Computational Linguistics 2014-2015

- **Walter Daelemans** (walter.daelemans@uantwerpen.be)
- **Guy De Pauw** (guy.depauw@uantwerpen.be)
- **Mike Kestemont** (mike.kestemont@uantwerpen.be)

http://www.clips.uantwerpen.be/cl1415

Practical

Location	P0.11 (Scribanihuis)
Reading material	• D. Jurafsky & J.H. Martin (2009) Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed). Pearson Education, USA. • Natural Language Processing with Python
Software	Python 3.4 and NLTK: Installation Instructions
Evaluation	Take-home assignments and oral examination
Lecturers	Walter Daelemans: walter.daelemans@uantwerpen.be Mike Kestemont: mike.kestemont@uantwerpen.be Guy De Pauw: guy.depauw@uantwerpen.be

Program

Session	Day	Date	Chapter	Topic	Reading Assignment	Slides	Take-home Assignment
1	Monday	29/9/2014	**Python**	Session 1 - Variables	See Github		
2	Thursday	2/10/2014	**Python**	Session 2 - Collections			
3	Monday	6/10/2014	**Python**	Session 3 - Conditions (and an introduction to loops)			
4	Thursday	9/10/2014	**Python**	Session 4 - Loops			
5	Monday	13/10/2014	**Python**	Session 5 - Reading and writing to files			
6	Thursday	16/10/2014	**Python**	Session 6 - Writing your own Functions and importing packages			
7	Monday	20/10/2014	**Python**	Session 7 - Regular Expressions in Python			
8	Thursday	23/10/2014	**Python**	Session 8 - Advanced looping in Python and list comprehensions			
9	Monday	27/10/2014	**Theory**	Introduction to Computational Linguistics	Jurafsky & Martin: Chapter 1		
10	Monday	3/11/2014	**Theory**	Regular Expressions and Finite State Automata & Transducers	Jurafsky & Martin: Chapter 2; Chapter 3		
	Monday	10/11/2014	**Remembrance day: no session**				
11	Monday	17/11/2014	**Theory**	Part-of-Speech Tagging	Jurafsky & Martin: Chapter 5 (not 5.5, 5.8 and 5.9)		
12	Monday	24/11/2014	**Theory**	Syntactic Analysis & Parsing	Jurafsky & Martin: Chapter 12 (not 12.7.2, 12.8); Chapter 13 (not 13.4.1, 13.4.2, 13.5.1)		
13	Monday	1/12/2014	**Theory**	Probabilistic Methods	Jurafsky & Martin: Chapter 4.1, 4.2 and 4.3; Chapter 5.5 and 5.9; Chapter 14.1, 14.3 and 14.4		
14	Monday	8/12/2014	**Theory**	Word Sense Disambiguation	Jurafsky & Martin: Chapter 19.1, 19.2, 19.3, Chapter 20 (20.1->20.5)		
15	Monday	15/12/2014	**Theory**	Sentence semantics and discourse; Information extraction	Jurafsky & Martin: Chapter 21; Chapter 22		

Introduction

Goals

- Get an overview of the most important techniques, approaches, problems, applications, ...

- Get hands-on experience with using these techniques (Python, NLTK)

Limitations

http://www.youtube.com/watch?v=LMkJuDVJdTw

Limitations

Limitations

Limitations

I have a spelling checker,
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea.

Eye ran this poem threw it,
Your sure reel glad two no.
Its vary polished in it's weigh.
My checker tolled me sew.

A checker is a bless sing,
It freeze yew lodes of thyme.
It helps me right awl stiles two reed,
And aides me when eye rime.

Each frays come posed up on my screen
Eye trussed too bee a joule.
The checker pours o'er every word
To cheque sum spelling rule.

Limitations

Limitations

But…

Natural Language processing is taking off

- Google Translate
- Apple SIRI
- IBM's Watson
- …

- Text analysis and generation
- Speech recognition and synthesis

Issues

- Possibilities
  - Most information is in unstructured data (text)
  - Most data is in digital form
  - Big Data (too big to handle with conventional means)

Issues

- Possibilities
  - Most information is in unstructured data (text)
  - Most data is in digital form
  - Big Data (too big to handle with conventional means)
    - >90% of currently available data was created in the last 2 years
      - Until 2002: 5 exabytes (5 billion gigabytes)
      - 2011: 5 exabytes per 2 days
      - 2013: 5 exabytes per 10 minutes
      - E.g. 6000 tweets per seconde(200 billion/year)
    - Theoretic storage capacity of human brain: 2.5petabytes (1000 petabytes = 1exabyte)

Issues

- Possibilities
  - Most information is in unstructured data (text)
  - Most data is in digital form
  - Big Data (too big to handle with conventional means)
- Problems
  - Accuracy levels
  - Speed
  - Fundamental problems
    - form-meaning relation, semantics, world knowledge

Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies

Who, what, where, when, ...

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

+ links to ontologies, e.g. Wikipedia

Who, what, where, when, ...

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

+ links to ontologies, e.g. Wikipedia

Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies
- Subjective
  - Sentiment, opinion, emotion
  - Modality, (un)certainty

Subjectivity

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies
- Subjective
  - Sentiment, opinion, emotion
  - Modality, (un)certainty
- Metaknowledge
  - Authorship, author attributes (educational level, age and gender, personality, region, illness), text attributes (date of writing, ...)

Metaknowledge

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

Male, adult, non-native author?

Figure 3. Words, phrases, and topics most highly distinguishing females and males.

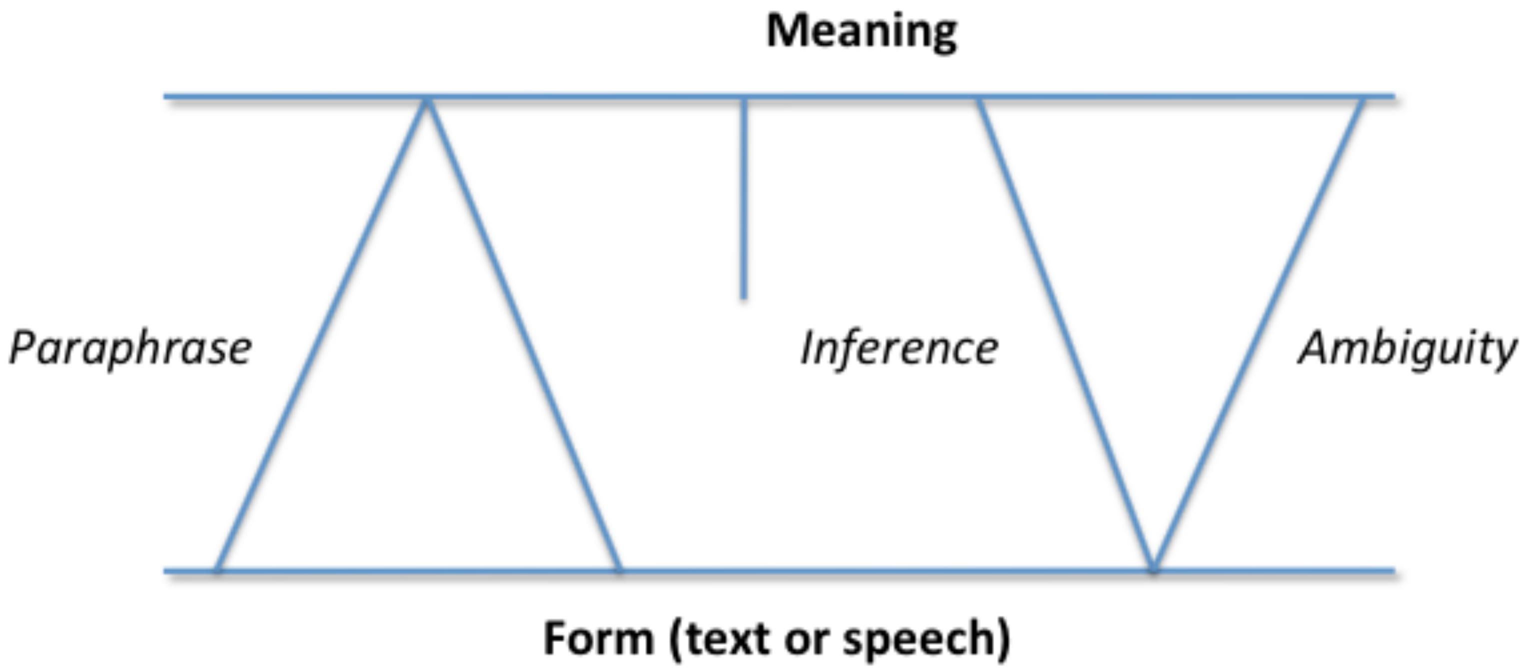
Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791

“Gender” is a matter of small words

- Women use more pronouns, men use more determiners and quantors

- Relational language use in women
- informative language use in men

The problem of natural language understanding: from form to meaning



Ambiguity

- Lexical / morphological
  - He can can the can
    - *Tekstverwerker* translated as *text far worker*
    - *Fremdzugehen* translated as *external train marriages*
- Syntactic
  - The prime minister reported his marriage to the king

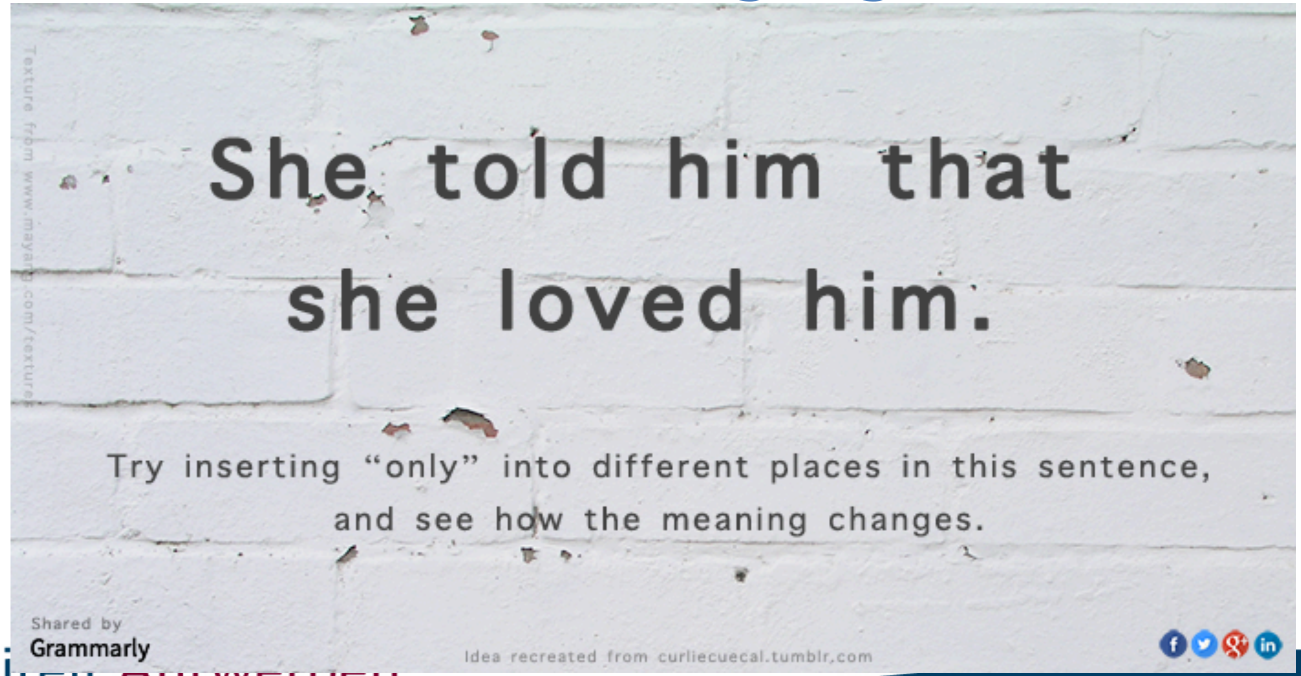
Ambiguity


21:08
24%
read the rest...
4
Patrick Stewart Surprises Fan With A Life Threatening Illness!
JamesBarker.com
Sir Patrick Stewart is a knight in shining armor.

Ambiguity

Ambiguity

- Scope of negation, modality and quantification
  - *It's not that it isn't improbable*
    - http://www.clips.ua.ac.be/cgi-bin/nespdemo.html
- – All students know *two languages*



Paraphrase

- Google acquires Microsoft
- The takeover of Microsoft by Google
- Google has obtained the majority of the shares of Microsoft
- ...

- Also synonyms:
  – E.g. biomedical text mining: protein names

Inference

- Why did John take the newspaper?
  - John was looking for a job. He took the newspaper
  - John was pestered by a fly. He took the newspaper

Inference

- Why did John take the newspaper?
  - John was looking for a job. He took the newspaper
    - Looking for job – job advertisements – newspaper

  - John was pestered by a fly. He took the newspaper
    - Catch fly – something to hit – newspaper

Inference

- What does the they refer to?
  - The mayors prohibited the students to demonstrate because they preached the revolution
  - The mayors prohibited the students to demonstrate because they feared violence

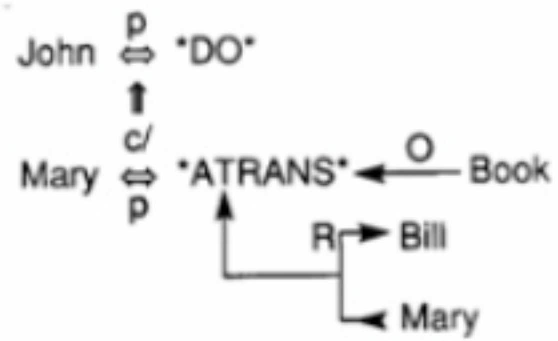
Brief History of Text Understanding

- 1970s: Knowledge Representation
  - Deep understanding (Roger Schank & students)
  - Scripts, plans, mops, universal semantic primitives
- 1980s: Logics and Parsing
  - non-monotonic reasoning, temporal logic, epistemic logic, deontic logic, …
  - Knowledge-based parsing methods
- From mid 1990s: Statistics and Shallow Understanding
  - Linguistic analysis pipeline
  - Scalable, efficient, “accurate”, robust, …
  - But: Scaling up by dumbing down? (Ray Mooney)

From form to meaning

- Language processing pipeline
  - Morphological analysis
  - Syntactic analysis
  - Lexical semantic analysis
  - Sentence semantic analysis
  - Discourse analysis
- **Result**: predicate logic or semantic network-like representation
- **Method**: Hand-Crafted or statistical / machine learning based

Deep Understanding (E.g. Schank's conceptual dependencies)



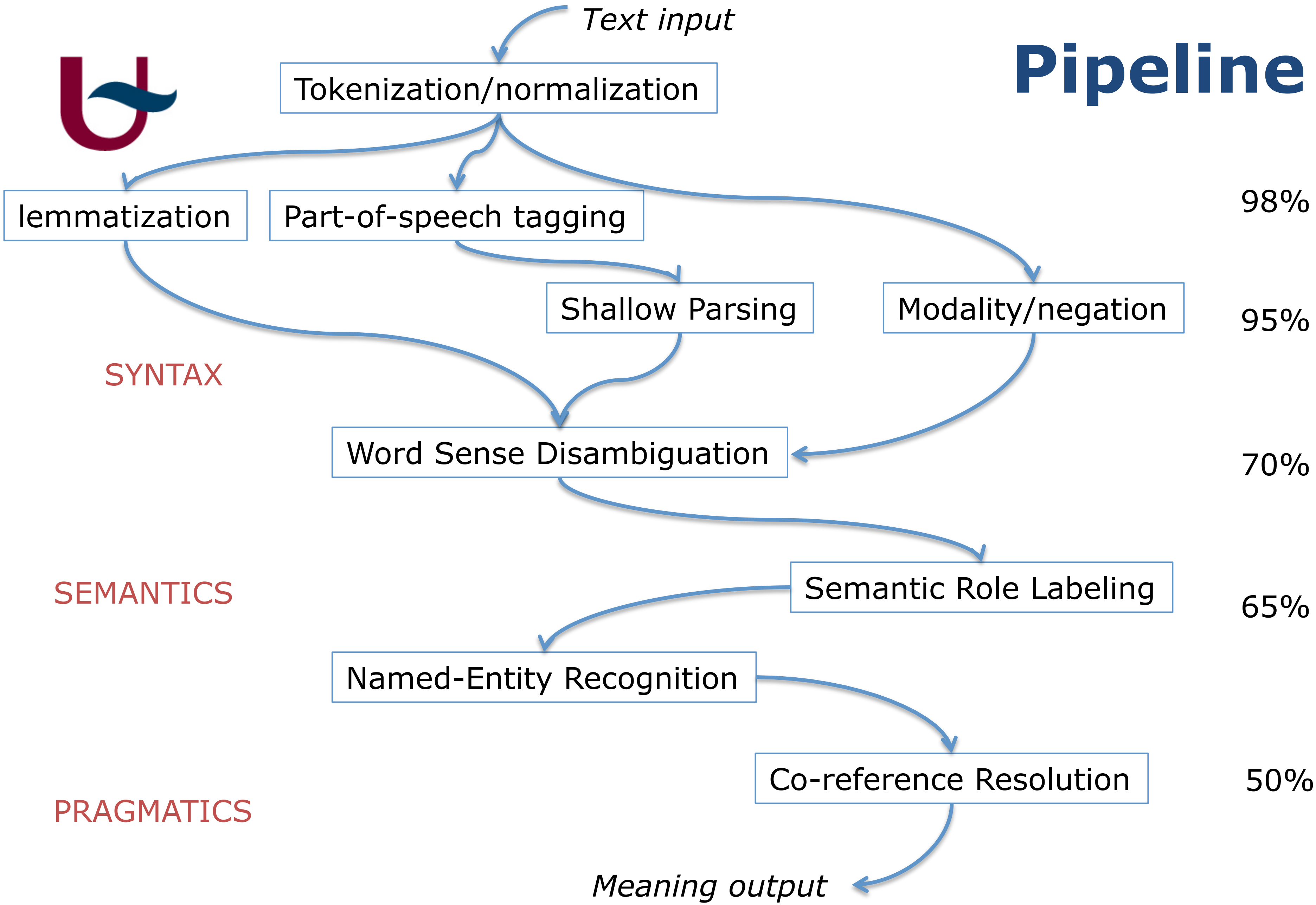
John prevented Mary from giving a book to Bill.

Pipeline

Text input

Meaning output

Pipeline



Text input
Tokenization/normalization
lemmatization
Part-of-speech tagging
Shallow Parsing
Modality/negation
SYNTAX
Word Sense Disambiguation
Semantic Role Labeling
SEMANTICS
Named-Entity Recognition
Co-reference Resolution
PRAGMATICS
Meaning output
98%
95%
70%
65%
50%

How can we represent meaning ???

Problems

- What would such a “language of thought” look like?
- How can such a representation ever be complete and universal (not language-specific)?
- It would never be possible to foresee all possible inferences needed
- It takes an enormous amount of work to model even a small domain

Text Mining (shallow understanding)

- Contents
  - Extract facts (concepts and relations between concepts) and opinions
- Meta-data
  - E.g. computational stylometry
  - Authorship attribution
  - Gender attribution
  - Personality from text

Example: Biograph (www.biograph.be)

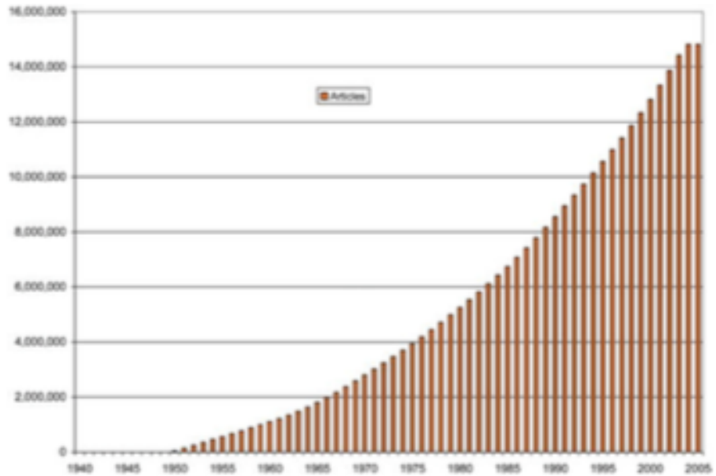
- Funded by University of Antwerp:
  - Text Mining: CLiPS CL Group
  - Graph Data Mining: ADReM, Department of Mathematics and Computer Science
  - Genetics: AMG, Department of Molecular Genetics

Goals of Biomedical Research

- Discover biomedical knowledge
- Apply this knowledge in
  - Prevention
  - Diagnosis
  - Treatment

Information overload in medical science

- Leads to
  - Fragmentation of the field
  - Poor communication between subfields



Rebholz-Schuhmann D, Kirsch H, Couto F (2005)
Facts from Text. Is Text Mining Ready to Deliver? PLoS Biol 3(2))

What can we do to help?

- Develop generic text mining tools that:
  - retrieve relevant documents from the biological literature (IR)
  - extract the required information (IE)
  - discover new knowledge
  - output the results in an intelligible way

- Two essential support services:
  - A curator's assistant: accelerating, by partially automating, the annotation and update of databases
  - A researcher's assistant: generating understandable reports in response to queries from biological researchers.

Text Mining

- Text Mining (Marti Hearst 2003)

"Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. 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"

Don Swanson 1981: medical hypothesis generation

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high levels of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability
- ...
- Conclusion: Magnesium deficiency implicated in migraine (?)

Can we automate this process and use it on a large scale?

Text Understanding!

Example event extraction

- In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain.

- Event structure:
  - Event 1: Phosphorylation (TRAF2)
  - Event 2: Binding (TRAF2, CD40)
  - Event 3: Negative_Regulation (Event1, Event2)

Example: Deception

- Intentional attempt to distort another person's beliefs about reality (e.g. by lying)

- Personality affects success in deception: not everyone is equally “successful” in deceiving others
  - Outgoing, expressive, energetic people are successful deceivers
    - Honest demeanor
  - Vein, aloof, distant people are less successful
    - Deceptive demeanor

- Cf. Riggio et al. 1987

Review SPAM

Elizabeth...
Chester
16 reviews
8 helpful votes

"A very welcoming hotel"

Reviewed July 15, 2011

Although the hotel overlooked a roundabout that was constantly busy, it was very quiet inside. The roundabout was like a little park with a children's play area. The roof terrace with the little pool had comfortable seats and was a popular place to relax and sunbathe.
The cafe on the ground floor served reasonably priced tea and coffee with a free pastry and was open until 11pm.
Our bedroom had good quality modern furniture with a television and a fridge, although the bed was very firm.
The bathroom was well appointed.
Breakfast was self service with a good choice of food.
All staff were friendly and helpful.

Stayed June 2011, traveled as a couple

Value | Cleanliness
Sleep Quality | Service

less

Was this review helpful? Yes

Ask ElizabethChester about Hotel Ciutat de Tarragona

This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC. Report problem with review

Ham or Spam

1. I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.
2. My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn’t ask for more!! We will definatly be back to Chicago and we will for sure be back to the James Chicago.

Ham or Spam

1. I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.
2. My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! [illegible] EAUTIFUL and the staff very attentive and wonde[illegible] the hotel is great, since I love to shop I couldn't ask [illegible] ill definatly be back to Chicago and we will for sure be back to the James Chicago.

SPAM

Meta-analysis

- Hauch et al. 2012
- Liars use
  - fewer exclusive wordsb
    - but, except, without, exclude
  - fewer self- and other-references (distance)
    - Fewer "I" "me" "my"
  - fewer time-related words
  - fewer tenta/ve words
  - more space-related words
  - more negative and postive emotion words
  - more motion verbs
  - more negations

Impressive Results

- Cornell University study (Ott et al. 2011)
- Data
  - Positive reviews only
  - Using *mechanical turk*, produced 400 fake positive reviews
  - Take 400 true positive reviews from TripAdvisor

- Classes
  – True (truthful) or False (deceptive)

- Features
  – LIWC, bigrams + unigrams of words

- Classifier – SVM, NB

Impressive Results

- • Human judges fail to make the distinction
  - Truth bias
  - Low inter-annotator agreement (kappa = 0.11)
  - 2 out of 3 perform at chance level
- Classifier succeeds (90% accurate)
- Bigrams+ better than LIWC
- SVM better than NB
- Cues:
  - More superlatives
  - Deceptive: imaginative rather than informative language (narrative about why they were there)
  - more V, Adv, Pro (I, me)

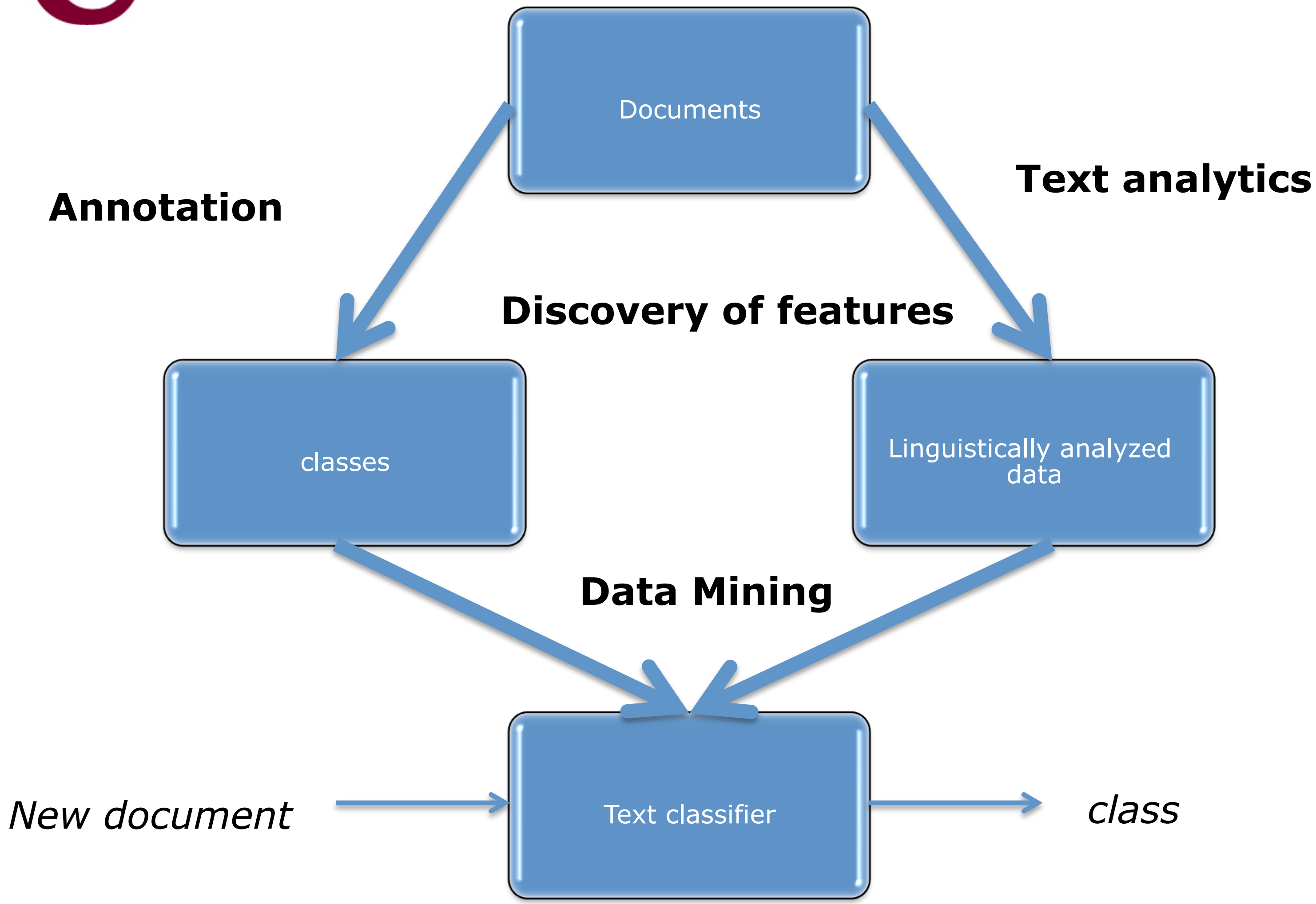
Explorative deception experiment

- Students write true review about object they like and false review about imaginary object
  - Object types: movies, books, musicians, smartphones and restaurants
  - 292 reviews, 100 words average
  - True like, True dislike, False like, False dislike

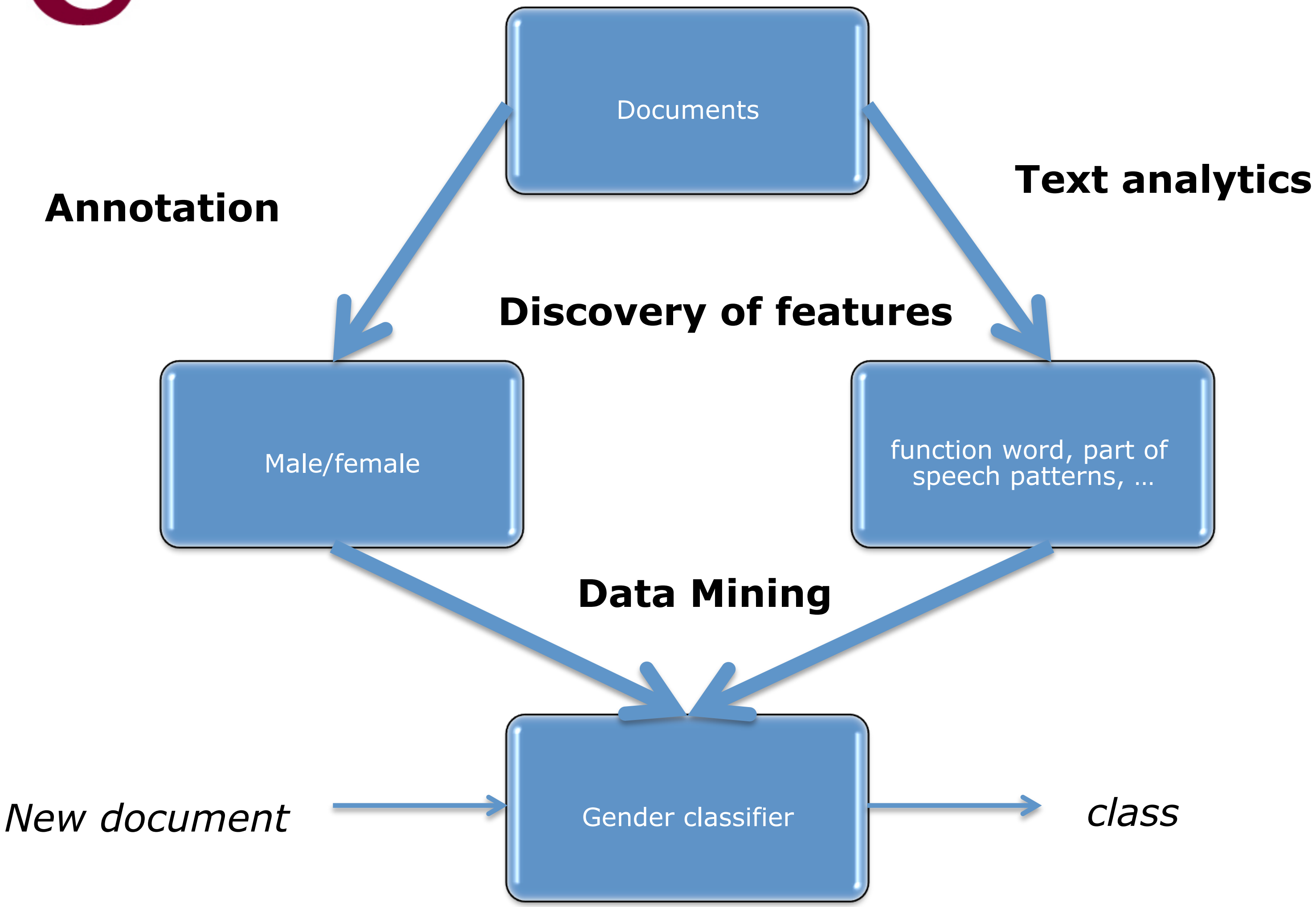
- Best results ~ 60-70% f-score

- Predictive features
  - Optimistic words, ‘perfect’, exclamation mark, ‘especially’, ‘we’, subjective words

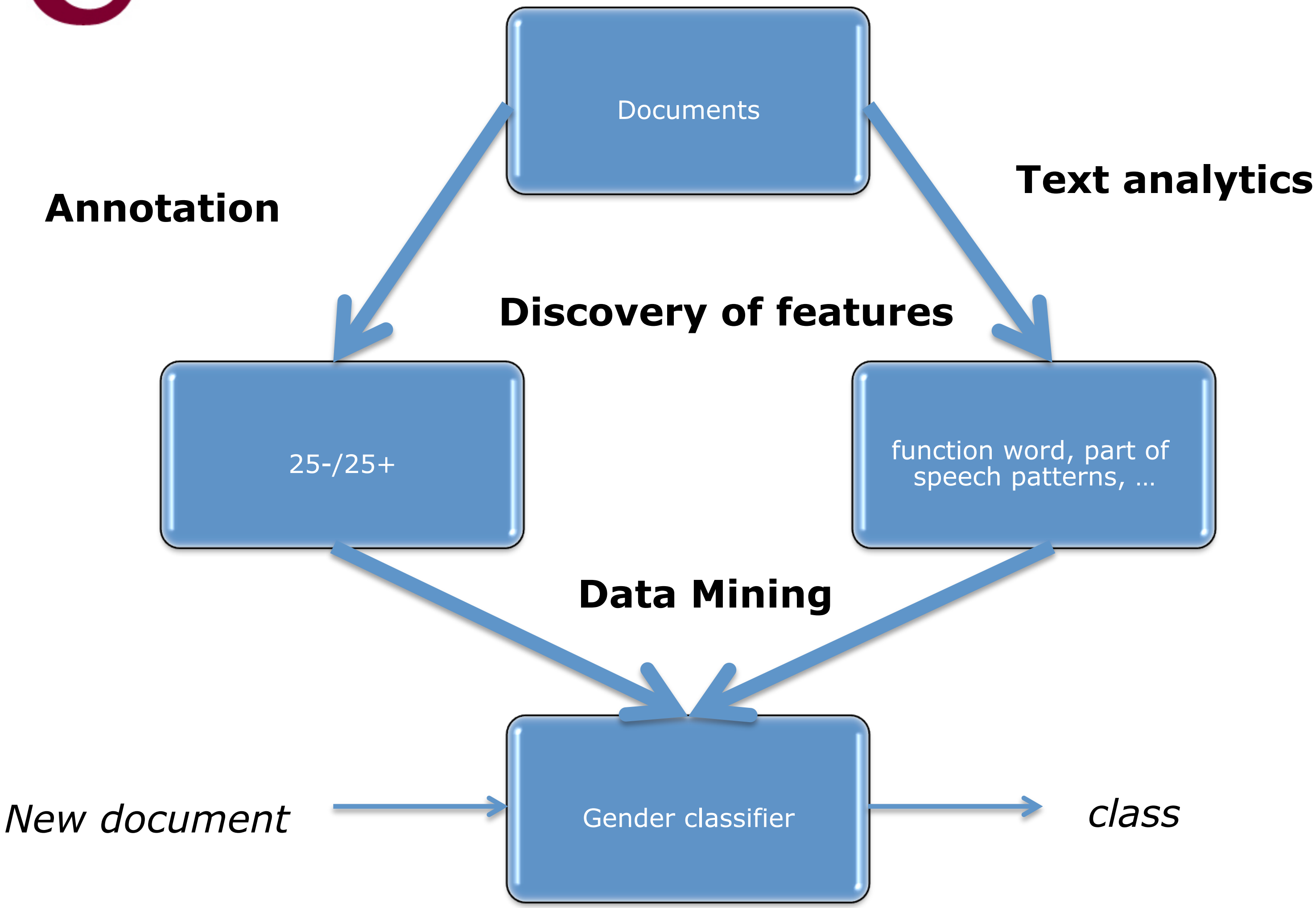
Text Categorization



Text Categorization



Text Categorization

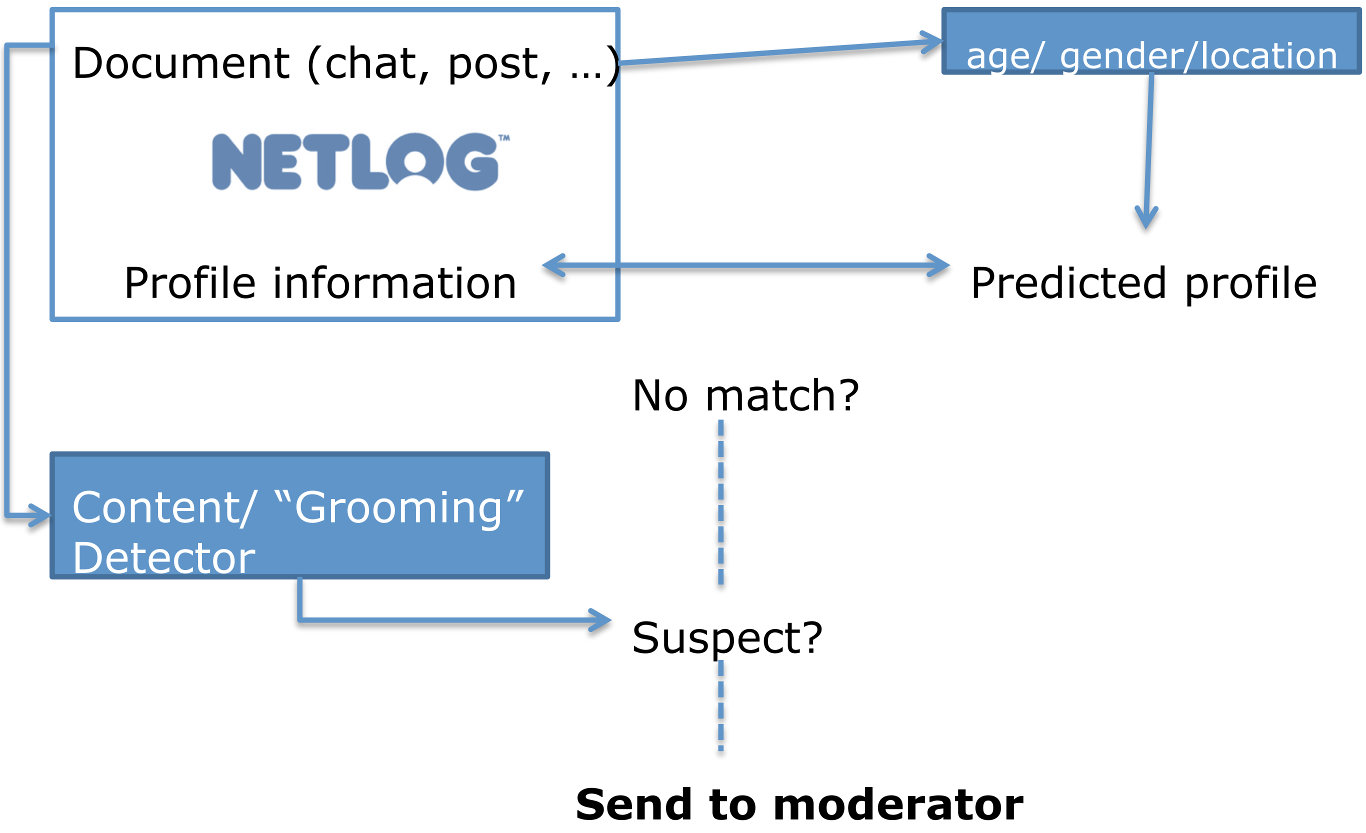


AMiCA

Automatic Monitoring for Cyberspace Applications

- Approach: text analytics, image and video analytics, data mining
- Case:
  - Sexual transgressive behavior
    - E.g. “grooming” by paedophiles
- Applications:
  - Action by moderators, police, parents, peers, social services, ...
  - Objective measurements, monitoring, trend analysis, …

AMiCA


Document (chat, post, …)
NETLOG
Profile information
age/ gender/location
Predicted profile
No match?
Content/ "Grooming" Detector
Suspect?
Send to moderator

World Knowledge needed

- Ontology
  - a domain model based on a consensus about the concepts and semantic relations in a domain. May include an inference component
  - Classes, instances, attributes, relations, events
  - Reflects conceptual structure of the domain
  - Semantic Web: OWL (ontology language)

Approaches to World Knowledge

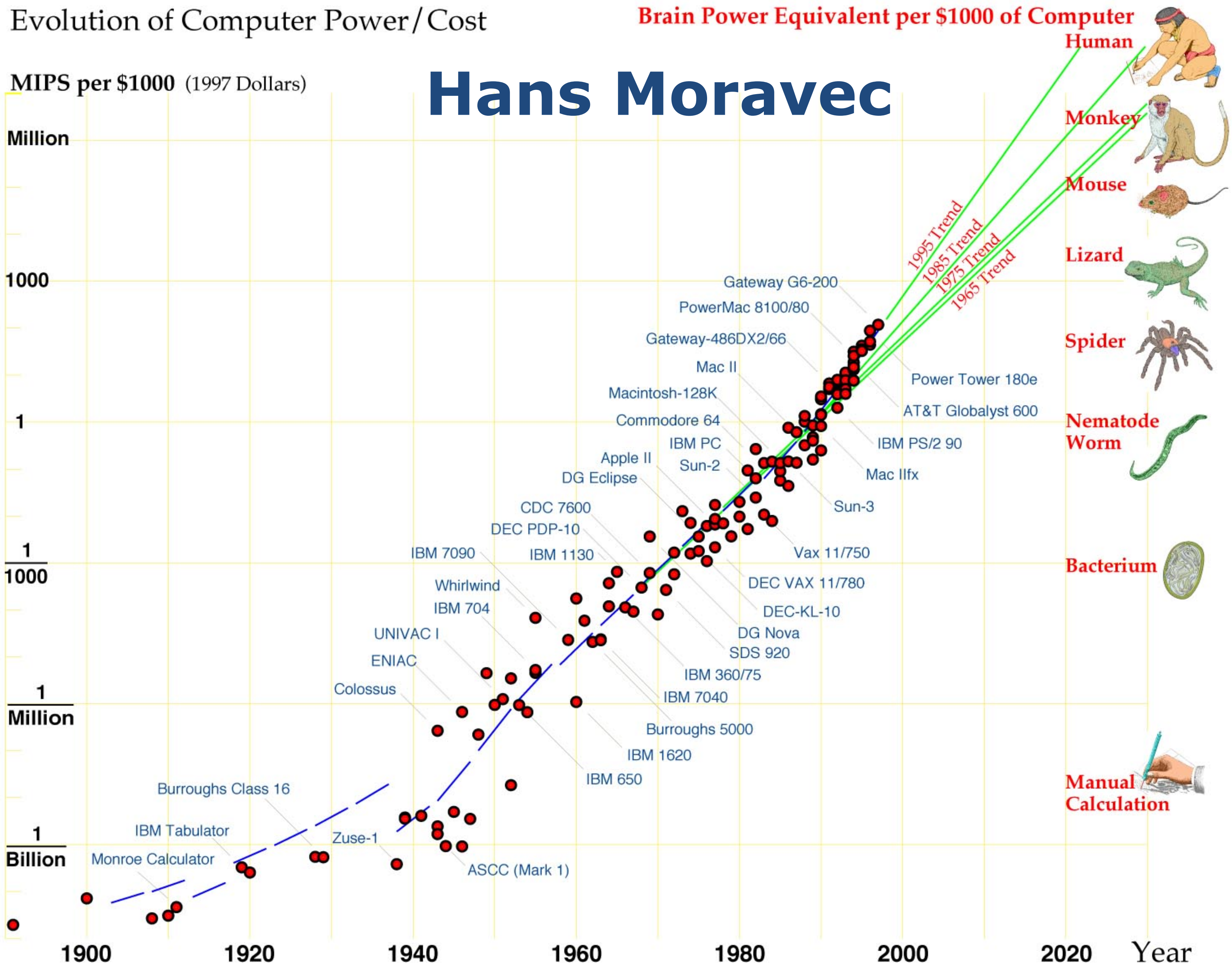
- Handcrafted ontologies
  - High quality but restricted coverage and size
  - Expensive
  - Examples: WordNet, UMLS, Cyc
- Large-scale analysis unstructured text
  - Easy and cheap
  - Limited quality
  - Domain-dependent

Semantics from Text?

- Is it possible to learn semantic concepts and semantic relations (world knowledge) automatically from text without annotation?
- In other words: a step towards a solution for the *AI-completeness* problem in natural language processing

Deep Blue optimism

Hans Moravec



Deep Blue optimism

- Exponential growth in computing power and storage capabilities of the hardware
- More is better in machine learning approaches
- Machine Learning can improve (better methods, better optimization, new algorithms, ...)
- Leap of faith: semantics and world knowledge are implicitly present in language use
  - Large (multilingual) corpora
  - Large lexical databases

Can we learn world knowledge?

- Hypothesis: syntactic context is sufficient to group semantically related words
- Words occurring in the same syntactic contexts are semantically related
  - distributional hypothesis, e.g. Zellig Harris, 1985
- Use text analysis pipeline to provide these syntactic contexts

Adjective-noun

- Nouns (concepts)
- Adjectives (properties)
- Can nouns be grouped semantically on the basis of the adjectives they are combined with in a corpus and vice versa

- Other possibilities:
  - Group verbs on the basis of the subjects and objects they co-occur with

Harris' hypothesis

- Examples
  - Green stromble
  - Ripe stromble
  - Low-calory stromble
  - ...

- => stromble = edible

- Problems
  - Polysemy
    - A good knight [person]
    - A pinned knight [chess]

Approach

- Start from a large analyzed corpus
- Look for all combinations of adjective and noun in the same nominal phrase
- Produce a matrix with nouns as rows and adjectives as columns and number of occurrences together as values
- Hierarchical clustering of the adjective vectors

Experiment

- Twente News Corpus
  - 5000 most frequent nouns
  - 20000 most frequent adjectives
- Some examples of results

straatje straat steeg gracht
boete schorsing straf sanctie celstraf gevangenisstraf
bosbrand droogte overstroming aardbeving epidemie
metafoor citaat vergelijking parallel verwijzing
hiphop blues popmuziek pop jazz rock
schoonzoon echtgenoot bruid echtgenote minnaar schoonvader

Clustering of adjectives

- In matrix: rows become adjectives, columns nouns
- Examples

  geel paars zwart groen blauw grijs oranje bruin roze wit rood
  zonovergoten herfstig winters druilerig zomers regenachtig zonnig

  cool tof lelijk stom dom brutaal geil
  tenger slank iel frêle schriel rijzig slank
  grondig zorgvuldig nauwgezet nadere nauwkeurig minutieus

Evaluation

- Compare the automatically computed clusters with handcrafted resources like Wordnet

- How many words clustered together appear in Wordnet relations (synonyms, antonyms, ...)?

  = precision (~ 43%)

- How many Wordnet relations appear in a cluster?

  = recall (~ 8%)

Program

Session	Day	Date	Chapter	Topic	Reading Assignment	Slides	Take-home Assignment
1	Monday	29/9/2014	**Python**	Session 1 - Variables	See Github		
2	Thursday	2/10/2014	**Python**	Session 2 - Collections			
3	Monday	6/10/2014	**Python**	Session 3 - Conditions (and an introduction to loops)			
4	Thursday	9/10/2014	**Python**	Session 4 - Loops			
5	Monday	13/10/2014	**Python**	Session 5 - Reading and writing to files			
6	Thursday	16/10/2014	**Python**	Session 6 - Writing your own Functions and importing packages			
7	Monday	20/10/2014	**Python**	Session 7 - Regular Expressions in Python			
8	Thursday	23/10/2014	**Python**	Session 8 - Advanced looping in Python and list comprehensions			
9	Monday	27/10/2014	**Theory**	Introduction to Computational Linguistics	Jurafsky & Martin: Chapter 1		
10	Monday	3/11/2014	**Theory**	Regular Expressions and Finite State Automata & Transducers	Jurafsky & Martin: Chapter 2; Chapter 3		
	Monday	10/11/2014	**Remembrance day: no session**				
11	Monday	17/11/2014	**Theory**	Part-of-Speech Tagging	Jurafsky & Martin: Chapter 5 (not 5.5, 5.8 and 5.9)		
12	Monday	24/11/2014	**Theory**	Syntactic Analysis & Parsing	Jurafsky & Martin: Chapter 12 (not 12.7.2, 12.8); Chapter 13 (not 13.4.1, 13.4.2, 13.5.1)		
13	Monday	1/12/2014	**Theory**	Probabilistic Methods	Jurafsky & Martin: Chapter 4.1, 4.2 and 4.3; Chapter 5.5 and 5.9; Chapter 14.1, 14.3 and 14.4		
14	Monday	8/12/2014	**Theory**	Word Sense Disambiguation	Jurafsky & Martin: Chapter 19.1, 19.2, 19.3, Chapter 20 (20.1->20.5)		
15	Monday	15/12/2014	**Theory**	Sentence semantics and discourse; Information extraction	Jurafsky & Martin: Chapter 21; Chapter 22		