What is semantics

- Semantics is many things to many people
- My definition of semantics (general public)
- Semantics in Search Engines
- Successful semantics
  - Entity extraction
  - Topic segmentation
  - Affect analysis, domain analysis
- Semantics in Products
  - Voxalead News
  - Urbanizer
- Future
  - LOD
Semantic Entities

Unified Access to Heterogeneous Audiovisual Archives
Y. Avrithis, G. Stamou, M. Wallace et al.

10.3217/jucs-009-06-0510
The Multidimensional Learning Model: A Novel Cognitive Psychology-Based Model for Computer Assisted Instruction in Order to Improve Learning in Medical Students

Tarek M. Abdelhamid, M.D., Formerly from the Medical Education Development Office, The University of Auckland School of Medicine and Health Science

Figure 3 According to the level of processing approach the semantic (meaningful) coding of information provides the deepest form of processing.
Major levels of linguistic structure

Source: National Visualization and Analytics Center: *Illuminating the Path: The Research and Development Agenda for Visual Analytics* p.110, 2005

Author: James J. Thomas and Kristin A. Cook (Ed.)
Syntactic and semantic level description of the surface text

Predicate-Argument relations

Argument categories

Syntactic relations

Syntactic categories

Surface Text

<table>
<thead>
<tr>
<th>Subject</th>
<th>Complement</th>
<th>Complement</th>
<th>Complement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase</td>
<td>Verb phrase</td>
<td>Prep. phrase</td>
<td>Prep. phrase</td>
</tr>
<tr>
<td>One exon</td>
<td>is spliced out</td>
<td>of the MLC3nm transcript</td>
<td>in smooth muscle</td>
</tr>
</tbody>
</table>

lost component | entity getting spliced | location referring to tissue |

IE core argument (purpose argument)

component in gene | RNA | tissue | alternative mRNA

to give an alternative product
Semantic Level

Open XML File Format Higher Level
- Schema Level Validation
- Additional Semantic Validation

Open XML File Format Base Level
- Stream Reading/Writing
- Open XML Low Level DOM
- Open XML Packaging API

System Support
- .NET Framework 3.5
- System.IO.Packaging
- Open XML Schemas

Announcing the Release of the August 2009 CTP for the Open XML SDK

Brian Jones
27 Aug 2009 9:21 AM
Semantic Map

LearningTip #50: Strategies Help Reluctant Silent Readers Read to Learn
By Joyce Melton Pagés, Ed.D.
Educator, President of KidBibs
Semantic Structure

Foundations of language: brain, meaning, grammar, evolution
By Ray Jackendoff

http://www.cse.iitk.ac.in/users/amit/books/jackendoff-2002-foundations-of-language.html/
Semantic Levels and Computer Evidence

http://www.anguas.com  Mar 18 2010
Semantic Gap

Original Image

Annotated Image

http://mklab.iti.gr/svmcf

Region Annotation

<table>
<thead>
<tr>
<th>Natural-Person: 0.181767</th>
<th>Waterfall: 0.182091</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sailing-Boat: 0.208791</td>
<td>Wave: 0.158562</td>
</tr>
<tr>
<td>Sand: 0.151650</td>
<td>Dried-Plant: 0.179367</td>
</tr>
<tr>
<td>Building: 0.178949</td>
<td>Dried-Plant-Snowed: 0.207337</td>
</tr>
<tr>
<td>Pavement: 0.222864</td>
<td>Foliage: 0.184398</td>
</tr>
<tr>
<td><strong>Road: 0.288616</strong></td>
<td>Grass: 0.233446</td>
</tr>
<tr>
<td>Body-Of-Water: 0.237975</td>
<td>Tree: 0.165337</td>
</tr>
<tr>
<td>Cliff: 0.156257</td>
<td>Trunk: 0.179618</td>
</tr>
<tr>
<td>Cloud: 0.177770</td>
<td>Snow: 0.243884</td>
</tr>
<tr>
<td>Mountain: 0.187080</td>
<td>Sunset: 0.180170</td>
</tr>
<tr>
<td>Sea: 0.150582</td>
<td>Car: 0.184918</td>
</tr>
<tr>
<td>Sky: 0.181040</td>
<td>Ground: 0.220350</td>
</tr>
<tr>
<td>Stone: 0.188128</td>
<td>Lamp-Post: 0.178783</td>
</tr>
<tr>
<td></td>
<td>Statue: 0.181745</td>
</tr>
</tbody>
</table>
Semantic Filtering

M. Naphade, I. Kozintsev and T. Huang, “A Factor Graph Framework for Semantic Video Indexing”, accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology
http://domino.research.ibm.com
Semantic Vocabulary

Semantic Needs

Internet Hierarchy of Needs

- **Actualization**
  the Web becomes a frictionless tool for personal growth and fulfilment

- **Semantic needs**
  the ability to derive meaning from language, content and context

- **Organization needs**
  the ability to sort and search, based on titles, metatags, and document contents

- **Connectivity needs**
  the ability to connect to and between documents and sites, and its subsequent implications

- **Existence needs**
  basic elements for the Internet to work: computers connecting to each other and volume of documents

© VortexDNA 2007 May be freely reproduced if attributed

Unified Semantic Ecosystem

Swoogle Semantic Web indexer
Semantic Web

The element of the Semantic Web
Can be encoded in XML
Simplicity and mathematical consistency
This is called Resource Description Framework (RDF)

http://www.w3.org/2007/Talks/1211-whit-tbl/
Tim Berners-Lee
MIT Computer Science & Artificial Intelligence Laboratory (CSAIL)
Southampton University School of Electronics and Computer Science
World Wide Web Consortium
Semantic Web

- RDF data...

http://www.w3.org/2007/Talks/1211-whit-tbl/
Tim Berners-Lee
MIT Computer Science & Artificial Intelligence Laboratory (CSAIL)
Southampton University School of Electronics and Computer Science
World Wide Web Consortium
Semantics

- Upper level classes
- Relations between concepts
- Predicates extracted from text
- Restrictions on content
- Labels on visual data
- « meaning »
- Shared master data space
- RDF triples
MY EXPERIENCE WITH SEMANTICS
My thinking in 2007

Language identification

Tokenization

Word segmentation

Morphological analysis

Part of speech tagging

Lemmatisation
  Spelling correction, synonym expansion

Phrase chunking

Dependency parsing
  Entity recognition
  Sentiment analysis
  Event analysis

Document analysis
  Summarization
  Topic identification
  Classification
  Clustering

Low level semantics

Semantics
In 2008, I joined Exalead

16 milliards de pages web
2 milliards d’images
50 millions de vidéos

www.exalead.com/search
Searching for known items with facets

ASIDE ABOUT ENTERPRISE SEARCH
Semantic Dimensions

Facets

Hunting Gathering
Shopping
END OF ASIDE
BACK TO CLIENTS DESIRE FOR SEMANTICS
I learned that this is what Clients think

Anything added to but not in the original text is semantics

Anything fancier than indexing original strings

Anything that has "language"
Semantics is three things to the ordinary person

Is this an example of that?

Are these two things the same?

What is the relation between these two things?
Semantics is three things to the ordinary person

Is this an example of that?
Language identification, entity extraction, word sense disambiguation, affect analysis

Are these two things the same?
Tokenization, spelling correction, morphological analysis, summarization, synonym expansion, paraphrase, anaphor

What is the relation between these two things?
Parsing, event extraction, relation extraction, classification, clustering
Semantics in a Search Engine

- **Text fields**
  - “... the certification test is documented in the report...”

- **Numerical fields**
  - 3.14159265

- **Date**
  - 16/11/1957

- **GPS coordinates, real world**
  - 48.451065619, 1.4392089

- **Categories**
  - Top/Animals/Pets/Dog

- **Value separated fields**
  - Color: outside red; interior: white; trimming: silver

- **Metadata**
  - Source: dailymotion
Exalead Fields capture most types, documents capture rows

- **Exalead Field types**
  - Text
  - Dates
  - Numbers
  - Categories

- **Exalead Documents**
  - Each doc an entity, database rows, or
  - Each doc a text, with entities added in fields
Semantics Processes in a Search Engine

- autosuggestion
- language identification
- related terms
- related queries
- spell checker
- faceted search
- local search

INDEX

- lemmatisation
- synonyms
- phonetic
- cross language
- Ontology Matcher

Data Sources

- QueryMatcher
- FastRules
- HTMLRelevantContextExtractor
- Categorizer
- Clusterer

Semantics Processes in a Search Engine
AutoSuggestion Example

Auto completion of user query as they type
Auto Suggestion Resource

```xml
<SuggestDictionary xmlns="exa:com.exalead.mot.suggest.v10" maxEntries="3" subExpr="false" subString="false">
    <SuggestDictEntry entry="airport" score="1"/>
    <SuggestDictEntry entry="air" score="2"/>
    <SuggestDictEntry entry="airlines" score="3"/>
    <SuggestDictEntry entry="aircraft" score="4"/>
    <SuggestDictEntry entry="airelles" score="1"/>
    <SuggestDictEntry entry="airpower" score="1"/>
    ....
    <SuggestDictEntry entry="zebra" score="1"/>
</SuggestDictionary>
```

Scores > administrator-set threshold are displayed

```xml
<SuggestDictionary xmlns="exa:com.exalead.mot.suggest.v10" maxEntries="3" subExpr="false" subString="false">
    <SuggestDictEntry entry="airport" score="1">
        <SuggestDictEntryExtraInfo info="http://www.c.com"/>
        <SuggestDictEntryExtraInfo info="http://www.d.com"/>
    </SuggestDictEntry>
    ...
</SuggestDictionary>
```

Format for URL suggestions
Synonyms, Example

DB administrator is defined as synonym of Database Administrator. This synonymy can be in one direction or both ways.
**Synonym Resource**

Example 1

```xml
<Synonyms xmlns="exa:com.exalead.mot.qrewrite.v10" equivalenceClass="false">
  <SynonymSet originalExpr="db administrator" lang="en">
    <Synonym alternativeExpr="database administrator"/>
    <Synonym alternativeExpr="data base administrator"/>
  </SynonymSet>
</Synonyms>
```

In this example the query "db administrator" generates query ("db administrator" or "database administrator" or "data base administrator") but "database administrator" and "data base administrator" queries do not perform expansion.

Example 2

```xml
<Synonyms xmlns="exa:com.exalead.mot.qrewrite.v10" equivalenceClass="true">
  <SynonymSet originalExpr="db administrator" lang="en">
    <Synonym alternativeExpr="database administrator"/>
    <Synonym alternativeExpr="data base administrator"/>
  </SynonymSet>
</Synonyms>
```

In this example any of the queries "db administrator", "database administrator" or "data base administrator" will generate the combined query ("db administrator" or "database administrator" or "data base administrator")
Phonetic Example

soundslike:Filipine

related terms:
- Cebu City
- Hotels in Philippines
- Metro Manila
- PDF
- Adobe Acrobat
- Philippine Business
- Philippine News
- Philippine Hotels
- Philippine Online
- Rights Reserved
- Travel Philippines

Languages:
- English (59%)
- French (29%)
- Tagalog (4%)
- Italian (3%)
- Others (5%)

Countries:
Phonetic Resource

contextualizable character sequence

Phonetic form

comment for linguist developer

# names and old spelling
"(oy)", "O.Y"
"(oy)", "O.Y"
"(ooy)", "O.Y"
"(oij)", "O.Y"
"(ooij)", "O.Y"

"(oeu)", "@"

# hoi (hi)
# Plockhoy, Roy, Van Royen
# Verlooy
#
# Van Ooijen

# oeuvre

## OU
# exceptions for foreign words
"r(ou)time", "OU"
"g(ou)rneur", "OU"
"<b(ou)llon", "OU"
"<r(ou)te", "OU"
"<s(ou)fleur", "OU"
"<b(ou)gie", "OU"

# routebeschrijving
# souffleurshokje
# bougiesleutel

# !! OU in names and old spelling
"(ouw)", "A.OU"

"(ou)", "A.OU"

# bouw, sjouwer, mouw, stouwen

# zout (sel), vrouw (femme), koud (froid), nou (eh bien)

For each language, phonetic rules are developed in house.

Currently available for 20 languages:

<table>
<thead>
<tr>
<th>CA</th>
<th>CS</th>
<th>DA</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>ES</td>
<td>ET</td>
<td>FI</td>
</tr>
<tr>
<td>FR</td>
<td>IT</td>
<td>NL</td>
<td>NO</td>
</tr>
<tr>
<td>PL</td>
<td>PT</td>
<td>RO</td>
<td>RU</td>
</tr>
<tr>
<td>SK</td>
<td>SL</td>
<td>SV</td>
<td></td>
</tr>
</tbody>
</table>
54 languages are currently detected
Lemmatisation

Cloning

Hopefully, this will turn into a full fledged tool to recompute the physical map and assess its reliability. At present, it does very little. This option is not linked in release 1.4. Select project. This will look for a binary file $ACEDB_DATA/subiston/whatever_you_enter.bnd. If it does not exist, the program will try.

Source: AceDB; public websites

TAR CLONING:

Two stages: (see figure 1): In the first stage a targeting vector is created, by cloning a 5000bp PCR product into the XhoI site of pEB2. The PCR product Telomere Capture URA3 pEB2 6548 bp XhoI TEL Clal 5000bp PCR product is cloned into XhoI Site. Correct orientation has: CEN TEL Plasmid is linearised by digestion.

Source: Library; public websites

S. scrofa

and also the name of the BAC from which each BES was produced. A BAC clone name e.g. CH242-196P11, provides a link to further information containing, where possible, as shown in the 'Pig BAC ends' track in human ensemble. d, porcine sequenced CH242-196P11 (CR9) contains the porcine homologue to GNMT. e, pig clone CH242-196P11 (CR9)56379 which
Cross Language Information Retrieval

Afghanistan

Show tags
Show Time Line

October 12 2009... 路透社 聯合報 委員會 主席 范文
说在阿富汗制定新战略以增加...

August 25 2009... 墨宇 阿富汗总统大选后的情况受到各方高度关注...

December 2009... 美国总统奥巴马公布了对阿富汗新战略在明年之前美国...

December 1 2009... 路透社 中国

December 1 2009... 计划于日本时间明天宣布新的阿富汗战略在此之前他一直是...

November 30 2009... 如果是友好关系更好他们有什么样的阿富汗恐怖活动他们只...
Spell Checker

Did you mean: **animal shelter**
Samples from Google N-gram corpus

- Sample of the 3-gram data in this corpus:

  ceramics collectables collectibles 55
  ceramics collectables fine 130
  ceramics collected by 52
  ceramics collectible pottery 50
  ceramics collectibles cooking 45
  ceramics collection , 144
  ceramics collection . 247
  ceramics collection </S> 120
  ceramics collection and 43
  ceramics collection at 52
  ceramics collection is 68
  ceramics collection of 76
  ceramics collection | 59
  ceramics collections , 66
  ceramics collections . 60
  ceramics combined with 46

- Sample of the 4-gram data in this corpus:

  serve as the incoming 92
  serve as the incubator 99
  serve as the independent 794
  serve as the index 223
  serve as the indication 72
  serve as the indicator 120
  serve as the indicators 45
  serve as the indispensables 111
  serve as the indispendible 40
  serve as the individual 234
  serve as the industrial 52
  serve as the industry 607
  serve as the info 42
  serve as the informal 102
  serve as the information 838
  serve as the informational 41
add a category to a document matching a compiled regular expression

Jones, G. et al. Non-hierarchic document clustering Willett Introduction Cluster analysis, or automatic classification, is a multivariate statistical technique ... henceforth a GA, for document clustering. 04 nov. 2003 - 20k

MyCategory: MachineLearning
Personnalité:
Gareth Jones, Peter Willett, Edward Arnold, Van Nostrand Reinhold
Lieu:
Berlin, Duran, Sheffield, New York, London
Organisation:
British Library, Wellcome, Everitt, American Society, ACM, Springer

FastRulesAdvantages
- Very fast rules categorization
- Support huge set of rules (> 100k rules)
- Regular expression on chunk (for example a regular expression on an url, title or text)
- Available as a standalone component (Exa and Java)

Drawbacks
- Limited syntax (only AND/OR/NOT operators are supported)
- Rules must be compiled before usage

100 000 queries with boolean large multi word expression matching 1 Mb / sec / server
RulesMatcher, Named Entities

Michelle Obama
See also: Michelle Robinson’s family tree […] 2008 Presidential election Campaigning for Barack Obama […] conservative columnist Michelle Malkin about …
en.wikipedia.org/wiki/Michelle_Obama

Oprah Winfrey’s endorsement of Barack Obama
Endorsing Barack Obama […] vote for Obama, if only because they thought her endorsement made Obama more […] “It made Obama more than a ‘Black’ candidate…
en.wikipedia.org/wiki/Oprah_Winfrey’s_endorsement_of_Barack_Obama

Marian Shields Robinson
[…] Lady of the United States Michelle Obama. She is the mother-in-law of President Barack Obama and the only surviving grandparent of Malia and Sasha Obama.
en.wikipedia.org/wiki/Marian_Shields_Robinson
SentimentAnalyzer

Overview

Review

Forget what the clue quibblers say. Blue Smoke is an ingenious Manhattan barbecue joint with a woody barroom full of red-vinyl booths, a skylit dining room, and a hot jazz club downstairs. True, some of the barbecue isn’t there yet, but each visit shows improvement, as pit master Kenny Callaghan breaks in those Missouri-made smokers. Salt-and-pepper beef ribs are terrific; so are sides and desserts, and as far as pit-smoked five grams and salmon goes, what’s the problem? This ain’t least tweestump, Texas, after all.

Most viewed comment

Went to Blue Smoke tonight. The waitstaff was friendly, even if the last sip of my beer was whisked away without so much as an "Are you finished with that?" Had the baby back ribs, with a side of fries, mac & cheese and hush-puppies. The remoulade that comes with the hush-puppies makes this dish a real specialty. The mac & cheese was creamy, to put it mildly, very tasty, but for the hint of chemically velveta-esque after-taste. The fries - excellent; they arrived fresh, hot, with a crispy exterior (no batter on these little wonders) and with the genuine taste of potato without...

Blogger entries about Blue Smoke

- Restaurant Week '08: Chase Away the Blues At Blue Smoke (New York City's Hottest Restaurant Deals)
- Summer Restaurant Week 2007 Pick: Blue Smoke (New York City's Hottest Restaurant Deals)
- Hill Country to Challenge Blue Smoke, Rubs on Their Own Butt (Grub Street)
- Restaurant Week Picks: Winter 2003 (Part 2) (New York City's Hottest Restaurant Deals)
- The Big Apple BBQ Block Party Draws a Lot From Its Vendors (Grub Street)
- Meyer Makes a Move on Citi Field (Grub Street)
- Robb Report and Hill Country Part Ways (Grub Street)
- B.K. Guests New BBQ: Will Wildwood Succeed? (Grub Street)
- Heroic Blogger Saves Burger Lovers From Winter of Discontent (Grub Street)
- Profile: Midtown Lunch 'Chili' (Midtown Lunch)
- Daisy May's BBQ Cart... the best-smash that- BBQ in Midtown (Midtown Lunch)

Details

Payment options: AMEX, Diners Club, Discover, MasterCard, Visa

People sentiments

reasonable okay spicy inspired hot baked overpriced juicy hard grilled amazing consistent negative expensive superb efficient endless tiny meaty base nasty available genuine ome type picky yummy
good popular outer tableless tasty traditional date soury exceptional delicious

10/29/2010
Categorizer

Directory

- Business and Economy
  - Consumer Goods and Services
  - Shopping
- Shopping
- Pets
- Recreation and Sports
- Pets
- Society and Culture

Documents D'apprentissage

Nouveau document
Semantics Processes in a Search Engine

- autosuggestion
- language identification
- related terms
- related queries
- spell checker
- faceted search
- local search

INDEX

- lemmatisation
- synonyms
- phonetic
- cross language
- Ontology Matcher

phonetic
lemmatisation
synonyms
stopwords
sentiment analysis
named entity

Data Sources

- QueryMatcher
- FastRules
- HTMLRelevantContextExtractor
- Categorizer
- Clusterer
• With documents as containers
• With rich semantic markup
• With database offloading
  • database queries generating documents
• Search engines can become part of a process

Search Based Applications
Search-Based Applications

Search Engines can provide a powerful search-based infrastructure...

Without Offloading
High complexity/costs, Low performance/reusability

With Offloading
Low complexity/costs, High performance/reusability
Affordance of Search Engines

Web Search Technology has evolved over the past decade to perfect distribution of information.

- Handle 100s of millions of database records
- Offer 99,999% availability
- Follow a 2-month innovation roadmap
- Support impossible to forecast traffic
- Provide sub-second response times
- Present up-to-date information
- Be usable without training
- Match low revenue model ($0,0001/session)
Databases
Search Based Applications
Search engines
Transfer from research to industry

SEMANTICS SUCCESS STORIES
Named Entities

"Named entities are phrases that contain the names of persons, organizations, locations, times, and quantities." (CoNLL 2002)


  <ENAMEX TYPE="PERSON">Jim</ENAMEX> bought
  <NUMEX TYPE="QUANTITY">300</NUMEX> shares of
  <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in
  <TIMEX TYPE="DATE">2006</TIMEX>.

http://cs.nyu.edu/cs/faculty/grishman/
VoxaleadNews Methodology

RSS feed

Video file

Stream splitter

Audio track .wav

Extract vignettes .jpg

Video re encoding .mp4

Meta data

Title
Date
Channel
...

Speech-to-Text

Text with timetracking .xml

Push API Index

Metadata
Named Entities
Cross Lingual

VECSYS research

EXALEAD
connect the dots

© DASSAULT SYSTEMES

10/29/2010
In this video:

On the broadcast tonight, politics and what's different this year. Hidden fortune being funneled into campaigns and is this the new face of the Republican party. We call from Toyota. What's the problem. This time it's a prominent media figure fired for what he said on the air. Is it fair punishment or political correctness gone and houses of worship. Americans finding a place to pray right in their own backyard. Also tonight something very exciting on the moon and lots of the nightly news begins now.

More news in New York this is NBC nightly News with Brian Williams.

Good evening. We are just 12 days away now from what could be a game changing election and tonight we have exclusive information about the campaign to present. If anyone's mind should -
cours sur le thème de la généralisation du Revenu de Solidarité Active
Transfer from research to industry

SEMANTICS SUCCESS STORY 2
## Topic Segmentation

<table>
<thead>
<tr>
<th>Sentence:</th>
<th>05</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
<th>55</th>
<th>60</th>
<th>65</th>
<th>70</th>
<th>75</th>
<th>80</th>
<th>85</th>
<th>90</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Distribution of selected terms from the *Stargazer* text, with a single digit frequency per sentence number (blanks indicate a frequency of zero).

http://people.ischool.berkeley.edu/~hearst/papers/acl94.pdf
Topic Segmentation

Figure 4: Results of the block similarity algorithm on the Stargazer text. Internal numbers indicate paragraph numbers, x-axis indicates token sequence, y number, y-axis indicates similarity between blocks centered at the sentence:

http://people.ischool.berkeley.edu/~hearst/papers/acl94.pdf

Professor Marti Hearst
She was fitted with a prosthetic nose like the kind they use in the movies for special effects. She puts it on herself with a special skin that he's at each morning, a kind of preview of what the surgery will make when she on the bed comfort. It was a benefit for the Grossman burn Foundation and the surgeons will operate on her for 3 he you should have for the favorite musician one in that former first Lady Laura Bush it was like California first Lady Muslim and
She was fitted with a prosthetic nose like the kind they use in the movies for special effects. She puts it on herself with a special skin that he's at each morning, a kind of preview of what the surgery will make when she on the bed comfort, it was a benefit for the Grossman burn Foundation and the surgeons will operate on her for 3 he you should have for the favorite musician one in that former first Lady Laura Bush it was like California first Lady Muslim and
She was fitted with a prosthetic nose like the kind they use in the movies for special effects. She puts it on herself with a special skin that he's at each morning, a kind of preview of what the surgery will make when she on the bed comfort. It was a benefit for the Grossman burn Foundation and the surgeons will operate on her for 3 he you should have for the favorite musician one in that former first Lady Laura Bush it was like California first Lady Muslim and
She was fitted with a prosthetic nose like the kind they use in the movies for special effects. She puts it on herself with a special skin that he's at each morning, a kind of preview of what the surgery will make when she on the bed comfort. It was a benefit for the Grossman burn Foundation and the surgeons will operate on her for 3 he you should have for the favorite musician one in that former first Lady Laura Bush it was like California first Lady Muslim and
She was fitted with a prosthetic nose like the kind they use in the movies for special effects. She puts it on herself with a special skin that he's at each morning, a kind of preview of what the surgery will make.

when she on the bed comfort. It was a benefit for the Grossman burn Foundation and the surgeons will operate on her for 3 he you should have

for the favorite musician one in that former first Lady Laura Bush

It was like California first Lady Muslim and
Transfer from research to industry

SEMANTICS SUCCESS STORY 3
Brief History of Semantic Fields and Affect Analysis

- **Deese (1965) semantic axes**
  - free association experiments
  - “big-small”, “hot-cold”, etc.


- **Levinson (1983) linguistic scale**
  - set of alternate or contrastive expressions
  - arranged on an axis by degree of semantic strength
Affect Lexicons

- **Lasswell Value Dictionary (1969)**
  - Eight dimensions:
    - WEALTH, POWER, RECTITUDE, RESPECT, ENLIGHTENMENT, SKILL, AFFECTION, AND WELLBEING with positive or negative orientation
    - e.g., *admire*: RESPECT *(positive)*

- **General Inquirer dictionary (Stone, *et al.* 1965)** 9051 headwords
  - 1,915 positive and 2,291 negative words (Pos/Neg)
  - also labels: Active, Passive, ... , Pleasure, Pain, ... Human, Animate, ..., Region, Route, ..., Fetch, Stay, ...

  [link](http://www.wjh.harvard.edu/~inquirer/inqdict.txt)
Clairvoyance Affect Lexicon

<table>
<thead>
<tr>
<th>lexical entry</th>
<th>POS</th>
<th>class</th>
<th>centrality</th>
<th>intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;arrogance&quot;</td>
<td>sn</td>
<td>&quot;superiority&quot;</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>&quot;gleeful&quot;</td>
<td>adj</td>
<td>“happiness”</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>&quot;gleeful&quot;</td>
<td>adj</td>
<td>“excitement”</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

84 pair affect classes (positive/negative)
Finding Emotive Words using Patterns

“appearing so…”
“feeling extremely…”
“was too…”
“looks almost…”

105 patterns with conjugated word forms
Most Common Words from All Patterns

- Using snippets from all 105 patterns:
  - 8957 good
  - 2906 important
  - 2506 happy
  - 2455 small
  - 2024 bad
  - 1976 easy
  - 1951 far
  - 1745 difficult
  - 1697 hard
  - 1563 pleased

... in all 15111 different, inflected word
### SO-PMI of emotive pattern words

<table>
<thead>
<tr>
<th>Score</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.5</td>
<td>knowlegeable</td>
</tr>
<tr>
<td>33.6</td>
<td>tailormade</td>
</tr>
<tr>
<td>32.9</td>
<td>eyecatching</td>
</tr>
<tr>
<td>29.0</td>
<td>huggable</td>
</tr>
<tr>
<td>26.2</td>
<td>surefooted</td>
</tr>
<tr>
<td>24.6</td>
<td>timesaving</td>
</tr>
<tr>
<td>22.9</td>
<td>personable</td>
</tr>
<tr>
<td>21.9</td>
<td>welldone</td>
</tr>
<tr>
<td>-17.1</td>
<td>underdressed</td>
</tr>
<tr>
<td>-17.5</td>
<td>uncreative</td>
</tr>
<tr>
<td>-18.1</td>
<td>disapproving</td>
</tr>
<tr>
<td>-18.5</td>
<td>meanspirited</td>
</tr>
<tr>
<td>-18.6</td>
<td>unwatchable</td>
</tr>
<tr>
<td>-22.7</td>
<td>discombobulated</td>
</tr>
</tbody>
</table>
Mood of a city

URBANIZER.COM
Urbanizer for iPhone
Discover your city by mood!

Urbanizer is an innovative interactive application that recommends places in your city based on mood. Whether you are planning a romantic dinner, or an evening out with friends, Urbanizer will help you discover the perfect place that suits the mood.

Vote for your city, enter your email, select city and vote:

[Enter email] [Select City] [Vote]
Select the city you want to search in:

Cities

- Calgary
- Montreal
- Ottawa
- Toronto
- Vancouver

Go to www.urbanizer.com to vote for the next city!
Urbanizer is powered by YellowPages.

Montreal

Restaurants

Business Lunch

Finding your city's mood!

An interactive application that recommends the best places to eat based on your mood. Whether you are planning a romantic dinner for two, celebrating a special occasion, or just hanging out with friends, Urbanizer will help you find that special place that suits the mood.

To vote on your favorite restaurant, select your city and vote.

13. Restaurant

14. Taj Restaurant

15. Taj Restaurant

Friends

Reviews
Restaurant Toqué

- **Price**: N/A
- **Cuisine**: French, European

**Mood Tags**:
- Quick
- Homey
- Casual
- Leisure
- Refined
- Upscale

**Friends**: 1

**Reviews**: 58

**Contact**: (514) 499-2084

**Address**: 900, Place Jean-Paul-Riopelle, Montreal
Listed in 1,000 Places To See Before You Die, Toque! is worth keeping on the Buck... via Zagat

Absolutely fantastic experience. Service, wine, and food was impeccable. Very sm... via Zagat

We paid as much here as at many Michelin starred restaurants in Europe; I only wish... via Zagat

Perfectly nice French restaurant, but not worth the money.. via Zagat

Horrible service. Way overrated. Not worth the money.. via Zagat

The decor was okay but the ambiance came off as a little business like, not romantic.... via Opentable
"...decor is not particularly elegant..."

ambiance $\downarrow 70$

Business Lunch $\approx$ service $\in [0, 40]$ V cuisine $\in [45, 100]$ V ambiance $\in [45, 100]$
URBANIZER Domain modeling

- Service
- Cuisine
- Ambiance

- Domain vocabulary
- Sentiment for domain
- Syntactic patterns
Domain analysis, same as Affect Analysis

- Deese (1965) semantic axes
  - free association experiments
  - “big-small”, “hot-cold”, etc.

...spent \{two, three\} hours...
The interactive application that recommends a restaurant based on mood. Whether you are planning a dinner out with friends, Urbanizer will help you find the perfect place that suits the mood.

Enter your email, select city and vote:
Urbanizer is powered by YellowPages.
Another example of domain modeling
USED CARS > PORSCHE IN BOSTON

Price
$30,000 $40,000 $50,000 $70,000 $200,000
Mileage
10,000 mi 40,000 mi > 100,000 mi Not specified
Year
Brand
porsche audi
Colour
black white silver gray
Fuel
gasoline
Picture
pictures only

Ads by Google

Porsche Of Boston
Porsche.SmartAutoSavings.com Get a great price on a new car at Porsche Of Boston.

Lamborghini
Kijiji.ca Buy/Sell Cars & Auto Parts Locally Free Classifieds at Kijiji Canada!

Porsche
www.Pfaff.porschedealer.com Visit Canada's Porsche Experts With Canada's Largest Porsche Inventory

Sort by: relevance | price | date

Porsche 911. 182436 mi. 2003
Boston MA porsche 911 silver 2003 182436 mi abs automatic power locks power windows cd player air conditioning ... with back leather, never raced, heated garage, no salt. annual porsche dealer service. full leather, clear armor protection, crest embossed ... or dents, all records. 40K miles. includes 1sets wheels sponges.

$42,000
Email to a friend
View on a map
Foreseeable, near term

FUTURE SUCCESSES
Dbpedia, RDF triples

About: Pulp Fiction
An Entity in Data Space: dbpedia.org

Pulp Fiction is a film directed by Quentin Tarantino, who cowrote its screenplay with Roger Avary. The film is known for its rich, eclectic dialogue, iconic mix of humor and violence, nonlinear storyline, and host of cinematic and pop culture references. The film was nominated for seven Oscars, including Best Picture. Tarantino and Avary won Best Original Screenplay. It was also the Palme d’Or at the Cannes Film Festival. A major commercial success, it revitalized the career of its leading man, John Travolta, who received an Academy Award nomination, as did costars Samuel L. Jackson and Uma Thurman. The film’s title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue. Pulp Fiction is self-referential from its opening moments, beginning with a title card that gives two dictionary definitions of “pulp.” The plot, in keeping with most of Tarantino’s other works, is presented out of chronological sequence. The picture’s self-reflexivity, unconventional structure, and extensive use of homage and pastiche have led critics to describe it as a prime example of postmodern film. Pulp Fiction is viewed as the inspiration for many later movies that adopted various elements of its style. The nature of its development, marketing, and distribution and its consequent profitability had a sweeping effect on the field of independent cinema. A cultural watershed, Pulp Fiction’s influence has been felt in several other popular media. Pulp Fiction is a 1994 American neo-noir, black comedy, gangster film directed by Quentin Tarantino. The film’s title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue. Pulp Fiction is self-referential from its opening moments, beginning with a title card that gives two dictionary definitions of “pulp.” The plot, in keeping with most of Tarantino’s other works, is presented out of chronological sequence. The picture’s self-reflexivity, unconventional structure, and extensive use of homage and pastiche have led critics to describe it as a prime example of postmodern film. Pulp Fiction is viewed as the inspiration for many later movies that adopted various elements of its style. The nature of its development, marketing, and distribution and its consequent profitability had a sweeping effect on the field of independent cinema. A cultural watershed, Pulp Fiction’s influence has been felt in several other popular media.

Pulp Fiction is a 1994 American neo-noir, black comedy, gangster film directed by Quentin Tarantino. The film’s title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue. Pulp Fiction is self-referential from its opening moments, beginning with a title card that gives two dictionary definitions of “pulp.” The plot, in keeping with most of Tarantino’s other works, is presented out of chronological sequence. The picture’s self-reflexivity, unconventional structure, and extensive use of homage and pastiche have led critics to describe it as a prime example of postmodern film. Pulp Fiction is viewed as the inspiration for many later movies that adopted various elements of its style. The nature of its development, marketing, and distribution and its consequent profitability had a sweeping effect on the field of independent cinema. A cultural watershed, Pulp Fiction’s influence has been felt in several other popular media.
Wikipedia InfoBoxes

```{Infobox Town AT |
    name = Innsbruck |
    image_coa = InnsbruckWappen.png |
    image_map = Karte-tirol-I.png |
    state = [[Tyrol]] |
    regbzk = [[Statutory city]] |
    population = 117,342 |
    population_as_of = 2006 |
    pop_dens = 1,119 |
    area = 104.91 |
    elevation = 574 |
    lat_deg = 47 |
    lat_min = 16 |
    lat_hem = N |
    lon_deg = 11 |
    lon_min = 23 |
    lon_hem = E |
    postal_code = 6010-6080 |
    area_code = 0512 |
    licence = I |
    mayor = Hilde Zach |
    website = [http://innsbruck.at] |
    }}
```

Country: Austria
State: Tyrol
Administrative region: Statutory city
Area: 104.91 km²
Population density: 1,119 /km²
Elevation: 574 m
Coordinates: 47°16' N 11°23' E
Postal code: 6010-6080
Area code: 0512
Licence plate code: I
Mayor: Hilde Zach
Website: www.innsbruck.at
Implicit tables

(a) After independent splitting phase

<table>
<thead>
<tr>
<th></th>
<th>What’s Opera Doc</th>
<th>Warner Bros</th>
<th>1957</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Duck Amuck</td>
<td>Warner Bros</td>
<td>1953</td>
</tr>
<tr>
<td>3</td>
<td>The Band Concert</td>
<td>Disney</td>
<td>1935</td>
</tr>
<tr>
<td>4</td>
<td>Duck Dodgers in the 24 1/2th Century</td>
<td>Warner Bros</td>
<td>1953</td>
</tr>
<tr>
<td>5</td>
<td>One Froggy Evening</td>
<td>Warner Bros</td>
<td>1956</td>
</tr>
<tr>
<td>6</td>
<td>Gentle The Dinosaur</td>
<td>McCoy</td>
<td>1955</td>
</tr>
<tr>
<td>7</td>
<td>Red Hot Riding Hood</td>
<td>MGM</td>
<td>1943</td>
</tr>
<tr>
<td>8</td>
<td>Porky In Wackyland</td>
<td>Warner Bros</td>
<td>1938</td>
</tr>
<tr>
<td>9</td>
<td>Gerald McBoing Boing</td>
<td>UPA</td>
<td>1951</td>
</tr>
<tr>
<td>10</td>
<td>King-Size Canary</td>
<td>MGM</td>
<td>1947</td>
</tr>
<tr>
<td>11</td>
<td>Three Little Pigs</td>
<td>Disney</td>
<td>1933</td>
</tr>
<tr>
<td>12</td>
<td>Rabbit of Seville</td>
<td>Warner Bros</td>
<td>1950</td>
</tr>
<tr>
<td>13</td>
<td>Steamboat Willie</td>
<td>Disney</td>
<td>1928</td>
</tr>
<tr>
<td>14</td>
<td>The Old Mill</td>
<td>Disney</td>
<td>1937</td>
</tr>
<tr>
<td>15</td>
<td>Bad Luck Blackie (MGM)</td>
<td>Disney</td>
<td>1949</td>
</tr>
<tr>
<td>16</td>
<td>The Great Piggy Bank Robbery</td>
<td>Warner Bros</td>
<td>1946</td>
</tr>
<tr>
<td>17</td>
<td>Popeye the Sailor Meets Sinbad the Sailor</td>
<td>Fleischer</td>
<td>1936</td>
</tr>
</tbody>
</table>

(b) After re-splitting records given the number of columns

<table>
<thead>
<tr>
<th></th>
<th>What’s Opera Doc</th>
<th>Warner Bros</th>
<th>1957</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Duck Amuck</td>
<td>Warner Bros</td>
<td>1953</td>
</tr>
<tr>
<td>3</td>
<td>The Band Concert</td>
<td>Disney</td>
<td>1935</td>
</tr>
<tr>
<td>4</td>
<td>Duck Dodgers in the 24 1/2th Century</td>
<td>Warner Bros</td>
<td>1953</td>
</tr>
<tr>
<td>5</td>
<td>One Froggy Evening</td>
<td>Warner Bros</td>
<td>1956</td>
</tr>
<tr>
<td>6</td>
<td>Gentle The Dinosaur</td>
<td>McCoy</td>
<td>1955</td>
</tr>
<tr>
<td>7</td>
<td>Red Hot Riding Hood</td>
<td>MGM</td>
<td>1943</td>
</tr>
<tr>
<td>8</td>
<td>Porky In Wackyland</td>
<td>Warner Bros</td>
<td>1938</td>
</tr>
<tr>
<td>9</td>
<td>Gerald McBoing Boing</td>
<td>UPA</td>
<td>1951</td>
</tr>
<tr>
<td>10</td>
<td>King-Size Canary</td>
<td>MGM</td>
<td>1947</td>
</tr>
<tr>
<td>11</td>
<td>Three Little Pigs</td>
<td>Disney</td>
<td>1933</td>
</tr>
<tr>
<td>12</td>
<td>Rabbit of Seville</td>
<td>Warner Bros</td>
<td>1950</td>
</tr>
<tr>
<td>13</td>
<td>Steamboat Willie</td>
<td>Disney</td>
<td>1928</td>
</tr>
<tr>
<td>14</td>
<td>The Old Mill</td>
<td>Disney</td>
<td>1937</td>
</tr>
<tr>
<td>15</td>
<td>Bad Luck Blackie (MGM)</td>
<td>Disney</td>
<td>1949</td>
</tr>
<tr>
<td>16</td>
<td>The Great Piggy Bank Robbery</td>
<td>Warner Bros</td>
<td>1946</td>
</tr>
<tr>
<td>17</td>
<td>Popeye the Sailor Meets Sinbad the Sailor</td>
<td>Fleischer</td>
<td>1936</td>
</tr>
</tbody>
</table>

(c) After alignment phase (initial table $T_1$)

(d) After refinement phase (final table $T$)

Figure 3: Applying the ListExtract technique on the cartoons list in Figure 1.
LOD² turns linked data into knowledge
Challenge: searching ad hoc, idiosyncratic spaces

http://www.imrc.kist.re.kr/wiki/LifeLog
Microsoft SenseCam


http://justgetthere.us/blog/plugin/tag/cameras
http://gadgets.infoniac.com/head-worn-camera-for-special-events.html
Bike My Way

View of Puget Sound. Olympic Mountains in the background.
Insta Mapper

Click on positions to select track endpoints and to see other options.
Challenge: searching ad hoc, idiosyncratic spaces


http://www.imrc.kist.re.kr/wiki/LifeLog
Conclusions

- **Semantics**
  - What is this? Are these the same? Who does what to whom?

- **Search Engines now handle rich semantics**
  - Search based applications, between DB and IR

- **Semantic Research**
  - Successful transfer to industry

- **Near Future**
  - Integration crowd source knowledge Wikipedia
  - Linked open data
Thank you