

Terms over LOAD:
Leveraging Named Entities for Cross-Document
Extraction and Summarization of Events

Andreas Spitz and Michael Gertz

Heidelberg University, Institute of Computer Science
Database Systems Research Group, Heidelberg
{spitz, gertz}@informatik.uni-heidelberg.de

SIGIR '16
Pisa, July 20, 2016





Mark Spitz

From Wikipedia, the free encyclopedia

Mark Andrew Spitz (born February 10, 1950) is an American former competition swimmer, nine-time Olympic champion, and former world record-holder in seven events. He won **seven gold medals** at the **1972 Summer Olympics** in Munich, an achievement surpassed only by Michael Phelps, who won eight golds at the **2008 Summer Olympics** in Beijing. Spitz set new world records in all seven events in which he competed in 1972, an achievement that still stands. Since the year 1900, no other swimmer has gained so great a percentage of all the medals awarded for Olympic events held in a single Games.



WIKIPEDIA
The Free Encyclopedia



Mark Spitz

From Wikipedia, the free encyclopedia

Mark Andrew Spitz (born February 10, 1950) is an American former competition swimmer, nine-time Olympic champion, and former world record-holder in seven events. He won **seven gold medals** at the **1972 Summer Olympics** in Munich, an achievement surpassed only by **Michael Phelps**, who won eight golds at the **2008 Summer Olympics** in Beijing. Spitz set new world records in all seven events in which he competed in 1972, an achievement that still stands. Since the year 1900, no other swimmer has gained so great a percentage of all the medals awarded for Olympic events held in a single Games.



WIKIPEDIA
The Free Encyclopedia

Olympia Schwimmhalle

From Wikipedia, the free encyclopedia

The **Olympia Schwimmhalle** is an aquatics centre located in the **Olympiapark** in Munich, Germany. It hosted the swimming, diving, water polo, and the swimming part of the modern pentathlon events at the **1972 Summer Olympics**. At the 1972 Olympics, the stadium had a 9000-seat capacity which was reduced to 1,500 soon after. During the 1972 Olympics, the Olympic Records in all 29 Olympic swimming events were broken as well as the World Records in 20 events.^[*citation needed*]

The Schwimmhalle is unique for its roof construction which is a lightweight stressed-skin structure. This curved structure bears loads through tension only, not compression. The double curvature in the roof design is what provides support which is further stabilized through pretensioned guy wires.

The Olympia Schwimmhalle is where swimmer **Mark Spitz** broke the record for most individual gold medals won in a single Olympics with seven gold medals. This record was not surpassed until fellow swimmer **Michael Phelps** won eight gold medals at the **2008 Summer Olympics** in Beijing.

1972 Summer Olympics

From Wikipedia, the free encyclopedia

The **1972 Summer Olympics** (German: *Olympische Sommerspiele 1972*), officially known as the **Games of the XX Olympiad**, was an **international multi-sport event** held in **Munich, West Germany**, from August 26 to September 11, 1972.

Steve Genter

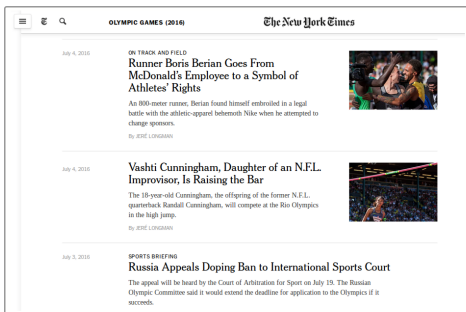
From Wikipedia, the free encyclopedia

Robert Steven Genter (born January 4, 1951) is an American former competition swimmer and three-time Olympic medalist. He was freestyle specialist who earned a gold medal as a member of the winning U.S. team in the 4x200-meter freestyle relay at the **1972 Summer Olympics** in Munich, Germany. He also won silver medals in the 200-meter and 400-meter freestyle events.

Swimming at the 1972 Summer Olympics

From Wikipedia, the free encyclopedia

The **1972 Summer Olympics** were held in Munich, West Germany, 29 events in **swimming** were contested. There was a total of 532 participants from 52 countries competing.



The screenshot shows a mobile view of The New York Times website. At the top, there is a navigation bar with a menu icon, a search icon, and the text "OLYMPIC GAMES (2016)" and "The New York Times". Below this, three news articles are listed. Each article includes a date, a category, a title, a short summary, and a byline. The first article is about a runner, the second about a high jumper, and the third is a sports briefing about a doping ban appeal.

July 4, 2016 **ON TRACK AND FIELD**
Runner Boris Berian Goes From McDonald's Employee to a Symbol of Athletes' Rights
An 800-meter runner, Berian found himself embroiled in a legal battle with the athletic-apparel behemoth Nike when he attempted to change sponsors.
By JEFF LONGMAN

July 4, 2016 **Vashti Cunningham, Daughter of an N.F.L. Improvisor, Is Raising the Bar**
The 18-year-old Cunningham, the offspring of the former N.F.L. quarterback Randall Cunningham, will compete at the Rio Olympics in the high jump.
By JEFF LONGMAN

July 3, 2016 **SPORTS BRIEFING**
Russia Appeals Doping Ban to International Sports Court
The appeal will be heard by the Court of Arbitration for Sport on July 19. The Russian Olympic Committee said it would extend the deadline for application to the Olympics if it succeeds.


OLYMPIC GAMES (2016) The New York Times

July 4, 2016 ON TRACK AND FIELD

Runner Boris Berian Goes From McDonald's Employee to a Symbol of Athletes' Rights

An 800-meter runner, Berian found himself embroiled in a legal battle with the athletic-apparel behemoth Nike when he attempted to change sponsors.


By JEFF LONGMAN



July 4, 2016 Vashti Cunningham, Daughter of an N.F.L. Improvisor, Is Raising the Bar

The 18-year-old Cunningham, the offspring of the former N.F.L. quarterback Randall Cunningham, will compete at the Rio Olympics in the high jump.

By JEFF LONGMAN



July 3, 2016 SPORTS BRIEFING

Russia Appeals Doping Ban to International Sports Court

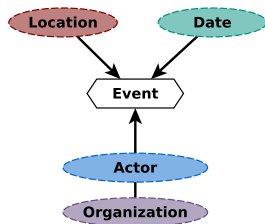
The appeal will be heard by the Court of Arbitration for Sport on July 19. The Russian Olympic Committee said it would extend the deadline for application to the Olympics if it succeeds.



Motivation

Definition: Event

“Something that happens at a given place and time between a group of actors.”
[CSG⁺02]



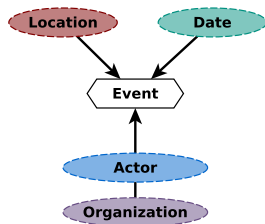
Motivation

Definition: Event

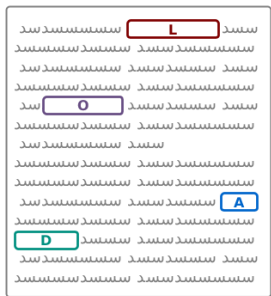
“Something that happens at a given place and time between a group of actors.”
[CSG⁺02]

For large document collections,
how can we...

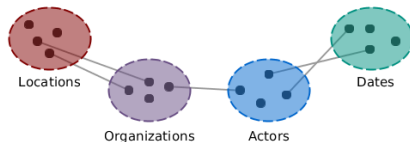
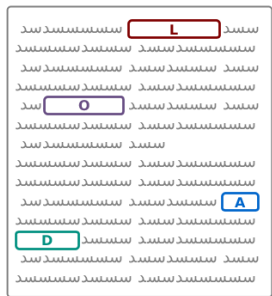
- obtain events from unstructured text?
- identify connections across documents?
- support ad-hoc event search?



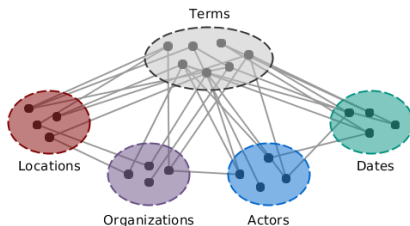
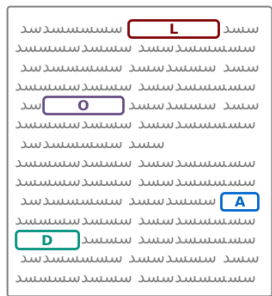
Graph Extraction from Unstructured Text



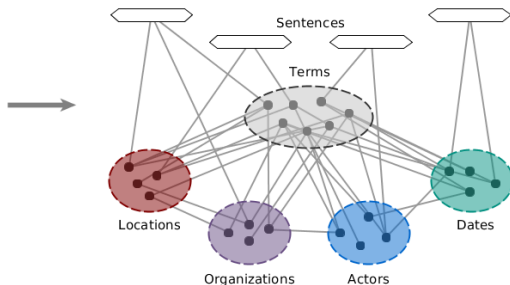
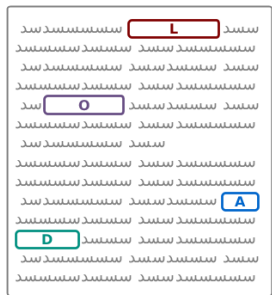
Graph Extraction from Unstructured Text



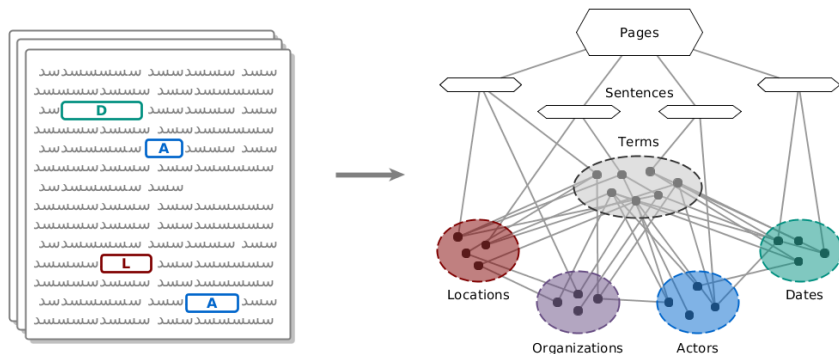
Graph Extraction from Unstructured Text



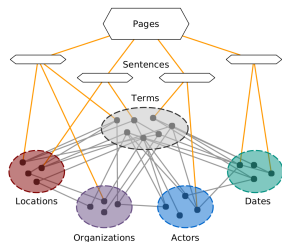
Graph Extraction from Unstructured Text



Graph Extraction from Unstructured Text



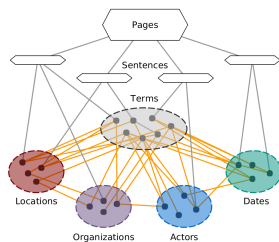
Edge Weight Generation



For edges (x, y) for which y is a page or sentence, count only (co-) occurrences:

$$\omega(x, y) = \begin{cases} 1 & \text{if } y \text{ contains } x \\ 0 & \text{otherwise} \end{cases}$$

Edge Weight Generation



For edges (x, y) for which y is a page or sentence, count only (co-) occurrences:

$$\omega(x, y) = \begin{cases} 1 & \text{if } y \text{ contains } x \\ 0 & \text{otherwise} \end{cases}$$

For edges (x, y) between entity types and terms, aggregate co-occurrence instances I : sum over similarities derived from sentence distances s .

$$\omega(x, y) := \sum_{i \in I} \exp(-s(x, y, i))$$

LOADing Wikipedia

For the entire English Wikipedia
(~ 4.5M articles with annotations):

- use only **unstructured** text.
- exclude pages of lists.
- exclude info boxes.
- exclude references.

Extract named entities with:

- Stanford NER for **locations**, **organizations** and **actors** [FGM05]
- Heideltime for **dates** [SG13]

The screenshot shows the Wikipedia article for the 1972 Summer Olympics. The article title is "1972 Summer Olympics" and it is categorized as "From Wikipedia, the free encyclopedia". The main text describes the event as the 1972 Summer Olympics (German: Olympische Sommerspiele 1972), officially known as the Games of the XX Olympiad, which was an international multi-sport event held in Munich, West Germany, from August 26 to September 11, 1972. It notes that the sporting nature of the event was largely overshadowed by the Munich massacre in which eleven Israeli athletes and coaches and a West German police officer were killed, five Black September Palestinian terrorists died. The article also mentions that the 1972 Summer Olympics were the second Summer Olympics to be held in Germany, after the 1936 Games in Berlin, which had taken place under the Nazi regime. Mindful of the connection, the West German Government was eager to take the opportunity of the Munich Olympics to present a new, democratic and optimistic Germany to the world, as shown by the Games' official motto, "Die Heiteren Spiele" (lit. "the cheerful Games"). The logo of the Games was a blue solar logo (the "Bright Sun") by Ott Aicher, the designer and director of the visual conception commission. The Olympic mascot, the dachshund "Waldi", was the first officially named Olympic mascot. The Olympic Fanfare^[H] was composed by Herbert Rehebin, a companion of Bert Kaempfert.

The article includes a sidebar with navigation links such as "Main page", "Current events", "Random article", "Donate to Wikipedia", "Wikipedia store", "Interaction", "Help", "About Wikipedia", "Community portal", "Recent changes", "Contact page", "Tools", "What links here", "Related changes", "Upload file", "Special pages", "Permanent link", "Page information", "Wikidata item", and "Cite this page".

On the right side, there is a section titled "Games of the XX Olympiad" featuring the Olympic rings logo and a blue solar logo. Below the logos, it lists the "Host city" as Munich, West Germany, and the "Motto" as "The Happy Games".

Wikipedia LOAD Graph

edges	<i>LOC</i>	<i>ORG</i>	<i>ACT</i>	<i>DAT</i>	<i>TER</i>	<i>SEN</i>	<i>PAG</i>
<i>LOC</i>	0						
<i>ORG</i>	91	0					
<i>ACT</i>	276	106	0				
<i>DAT</i>	83	46	128	0			
<i>TER</i>	183	94	317	57	0		
<i>SEN</i>	71	21	84	38	412	0	
<i>PAG</i>	0	0	0	0	0	54	0
nodes	2.7	3.4	7.1	0.2	4.9	53.5	4.5

Number of edges and nodes (in millions) of the LOAD graph of the English Wikipedia. $\sim 2\text{B}$ edges and $\sim 76\text{M}$ nodes in total.

Single Entity Queries

How can we rank nodes in one set Y by their neighbours in set X ?
Adapt *tf-idf* scores to the graph [RV13]!

- Term frequency:
edge weights
 $tf(x, y) \approx \omega(x, y)$
- Inverse document frequency:
number of neighbours
 $idf(x) \approx \frac{|Y|}{deg_Y(x)}$

$$r(x, y) \approx \omega(x, y) \log \frac{|Y|}{deg_Y(x)}$$

Single Entity Queries

How can we rank nodes in one set Y by their neighbours in set X ?
Adapt *tf-idf* scores to the graph [RV13]!

- Term frequency:
edge weights
 $tf(x, y) \approx \omega(x, y)$
- Inverse document frequency:
number of neighbours
 $idf(x) \approx \frac{|Y|}{deg_Y(x)}$

$$r(x, y) \approx \omega(x, y) \log \frac{|Y|}{deg_Y(x)}$$

$\langle LOC : (ACT, \text{Mark Spitz}) \rangle$

location	score
munich	1.00000
us	0.70651
states	0.49010
united states	0.46918

Query: $\langle Y : (X, \text{value}) \rangle$

Multi-Entity Queries

How can we rank nodes in Y by neighbours in multiple sets X^n ?
Combine individual set scores:

$$r(\vec{x}, y) := \frac{1}{n} \eta(\vec{x}, y) \sum_{i=1}^n r(x_i, y)$$

Multi-Entity Queries

How can we rank nodes in Y by neighbours in multiple sets X^n ?
Combine individual set scores:

$$r(\vec{x}, y) := \frac{1}{n} \eta(\vec{x}, y) \sum_{i=1}^n r(x_i, y)$$

Ensure triangular cohesion when combining results:

$$\eta(\vec{x}, y) := \begin{cases} 1 & \text{if } \sum_{i=1}^n \sum_{j>i}^n M_{yx_i} M_{yx_j} > 1 \\ 0 & \text{otherwise} \end{cases}$$

Where M is the adjacency matrix of the graph.

Multi-Entity Query Examples

$\langle DAT : (ACT, Mark Spitz), (LOC, Munich) \rangle$

date	score
1972-08-29	0.50851
1972-08-31	0.48217
1972-09-05	0.22738
1947-03-10	0.10511
2006-09-07	0.09226

Multi-Entity Query Examples

$\langle DAT : (ACT, \text{Mark Spitz}), (LOC, \text{Munich}) \rangle$

date	score
1972-08-29	0.50851
1972-08-31	0.48217
1972-09-05	0.22738
1947-03-10	0.10511
2006-09-07	0.09226

$\langle TER : (ACT, \text{Mark Spitz}), (LOC, \text{Munich}), (DAT, 1972) \rangle$

term	score
olymp	0.89630
medal	0.54205
gold	0.43211
won	0.38904
record	0.34548

Summarization: Sentence Queries

How can sentences in S be used to describe combinations of entities in X^n ?

Find a sentence that contains them:

$$r(\vec{x}, s) := \sum_{i=1}^n M_{sx_i}$$

Summarization: Sentence Queries

How can sentences in S be used to describe combinations of entities in X^n ?

Find a sentence that contains them:

$$r(\vec{x}, s) := \sum_{i=1}^n M_{sx_i}$$

$\langle SEN : (ACT, \text{Mark Spitz}) \rangle$

Mark Spitz of the United States had a spectacular run, lining up for seven events, winning seven Olympic titles and setting seven world records.

Entity Linking: Document Queries

Since we created the LOAD graph from Wikipedia, can we link entities in X^n to pages P ?

Use sentences to find the page that contains them most frequently:

$$r(\vec{x}, p) := \sum_{s \in S} \sum_{i=1}^n M_{sx_i} M_{sp}$$

Entity Linking: Document Queries

Since we created the LOAD graph from Wikipedia, can we link entities in X^n to pages P ?

Use sentences to find the page that contains them most frequently:

$$r(\vec{x}, p) := \sum_{s \in S} \sum_{i=1}^n M_{sx_i} M_{sp}$$

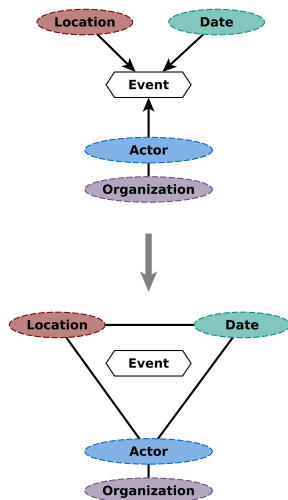
$\langle \text{PAG} : (\text{ACT}, \text{Mark Spitz}) \rangle$

Wiki page ID 66265: Mark Spitz

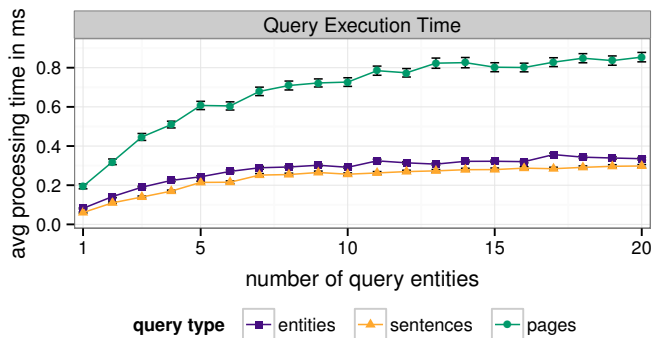
Event Extraction and Completion

Intuition:

- Events correspond to triangular structures in the network
- Participating entities can be used to complete events



Query Answering Speed

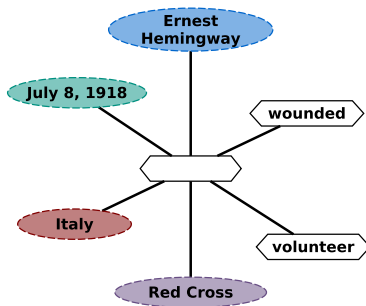


Asymptotic complexity of entity queries: $\mathcal{O}(deg_X(y) deg_Y(x))$

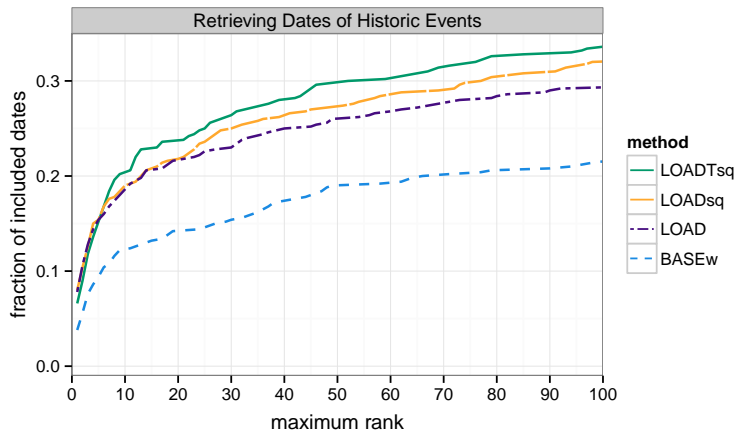
Historic Event Evaluation Data

Evaluation data set from a “This Day in History” website [Gui95]

- old enough to not contain Wikipedia data
- exactly one date per sentence
- 500 hand-annotated historic events
- example: Ernest Hemingway, Red Cross volunteer, wounded in Italy on 1918-07-08.



Evaluation on Historic Event Data



NER based on Wikipedia & Wikidata



WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

Interaction

- Help

Article [Talk](#)

1984 Summer Olympics

From Wikipedia, the free encyclopedia

The **1984 Summer Olympics**, officially known as the **Games of the XXIII Olympiad**, was an international **multi-sport event** held in **Los Angeles, California, United States** in 1984. When **Tehran**, the only other interested city on the international level, declined to bid due to the concurrent Iranian political and social changes, the IOC awarded Los Angeles the Games for the second occasion Los Angeles hosted the Olympics, having first hosted in 1932.

Los Angeles

From Wikipedia, the free encyclopedia

Coordinates: 34°03′N 118°15′W

This article is about the U.S. city. For the county in California, see *Los Angeles County*. For other uses, see *Los Angeles (disambiguation)*.
 "L.A." and "City of Los Angeles" redirect here. For other uses, see *L.A. (disambiguation)* and *City of Los Angeles (disambiguation)*.

Los Angeles (/loʊˈlɑːndʒələs/ *and* /təˈlɑːnjəs/ *less* AW-jə-lɑːs or AW-jə-lɪs^[d]) officially the **City of Los Angeles** and often known by its initials **L.A.**, is the second-largest city in the United States, the most populous city in the U.S. state of California, and the county seat of Los Angeles County. Situated in Southern California, Los Angeles is known for its Mediterranean climate, ethnic diversity, sprawling metropolis, and as a major center of the American entertainment industry. Los Angeles lies in a large coastal basin surrounded on three sides by mountains reaching up to and over 10,000 feet (3,000 m).

Los Angeles, California

Charter city

City of Los Angeles



Tehran

From Wikipedia, the free encyclopedia

Coordinates: 35°41′46″N 51°29′23″E

This article is about the Iranian capital city. For other uses, see *Tehran (disambiguation)*.

Tehran (/tɛhˈrɑːn/ pronunciation تهران; Persian: تهران - Tēhrān) is the capital of Iran and Tehran Province. With a population of around 9 million in the city and 16 million in the wider metropolitan area,^[d] Tehran is the largest city and urban area of Iran, the second largest city in Western Asia, and the 3rd largest in the Middle East. It is ranked 29th in the world by the population of its metropolitan area.^[d]

In the Classical era, part of the present-day city of Tehran was occupied by a Median city which in the *Avesta* occurs as *Rhaga*.^[d] It was destroyed by the Mongols in the early 13th century, and remains now as a city in Tehran Province, located towards the south end of the modern-day city of Tehran.

Tehran

تهران

Metropolis

کلانشهر تهران · Tehran Metropolis





Summary

Ongoing work:

- online search and query interface for Wikipedia
- streaming model for online news
- inclusion of parts-of-speech

LOAD summary:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and fast
- language-agnostic with entity linking

Summary

Ongoing work:

- online search and query interface for Wikipedia
- streaming model for online news
- inclusion of parts-of-speech

LOAD summary:

- fast entity and event exploration
- can support most entity-related IE tasks
- can be extended to any kind of entity
- scalable and fast
- language-agnostic with entity linking

LOAD your data before you do entity-based analyses.

Available for download:

- Wikipedia LOAD network (Stanford NER)
- Wikipedia LOAD network (Wikidata)
- Code for generating LOAD networks
- Code for LOAD query interface



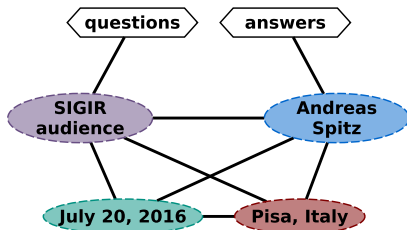
<http://dbs.ifi.uni-heidelberg.de/index.php?id=load>

Available for download:

- Wikipedia LOAD network (Stanford NER)
- Wikipedia LOAD network (Wikidata)
- Code for generating LOAD networks
- Code for LOAD query interface



<http://dbs.ifi.uni-heidelberg.de/index.php?id=load>



Bibliography



Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman.

Corpora for topic detection and tracking.

In *Topic Detection and Tracking*. Springer, 2002.



Jenny Rose Finkel, Trond Grenager, and Christopher Manning.

Incorporating non-local information into information extraction systems by Gibbs sampling.

In *ACL*, 2005.



Robert A Guiseppi.

History world: On this day in history, 1995.

<http://history-world.org/ontd.htm> (2015-10-02).



François Rousseau and Michalis Vazirgiannis.

Graph-of-word and TW-IDF: new approach to ad hoc IR.

In *CIKM*, 2013.



Jannik Strötgen and Michael Gertz.

Multilingual and cross-domain temporal tagging.

Language Resources and Evaluation, 47(2):269–298, 2013.