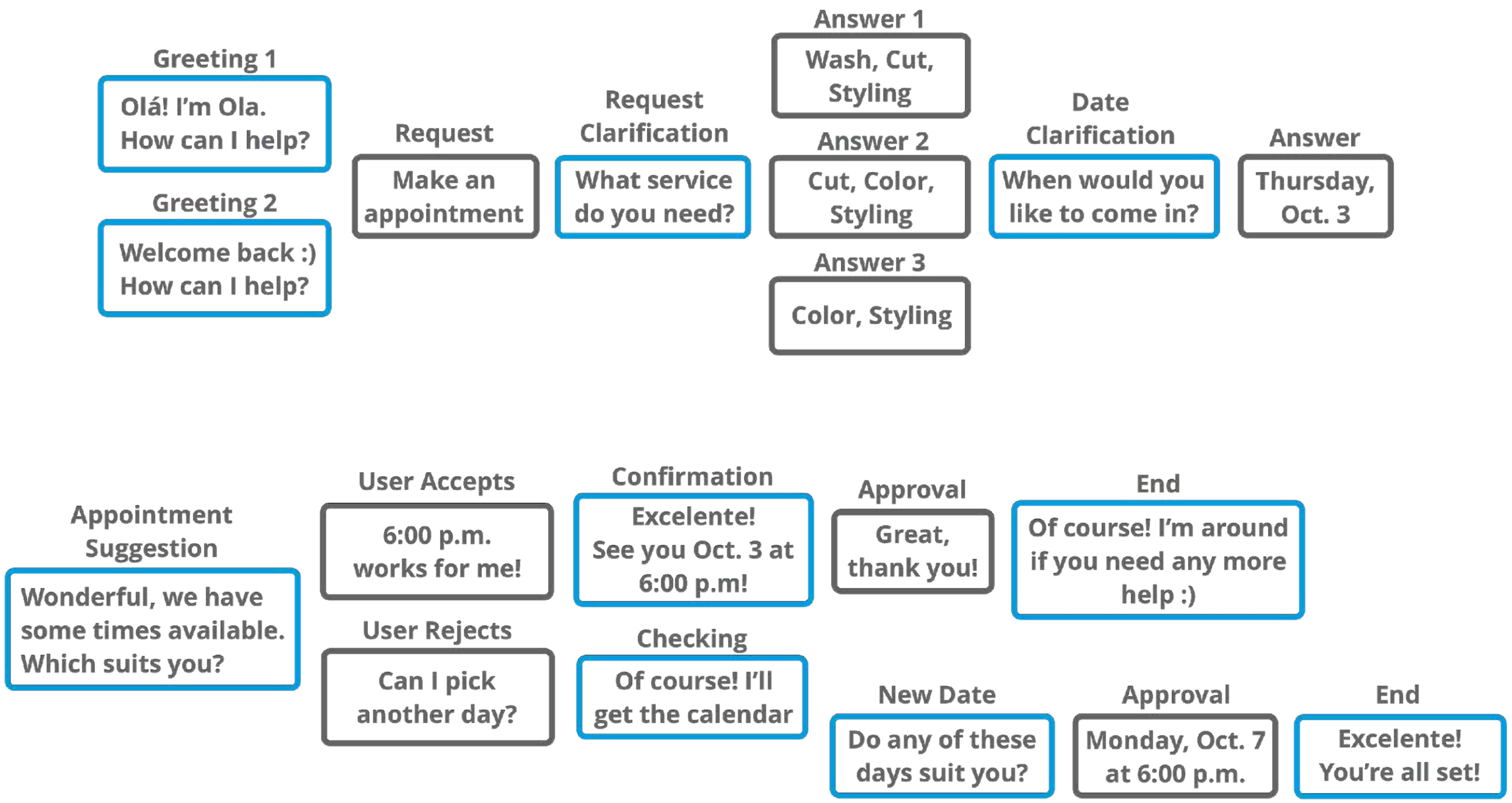
Task: Create a conversation flow diagram for this. Can you identify conversation elements that could be generally useful, beyond your specific scenario?



https://www.userlike.com/en/blog/chatbot-conversation-flow

Conversation Design

Conversation flow



https://www.userlike.com/en/blog/chatbot-conversation-flow

Conversation Design

Typical elements of a conversation

- **Start**
  - Establish relationship, provide information
  - E.g. a greeting
- **Suggestions / prompts**
  - Present the user with possible / relevant actions or options.
  - E.g. *“I can help you to create a new order or check on an existing one.”*
- **Progress indicators**
  - Reveal progress through a conversation towards a goal
  - E.g. *“Almost there! To complete this I also need…”*
- **Confirmation**
  - Confirm actions, restate information
  - E.g. *„Ok, I will update your address to: …“,*
- **Error handling**
  - Acknowledge limitations or breakdowns
  - *„Sorry, I do not understand this.“*
- **End:** provide a clear end to the conversation

https://www.userlike.com/en/blog/chatbot-conversation-flow

Conversation Design

Handling conversation breakdowns

[Ashktorab et al., 2019]

Example: Bot doesn‘t understand this.

Bot: Hello, I'm your digital assistant. How can I help you with your banking needs today?

User: hi I'd like to add my daughter to my card

Repair strategies:

Repeat	**Bot**: I don’t quite understand. How can I help you with your banking needs today?
Give options	**Bot**: I am not quite sure if I understood. Which of these are you trying to do today: 1) Add my daughter as an authorized user to my credit card. 2) Add a credit card to my account, or 3) Apply for a credit card for my daughter, 4) None of the above?
Confirm	**Bot**: Sounds like you want me to add a credit card to your account. Is that right?
Defer	**Bot**: Sorry, I do not understand. For this request, I will need to refer you to a human agent who will address your request in the next few minutes.
Explain e.g. highlight words	**User**: hi i’d like to add my **daughter** to my card **Bot**: Sorry I don’t understand. I’ve highlighted words I am unable to process. Can you please rephrase?

 Acknowledge breakdowns, assist user repair, proactively suggest solutions

Personification of a CUI

- In the conversation: E.g. should the voice assistant / chatbot portrait a person with specific characteristics?
- Beyond conversation: E.g. visual avatar?
- Conceptual framework: Analogy to human personality?

Summary: CUI Design Factors

Input

- Modality: Voice vs text
- Options: Free vs fixed choices

Conversation

- Dialogue characteristics: e.g. short vs long replies/chunks
- Conversation flow: e.g. branching, boundaries
- Error handling: e.g. repair strategies

Personification of the System

- In the conversation: e.g. should the assistant / chatbot portrait a specific personality
- Beyond conversation: e.g. visual avatar, physical device design

Output

- Modality: Voice vs text
- Content & style: e.g. prosody, level of formality, slang, …
- Non-conversation UI elements: e.g. status, progress, images, shortcuts

Integration

- E.g. of a chatbot in the surrounding GUI
- In context…

Design Principles

Comparison of graphical UIs and voice UIs

- Some design principles from GUIs are applicable with voice
  - Consistency (especially lexical)
  - Feedback (visual, acoustic or spoken)
  - Metaphors (when expressed in language)
- Other principles are different with voice
  - Constraints (only logical and cultural, no physical)
  - Error tolerance (homophones, ambiguities, …)
- Some are not available with voice
  - Spatial mappings
  - Visual affordances

Some Guidelines

For voice UIs

- **Tell users what they can do**
  e.g. „You can ask for today's weather or a weekly forecast."
- **Tell users where they are**
  e.g. „Today's weather forecast is mostly sunny and dry"
  (rather than just „sunny and dry.")
- **Give examples rather than instructions**
  e.g. in the help function or in the greeting
- **Limit the amount of information**
  - e.g. not more than three different options for an interaction
  - tradeoff between efficiency and short term memory!
- **Use visual feedback** (if possible)

CUI Design Process

- Special challenge with CUIs: Handling UI, conversation, technical aspects with mutual influences
- Think of possible ways a conversation could go for your use case
- Try it out! User-centred approach
- Example

https://uxknowledgebase.com/conversational-ui-designing-part-2-52bad56f005c

Challenges & Limitations of CUIs

Challenges

Semantics and pragmatics of natural language

- **Semantics:** The meaning of the words

or

„time flies“

- **Pragmatics:** How we actually use language
  - A: „Have you got coffee to go?“
  - B: „Milk and sugar?“
  - A: „Black as my soul…“

Try to teach this to a computer…

Limitations of CUIs

Questioning assumptions around CUIs

- Is it really a **conversation**? Is it **natural**?
  - *“Conversational interaction […] confuses interaction with a device* within *conversation with an* actual *conversation.“*
  - People use voice in daily life - but the system is not a conversation partner like a human

[Porcheron et al., 2018]

- Incorrect expectations about **capabilities**, i.e. what the system can or cannot do
- Voice promises **hands-free use** but often shifts focus away from main task
  - E.g. due to bad error-handling
  - Splits attention (e.g. driving, in the kitchen, …)

[Luger and Sellen, 2016]

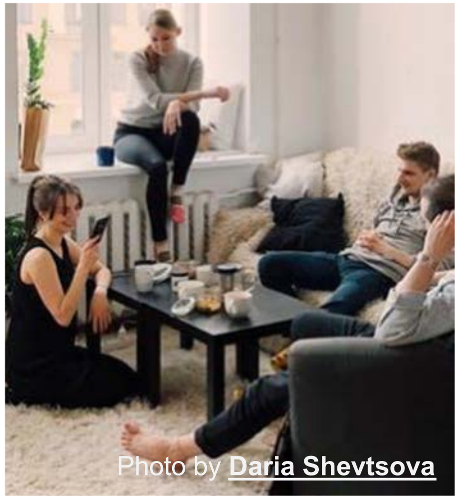


Photo by Daria Shevtsova


Photo by C Technical

Limitations of CUIs

Managing user expectations

- Users tend to **overestimate** machine capabilities
  - Normally we use voice to communicate with humans
  - Humans are intelligent
  - …hence, the thing we talk to must be intelligent

- Some **guidelines** for managing limitations:
  - Don’t assume your CUI can fully understand the user’s context
  - If in doubt, go for limited scenarios
  - Make options and suggestions easily available to users
  - …but never interrupt to provide them

Conversational UIs in HCI Research

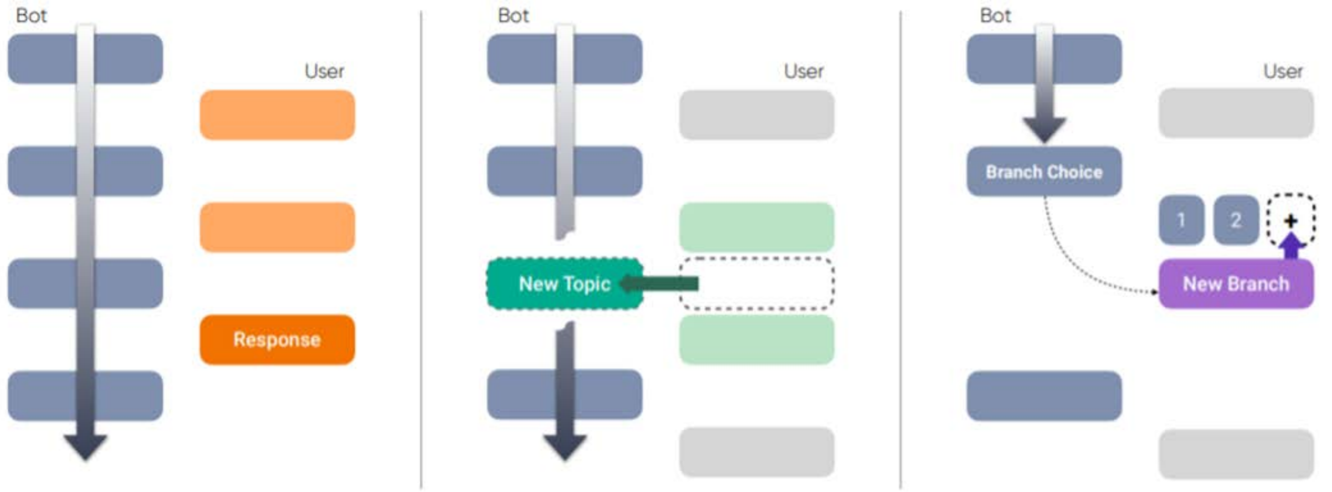
Selected examples

Conversation Design Methods & Tools

Example: „ProtoChat“

Idea: Crowdsourcing conversation design

1. Crowdworkers test a conversation
2. They can add missing but desired responses for the bot
3. They can add missing but desired branches for the user

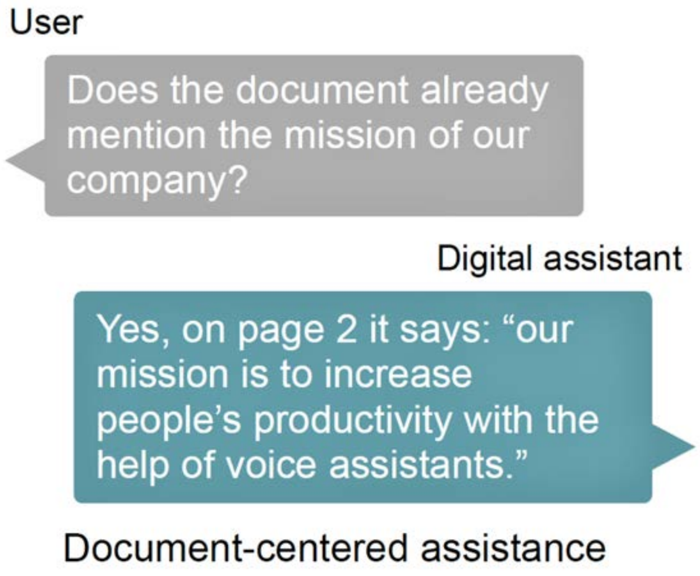


1. 2. 3. [Choi et al., 2021]

Exploring new Use-Cases for CUIs

Example: CUIs for text documents

- Idea: Use a chatbot to query a text document
- What‘s the benefit over simple text-based search?
  - E.g. Context: usable in the car
  - Bridge the „Semantic gap“: User has an idea what they want but does not know how to formulate a query
    → use natural language



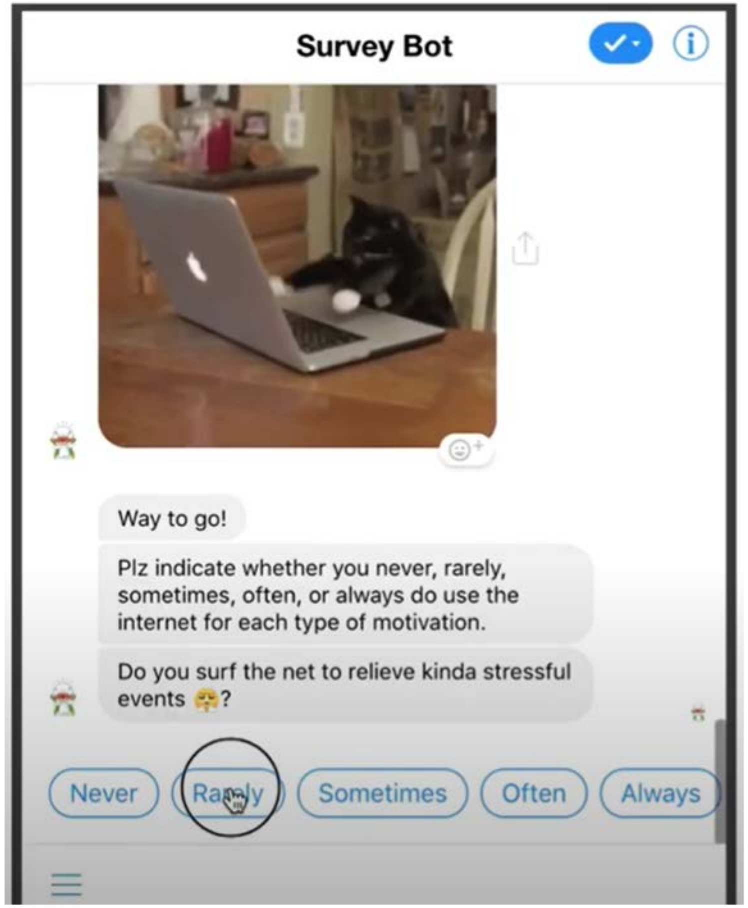
[ter Hoeve et al., 2019]

CUIs as a Research Tool

Example: A chatbot as an interviewer

- Idea: Use a chatbot instead of a questionnaire
- E.g. for user research on smartphones
- Potential benefits?

"We found that the participants in the chatbot survey, as compared to those in the web survey, were more likely to produce differentiated responses and were less likely to satisfice; the chatbot survey thus resulted in higher-quality data."* [Kim et al., 2019]



https://www.youtube.com/watch?v=A2AzbKFozc8

*satisfice - here: just answering to complete the questionnaire, not necessarily giving a „true" response

References

- Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction*, *26*(3), 17:1-17:28. https://doi.org/10.1145/3311956
- Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300484
- Choi, Y., Monserrat, T.-J. K. P., Park, J., Shin, H., Lee, N., & Kim, J. (2021). ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 225:1-225:27. https://doi.org/10.1145/3432924
- Kim, S., Lee, J., & Gweon, G. (2019). Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300316
- Luger, E., & Sellen, A. (2016). „Like Having a Really Bad PA“: The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3174214
- Ter Hoeve, M., Sim, R., Nouri, E., Fourney, A., Rijke, M. de, & White, R. W. (2020). Conversations with Documents: An Exploration of Document-Centered Assistance. *The 5th ACM SIGIR Conference on Human Information Interaction and Retrieval*. CHIIR. https://doi.org/10.1145/3343413.3377971

License