Information Retrieval (IN4325)

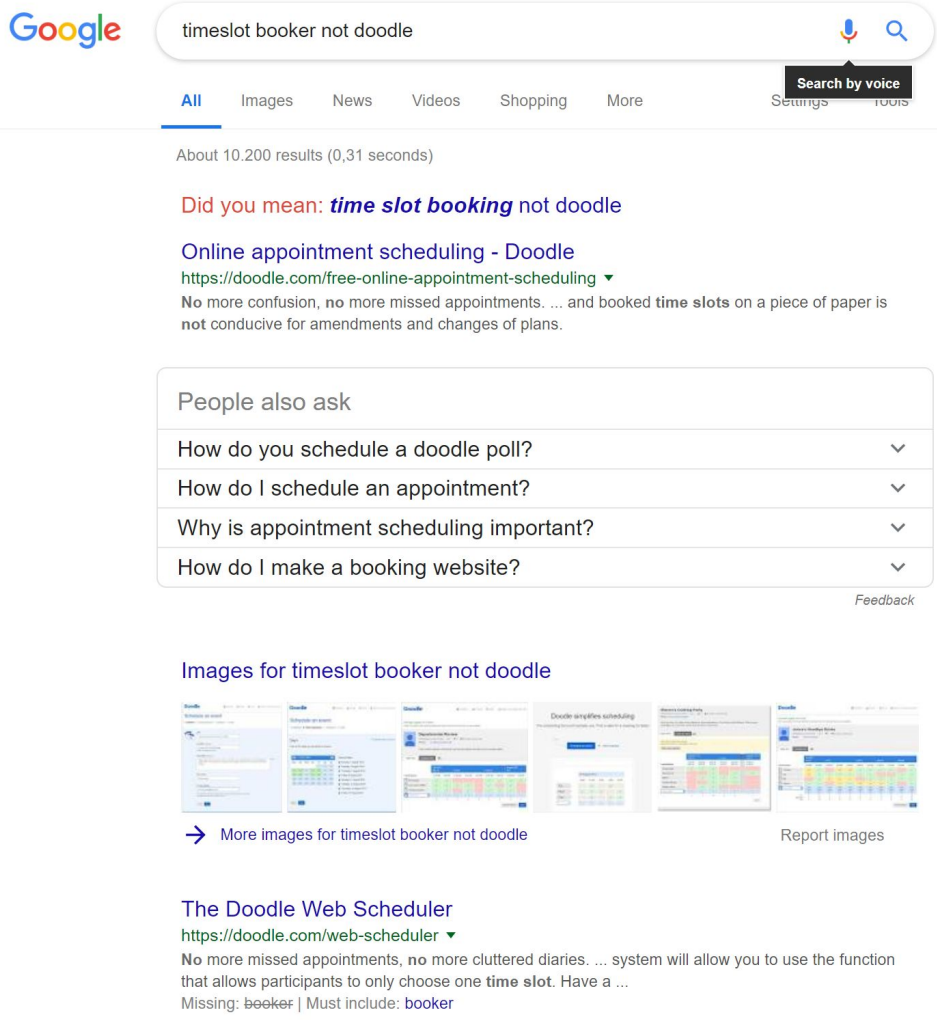
Introduction

Dr. Nava Tintarev
Assistant Professor, TU Delft

- IR deals with the representation, storage, organization of, and access to largely **unstructured** information

- Central notions are **information needs** and **relevance**

- IR has its roots in the **library and information sciences**



crawling and indexing

vertical selection

query suggestions

result ranking

snippet generation

implicit feedback

entity cards

Google

time slot booking but not doodle

Search by voice

All Images News Videos Shopping More Settings Tools

About 9.150.000 results (0,48 seconds)

Online appointment scheduling - Doodle
https://doodle.com/free-online-appointment-scheduling

No more confusion, **no** more missed appointments. ... **and booked time slots** on a piece of paper is **not** conducive for amendments **and** changes of plans.

quick and easy online booking system - Doodle
https://doodle.com/online-booking-system

Use **Doodle** to **schedule** events with friends **and** colleagues. ... **Doodle** is an online **booking** system where you can quickly **and** without hassle arrange a **time** to ... most fun **and** relaxed (if **not** the most fun **and** relaxed!) online **booking** system on ...

Staff scheduling made easier with Doodle
https://doodle.com/staff-scheduling

Doodle already has a staff **schedule** template ready for you to adapt **and** send ... You can be confident that your staff scheduling will be done on **time and** in a ... **No** more double-**booked** meetings **and no** more double-fudge sundaes for you!

Free Online Booking Software from Doodle
https://doodle.com/booking-software

Doodle Booking Software lets you easily find **time** to meet up with friends **and** colleagues. Connect your online calendar **and schedule** your week all in one place. ... **and** 'no', when the participants **are** marking their availability for each **time**.

Easily arrange meetings with Doodle's appointment calendar
https://doodle.com/appointment-calendar

Arrange meetings **and** organise your **schedule** with **Doodle's** simple ... This is a particular challenge when colleagues or friends **are not** immediately ... Once created, you can invite your friends or colleagues **and** decide upon which **time** is best ...

Organise your work schedule with Doodle
https://doodle.com/work-schedule

On **Doodle**.com you can set up an event poll with various dates **and times and** send ... with the rest of your office team by showing when you **are not** available ... **schedule** software applications can easily

Klantenservice

Hulp nodig? We zijn er elke dag en nacht speciaal voor jou!

Hallo. Waarmee kan ik je van dienst zijn?

wie ben je?

Mijn naam is Billie, ik ben de digitale hulp van bol.com. Ik help je graag met het beantwoorden van vragen over bol.com. Waarmee kan ik je helpen?

hoe oud ben je?

Ik zag het levenslicht in december 2008. Maar goed, volgens mij kunnen we beter gewoon verder gaan met ons gesprek. Waarmee kan ik je helpen?

Je vraag of bericht

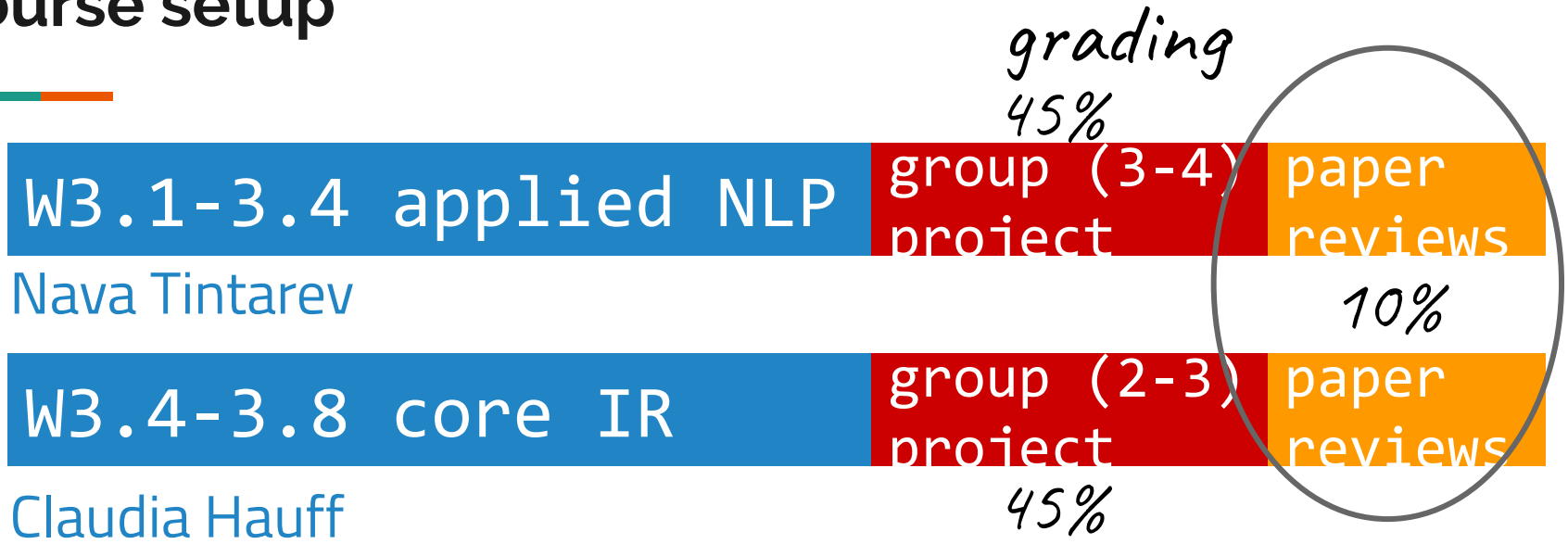
Bijvoorbeeld "Wat moet ik doen met een beschadigd artikel?"

Versturen

applied NLP module

core IR module

Course setup



To pass: grade of **5+** in both group projects and **6** out of 8 offered paper reviews completed with a *sufficient*.

Weekly support hours, starting in week 3.2.

github.com/chauff/IN4325

detailed

Applied NLP project

There are three expected outputs:

1. **Project proposal** (mandatory, but ungraded - you will receive feedback).
2. **Intermediate project report** (mandatory, but ungraded - you will receive feedback).
3. **Final project report** (mandatory, graded).

Group projects are conducted in groups of 3-4 students. Please enroll together with your team members in group on Bright

You can c
own.

Restrictions

Based on prior experience we put the following restrictions on your choice of project, no matter if you reproduce a paper or follow your own research idea:

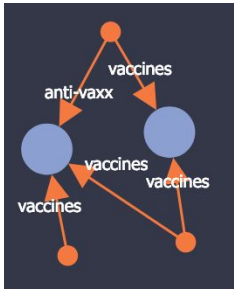
- **You can only conduct a project on neural NLP if you have successfully completed the Deep Learning course beforehand.**
- The main focus of your project is *NLP*. If you build a classifier, the features you study should be motivated by linguistic theory/previous findings in NLP.
- If you use off the shelf solutions: be aware of the defaults and motivate why they are suitable for your problem.
- Not all data is created equal -- what kind of biases might come from the dataset you are using?
- The project focuses on **textual data** (not video/audio/genomics/.... data).

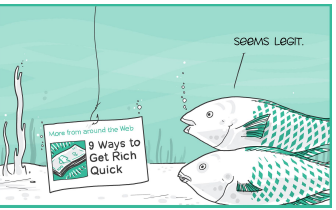
We have a first deadline in week 3.2, so you will get early feedback on the suitability of your proposal.

- NLP resources
  - Books
  - Software
  - Open Source Natural Language Processing Tools
  - Datasets
- Lecture 1: NLP Introduction
  - Recommended readings
- Lecture 2: Syntax
  - Recommended readings
  - Recommended resources:
  - ⚠️ Paper P1 to review
- Lecture 3: Semantics
  - Recommended readings
  - Recommended resources:
- Lecture 4: Evaluation NLP
  - Recommended readings
  - Recommended resources:
  - ⚠️ Paper P2 to review
- Lecture 5: ML for NLP
  - Recommended readings
  - Recommended resources:
- Lecture 6: Natural Language Generation
  - Recommended readings
  - Recommended resources
  - ⚠️ Paper P3 to review
- Lecture 7: NLP annotations
  - Recommended readings
- Lecture 8: Word embeddings
  - Recommended readings
  - ⚠️ Paper P4 to review

Group project

#find_a_team_member





- **What:** Design, develop, and evaluate a system
- **Group size:** 3-4
- **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

Optional support hours:

Fridays 9am-11:30pm (W3.2 - 3.4)

Sign up for one 15 minute slot!

Team: Oana Inel (postdoc), Mesut Kaya (PhD), Priya Sarkar (TA)

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

What will I learn?

Week	Thursday lecture	Friday lecture	Group project	No. of reviews
3.1	NLP introduction (`Pathe Zal 6` , N. Tintarev)	Text analysis (`BK-CZ B` , N. Tintarev)	NLP Project group settled	1
3.2	Semantics (`Pathe Zal 6` , N. Tintarev)	Evaluation NLP (`BK-CZ B` , N. Tintarev)	NLP project settled	1
3.3	ML for NLP (`Pathe Zal 6` , N. Tintarev)	Language generation (`BK-CZ B` , N. Tintarev)		1
3.4	NLP annotations (`Pathe Zal 6` , O. Inel)	Word embeddings (`BK-CZ B` , N. Tintarev)	Intermediate applied NLP report due	1

Brightspace

- Forming project groups
- Submission of reviews / project reports
- Grading

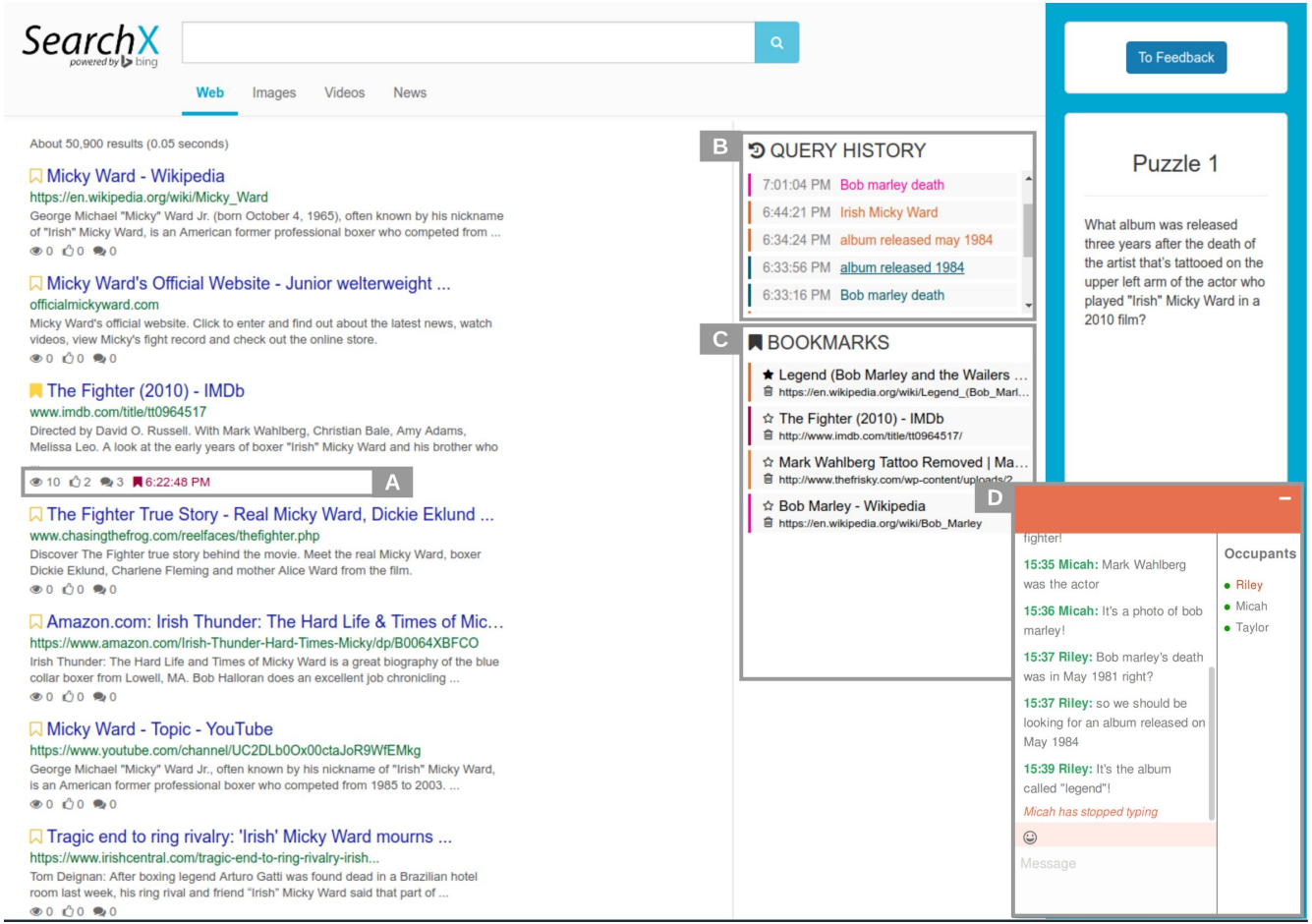
Slack: join.slack.com/t/in4325/signup

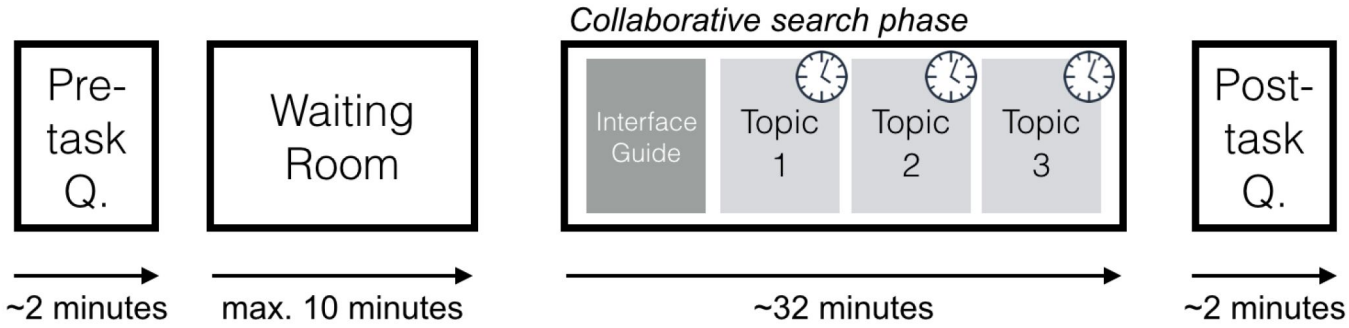
- Questions to the course team

Email: in4325-ewi@tudelft.nl

- Responsible instructors only

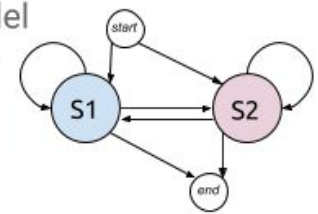
Quickly: our related research


SearchX
powered by bing
Web
Images
Videos
News
About 50,900 results (0.05 seconds)
Micky Ward - Wikipedia
https://en.wikipedia.org/wiki/Micky_Ward
George Michael "Micky" Ward Jr. (born October 4, 1965), often known by his nickname of "Irish" Micky Ward, is an American former professional boxer who competed from ...
0 0 0
Micky Ward's Official Website - Junior welterweight ...
officialmickyward.com
Micky Ward's official website. Click to enter and find out about the latest news, watch videos, view Micky's fight record and check out the online store.
0 0 0
The Fighter (2010) - IMDb
www.imdb.com/title/tt0964517
Directed by David O. Russell. With Mark Wahlberg, Christian Bale, Amy Adams, Melissa Leo. A look at the early years of boxer "Irish" Micky Ward and his brother who
10 2 3 6:22:48 PM
A
The Fighter True Story - Real Micky Ward, Dickie Eklund ...
www.chasingthefrog.com/reelfaces/thefighter.php
Discover The Fighter true story behind the movie. Meet the real Micky Ward, boxer Dickie Eklund, Charlene Fleming and mother Alice Ward from the film.
0 0 0
Amazon.com: Irish Thunder: The Hard Life & Times of Mic...
https://www.amazon.com/Irish-Thunder-Hard-Times-Micky/dp/B0064XBFCO
Irish Thunder: The Hard Life and Times of Micky Ward is a great biography of the blue collar boxer from Lowell, MA. Bob Halloran does an excellent job chronicling ...
0 0 0
Micky Ward - Topic - YouTube
https://www.youtube.com/channel/UC2DLb0Ox00ctaJoR9WfEMkg
George Michael "Micky" Ward Jr., often known by his nickname of "Irish" Micky Ward, is an American former professional boxer who competed from 1985 to 2003. ...
0 0 0
Tragic end to ring rivalry: 'Irish' Micky Ward mourns ...
https://www.irishcentral.com/tragic-end-to-ring-rivalry-irish...
Tom Deignan: After boxing legend Arturo Gatti was found dead in a Brazilian hotel room last week, his ring rival and friend "Irish" Micky Ward said that part of ...
0 0 0
B
QUERY HISTORY
7:01:04 PM Bob marley death
6:44:21 PM Irish Micky Ward
6:34:24 PM album released may 1984
6:33:56 PM album released 1984
6:33:16 PM Bob marley death
C
BOOKMARKS
Legend (Bob Marley and the Wailers ...
https://en.wikipedia.org/wiki/Legend_(Bob_Marl...
The Fighter (2010) - IMDb
http://www.imdb.com/title/tt0964517/
Mark Wahlberg Tattoo Removed | Ma...
http://www.thefrisky.com/wp-content/uploads/2...
Bob Marley - Wikipedia
https://en.wikipedia.org/wiki/Bob_Marley
To Feedback
Puzzle 1
What album was released three years after the death of the artist that's tattooed on the upper left arm of the actor who played "Irish" Micky Ward in a 2010 film?
D
fighter!
15:35 Micah: Mark Wahlberg was the actor
15:36 Micah: It's a photo of bob marley!
15:37 Riley: Bob marley's death was in May 1981 right?
15:37 Riley: so we should be looking for an album released on May 1984
15:39 Riley: It's the album called "legend"!
Micah has stopped typing
Message
Occupants
Riley
Micah
Taylor



Conversational search goals model

S1 Information-need elucidation

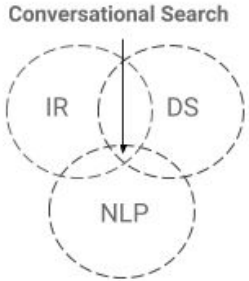


S2 Information presentation

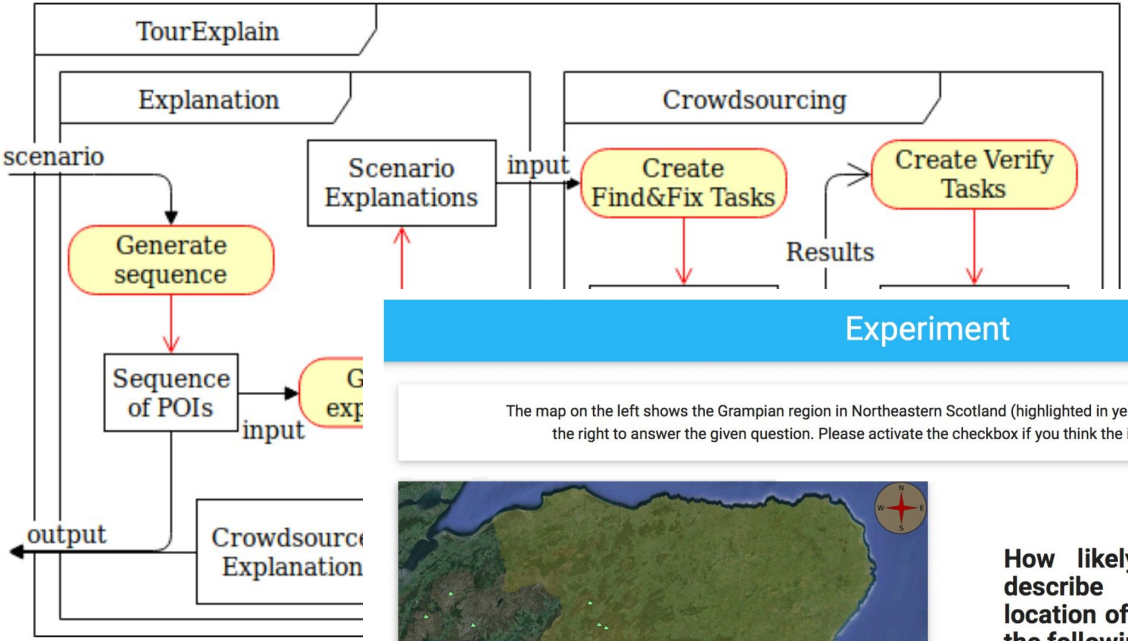
Related tasks and fields

IR	Query suggestion	
IR	Query disambiguation	
IR	Rank/generate clarification questions	
DS	Belief/Dialog state tracking	
DS	Slot-filling	
DS	Intent/Domain prediction	
DS	Slot tagging	NLP
	Word sense disambiguation	NLP

DS Dialog policy learning

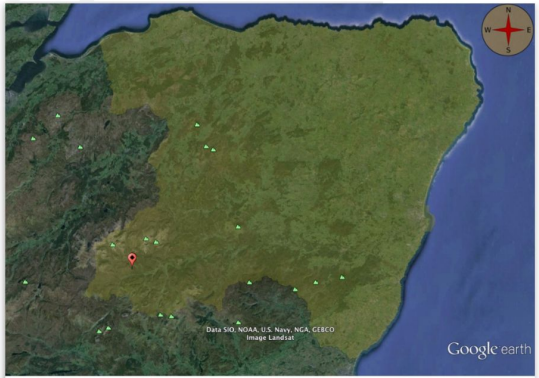


IR	Adhoc retrieval
IR	Document re-ranking
IR	Recommendation
NLP	Text summarization
NLP	Answer ranking/generation
NLP	Machine Reading comprehension



Experiment

The map on the left shows the Grampian region in Northeastern Scotland (highlighted in yellow). Please adjust the slider on the right to answer the given question. Please activate the checkbox if you think the initial middle value is fine.



Location marker Mountain

How likely is it for you to describe the geographical location of the red marker using the following term?

Coastal Grampian

Very unlikely — Very likely

I agree with the original value

Progress: 1/66

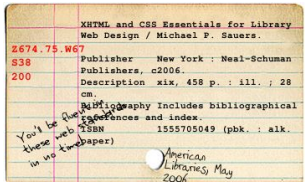
NEXT QUESTION

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

Hi. I'm Cortana.

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

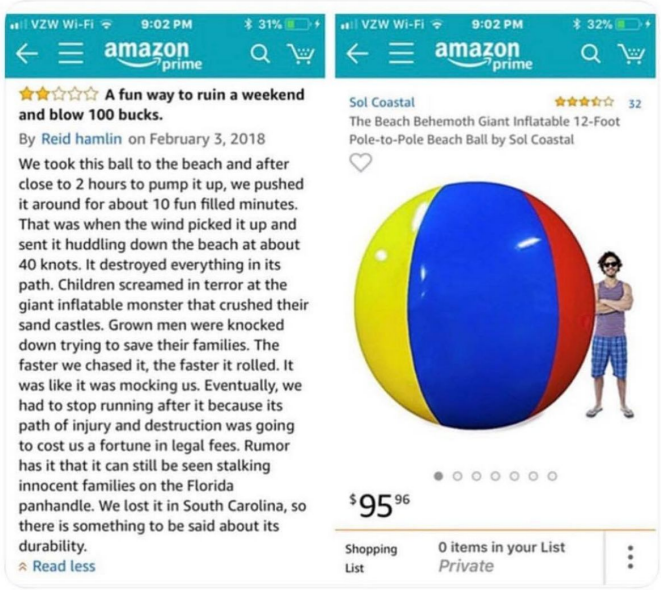
Capital and largest city (2011 est.): Buenos Aires, 13.528 million

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
ıdes
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**

Natural language processing

- From user to system
- From system to user



Part 1:
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

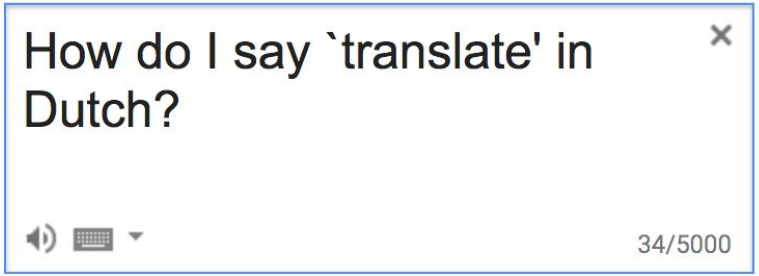
"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."

CORRECT CLASSIFICATION: DISAGREE

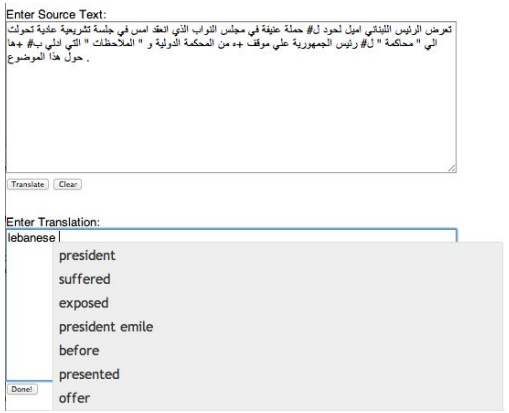
- Input: Headline + text
- Output: Classify stance (e.g., agrees, disagrees, discusses, unrelated)

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.

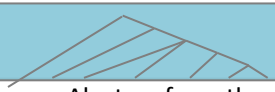
Coreference resolution

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?

Questions so far?

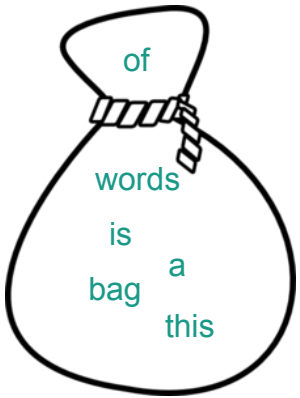
Part 1:

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

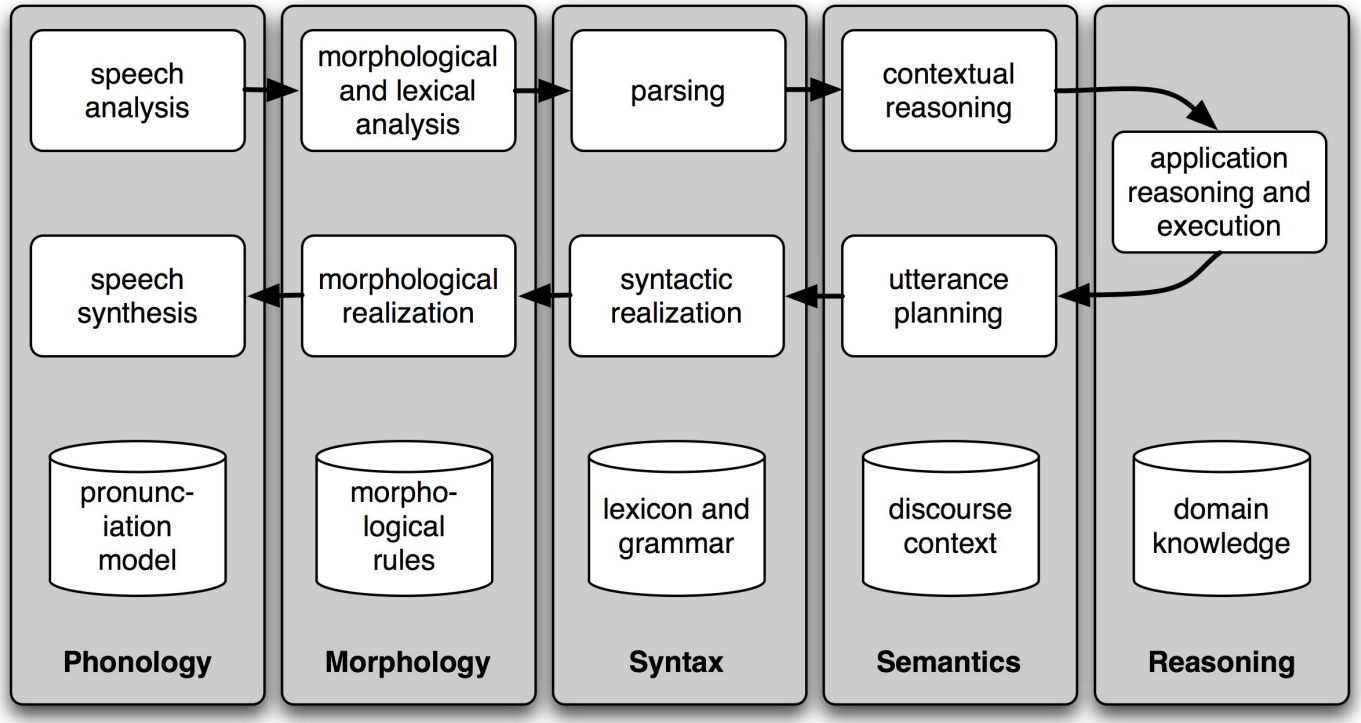
Where is *A Bug's Life* playing …

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

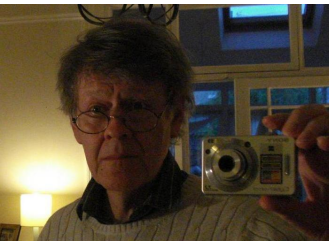
- Reduce terms to their stems in information retrieval
- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

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
…
```

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

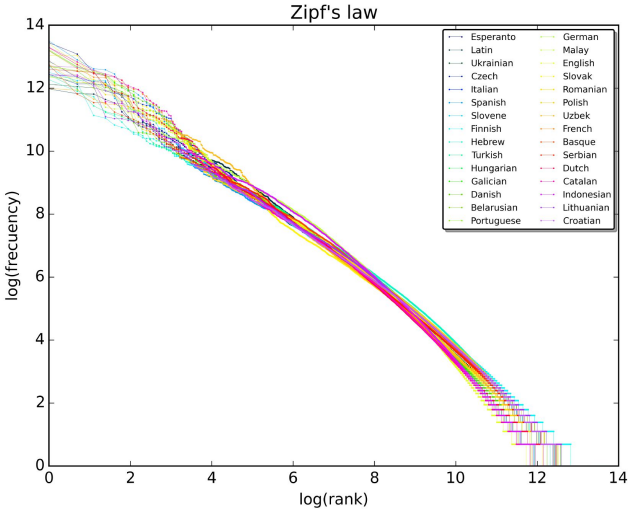
- Reduce inflections or variant forms to base form
  - *am, are, is → be*
  - *car, cars, car's, cars' → car*
- *the boy's cars are different colors → the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero (‘I want’), quieres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

(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

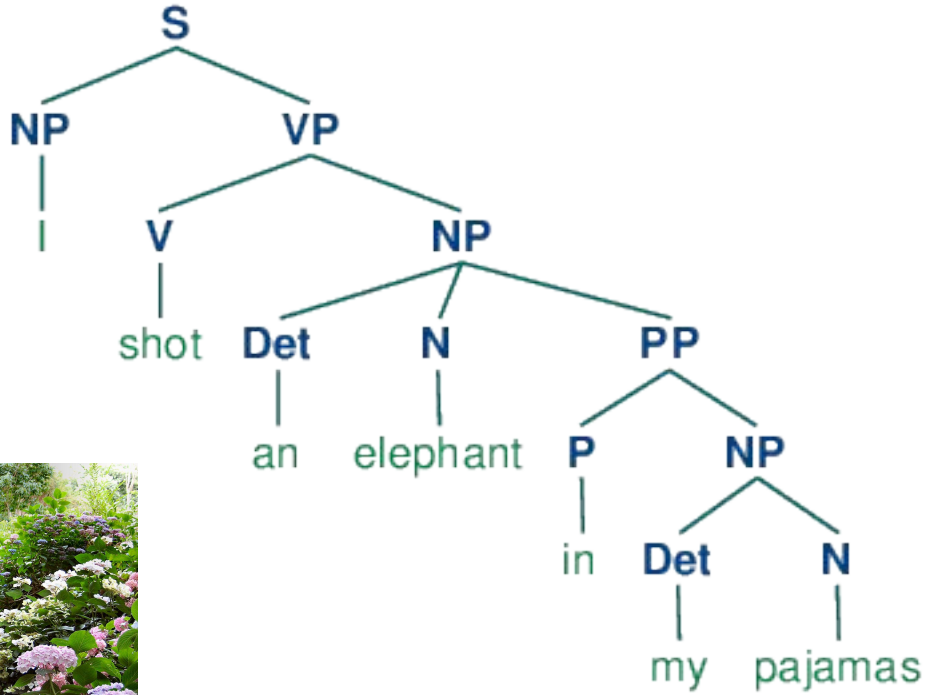
"`When I'M a Duchess, she said to herself,' (not in a very hopeful tone though), 'I won't have any pepper in the my kitchen AT ALL. Soup does very well without—Maybe it's always pepper that makes people hot-tempered,'…"

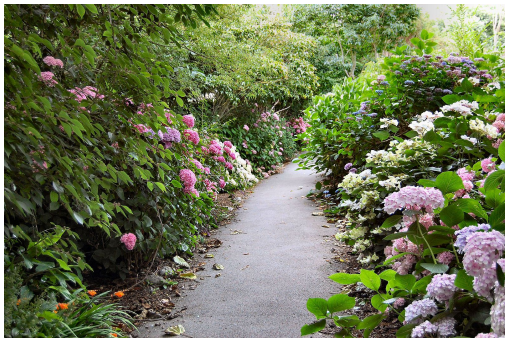
- Split on non-word character?
  - Space beginning and end
- How about punctuation? E.g., "(.,"
- How about numbers?

dark horse
get cold feet
lose face
throw in the towel

Questions so far?

Syntax





Ambiguity makes NLP hard: “Crash blossoms”

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

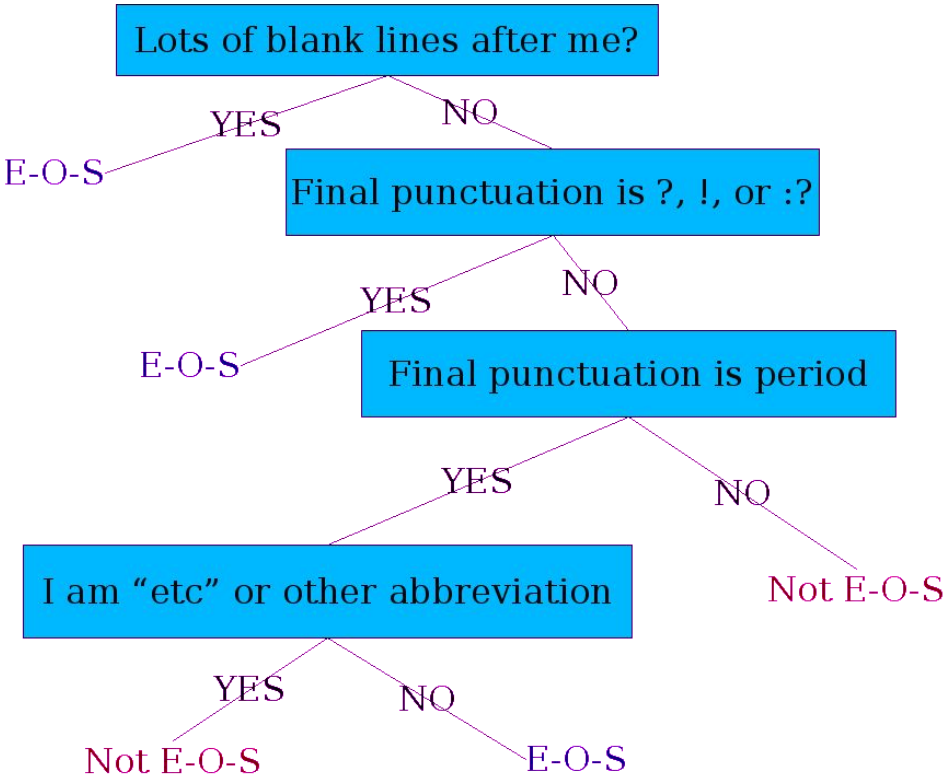
Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

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)

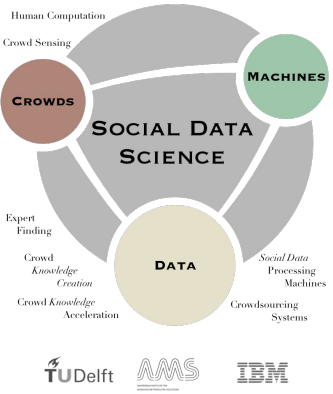
After today you should be able to…

- Explain the function of Natural Language Processing
- Identify NLP applications
  - dialog systems, search, chatbots
- Identify typical natural processing tasks
  - Sentiment analysis, spell checking
- Recognize typical components/sub-tasks
  - Tokenization, Stemming, Lemmatization

Next deadlines:

- Decide on a group: due Friday **Feb. 14th**
- Review P1: handed out Feb 14th, due **Feb 21st.**
- NLP project proposal: due **Friday Feb. 21st**.

If you liked this course you'll also like…

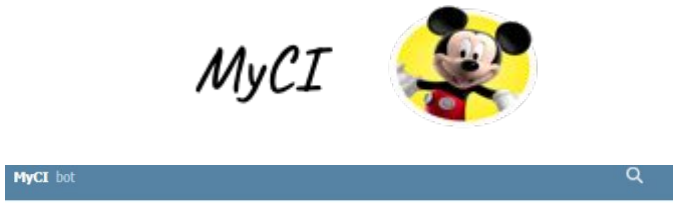


CS4145 - Crowd Computing

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



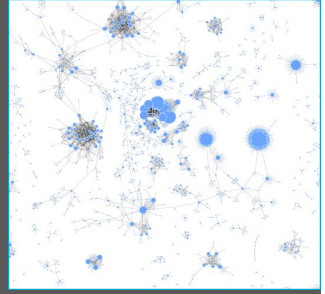
Background: A new emerging hybrid chatbot system has spawned using human-based computation to improve response flexibility and request comprehension of chatbots.

- 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.

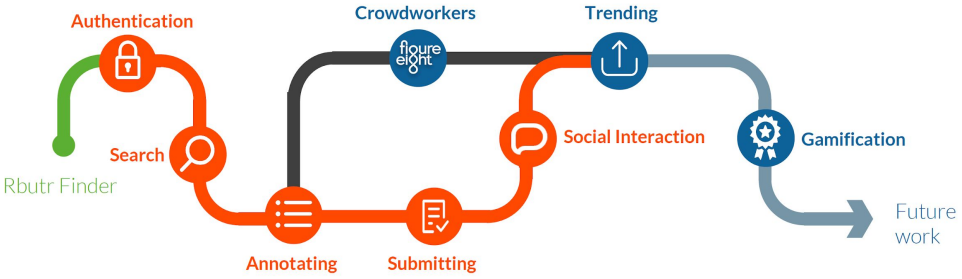
- In this project, we will test our crowdsensing platforms based on street panorama viewers (panorama data comes from Google and Amsterdam Data Science). Students are supposed to use this platform to find dirty/clean spots and trash bins in Noord-Zuid Lijn area (in Amsterdam). We will then compare the 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.

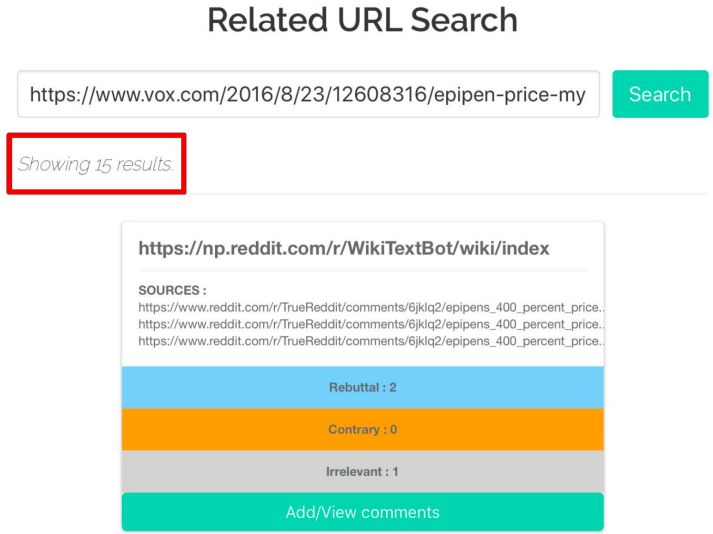


RBUTR

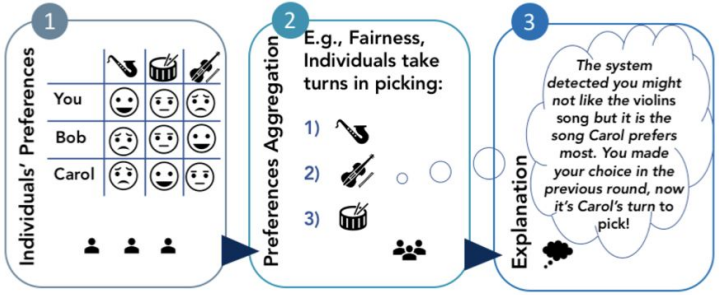
- Created 2012
- 2K active users per week
- 60K total users
- 28K edges, 40K nodes

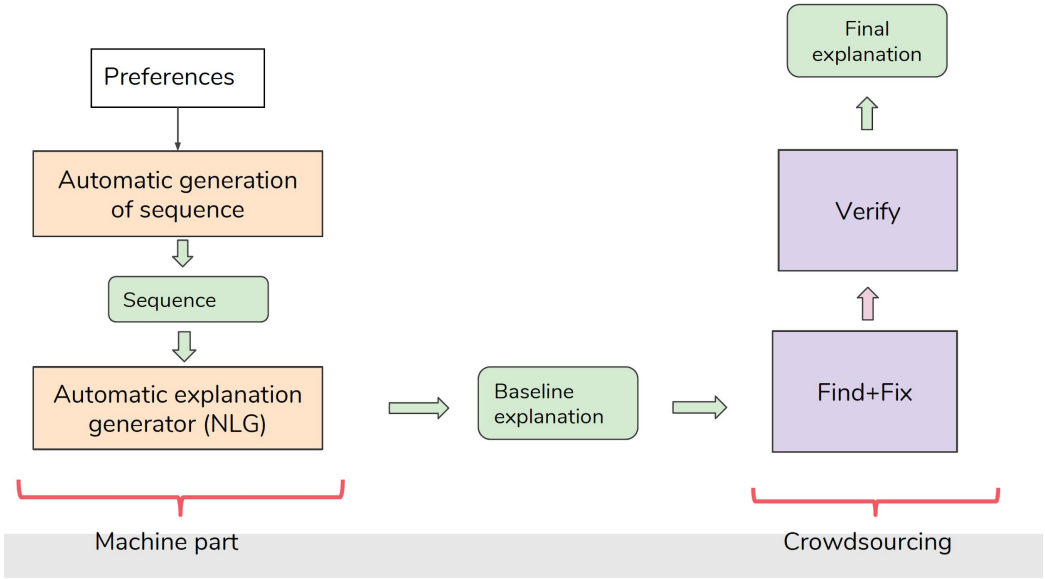
Encourage users to





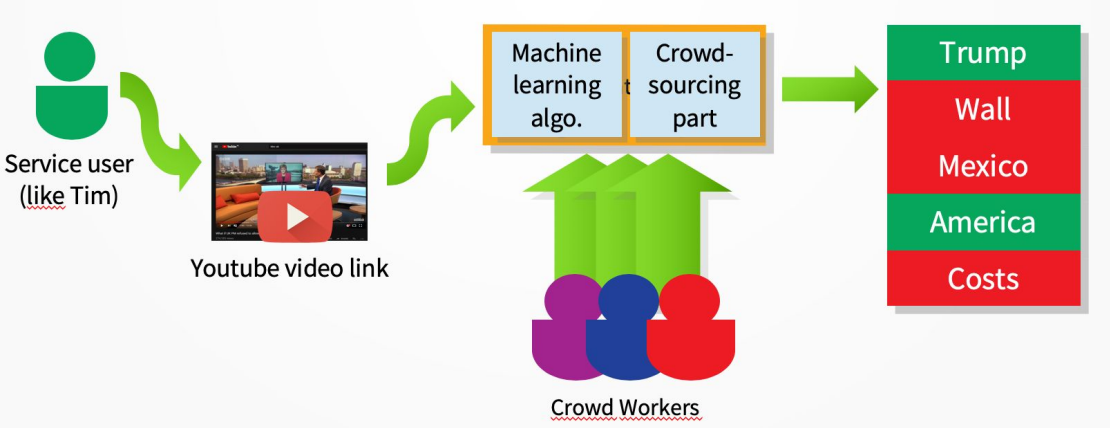
EXPLAINING SEQUENCES OF RECOMMENDATIONS





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.

Tomorrow

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@tudelft.nl

Credits: Many of these slides are modified from the Stanford NLP course:

https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html