Information Retrieval (IN4325)

Introduction to Natural Language Processing

Dr. Nava Tintarev
Assistant Professor, TU Delft

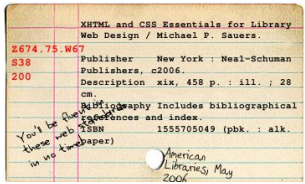
TU Delft

core information retrieval

natural language processing

IR ...

- deals with the representation, storage, organization of, and access to largely unstructured information items
- is centered around *information needs* and the concept of *relevance*
- has its roots in library and information sciences



Natural Language Processing

- Focused on text.

- Language - signs, meanings, and a code connecting signs with their meanings.
- Natural – human communication, unlike e.g., programming languages.
- Processing – computational methods to allow computers to `understand'…

- In order to perform certain information need tasks.

Applications

Star Wars: The Last Jedi - Wikipedia
https://en.wikipedia.org/wiki/Star_Wars:_The_Last_Jedi
Star Wars: **The Last Jedi** (also known as Star Wars: Episode VIII – The Last Jedi) is a 2017 American epic space opera film written and directed by Rian Johnson. It is the second installment of the Star Wars sequel trilogy and the eighth main installment of the Star Wars franchise, following Star Wars: The Force Awakens ...
Supreme Leader Snoke · Star Wars sequel trilogy · Rian Johnson · Kelly Marie Tran

Star Wars: Episode VIII - The Last Jedi (2017) - IMDb
73,999 votes
wly discovered abilities with the guidance of Luke Skywalker, who is
prepares for battle with the First

===> List some large cities in Argentina.

Argentina

Capital and largest city (2011 est.): Buenos Aires, 13.528 million

Other large cities: Córdoba, 1.556 million; Rosario 1.283 million; Mendoza 957,000; San Miguel de Tucuman 868,000; La Plata 759,000 (2011)

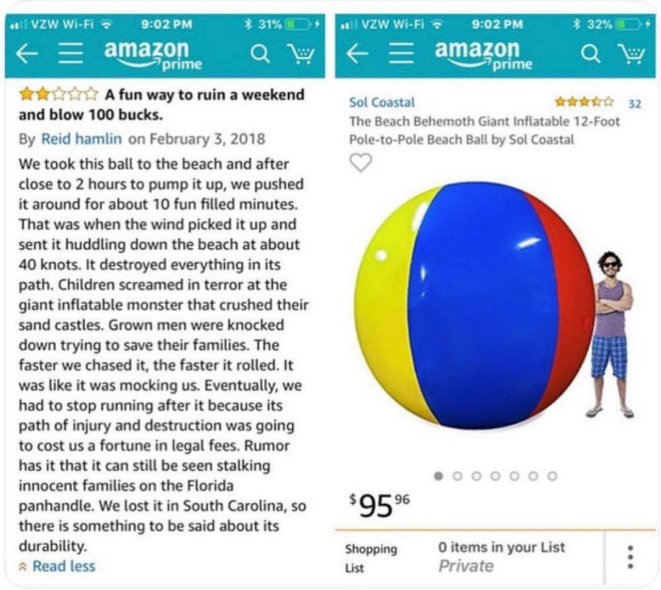
e
ides
little

/ do professors **do research**
/ do professors **assign group projects**
/ do professors **require new textbooks**
/ do professors **take attendance**
/ do professors **hate wikipedia**
/ do professors **curve**

TU Delft

Natural language processing

- From user to system
- From system to user



Part 1:
Administrative Issues
Natural Language Processing Tasks

Part 2:

Terminology: Components

Paper reviews

- 7 for NLP as well, already online
- Same template
- Thursday ->Thursday and Friday-> Friday

- Review P8: handed out March 15, **due March 22**.
- NLP project proposal: **due March 22**.
- Review P9: handed out March 21, **due March 28**.
- Review P10: handed out March 22, **due March 29**.
- Review P11: handed out March 28, **due April 4**.
- Review P12: handed out March 29, **due April 5**.
- NLP intermediate project report: **due April 5**.
- NLP final project report: **due April 10**.
- Review P13: handed out April 4, **due April 11**.
- Review P14: handed out April 5, **due April 12**.
- NLP project interviews: **April 11** and **April 12**.

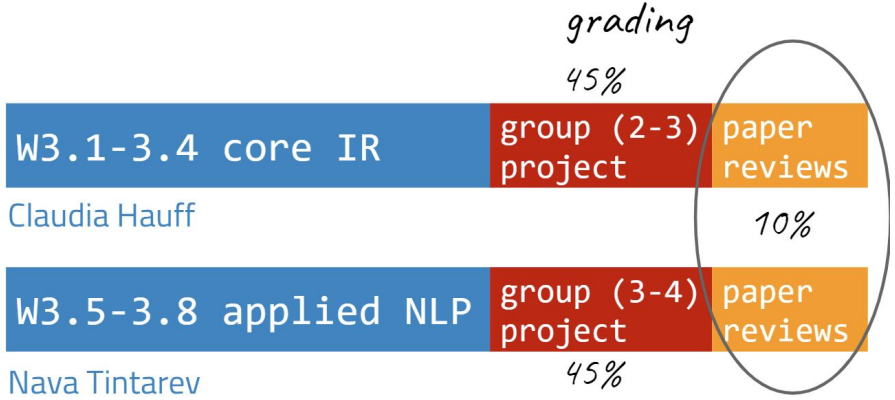
No reading today!

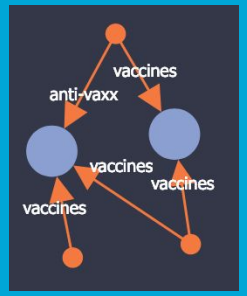
TU Delft

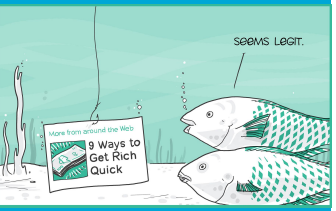
Grading

In order to pass this course, you need to fullfil **all** of the following:

1. Receive an overall grade of 5.8 or higher (in alignment with TU Delft's exam regulations).
2. Complete both project phases with at least a grade of 5.0.
3. Receive a *sufficient* for at least 9 of the 14 reviews.







Group project

#find_a_team_member

- **What:** Design, develop, and evaluate a system (NLP)
- **Group size:** ~3 (NLP) ← Larger groups! More expected!
- **Due**: Week 3.9 *includes a group interview*
- **Intermediate deadlines** to get feedback: proposal & intermediate draft

- **Project options:**
  - Reproducibility (SemEval etc.)
  - Original research

Example topics: image descriptions, Fake news, stance detection, sentiment analysis, click-bait detection,

Group project guidelines

- **Neural NLP ONLY if you have successfully completed the Deep Learning course beforehand.**
- If you build a classifier, the features you study should be motivated by linguistic theory.
- If you use off the shelf solutions: be aware of the defaults and motivate why they are suitable for your problem.
- What kind of biases might come from the dataset?
- The project focuses on **textual data**.
- Full details: https://github.com/chauff/IN4325/blob/master/projectAppliedNLP.md

TU Delft

Optional support hours:

Fridays 9am-11:30pm (W3.6 - 3.8)

Sign up for one 15 minute slot!

Team: Oana Inel (postdoc), Shabnam Najafian (PhD), Nirmal Roy (TA)

Online here: https://queue.ewi.tudelft.nl/

What will I learn?

3.5	NLP introduction (`Aula CZ C` , N. Tintarev)	Text analysis (`3mE-CZ B` , N. Tintarev)	Final core IR report due / interviews. Project group settled (NLP)
3.6	Semantics (`Aula CZ C` , N. Tintarev)	Evaluation NLP (`3mE-CZ B` , N. Tintarev)	Applied NLP project settled
3.7	ML for NLP (`Aula CZ C` , N. Tintarev)	Language generation (`3mE-CZ B` , N. Tintarev)	
3.8	NLP annotations (`Aula CZ C` , O. Inel)	Word embeddings (`CT-CZ E` , N. Tintarev)	Intermediate applied NLP report due
3.9			applied NLP report due/interviews

Part 1:

Administrative Issues

Natural Language Processing Tasks

Part 2:

Terminology: Components

After today you should be able to…

- Explain the function of Natural Language Processing
- Identify NLP applications
- Identify typical natural processing tasks
- Recognize typical components/sub-tasks

Natural language processing tasks

- **Easy:** spell checking, keyword search, finding synonyms
- **Medium:** information extraction, summarization, stance detection
- **Hard:** sentiment analysis, machine translation, co-reference, question answering

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

William Wilkinson's
"An account of the principalities of Wallachia and Moldovia"
inspired this author's most famous novel

Bram Stoker

Information Extraction

Subject: **curriculum mee**
Date: January 15, 2019
To: Claudia Hauff

Event: Curriculum mtg
Date: Jan-16-2019
Start: 10:00am
End: 11:30am
Where: 4.900

Hi Claudia, we've now scheduled the curriculum meeting.
It will be in 4.900 tomorrow from 10:00-11:30.
-Nava

Create new Calendar entry

Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

✓
- nice and compact to carry!

✓
- since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!

✗
- the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Automated Summarization

- Single document vs multiple document
- Generic summarization versus query/focused/topic-based summarization
- Extract versus abstract

- http://newsblaster.cs.columbia.edu/

Stance detection

EXAMPLE HEADLINE

"Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"

EXAMPLE SNIPPETS FROM BODY TEXTS AND CORRECT CLASSIFICATIONS

"... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. ..."

CORRECT CLASSIFICATION: AGREE

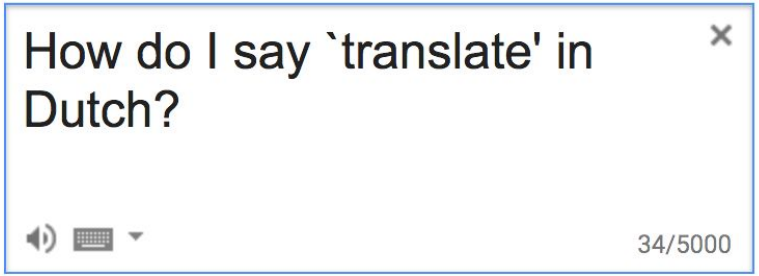
"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."

CORRECT CLASSIFICATION: DISAGREE

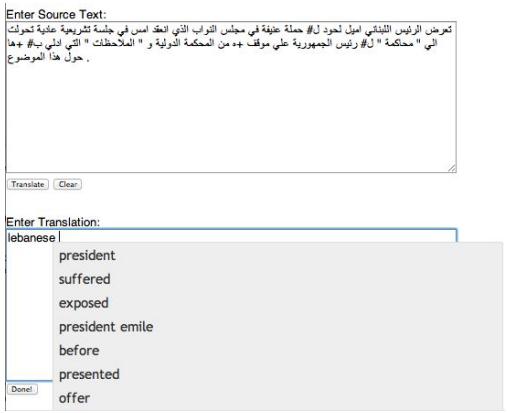
- Input: Headline + text
- Output: Classify stance (e.g., agrees, disagrees, discusses, unrelated)

Machine Translation

- Fully automatic



- Helping human translators
- "The flesh was weak, but the spirit was willing" → "The meat was off, but the vodka was fine".



Tasks in NLP

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

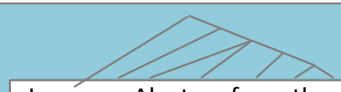
Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

TU Delft

Questions so far?

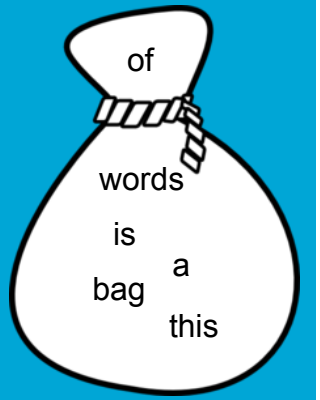
Part 1:
Administrative Issues

Applications of Natural Language Processing

Part 2:

Terminology: Components

Bag-of-words model



- Ignore order of words
- Ignore morphology/syntax (cats vs cat)
- No advanced semantics
- Just count matches between words
- Works pretty well!
- We know how to do better…

Some reasons why bag of words are limited…

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

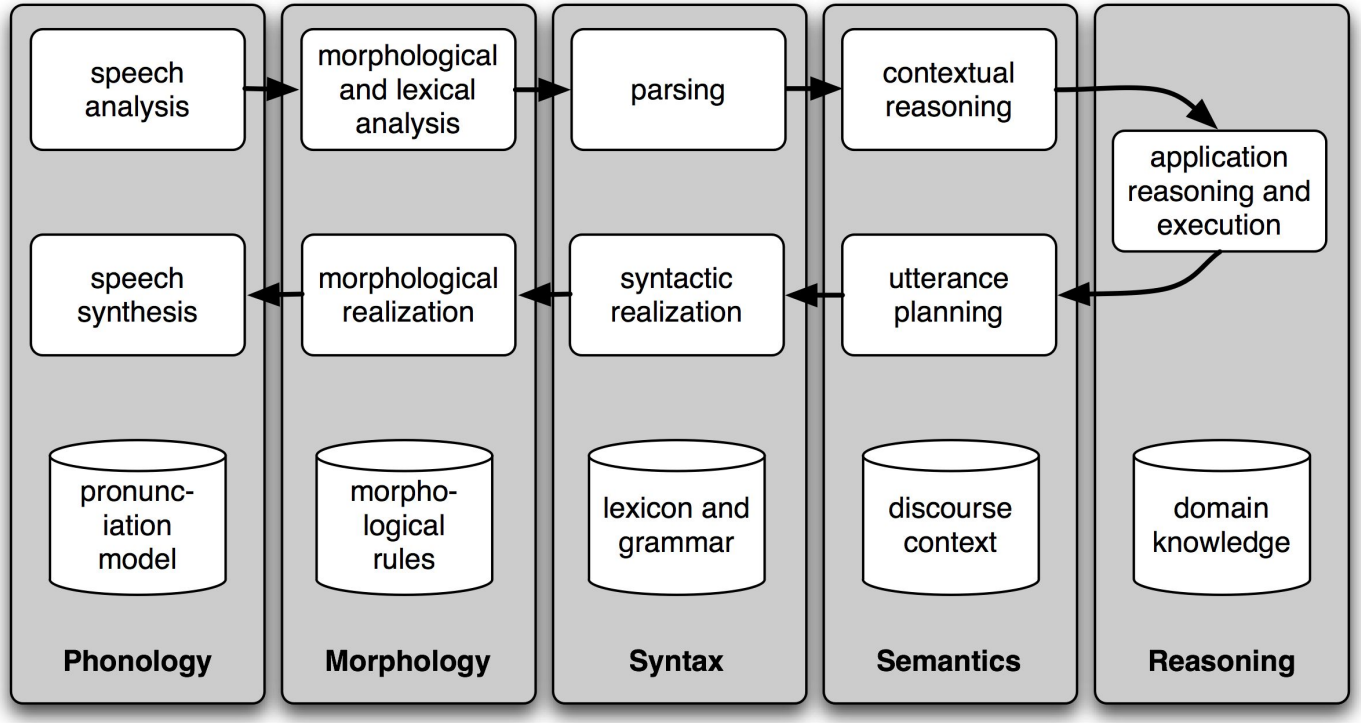
Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene

But that's what makes it fun!

Sub-tasks


speech analysis
morphological and lexical analysis
parsing
contextual reasoning
application reasoning and execution
speech synthesis
morphological realization
syntactic realization
utterance planning
pronunc-iation model
morpho-logical rules
lexicon and grammar
discourse context
domain knowledge
Phonology
Morphology
Syntax
Semantics
Reasoning

Sub-tasks

- **Phonology:** speech analysis and synthesis
- **Morphology:** Normalization, Stemming, Tokenization
- **Syntax:** Part-of-speech tagging, Parsing
- **Semantics:** (advanced) similarity, ontologies, dialog analysis
- **Reasoning:** domain and application knowledge

Morphology

- **Morphemes**:
  - The small meaningful units that make up words
- Sub-steps:
  - Word normalization, case folding
  - Word tokenization
  - Word lemmatization
  - Stemming
  - Sentence Segmentation

Normalization

- Need to “normalize” terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match ***U.S.A.*** and ***USA***
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: ***window*** Search: ***window, windows***
  - Enter: ***windows*** Search: ***Windows, windows, window***
  - Enter: ***Windows*** Search: ***Windows***
- Potentially more powerful, but less efficient

Case folding

- Applications like IR: reduce all letters to lower case
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
- For sentiment analysis, machine translation, Information extraction
  - Case is helpful (***US*** versus ***us*** is important)

Morphology

- **Morphemes**:
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems
    - Often with grammatical functions

Stemming

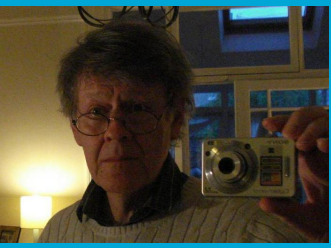
- Reduce terms to their stems in information retrieval
- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

Porter stemmer

M.F. Porter, 1980, An algorithm for suffix stripping, *Program*,**14**(3) pp 130−137.



- Good to use for indexing and want to support search using alternative forms of words
- Simple cascade rules
- Lexicon-free FST stemmer
- FST – finite-state transducer
  - Type of finite automaton which maps between two sets of symbols
- Can figure out exceptions e.g., Lying -> Lie
- Rules, common errors:
  - https://tartarus.org/martin/PorterStemmer/

Porter's algorithm
The most common English stemmer

Step 1a

sses	→ ss	caresses	→ caress
ies	→ i	ponies	→ poni
ss	→ ss	caress	→ caress
s	→ ø	cats	→ cat

Step 1b

(*v*)ing	→ ø	walking	→ walk
		sing	→ sing
(*v*)ed	→ ø	plastered	→ plaster

…

Step 2 (for long stems)

ational	→ ate	relational	→ relate
izer	→ ize	digitizer	→ digitize
ator	→ ate	operator	→ operate

…

Step 3 (for longer stems)

al	→ ø	revival	→ reviv
able	→ ø	adjustable	→ adjust
ate	→ ø	activate	→ activ

…

TU Delft

Errors in Porter

doing something you should not have done

NOT doing something you should have done

Errors of commissio		Errors of omission	
organization	organ	european	(europe)
doing	doe	analysis	(analyzes)
numerical	numerous	noise	(noisy)
policy	police	sparse	(sparsity)

Improvements

- Porter2
- Lancaster
- Snowball
  - A language for stemming algorithms
  - http://www.nltk.org/howto/stem.html

Lancaster stemmer

- Significantly more aggressive than the porter stemmer
- Faster
- Short words obfuscated
- More “exact” matching

Lemmatization

- Reduce inflections or variant forms to base form
  - *am, are, is → be*
  - *car, cars, car's, cars' → car*
- *the boy's cars are different colors → the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero (‘I want’), quieres (‘you want’) same lemma as querer ‘want’

Stopwords

- Stop list – high frequency words that may not contain much information
- E.g., `it', `and', `a', `the'…
- Fiction titles published between 1660 and 1799
  - https://dlsatbsu.wordpress.com/

the, of, and, in, or,

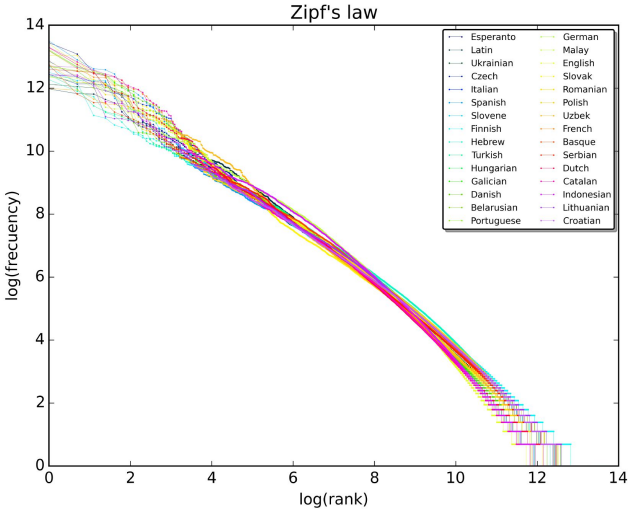
a, by, his, history, volumes

his, two, from, history, miss,

papers, adventures, mr, with, de, s

(George Kingslev) Zipf's law

- The frequency of any word is inversely proportional to its rank in the frequency table.
- *"Popular words are mentioned a lot more than unpopular words".*
- Holds for most languages.



Rank	Word	Frequency
5	a	10144200
2201	abandon	15323
783	ability	51476

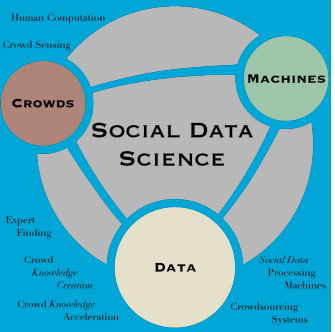
Tokenization

"`When I'M a Duchess, she said to herself,' (not in a very hopeful tone though), 'I won't have any pepper in the my kitchen AT ALL. Soup does very well without—Maybe it's always pepper that makes people hot-tempered,'…"

- Split on non-word character?
  - Space beginning and end
- How about punctuation? E.g., "(.,"
- How about numbers?

dark horse
get cold feet
lose face
throw in the towel

Questions so far?



TU Delft

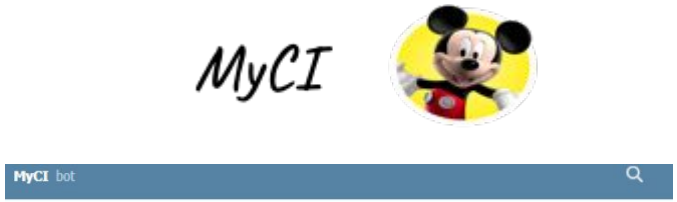
If you liked this course you'll also like…

CS4145 - Crowd Computing

How large groups of people can solve complex tasks that are currently beyond the capabilities of AI, and that cannot be solved by a single person alone.

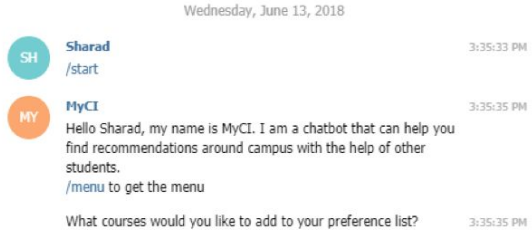
Gamification, conversational systems, human computation, and much much more...

HUMAN AIDED CHATBOT FOR CAMPUS INFORMATION RETRIEVAL



Background: A new emerging hybrid chatbot system has spawned using human-based computation to improve response flexibility and request comprehension of chatbots.

In this project, we will test our human aided chatbot based on telegram. Students will use this chatbot to request or provide campus information. Through this project, we will find an effective user interface to improve the quality of results provided by workers. We'll also compare the effectiveness and efficiency of traditional chatbot, hybrid chatbot and webpage-based crowdsourcing platform, by conducting several experiments among students.

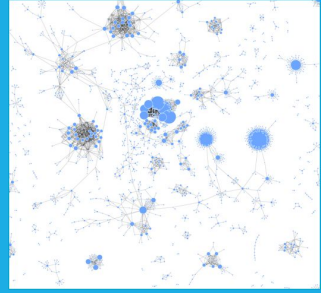


Trasherhunt

Background: Crowdsensing based on street view means the way to collect data by a group of people using web-based street panorama viewer.

In this project, we will test our crowdsensing platforms based on street panorama viewers (panorama data comes from Google and Amsterdam Data Science). Students are supposed to use this platform to find dirty/clean spots and trash bins in Noord-Zuid Lijn area (in Amsterdam). We will then compare the trash bin information collected by our platforms (Google and ADS) with OpenStreetMap (and ADS map). We'll also analyse the relationship between "cleanness/dirtiness" and the density of trash bins.

RBUTR

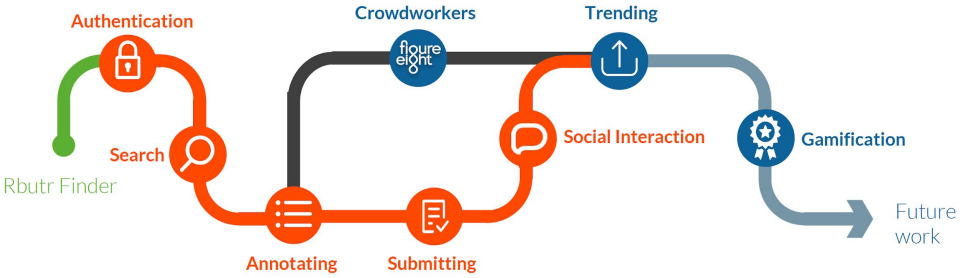


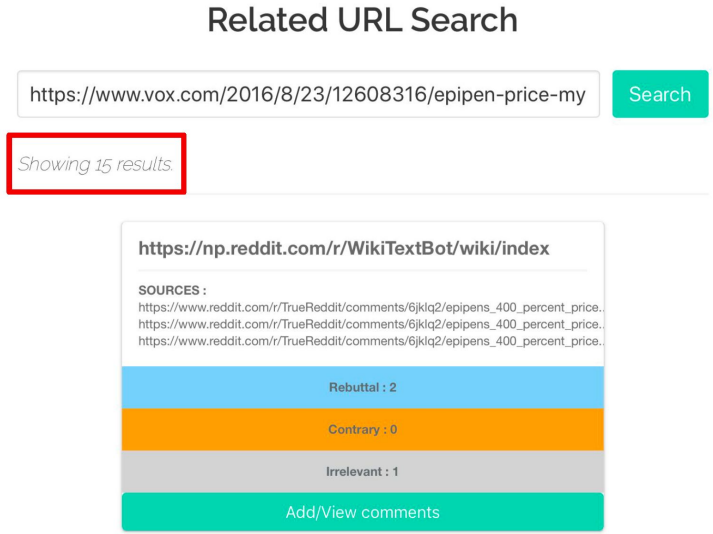
- Created 2012
- 2K active users per week
- 60K total users
- 28K edges, 40K nodes

Encourage users to annotate relationship between nodes:

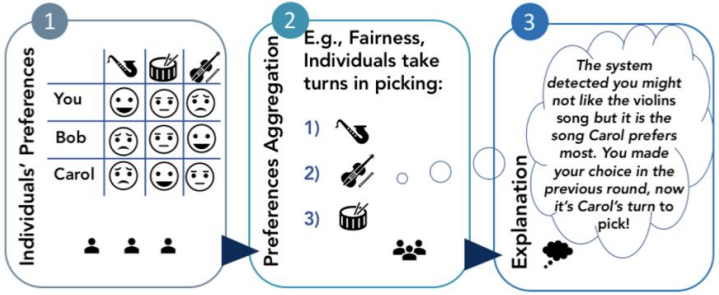
E.g., rebuttal, disagree, irrelevant

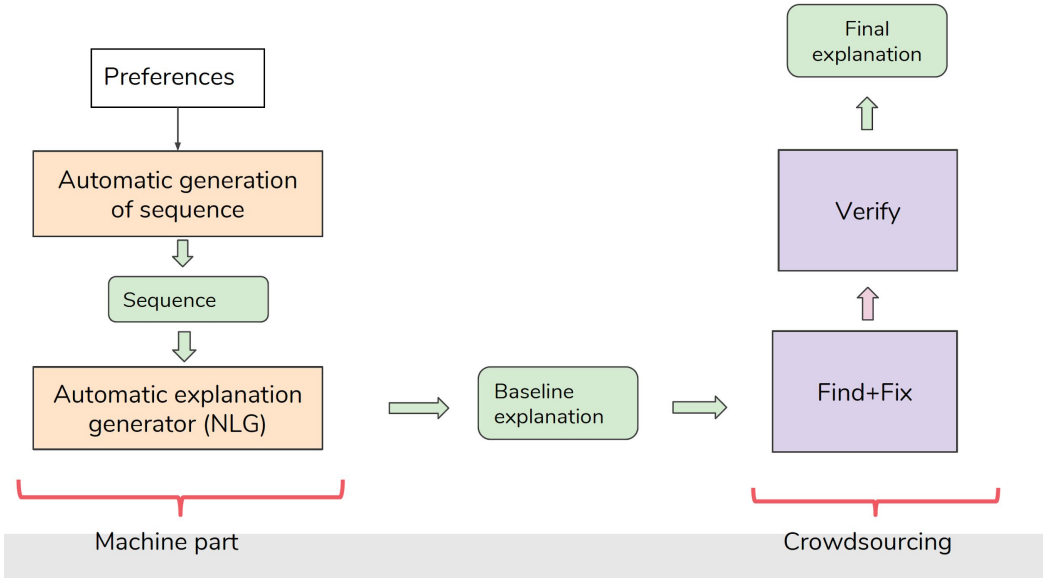
Use this information for an information seeking task.





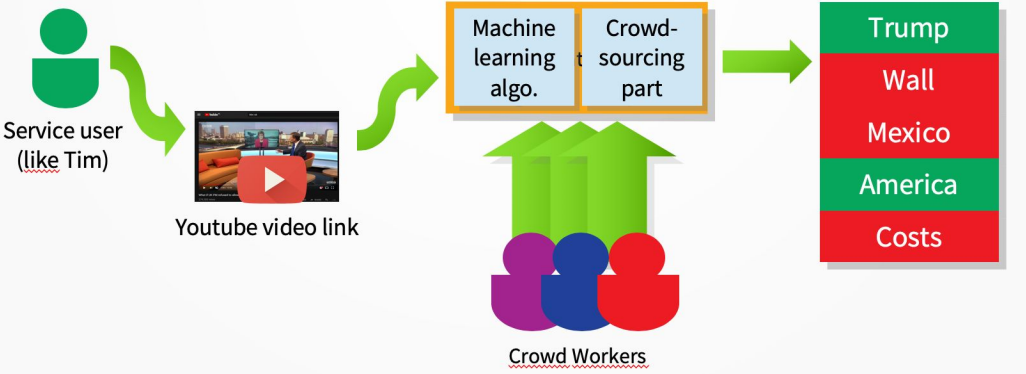
EXPLAINING SEQUENCES OF RECOMMENDATIONS





Use human computation to generate and evaluate explanations.

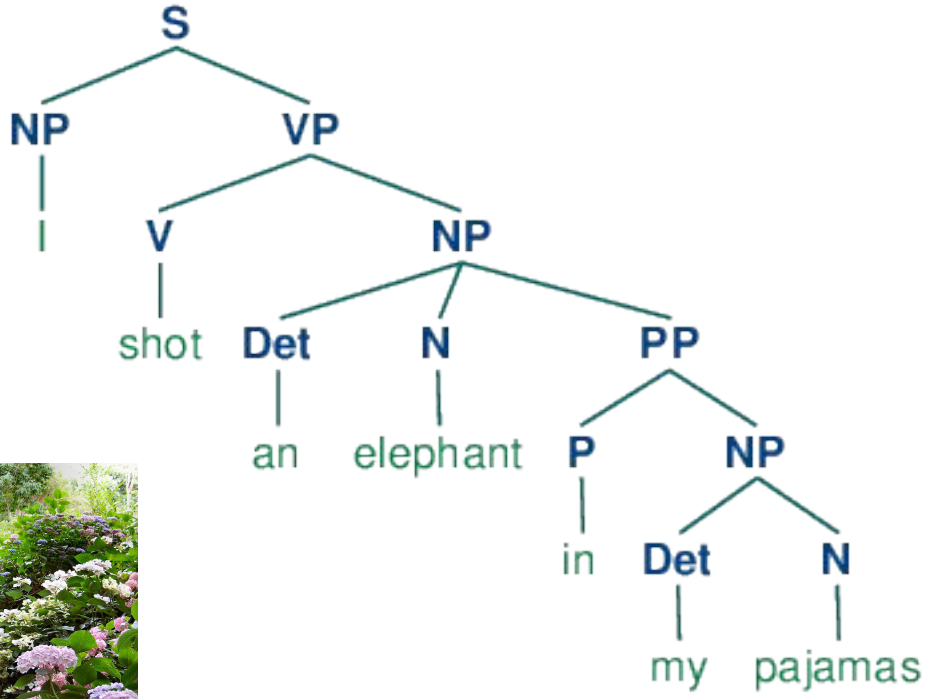
CAPTURING BIAS IN VISUAL NEWS SOURCES

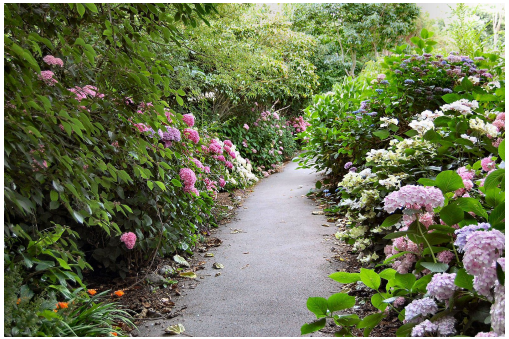


Background: Using experts to harness ground truth can be tedious, time consuming and expensive. That is why with the help of crowdsourcing platforms and computer interfaces we could train a crowd for task that we would normally need several experts for. Additionally we could assist experts with their job.

In this project we want to help humanities and social science experts to codify videos with the help of the crowd. To be able to help the experts we will need to be able to cut the videos into smaller chunks and assign them to the Crowd. A useful library and tool to cut a complex task into simpler ones could be CrowdForge from Google Research. We will experiment with a specific example on coding bias with the use of an expert example and will scale it up with the use of the crowd in Figure8 platform.

Syntax





Ambiguity makes NLP hard: “Crash blossoms”

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Ambiguity is pervasive

New York Times headline (17 May 2000)

Fed raises interest rates

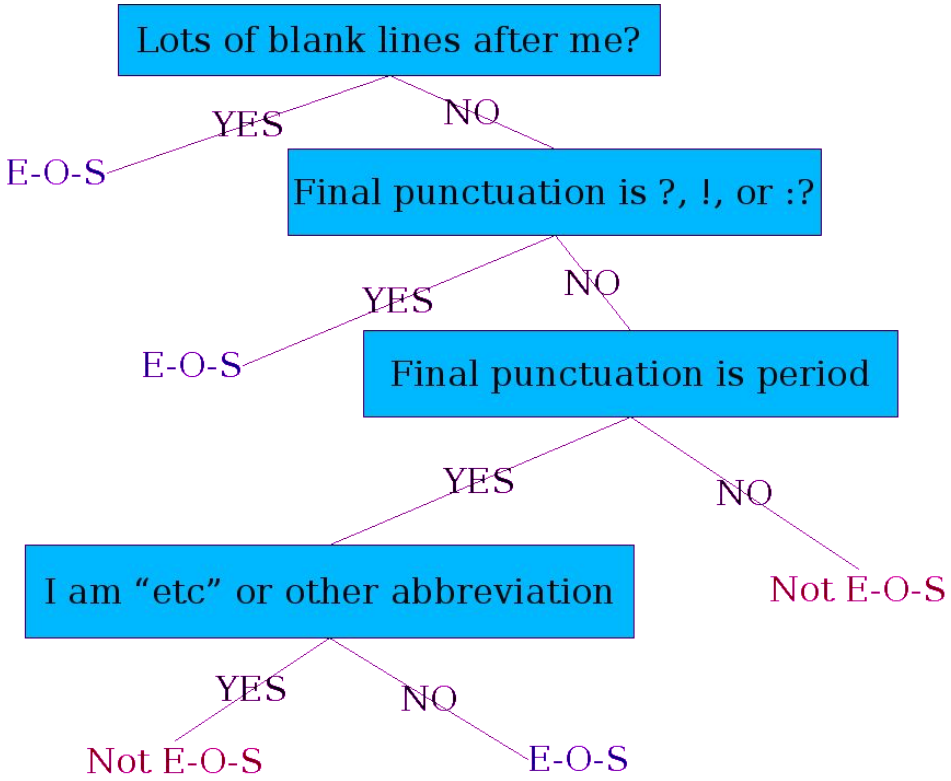
Fed raises interest rates

Fed raises interest rates 0.5%

Sentence Segmentation

- !, ? are relatively unambiguous
- Period “.” is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a “.”
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning
- NLTK uses Punkt (Kiss & Strunk, 2006)

Determining if a word is end-of-sentence: a Decision Tree



More sophisticated decision tree features

- Case of word with ".": Upper, Lower, Cap, Number
- Case of word after ".": Upper, Lower, Cap, Number

- Numeric features
  - Length of word with "."
  - Probability (word with "." occurs at end-of-s)
  - Probability (word after "." occurs at beginning-of-s)

Implementing Decision Trees

- A decision tree is just an if-then-else statement
- The interesting research is choosing the features
- Setting up the structure is often too hard to do by hand
  - Hand-building only possible for very simple features, domains
    - For numeric features, it's too hard to pick each threshold
  - Instead, structure usually learned by machine learning from a training corpus

Decision Trees and other classifiers

- We can think of the questions in a decision tree
- As features that could be exploited by any kind of classifier
  - Naive Bayes
  - Logistic regression
  - SVM
  - Neural Nets
  - etc.

Models and algorithms

- **Models**
  - State machines (e.g., automata)
  - Rule systems (i.e., grammars),
  - Logic (e.g., first order logic/predicate calculus),
  - Probabilistic models (e.g., hidden Markov models)
  - Vector-space models (i.e., linear algebra)
- **Algorithms**
  - State space search (for e.g., speech recognition, syntactic parse, translation hypothesis)
  - Machine learning (e.g. Classifiers, Expectation-Maximization, and sequence models)

After today you should be able to…

- Explain the function of Natural Language Processing
- Identify NLP applications
- Identify typical natural processing tasks
- Recognize typical components/sub-tasks

Next deadlines:

- (No reading today)
- Review P8: handed out March 15, **due March 22**.
- NLP project proposal: **due March 22**.

TU Delft

Next week...

Syntax (cont.)

- Part-of-speech (POS) tagging
- MM POS tagging: N-grams

- Sentiment analysis
- Named-entity recognition

Questions?

Room: 4.900 VMB 6 (Van Mourik Broekmanweg)

Office hours: Fridays 9-11:30pm (by appointment)

Email: **ewi-4325@tudelft.nl**

Or slack: in4325-ewi2019@tudelft.nl

Credits: Many of these slides are modified from the Stanford NLP course: https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html

TU Delft