Data-driven Intelligent Systems

Lecture 23
Text Mining I/II

KNOWLEDGE
TECHNOLOGY

http://www.informatik.uni-hamburg.de/WTM/

Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

Goal and Definition of Text Mining

- Text mining is the process of compiling, organizing, and ***analyzing large document*** collections
- Goal is to support the delivery of ***targeted types of information*** to analysts and decision makers
- ***Discovery of relationships*** between related facts that span wide domains of inquiry

Mining Text Data Comes with Different Names

- Data mining from text, text mining
- Natural language processing
- Information extraction
- Information retrieval from text
- Text categorization methods

Material based on book by Han and Kamber, 2006 and additional slides from Cheng Xiang Zhai, Mooney, Volinsky

Free Text versus Structured Data

Data Mining / Knowledge Discovery

Structured Data

HomeLoan (
Loanee: *Frank Rizzo*
Lender: *MWF*
Agency: *Lake View*
Amount: *$200,000*
Term: *15 years*
)

Multimedia

Loans(*$200K*,[*map*],...)

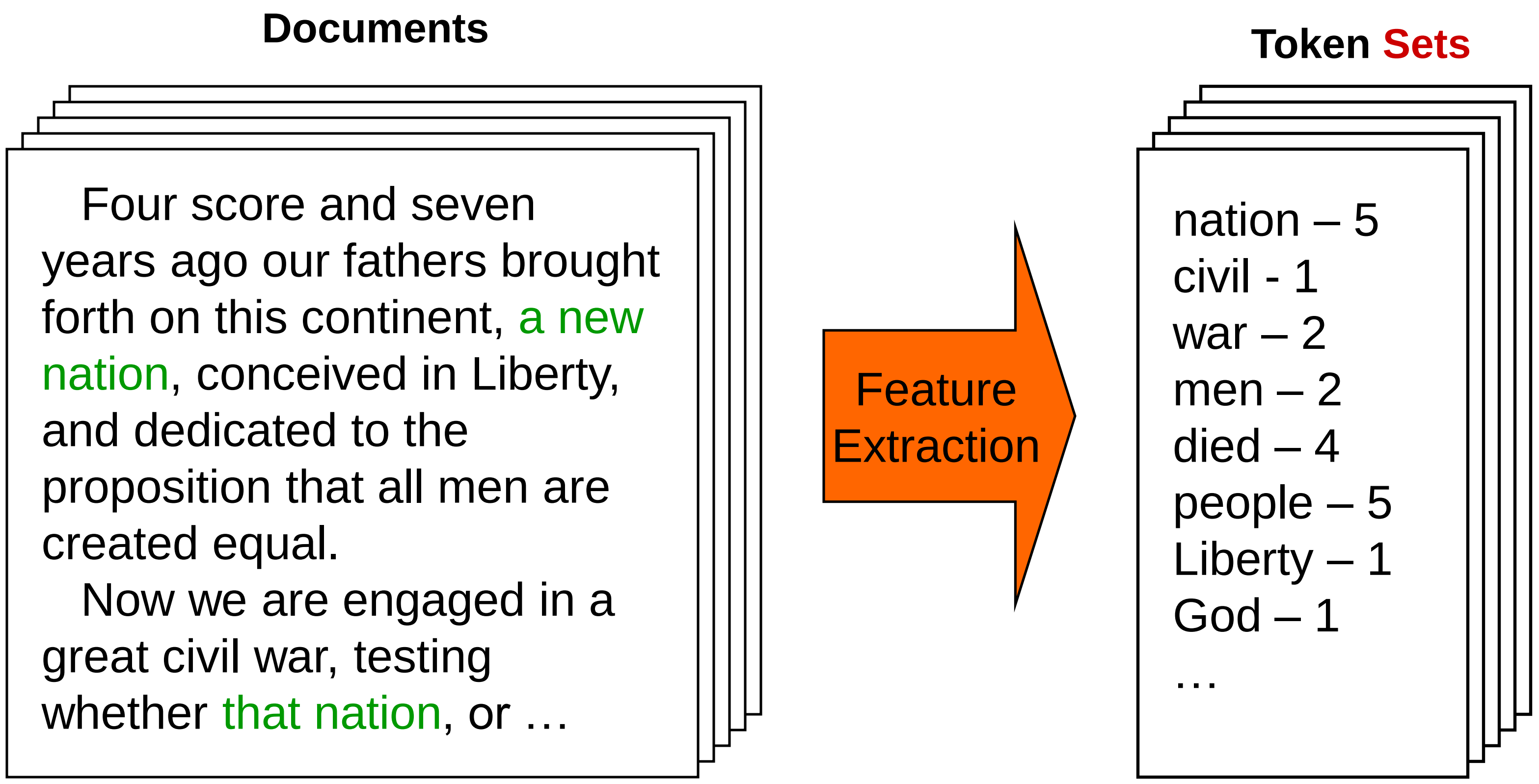
Free Text

Frank Rizzo bought his home from Lake View Real Estate in 1992.
He paid $200,000 under a 15-year loan from MW Financial.

Hypertext

<a href>*Frank Rizzo* *bought*
<a hef>*this home*
from <a href>*Lake View Real Estate*
In *1992*.
<p>...

Bag-of-Tokens Approaches



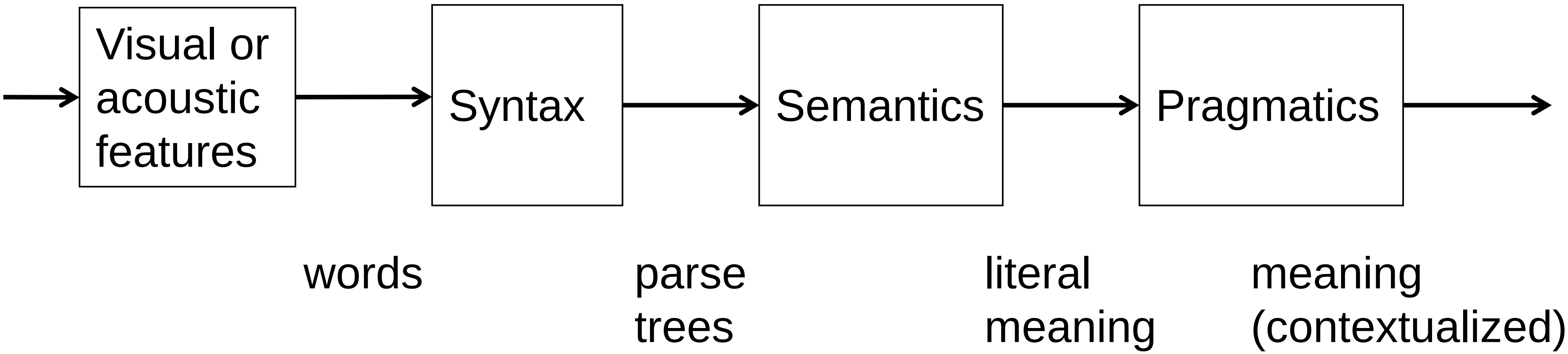
Looses all order-specific information!
Reduces context information.

a.k.a. "bag of words"

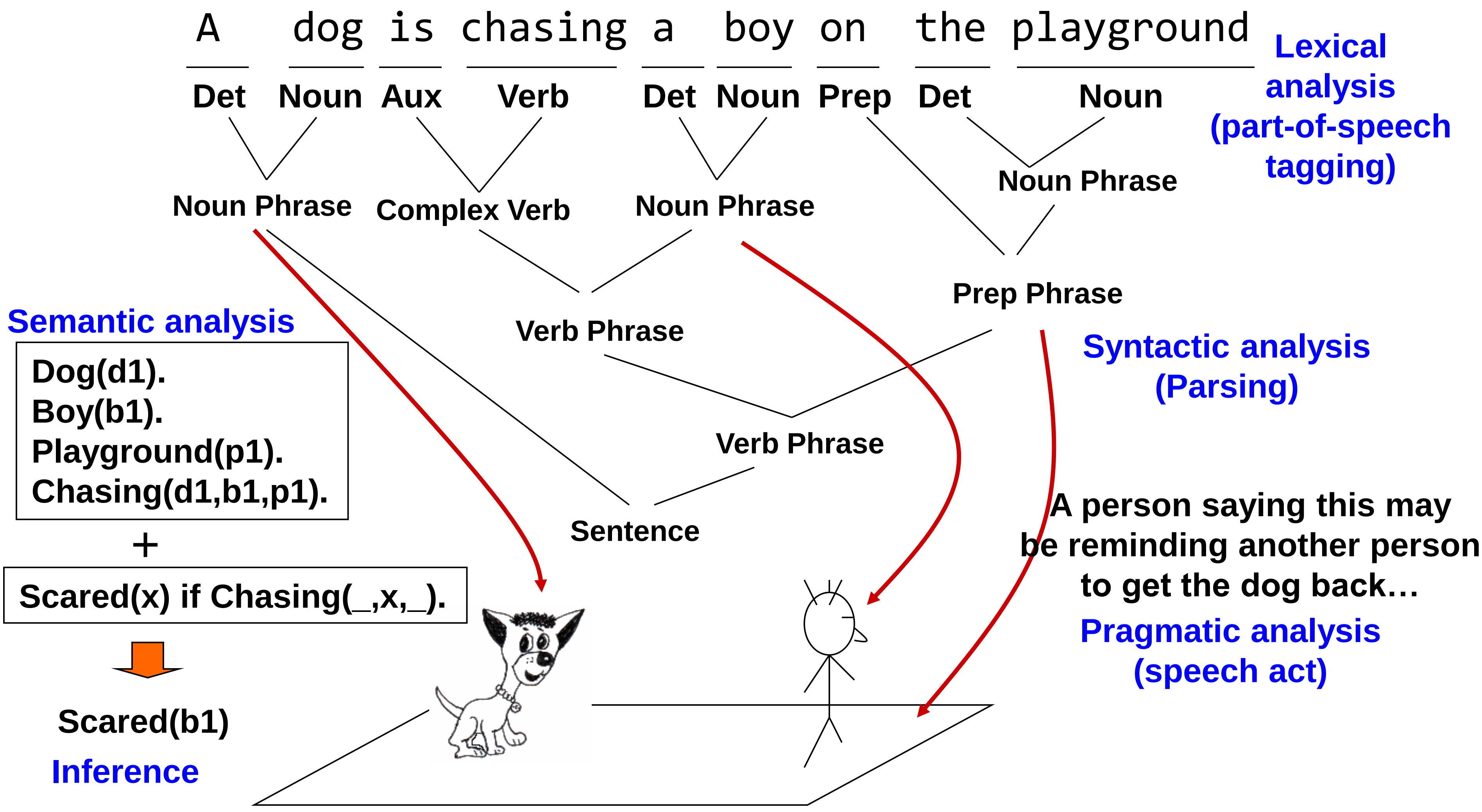
Syntax, Semantic, Pragmatics

- ***Syntax***: ordering of words and its possible effect on meaning.
  - `The dog bit the boy.`
  - `The boy bit the dog.`
  - `* Bit boy dog the the.`
  - `* Colorless green ideas sleep furiously.`
- ***Semantics:*** concerns the (literal) meaning of words, phrases, and sentences.
  - `“plant” as a photosynthetic organism`
  - `“plant” as a manufacturing facility`
  - `“plant” as an act of sowing`
- ***Pragmatics:*** concerns the overall communicative and social context and its effect on interpretation.
  - `The ham sandwich wants another beer.`
  - `John thinks vanilla.`

Comprehension as a Simplified Sequential Model



From Flat Text to Structure and Meaning



Language is full of Ambiguities

- Word-level ambiguity
  - "design" can be a noun or a verb (Ambiguous Part of Speech)
  - "root" has multiple meanings (Ambiguous semantic sense)
- Syntactic ambiguity
  - "natural language processing" (Modification/Bracketing)
  - "`A man saw a boy with a telescope.`" (Prepositional Phrase Attachment)
- Semantics and Anaphora resolution
  - "`John persuaded Bill to buy a TV for himself.`"
    (himself = John or Bill?)
- Presupposition and pragmatic inferences
  - "`He has quit smoking.`"
    implies that he smoked before.

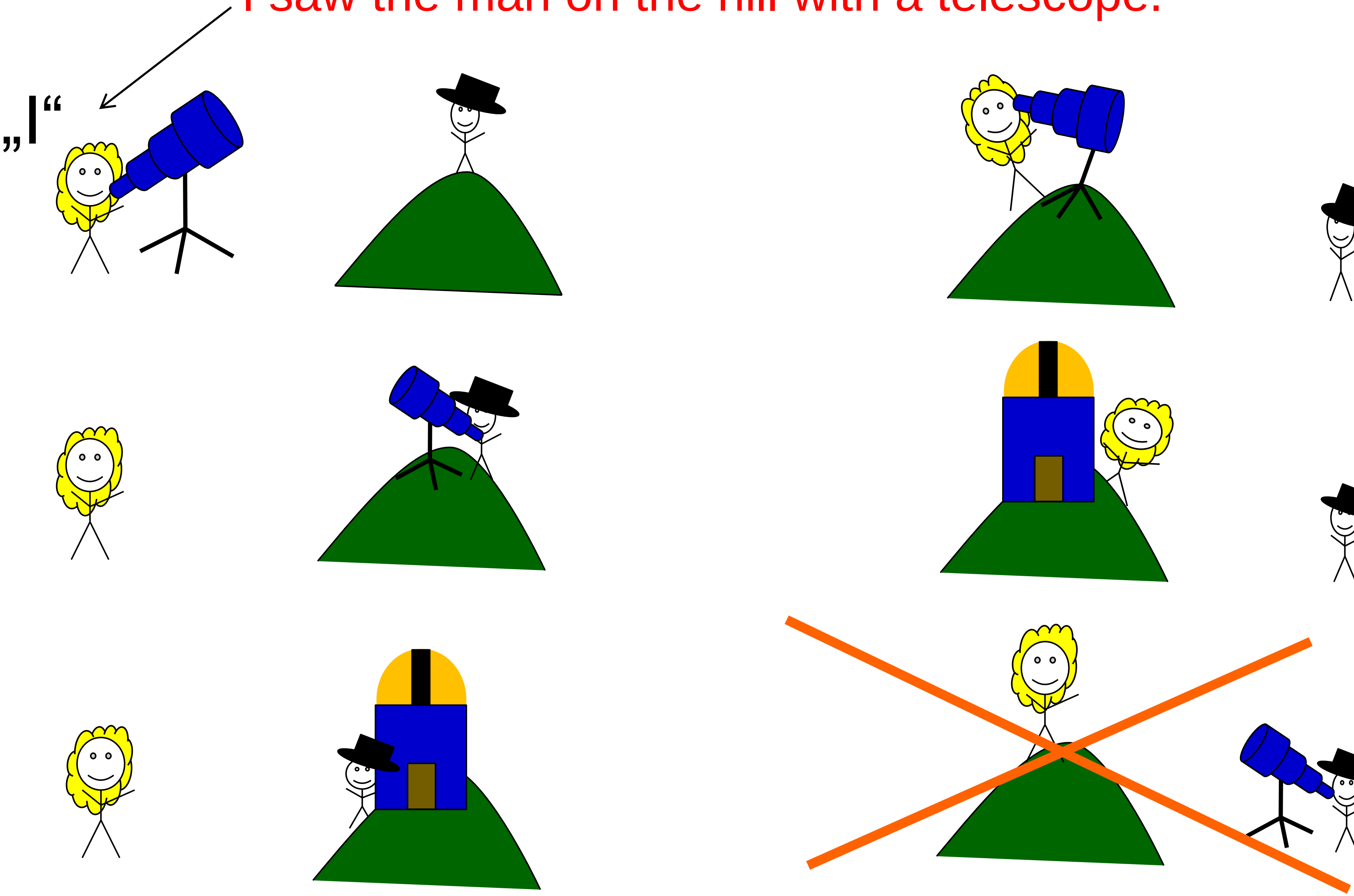
Humans rely on *context* to interpret (when possible).
This context may extend beyond a given document!

Ambiguity: Different Interpretations?

- Natural language can be highly ambiguous
- Can you find ambiguities?
  - `I saw the Grand Canyon flying to LA.`
  - `Time flies like an arrow.`
  - `I saw the man on the hill with a telescope.`

Ambiguity

I saw the man on the hill with a telescope.



Ambiguity is Ubiquitous but we may not Notice

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."
- Semantic Analysis
  - "I put the plant in the window" vs. "Ford put the plant in Mexico"
- Pragmatic Analysis
  - **Example** from "The Pink Panther Strikes Again":

Clouseau: Does your dog bite?
Hotel Clerk: No.
Clouseau: [*bowing down to pet the dog*] Nice doggie.
[*Dog barks and bites Clouseau in the hand*]
Clouseau: I thought you said your dog did not bite!
Hotel Clerk: That is not my dog.

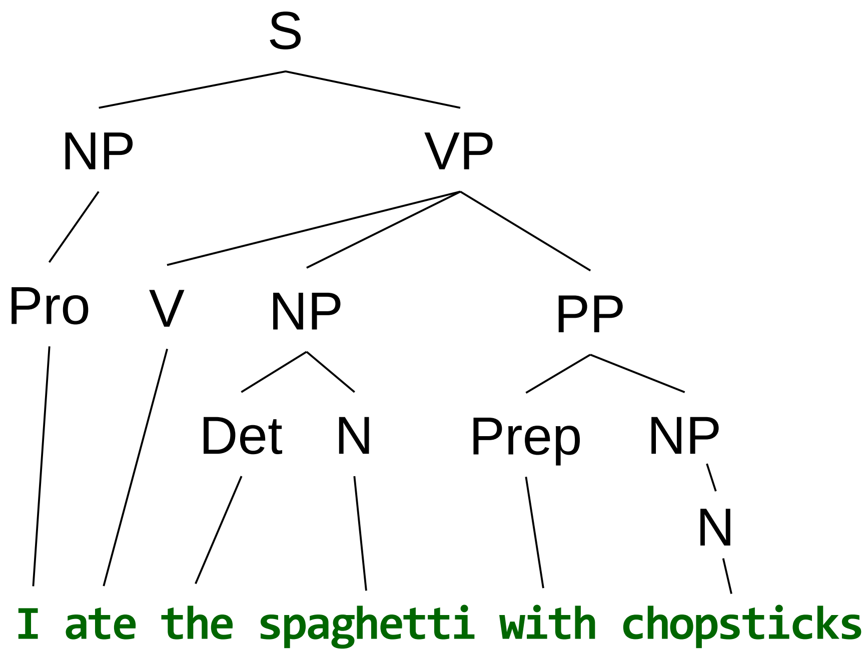
Overview

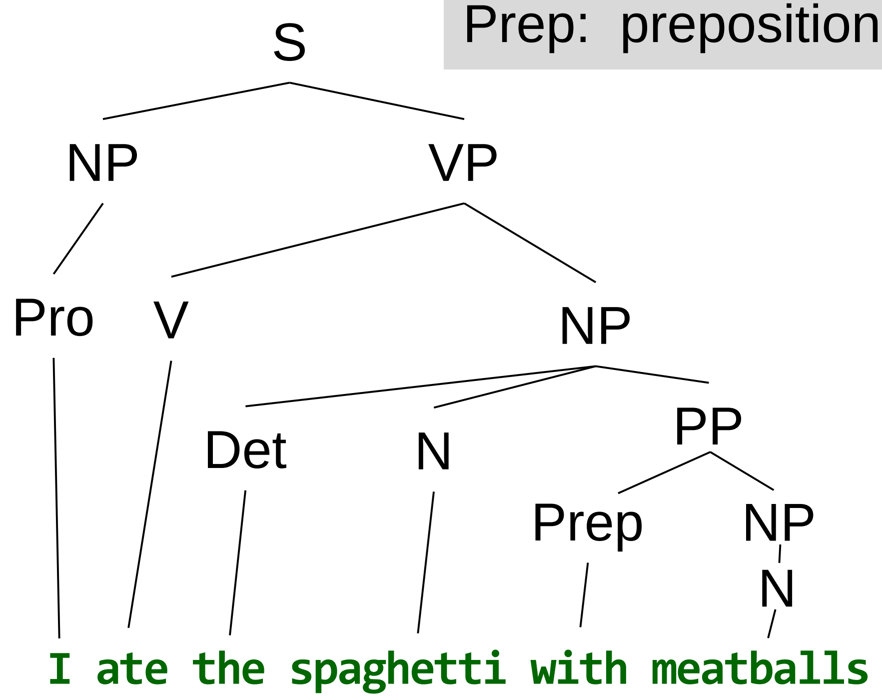
- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

Syntactic Parsing

S: sentence
NP: noun phrase
VP: verb phrase
PP: prep. phrase
N: noun
V: verb
Pro: pronoun
Det: determinant
Prep: preposition

- Produce the correct syntactic parse tree for a sentence





Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in *n* prepositional phrases has *over* 2^n syntactic interpretations.
  - "`I saw the man with the telescope.`": 2 parses
  - "`I saw the man on the hill with the telescope.`": 5 parses
  - "`I saw the man on the hill in Texas with the telescope.`": 14 parses
  - "`I saw the man on the hill in Texas with the telescope at noon.`": 42 parses
  - "`I saw the man on the hill in Texas with the telescope at noon on Monday.`" 132 parses

Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

Generative Models: Formal Grammars

- A grammar is a set of *production rules* which generates a set of strings (a language) by *rewriting* the top symbol S.
- *Nonterminal* symbols are intermediate results that are not contained in strings of the language.
  - S -> NP VP
  - NP -> Det N
  - VP -> V NP
- *Terminal* symbols are the final symbols (words) that compose the strings in the language.
- Production rules for generating words from part of speech categories constitute the lexicon.
  - N -> boy
  - V -> eat

Context-Free Grammars

- A context-free grammar (CFG) only has productions with a *single symbol* on the left-hand side.
- CFG:
  - S -> NP V
  - NP -> Det N
  - VP -> V NP
- not CFG:
  - A B -> C
  - B C -> F G

Probabilistic Structure Parsing to Reduce Ambiguity

Choose ***most likely*** parse tree…

<u>Probabilistic CFG</u>

Grammar

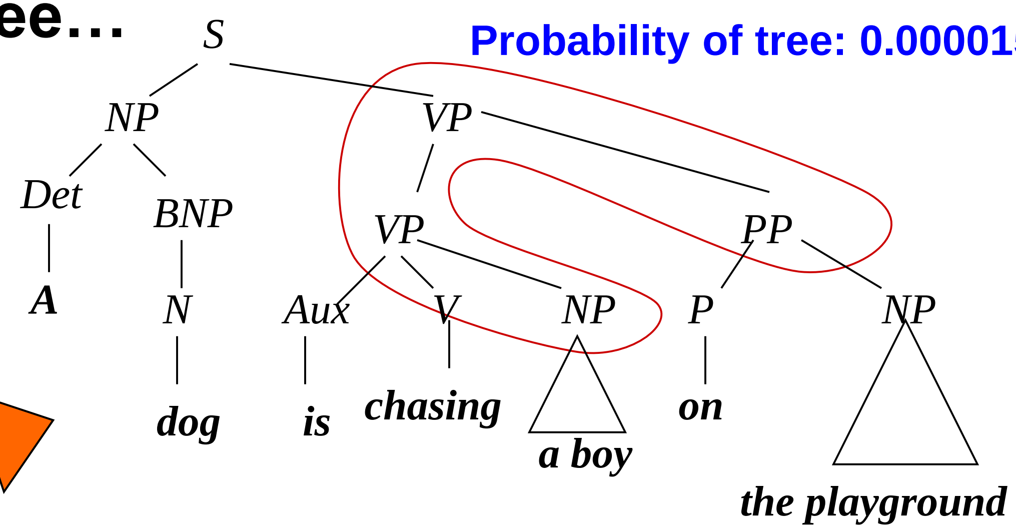
$S \rightarrow NP\ \ VP$	1.0
$NP \rightarrow Det\ BNP$	0.3
$NP \rightarrow BNP$	0.4
$NP \rightarrow NP\ \ PP$	0.3
$BNP \rightarrow N$	…
$VP \rightarrow V$	
$VP \rightarrow Aux\ V\ NP$	
$VP \rightarrow VP\ PP$	…
$PP \rightarrow P\ NP$	1.0

Lexicon

$V \rightarrow chasing$	0.01
$Aux \rightarrow is$	
$N \rightarrow dog$	0.003
$N \rightarrow boy$	
$N \rightarrow playground$	…
$Det \rightarrow the$	
$Det \rightarrow a$	…
$P \rightarrow on$	

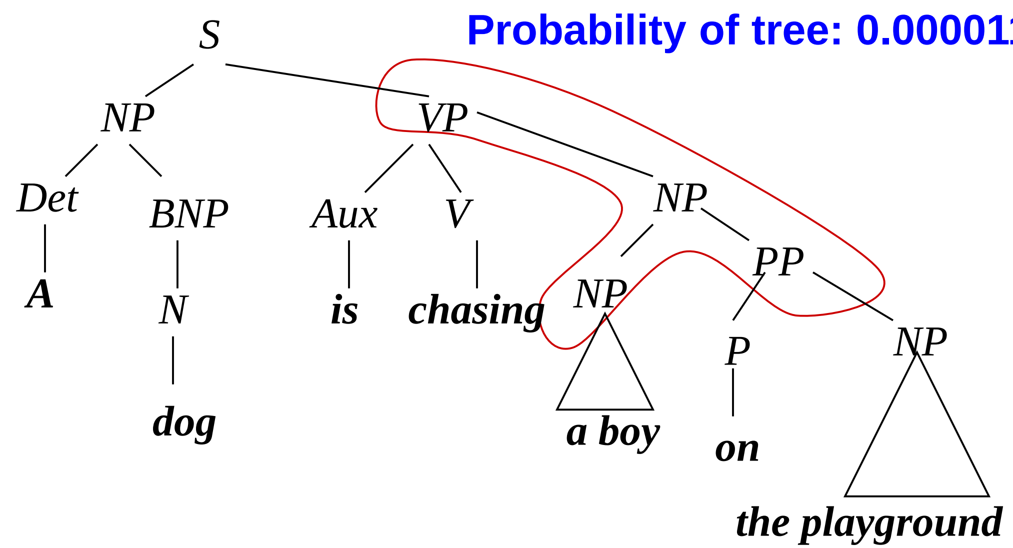
BNP = Base Noun Phrase

Probability of tree: 0.000015



⋮

Probability of tree: 0.000011



Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

How can we Deal with Mining from Text at all? Shallow Natural Language Processing

- Progress on useful ***Sub-Goals***:
  - English Lexicon
  - Part-of-Speech Tagging
  - Word Sense Disambiguation
  - Phrase Detection / Parsing

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words.
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning
  - **E.g.** “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried $\Rightarrow$ carry + ed (past tense)
  - independently $\Rightarrow$ in + (depend + ent) + ly
  - Googlers $\Rightarrow$ (Google + er) + s (plural)
  - unlockable $\Rightarrow$ un + (lock + able) ?

    $\Rightarrow$ (un + lock) + able ?

Part-of-Speech (POS) Tagging

Training data (Annotated text)

This	*sentence*	*serves*	*as*	*an*	*example*	*of*	*annotated*	*text…*
Det	N	V1	P	Det	N	P	V2	N

"This is a new sentence." → **POS Tagger** →

This	*is*	*a*	*new*	*sentence.*
Det	Aux	Det	Adj	N

Pick the *most likely* tag sequence.

$$p(w_1,\ldots,w_k,t_1,\ldots,t_k) = \begin{cases} p(t_1 \mid w_1)\ldots\underline{p(t_k \mid w_k)}\,p(w_1)\ldots p(w_k) \\ \prod_{i=1}^{k} \underline{p(w_i \mid t_i)\,p(t_i \mid t_{i-1})} \end{cases}$$

Independent assignment
Most common tag

Partial dependency
(HMM)

Phrase Chunking rather than Full Parsing

- Find all non-recursive noun phrases (**NPs**) and verb phrases (**VPs**) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only $ 1.8 billion] [PP in] [NP September]

Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

From Structure to Semantics: Word Sense Disambiguation

- Words in natural language usually have a fair number of different possible meanings.
  - `Ellen has a strong interest in computational linguistics.`
  - `Ellen pays a large amount of interest on her credit card.`
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

agent patient source destination instrument

  - `John drove Mary from Austin to Dallas in his Toyota Prius.`
  - `The hammer broke the window.`

- Also referred to as:
  - "***case role analysis***"
  - "***thematic analysis***"
  - "***shallow semantic parsing***"

Semantic Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- ***Named entity recognition*** for identifying names of people, places, organizations, etc. in text.

people organizations places

- `Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.`

- ***Relation extraction*** identifies specific relations between entities.

Question Answering

- Directly answer natural language questions based on information in a corpus of textual documents (e.g. the web).
  - When was Barack Obama born? (***factoid***)
    - ⇨ August 4, 1961
  - Who was president when Barack Obama was born?
    - ⇨ John F. Kennedy
  - How many presidents have there been since Barack Obama was born? (***towards more inferences***)
    - ⇨ 9

⇨ Much but not all information may be directly available

Text Summarization

- Produce a short summary of a longer document or article.
  - ***Article***: With a split decision in the final two primaries and a flurry of super-delegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
  - ***Summary***:
    Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

Mining Text Data in Internet (Video)

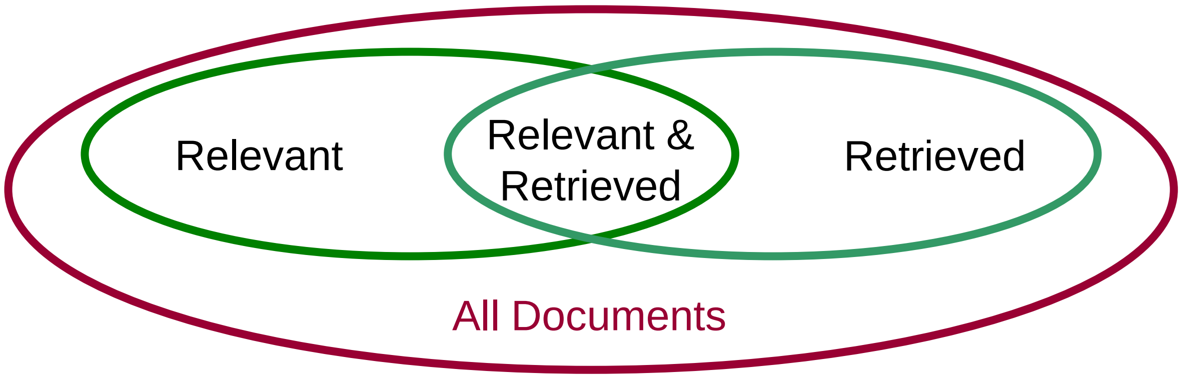
Information Retrieval as Start for Text Mining

- Typical traditional IR systems
  - Online library catalogs
  - Online document management systems
- Information retrieval vs. database management systems
  - Some IR problems are not addressed well in DBMS
    - **E.g.**, unstructured documents, approximate search using keywords and relevance
  - Some DB problems are not present in IR
    - **E.g.**, update, transaction management, complex objects

Information Retrieval vs Information Extraction

- ***Information Retrieval***
  - Given a set of query terms and a set of document terms select only
    - highly relevant documents [***precision***], and
    - preferably all the relevant [***recall***].
- ***Information Extraction***
  - Extract what the document contains from the text
- IR systems can FIND documents but do not need to "understand" them

Basic Measures for Text Retrieval



- ***Precision***: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- ***Recall***: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Precision vs. Recall

- In other words (we have been here before!)
  - Precision = TP/(TP+FP)
  - Recall = TP/(TP+FN)

	Truth: **Relevant**	Truth: **Not Relevant**
Algorithm: **Relevant**	**TP**	**FP**
Algorithm: **Not Relevant**	**FN**	**TN**

  - Trade-off:
    - If algorithm is ‘picky’: precision high, recall low
    - If algorithm is ‘relaxed’: precision low, recall high

BUT: Recall, specifically FN, often hard
if not impossible to calculate

Overview

- Structure, grammar and meaning; ambiguity in language
  - Parsing & part-of-speech tagging
  - Grammars
- Shallow NLP
  - Semantic role labeling
- Information retrieval
  - Vector space model, TF-IDF weighting

Information Retrieval Techniques

- Basic Concepts
  - A document can be described by a set of representative keywords called ***index terms***.
  - Different index terms have varying *relevance* when used to describe document contents.
  - This effect is captured through the assignment of ***numerical weights*** to each index term of a document, e.g.:
    - frequency
    - tf-idf: term frequency - inverse document frequency
- Information Retrieval Models
  - Boolean Model
  - Vector Model

Boolean Model

- Consider that ***index terms are either present or absent*** in a document

  → the index term weights are assumed to be all ***binaries***

- A query is composed of index terms linked by three connectives: *not*, *and*, and *or*
  - **E.g.**: car *and* repair, plane *or* airplane

- The Boolean model predicts that ***each document is either relevant or non-relevant*** based on the match of a document to the query

- Think about advantages / disadvantages …

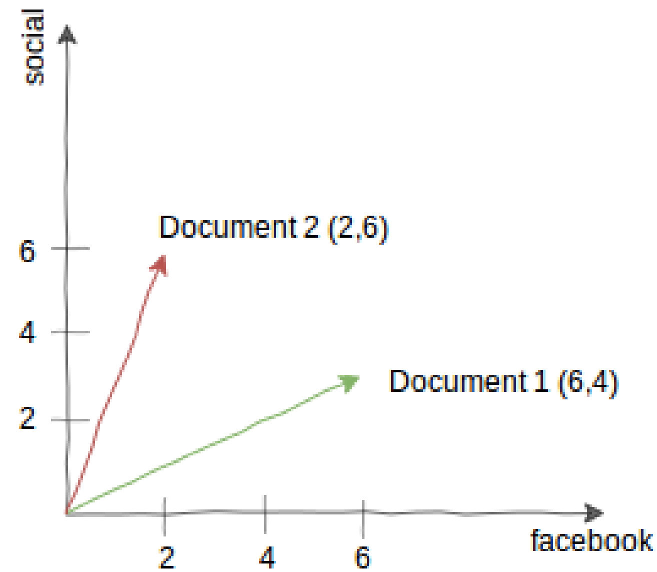
Vector Space Model

- ***Represent a document by a term vector***
  - Term: basic concept, e.g., word or phrase
  - Each term defines one dimension
  - *N* terms define an *N*-dimensional space
  - Element of vector corresponds to term weight
  - E.g., $d = (x_1, \ldots, x_N)$, x_i is ***importance*** of term *i*
- New document is assigned to the most likely category based on ***vector similarity***.

Vector Space Model

- Documents & user queries represented as N-dimensional vectors
  - *N* = # index terms in document collection

	doc1	doc2
facebook	6	2
social	3	6



- Degree of ***similarity*** of the document *d* with regard to the query *q*:
  - Calculated as the ***correlation*** between the vectors that represent them
  - Using measures such as the *Euclidian distance* or the *cosine* of the angle between these two vectors

Word Clouds

- Find most frequent/interesting words in text, and display them graphically
  - Summarize a document
  - Blogs do this
  - → wordle.net

blue: Harry Truman’s 1948 inaugural address

red: words absent from Truman’s speech, but in those of his contemporaries

The Vector Space Model does not Specify:

- How to ***select terms*** to capture "basic concepts"
  - Stop words (i.e., words that can be ignored)
    - **E.g.** "`a`", "`the`", "`always`", "`along`"
  - Word stemming (to reduce # terms)
    - **E.g.** "`computer`", "`computing`", "`computerize`" => "`compute`"
- How to ***assign weights***
  - Not all words are equally important: Some are more indicative than others
    - **E.g.** "`barracuda`" vs. "`fish`"
- How to measure the similarity?

How to assign Weights

- Two-fold heuristics based on frequency
  - TF (Term frequency)
    - More frequent ***within*** a document → more relevant to semantics
    - e.g., “algebra” vs. “trigonometry”
      (distinguish fields in maths)
  - IDF (Inverse document frequency)
    - Less frequent ***among*** documents → more discriminative
    - e.g. “algebra” vs. “science”
      (“algebra” is more specific & rare)

Term Frequency (TF)

- ***Weighting***:
  - More frequent ⇨ more relevant to topic
    - Raw_TF = *f(t,d)*: how many times term *t* appears in doc *d*
- ***Normalization***:
  - Document length varies ⇨ relative frequency preferred

$$\mathrm{TF}(t,d) = 0.5 + \frac{0.5 \cdot f(t,d)}{Length(d)}$$

  - After normalization: values between 0.5 and 1
  - Normalized frequency prevents bias for longer documents

Inverse Document Frequency (IDF)

- Measure of how much information the word provides
- Less frequent ***among*** documents → more discriminative

$$\mathrm{IDF}(t) = \log\left(\frac{n}{1+k}\right)$$

n — total number of docs

k — number of docs with term t appearing

(the DF document frequency)

1 — avoid division by 0

- Other weighting schemes for TF and IDF exist

https://en.wikipedia.org/wiki/Tf-idf

TF-IDF Weighting

- Combine term frequency and inverse document frequency:

$$\mathrm{TF\,I\,DF}(t, d, D) = \mathrm{TF}(t, d) \cdot IDF(t, D)$$

- High weighting values for

  1) high term frequency in a document, and

  2) a low frequency of term *t* in all documents *D*

- TF-IDF weighting useful for:
  - Better vector space model
    - e.g. for document classification
    - e.g. to compute cosine similarity between documents
  - Identify stop words in various subject fields

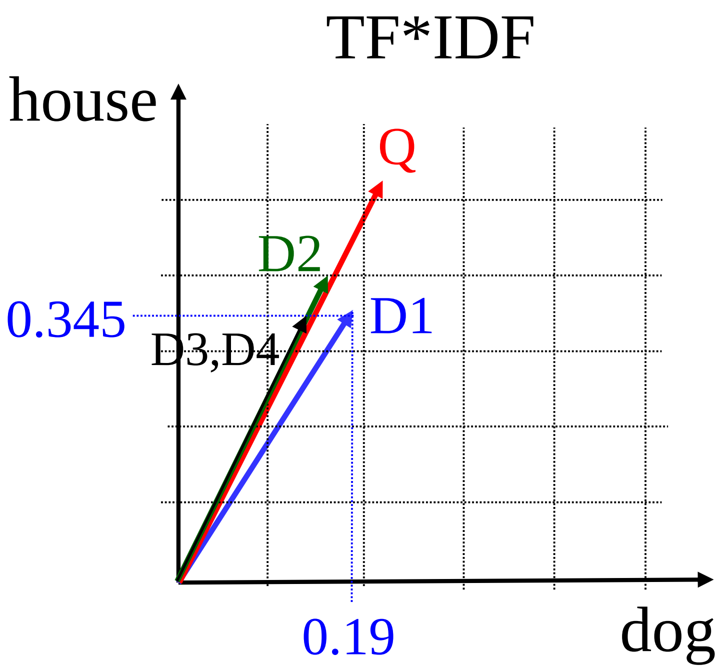
TF-IDF Weighting - Example

Query:

- Q: "dog house"

Documents:

- D1: "A dog plays with another dog."
- D2: "A dog chases a boy in the house."
- D3: "Good boy."
- D4: "Well done."



TF(dog,D1) = 0.5+0.5*2/6 = 0.67
TF(house,D1) = 0.5+0

IDF(dog) = log(4/(1+2)) = 0.29
IDF(house) = log(4/(1+1)) = 0.69

TFIDF(dog,D1) = 0.67*0.29 = 0.19
TFIDF(house,D1)=0.5*0.69 = 0.345

Use Euclidean distance!

Summary

- Much available information is stored in text databases
- Extract structure
  - POS tagging, parsing
- Extract meaning
  - Semantic role labelling
  - Understanding word meaning suffers from ambiguities
  - Pragmatics requires context
- Information retrieval
  - Vector Space Model with TF-IDF weighting