
Algorithms and Applications for Web-Scale Knowledge Graphs



Marco Ponza

Supervisor

Prof. Paolo Ferragina

Menu

1. Entity Annotation

- The Modeling of Knowledge
- Terminology
- The Annotation Pipeline
- Applications
- A New Text Representation

2. Work done in the first year

- Entity Relatedness
- Document Aboutness

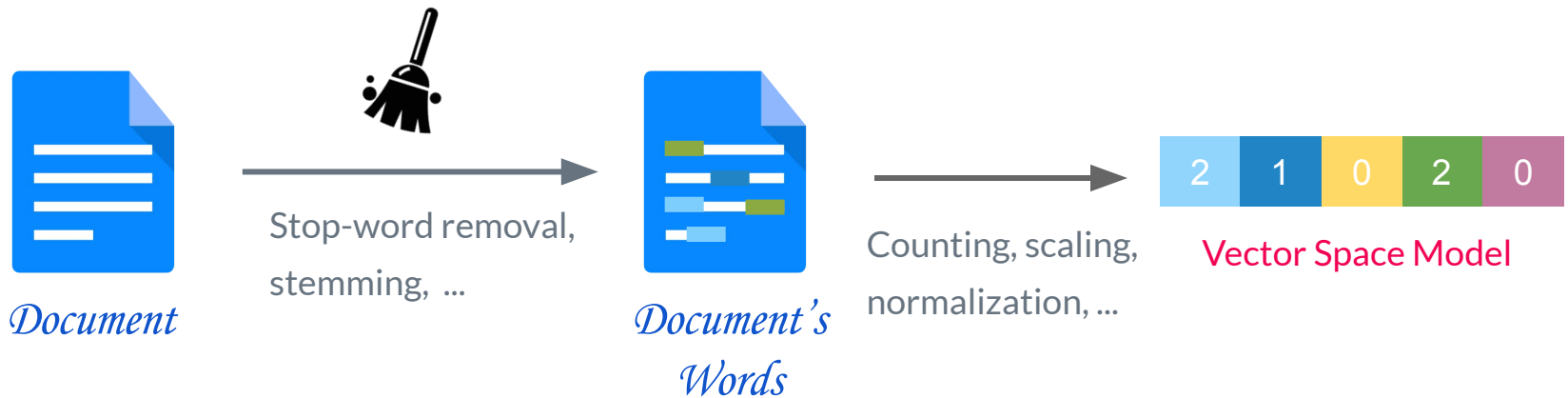
3. Future Work

1.

Entity Annotation

The Modeling of Knowledge

- ▷ Classical approaches
 - Document Knowledge = Words
 - Bag-of-words (aka BoW)
 - **Vector Space Model** (aka VSM) (Salton, 1971)



The Modeling of Knowledge

- ▷ Well-known **issues** (Jurafsky, '00)
 - **Ambiguity (Polysemy and Synonymy)**

Jaguar



Jaguar (felin)

or



Jaguar_Cars



The Modeling of Knowledge

- ▷ Well-known **issues** (Jurafsky, '00)
 - **Ambiguity (Polysemy and Synonymy)**
 - **Semantic Connections**



Barack_Obama



United_States

The Modeling of Knowledge

▷ Well-known **issues** (Jurafsky, '00)

- **Ambiguity (Polysemy and Synonymy)**
- **Semantic Connections**



▷ Algorithmic solutions

- **Latent Approaches** (e.g. LSI/LSA, Word2Vec)
 - **Unintelligible** for humans (Gabrilovich IJCAI '07)
- **“Knowledge is Power” Hypothesis** (Lenat, '91; Gabrilovich SIGIR '16)
 - **Semantic** and **unambiguous** concepts
 - Depend on the design of **Entity Annotators**

Entity Annotation

Terminology

- ▷ Wikipedia Knowledge Graph
- ▷ Node?





WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[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](#)
[Cite this page](#)

[Print/export](#)
[Create a book](#)
[Download as PDF](#)
[Printable version](#)

[In other projects](#)
[Wikimedia Commons](#)
[Wikiquote](#)
[Wikisource](#)

[Languages](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#)

[Read](#) [View source](#) [View history](#)

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see [Da Vinci \(disambiguation\)](#).

*This is a [Renaissance Florentine](#) name. The name *da Vinci* is an indicator of birthplace, not a family name; this person is properly referred to by the given name [Leonardo](#).*

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ictionary, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,^{[1][2][3]} he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "[Universal Genius](#)" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".^[4] According to art historian [Helen Gardner](#), the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".^[4] Marco Rosci notes that while there is much speculation regarding his life and personality, his view of the world was logical rather than mysterious, and that the empirical methods he employed were unorthodox for his time.^[5]

Born out of wedlock to a notary, Piero da Vinci, and a peasant woman, Caterina, in Vinci in the region of Florence, Leonardo was educated in the studio of the renowned Florentine painter [Andrea del Verrocchio](#). Much of his earlier working life was spent in the service of [Ludovico il Moro](#) in Milan. He later worked in Rome, [Bologna](#) and [Venice](#), and he spent his last years in France at the home awarded to him by [Francis I of France](#).

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona Lisa* is the most famous and most parodied portrait^[6] and *The Last Supper* the most reproduced religious painting of all time, their fame approached only by [Michelangelo's](#) *The Creation of Adam*.^[4]

Leonardo da Vinci



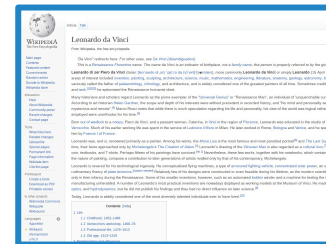
Portrait by Francesco Melzi.

Born	<div>Leonardo di ser Piero da Vinci</div> 15 April 1452 <div>Vinci, Republic of Florence (present-day Italy)</div>
Died	<div>2 May 1519 (aged 67)</div> Amboise, Kingdom of France
Known for	Art, science
Notable work	<i>Mona Lisa</i> <i>The Last Supper</i>

Entity Annotation

Terminology

- ▷ Wikipedia Knowledge Graph
- ▷ Node: Wikipedia Page (Entity)
- ▷ Link?





WIKIPEDIA
The Free Encyclopedia

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

- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page

- Tools
 - What links here

Science

From Wikipedia, the free encyclopedia

This article is about the general term. For other uses, see Science (disambiguation).

Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social sciences which study people and societies; and the formal sciences, such as mathematics. The formal sciences are often excluded as they do not depend on empirical observations. Disciplines which use science like engineering and medicine may also be considered to be applied sciences.

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in his *Book of Optics*. From classical antiquity through the 19th century, science as a type of knowledge was more closely linked to philosophy than it is now and, in fact, in the Western world, the term "natural philosophy" encompassed fields of study that are today associated with science, such as astronomy, medicine, and physics. While the classification of the material world was used by the ancient Indians and Greeks into air, earth, fire and water was more philosophical, medieval Middle Eastern scientists used practical, experimental observation to classify materials.

In the 17th and 18th centuries, scientists increasingly sought to formulate knowledge in terms of laws of nature. Over the course of the 19th century, the word "science" became increasingly associated with the scientific method itself, as a disciplined way to study the natural world. It was in the 19th century that scientific disciplines such as biology, chemistry, and physics reached their modern shapes. The same time period also included the origin of the terms "scientist" and "scientific community," the founding of scientific institutions, and increasing significance of the interactions with society and other aspects of culture.

- Wikimedia Commons
- Wikiquote
- Wikisource

- Languages
 - Адыгэбзэ

Article **Talk**

Read **View source** **View history**

Search

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).

This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,^{[1][2][3]} he epitomised the Renaissance humanist ideal.

Many historians and scholars regard him as a "Genius" or "Renaissance Man".

Invention

From Wikipedia, the free encyclopedia

"Inventor" and "Invented" redirect here. For other uses, see Invention (disambiguation).

For more details on inventions throughout history, see Timeline of historic inventions.

For the CAD design software, see Autodesk Inventor.

An **invention** is a unique or novel device, method, composition or process. The invention process is engineering and product development process. It may be an improvement upon a machine, process or system, or a completely new object or a result. An invention that achieves a completely unique function or result may be novel and not obvious to others skilled in the same field. An inventor may be taking a big step by improving upon a pre-existing idea or by radically departing from pre-existing ideas. Some inventions can be patented. A patent legally protects the intellectual property rights of an inventor. The process of obtaining a patent is often expensive.

Another meaning of invention is **cultural invention**, which is an innovative set of useful social practices or ideas passed on to others.^[1] The Institute for Social Inventions collected many such ideas in magazines and books. An important component of artistic and design creativity. Inventions often extend the boundaries of human capability.



Astronomy

From Wikipedia, the free encyclopedia

This article is about the scientific study of celestial objects. For other uses, see Astronomy (disambiguation).

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets, moons, and nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic microwave background radiation), and evolution of such objects and processes, and more generally all phenomena in the universe. A related but distinct subject, physical cosmology, is concerned with studying the universe as a whole. Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such as the Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the night sky. The disciplines included as diverse as astrometry, celestial navigation, observational astronomy and professional astronomy is nowadays often considered to be synonymous with astrophysics.^[2]

During the 20th century, the field of professional astronomy split into observational and theoretical astronomy. Theoretical astronomy is focused on acquiring data from observations of astronomical objects, which is then analyzed using mathematical models. Theoretical astronomy is oriented toward the development of computer or analytical models to study astronomical phenomena. The two fields complement each other, with theoretical astronomy seeking to explain observations being used to confirm theoretical results.

Astronomy is one of the few sciences where amateurs can still play an active role, especially in the discovery of transient phenomena. Amateur astronomers have made and contributed to many important astronomical discoveries, including new comets.

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona Lisa* is his most famous and most parodied portrait^[6] and *The Last Supper* the most reproduced religious painting of all time, their fame approached only by Michelangelo's *The Creation of Adam*. Leonardo's drawing of the *Vitruvian Man* is also regarded as a cultural icon,^[7] being

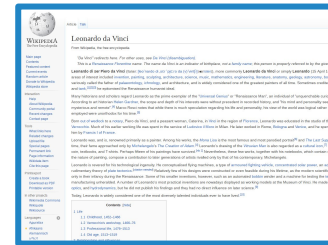
The Vitruvian Man

Entity Annotation

Terminology



- ▷ Wikipedia Knowledge Graph
- ▷ Node: Wikipedia Page (Entity)
- ▷ Link: Wikipedia Hyperlink



Enrich a text \mathcal{T} with proper annotations

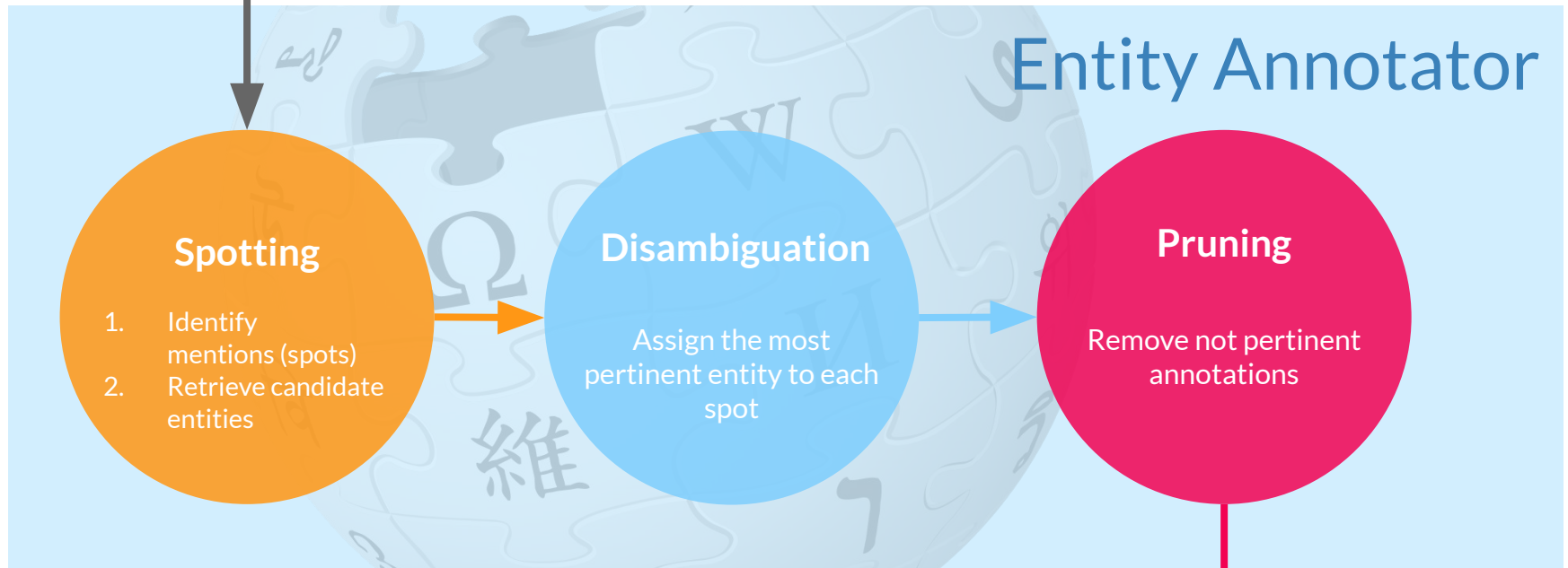
Annotation = (mention, entity)

Goal

Entity Annotation

The Annotation Pipeline

Input Text



Entity Annotator

Spotting

1. Identify mentions (spots)
2. Retrieve candidate entities

Disambiguation

Assign the most pertinent entity to each spot

Pruning

Remove not pertinent annotations

Annotated Text



Entity Annotation

The Annotation Pipeline



Spotting

Yesterday *Maradona* won against *Mexico*.

Mention Detection

Candidate Generation



Yesterday_(Time)



Yesterday_(Beatles_song)



Yesterday_(Guns_N_Roses_song)

...



Diego_Maradona



Diego_Sinagra



Maradona_by_Kusturica

...



Mexico



Mexico,_New_York



Mexico_national_football_team

...

Entity Annotation

The Annotation Pipeline



Spotting

1. Mention Detection

- Named Entity Recognition (aka NER)
- N-gram generation

2. Candidate Generation

- Gazetteer: { mention \rightarrow entities }
 - How?

Florence: Leonardo's artistic and social background



Lorenzo Ghiberti's *Gates of Paradise* (1425–52) were a source of communal pride. Many artists assisted in their creation

Florence at the time of Leonardo's youth was the centre of Christian **Humanist** thought and culture.^[22] Leonardo commenced his apprenticeship with Verrocchio in 1466, the year that Verrocchio's master, the great sculptor **Donatello**, died. The painter **Uccello**, whose early experiments with perspective were to influence the development of landscape painting, was a very old man. The painters **Piero della Francesca** and **Filippo Lippi**, sculptor **Luca della Robbia**, and architect and writer **Leon Battista Alberti** were in their sixties. The successful artists of the next generation were Leonardo's teacher Verrocchio, **Antonio del Pollaiuolo** and the portrait sculptor **Mino da Fiesole**, whose lifelike busts give the most reliable likenesses of Lorenzo Medici's father Piero and uncle Giovanni.^{[53][54][55][56]}

Leonardo's youth was spent in a Florence that was ornamented by the works of these artists and by Donatello's contemporaries, **Masaccio**, whose figurative **frescoes** were imbued with realism and emotion, and **Ghiberti**, whose **Gates of Paradise** gleaming with **gold leaf**, displayed the art of combining complex figure compositions with detailed architectural backgrounds. Piero della Francesca had made a detailed study of perspective,^[57] and was the first painter to make a scientific study of light. These studies and **Alberti's** treatise *De Pictura*^[58] were to have a profound effect on younger artists and in particular on Leonardo's own observations and artworks.^{[53][55][56]}

Massaccio's "Expulsion from the Garden of Eden" depicting the naked and distraught **Adam and Eve** created a powerfully expressive image of the human form, cast into three dimensions by the use of **light and shade**, which was to be developed in the works of Leonardo in a way that was to be influential in the course of painting. The humanist influence of Donatello's "David" can be seen in Leonardo's late paintings, particularly *John the Baptist*.^{[53][54]}

A prevalent tradition in Florence was the small altarpiece of the **Virgin and Child**. Many of these were created in **tempera** or glazed **terracotta** by the workshops of Filippo Lippi, Verrocchio and the prolific della Robbia family.^[53] Leonardo's early Madonnas such as *The Madonna with a carnation* and the



Article Talk

Fresco

From Wikipedia, the free encyclopedia

For other uses, see Fresco (disambiguation).

Fresco (plural **frescos** or **frescoes**) is a technique of mural painting executed on plaster, the painting becomes an integral part of the wall. The word fresco (Italian painting techniques, which are applied to dried plaster, to supplement painting

Contents [hide]

- 1 Technology
- 2 Other types of wall painting
- 3 History
 - 3.1 Egypt and Ancient Near East

Article Talk

Lorenzo Ghiberti

From Wikipedia, the free encyclopedia

Lorenzo Ghiberti (Italian: [loˈrentso ɡiˈbɛrti]; 1378 – 1 December 1455), born **Florence Baptistery**, called by **Michelangelo** the *Gates of Paradise*. Trained a important writing on art, as well as what may be the earliest surviving autobio

Contents [hide]

- 1 Life
 - 1.1 Early life
 - 1.2 Florence Baptistery doors

Article Talk

Florence Baptistery

From Wikipedia, the free encyclopedia

The **Florence Baptistery** (Italian: *Battistero di San Giovanni*), also known as **t baptistry** stands in both the **Piazza del Duomo** and the **Piazza San Giovanni**.

The Baptistery is one of the oldest buildings in the city, constructed between 1 the Pisan Romanesque or Lombard styles, its influence was decisive for the st **Brunelleschi**, and the other architects created Renaissance architecture. In the the late antique architectural tradition in Italy, as in the cases of the **Basilica of**

The Baptistry is renowned for its three sets of artistically important **bronze** doc Ghiberti.^[1] The east doors were dubbed by **Michelangelo** the *Gates of Paradise*

The Italian poet **Dante** and many other notable **Renaissance** figures, including Florentines were baptized here.

Entity Annotation

The Annotation Pipeline



Spotting

1. Mention Detection

- Named Entity Recognition (aka NER)
- N-gram generation

2. Candidate Generation

- Gazetteer: { mention \rightarrow entities }
 - How? Wikipedia anchor texts!
 - *Ranking (+ Thresholding)*
 - Commonness (Ferragina, CIKM '10; Guo, CIKM '14)
 - Entity-context Similarity (Zwicklbauer, SIGIR '16)
 - ...

Entity Annotation

The Annotation Pipeline



Disambiguation

Yesterday *Maradona* won against *Mexico*.



Yesterday_(Time)



Diego_Maradona



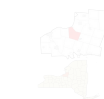
Mexico



Yesterday_
(Beatles_song)



Diego_Sinagra



Mexico_
New_York



Yesterday_
(Guns_N_Roses_
song)



Maradona_by_
Kusturica



Mexico_national_
football_team

...

...

...

Entity Annotation

The Annotation Pipeline



Disambiguation



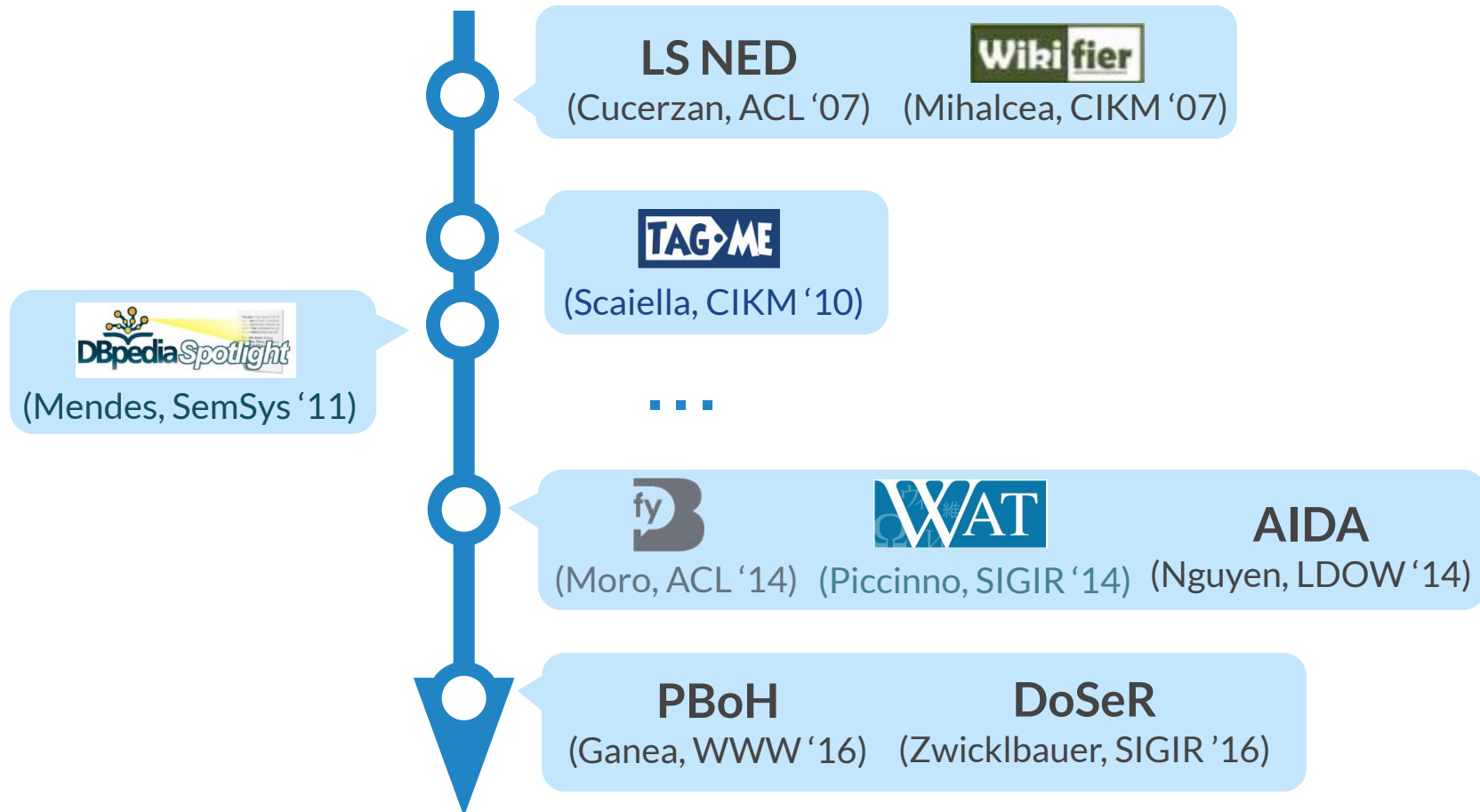
- ▷ Spots have been **disambiguated**
 - **Ambiguous lexical** elements (words) are now labeled with **unambiguous concepts**
- ▷ Finally, **coherence** scores are assigned

Entity Annotation

The Annotation Pipeline



Disambiguation



Entity Annotation

The Annotation Pipeline

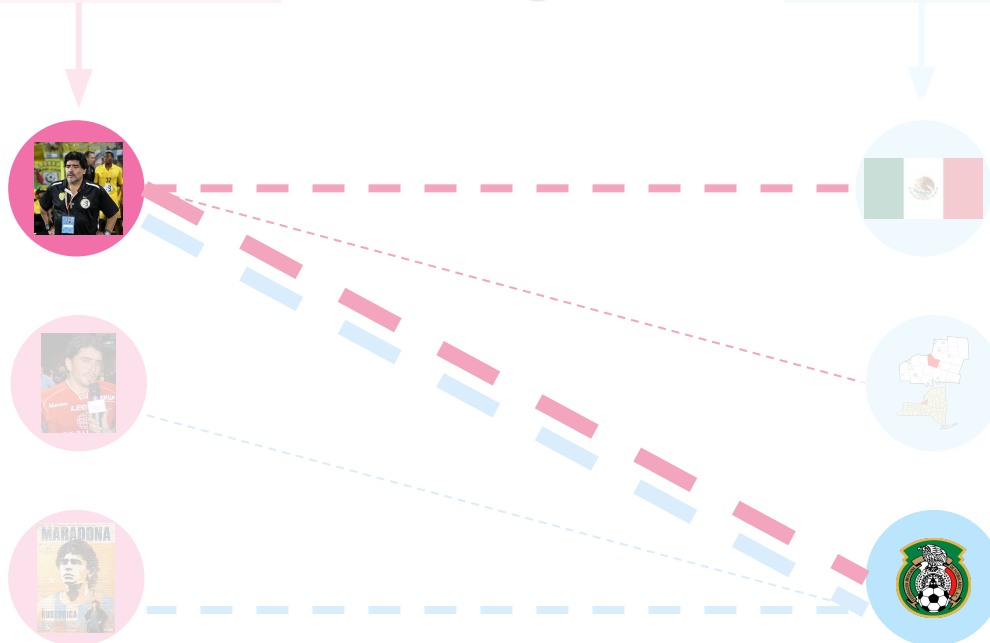


Disambiguation

Algorithm: **TAG-ME** **WAT** (Scaiella, CIKM '10; Piccinno, SIGIR '14)

[...] *Maradona* won against *Mexico*.

- Voting Scheme
- M&W / Jaccard Relatedness



Entity Annotation

The Annotation Pipeline

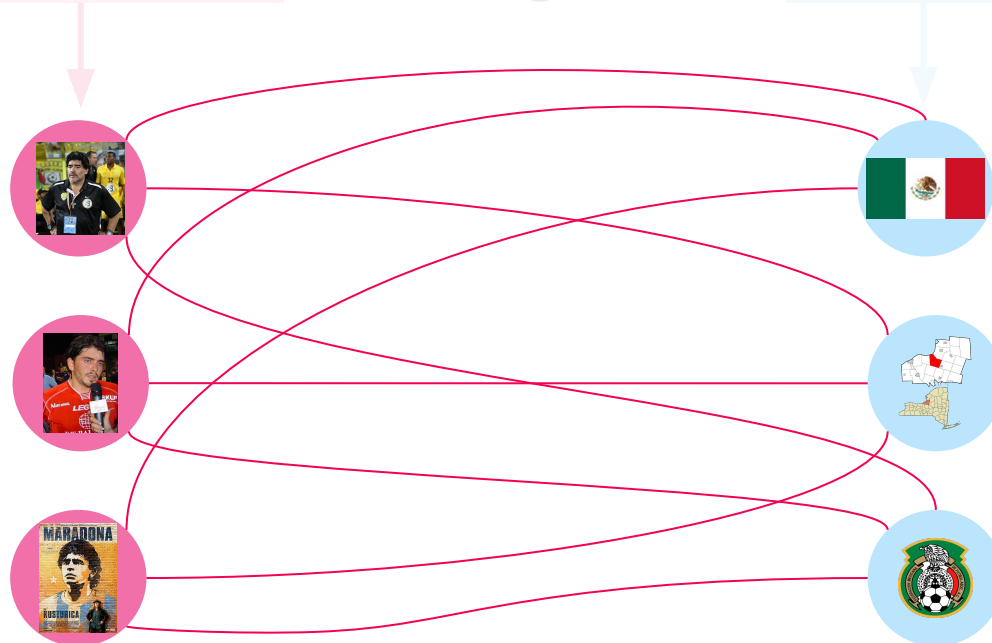


Disambiguation

Algorithm: **DoSeR** (Zwicklbauer, SIGIR '16)

[...] *Maradona* won against *Mexico*.

- Graph of candidates



Entity Annotation

The Annotation Pipeline

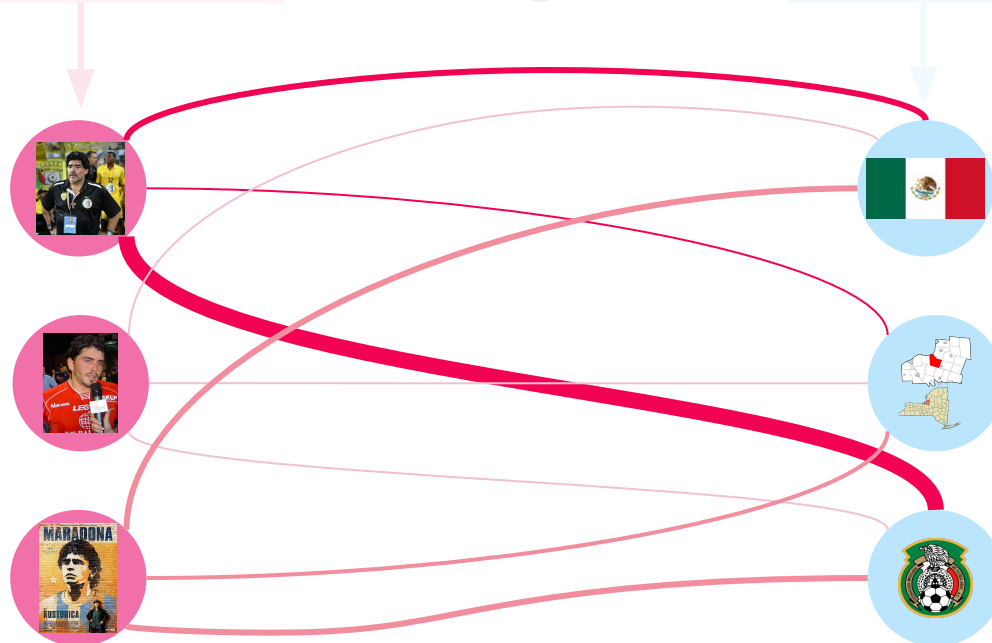


Disambiguation

Algorithm: **DoSeR** (Zwicklbauer, SIGIR '16)

[...] *Maradona* won against *Mexico*.

- Graph of candidates
- Entity2Vec Relatedness



Entity Annotation

The Annotation Pipeline

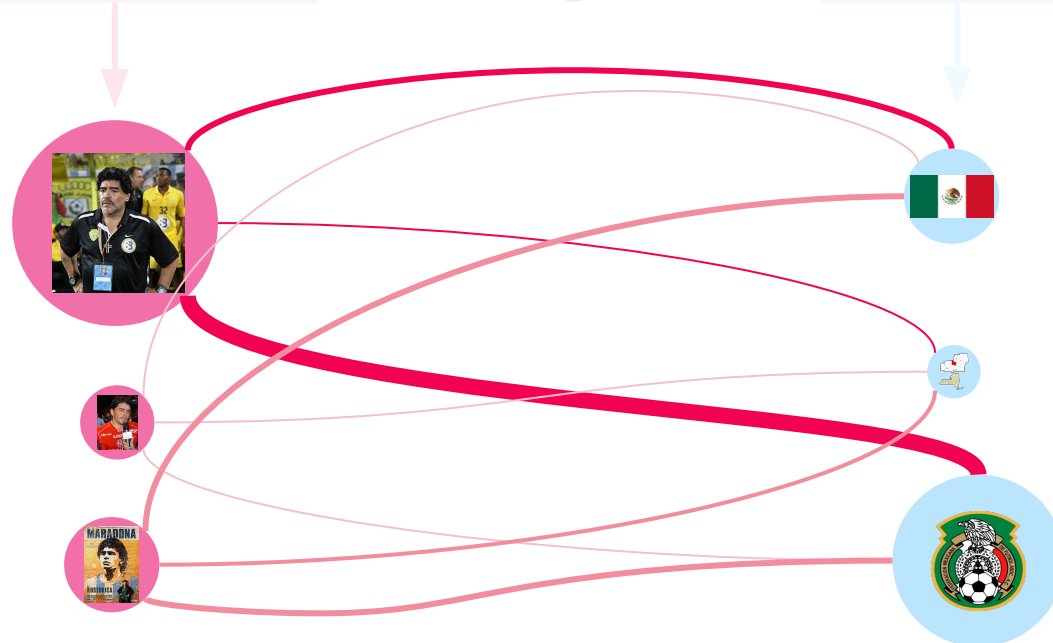


Disambiguation

Algorithm: **DoSeR** (Zwicklbauer, SIGIR '16)

[...] *Maradona* won against *Mexico*.

- Graph of candidates
- Entity2Vec Relatedness
- PageRank



Entity Annotation

The Annotation Pipeline



Pruning



- ▷ Remove **not pertinent** annotations
- ▷ Clear text from **erroneous** annotations
- ▷ Coherence thresholding

Applications

Web Search Results (Gabrilovich, SIGIR '16)

Google

All Images Maps News Videos More Search tools

About 87,600,000 results (0.74 seconds)

Pisa - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Pisa>
Pisa is a city in Tuscany, Central Italy, straddling the River Arno just before it empties into the Tyrrhenian Sea. It is the capital city of the Province of Pisa.
Pisa Charterhouse - Orto botanico di Pisa - Province of Pisa - Knights' Square

Virtual tour of Pisa Italy - History, facts, top attractions & things to do ...
www.italyguides.it > Italy > Tuscany
★★★★☆ Rating: 4 - 5,520 votes
Travel guide of Pisa Italy. Maps, articles, photos and destination guides about Pisa major attractions.

Pisa, Italy: Tourist Guide to Visiting the Leaning Tower of Pisa and ...
<https://www.discovertuscany.com/pisa/>
The Leaning Tower has made Pisa famous all over the world, and in addition to the tower, the city offers many other interesting things to see worth at least an ...

Pisa 2016: Best of Pisa, Italy Tourism - TripAdvisor
www.tripadvisor.com > Europe > Italy > Tuscany > Province of Pisa
Pisa Tourism: TripAdvisor has 164424 reviews of Pisa Hotels, Attractions, and Restaurants making it your best Pisa resource.


Comune di Pisa | Home
www.comune.pisa.it/ > Translate this page
Gran finale per la mostra Kan Pisa 2016. Centinaia di migliaia di visitatori per la mostra en plein air "Toccare il Tempo". La candidatura al Premio Italia 2016.

Aeroporto Galileo Galilei - Sito ufficiale - Aeroporto di Pisa - The ...
www.pisa-airport.com/ > Translate this page
Aeroporto Internazionale Galileo Galilei. Include le informazioni, gli orari dei voli, le infrastrutture.

PISA - OECD
<https://www.oecd.org/pisa/>
Webinar: PISA 2015: A Sneak Preview (October 25, 2016 9:30 am – 10:30 am EDT) Presented by the Alliance for Excellent Education and OECD Register here ...

Pisa - Lonely Planet
<https://www.lonelyplanet.com/italy/tuscany/pisa>
Once a maritime power to rival Genoa and Venice, Pisa now draws its fame from an architectural project gone terribly wrong. But the world-famous...

Official Tourism Website for Pisa Province - PisaUnicaTerra - Portale ...
www.pisaunicaterra.it/en/
According to its tradition as a former Maritime Republic, Pisa celebrates the New Year twice – not only on the 1st of January along with the rest of the world, but ...



Pisa

City in Italy

Pisa is a city in Italy's Tuscany region best known for its iconic Leaning Tower. Already tilting when it was completed in 1372, the 56m white-marble cylinder is the bell tower of the Romanesque, striped-marble cathedral that rises next to it in the Piazza dei Miracoli. Also in the piazza is the Baptistry, whose renowned acoustics are demonstrated by amateur singers daily, and the Caposanto Monumentale cemetery.

Plan a trip

- Pisa travel guide
- 3-star hotel averaging €121, 5-star averaging €178
- Upcoming Events

Weather: 12°C, Wind E at 18 km/h, 67% Humidity
Local time: Thursday 9:19 AM
Number of airports: 1

Points of interest

View 10+ more

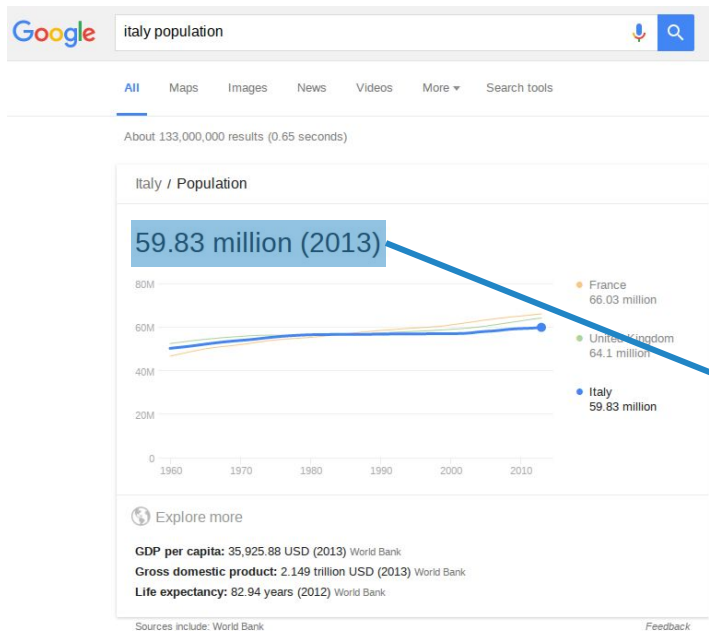
- Leaning Tower of Pisa
- Piazza dei Miracoli
- Camposanto Monumentale
- Santa Maria della Spina
- Knights' Square

More about Pisa

Feedback

Applications

Web Search Results (Gabrilovich, SIGIR '16)



Italy

Country in Europe

Italy, a European country with a long Mediterranean coastline, has left a powerful mark on Western culture and cuisine. Its capital, Rome, is home to the Vatican as well as landmark art and ancient ruins. Other major cities include Florence, with Renaissance masterpieces such as Michelangelo's "David" and Brunelleschi's Duomo; Venice, the city of canals; and Milan, Italy's fashion capital.

Capital: Rome
Dialing code: +39
Population: 59.83 million (2013) World Bank
President: Sergio Mattarella
Prime minister: Matteo Renzi

Destinations

View 15+ more

- Rome
- Venice
- Florence
- Sicily
- Milan

Points of interest

View 15+ more

- Colosseum
- Pantheon
- St. Peter's Basilica
- Roman Forum
- Amalfi Coast

Feedback

Applications

Question Answering (Gabrilovich, SIGIR '16)

Search: Barack Obama place of birth

ALL NEWS IMAGES VIDEOS MAPS SHOP

Barack Obama / Place of birth

Honolulu, HI

More about Honolulu

Barack Obama citizenship conspiracy theories - Wikipedia, the free encyclopedia
Wikipedia > wiki > Barack_Obama_citize...

Mobile-friendly - During Barack Obama's campaign for president in 2008 and in the Birth notices for Barack Obama were published in the Honolulu Advertiser on August ... in a biography in place until April 2007) which

I thought you asked...
Where is born Barack Obama

Barack Obama was born in Honolulu, Hawaii.

attribution >

Where is born Barack Obama

Today 10:38

Barack Obama was born in Honolulu, Hawaii.

attribution >

Good answer Bad answer

Barack Obama was born in Honolulu.

Wikipedia
Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of



Applications

Implicit Questions (Gabrilovich, SIGIR '16)

Search results for "type 2 diabetes". The page features a header with the search term and a microphone icon. Below is a teal bar with "Type 2 diabetes" and "Also called: adult onset diabetes". Navigation tabs for "ABOUT", "SYMPTOMS", and "TREATMENTS" are present. The main content includes an image of a man testing his blood sugar, a diagram comparing "Insulin resistance" (where glucose is blocked from cells) and "Normal (no resistance)" (where insulin moves glucose into cells), and a definition: "A chronic condition that affects the way the body processes blood sugar (glucose)." It also notes "Very common" with "More than 3 million US cases per year" and lists key facts: "Treatable by a medical professional", "Requires a medical diagnosis", and "Lab tests or imaging always required".

Search results for "chest pain radiating to back". The page features a header with the search term and a microphone icon. Below is a teal bar with "Health conditions related to this search". The main content is divided into two columns. The left column is titled "Chest pain" and describes causes like heavy lifting or trauma. The right column is titled "Heart attack" with a warning icon, advising to call 911 and describing a blood flow blockage. It lists risk factors: "Age: later in life" and "Male", and notes it is "Very common". At the bottom, it includes a disclaimer: "Individual cases may vary. Consult a doctor for medical advice." and sources: "Sources: Mayo Clinic and others. [Learn more](#)". A "Feedback" link is at the bottom right.

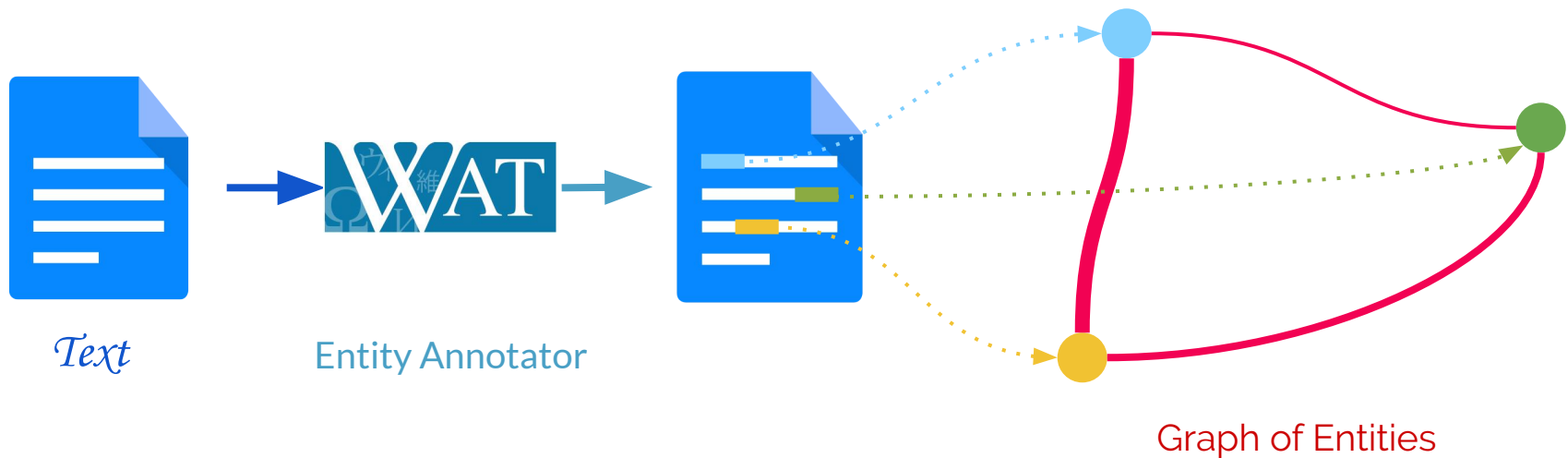
Search results for "knee pain when squatting down". The page features a header with the search term and a microphone icon. Below is a teal bar with "Health conditions related to this search". The main content is divided into two columns. The left column is titled "Knee pain" and describes causes like heavy physical activity or sitting on knees. The right column is titled "Patellofemoral pain syndrome" and describes it as a condition where cartilage is damaged. It lists risk factors: "Female" and "Running", and notes it is "Very common". At the bottom, it includes a disclaimer: "Individual cases may vary. Consult a doctor for medical advice." and sources: "Sources: Mayo Clinic and others. [Learn more](#)". A "Feedback" link is at the bottom right.

Condition → What does it mean?

Symptoms → What do they indicate?

A New Text Representation

- ▷ Originally introduced by (Scaiella, WSDM '12)
 - Widely deployed (Dunietz, EACL '14; Schuhmacher, WSDM '14; Ni, WSDM '15), ...
- ▷ *Text* = Graph of Entities
- ▷ What about...



A New Text Representation

- ▷ Originally introduced by (Scaiella, WSDM '12)
 - Widely deployed (Dunietz, EACL '14; Schuhmacher, WSDM '14; Ni, WSDM '15), ...
- ▷ *Text* = Graph of Entities
- ▷ What about...
 - ...edge weights?
 - ...node weights? } Work done in the first year

2.

Work done in the first year

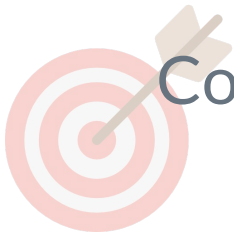
Entity Relatedness & Document Aboutness



Entity Relatedness



Entity Relatedness



Compute how much two **entities** are **related**

Relatedness : **Entities** × **Entities** → **Real**

Goal

- ▷ How much related are...
 - ...Bank with Money?
 - ...Wood with Book?

- ▷ Semantic Reasoning:
 - **Human**: Background Knowledge
 - **Machines**: Knowledge Graph

Entity Relatedness

(A brief list of) Algorithms and Applications

▷ Document/Word Similarity

- WikiRelate (Strube, AAI '06)
- Explicit Semantic Analysis (Gabrilovich, IJCAI '07)
 - WikiWalk (Yeh, ACL '09)
 - Temporal Semantic Analysis (Radinsky, WWW '11)
 - Concept Graph Representation (Ni, WSDM '16)
- Milne & Witten (Witten, AAI '08)
- Salient Semantic Analysis (Hassan, AAI '11)

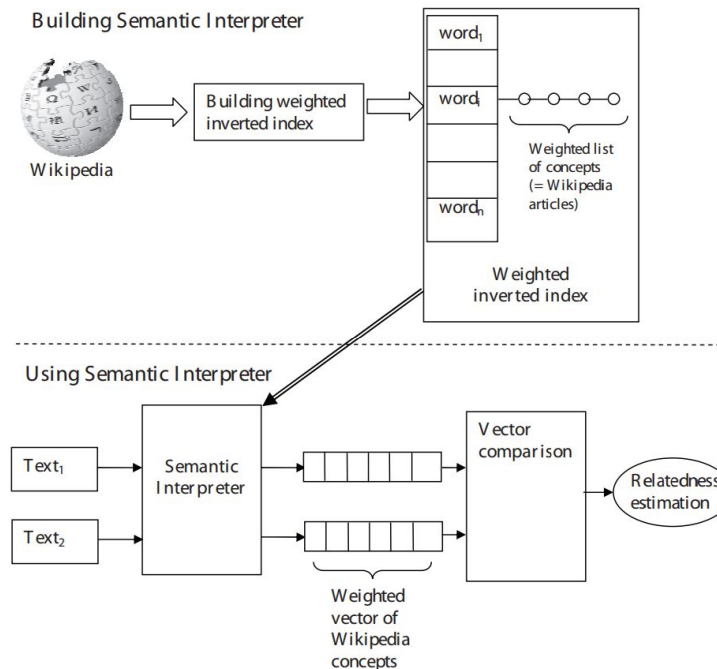
▷ Machine Translation (Agirre, NAACL '09; Rothe, ACL '14)

▷ Document Classification (Perozzi, WWW '14; Tang, WWW '15)

▷ ...

Entity Relatedness

- ▷ Two entities are related whether...
 - ...they are **described** by related texts (**Corpus-based**)
 - Example: ESA (Gabrilovich, IJCAI '07)
 - Concepts grounded in **human** cognition
 - Opposite to *latent* concepts

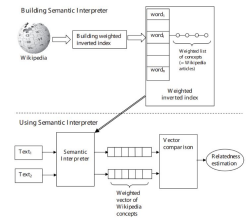


Entity Relatedness

- ▷ Two entities are related whether...
 - ...they are **described** by related texts (**Corpus-based**)

- Example: ESA (Gabrilovich, IJCAI '07)

- Concepts grounded in **human** cognition
 - Opposite to *latent* concepts



- ...they are **referenced** by related entities (**Graph-based**)
 - Example: CoSimRank (Rothe, ACL '14)

Entity Relatedness

CoSimRank (Rothe, ACL '14)

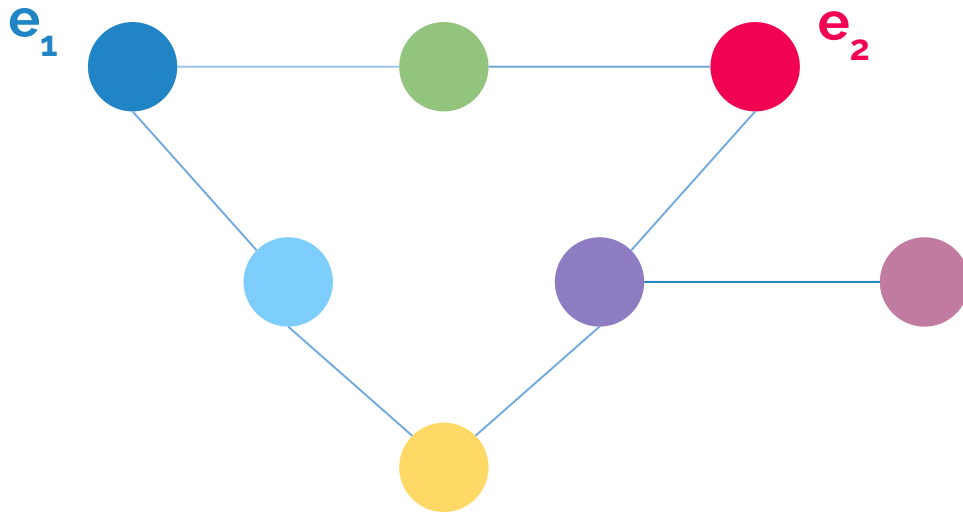
- ▷ **Graph**-based approach
- ▷ Relatedness algorithm for **nodes** in a graph
- ▷ Exploits Random Walks
- ▷ Algorithm (in brief)

1. Sets damping vectors for e_1 and e_2
2. Runs an iteration of PageRank
3. Updates relatedness score

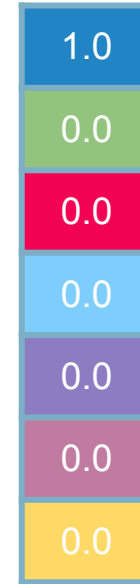
$e_1, e_2 \in \text{Entities}$

Entity Relatedness

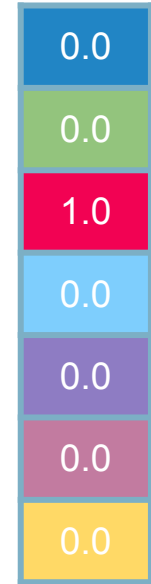
CoSimRank (Rothe, ACL '14)



$p^0(e_1)$



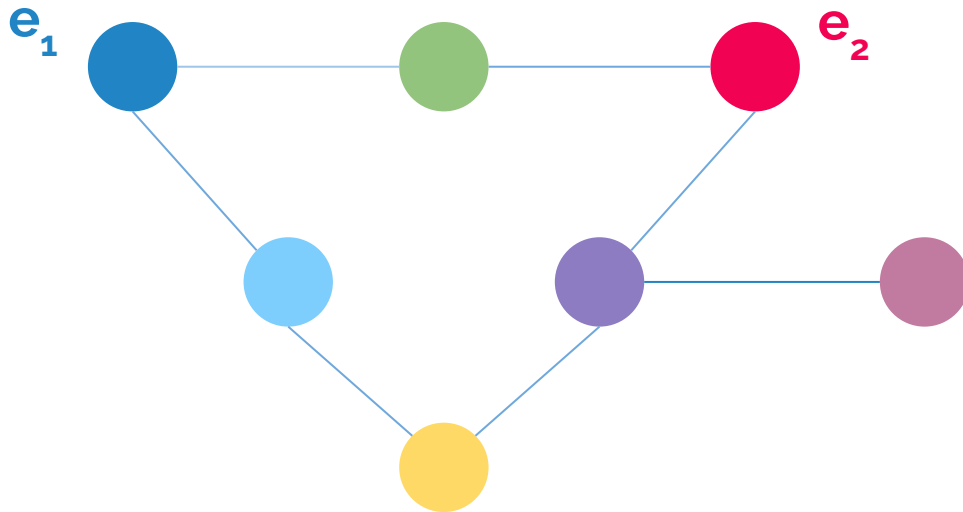
$p^0(e_2)$



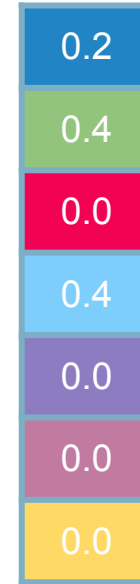
$$\text{Relatedness}^0(e_1, e_2) = 0.0$$

Entity Relatedness

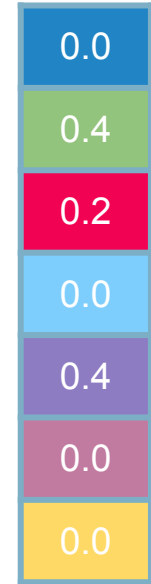
CoSimRank (Rothe, ACL '14)



$p^1(e_1)$



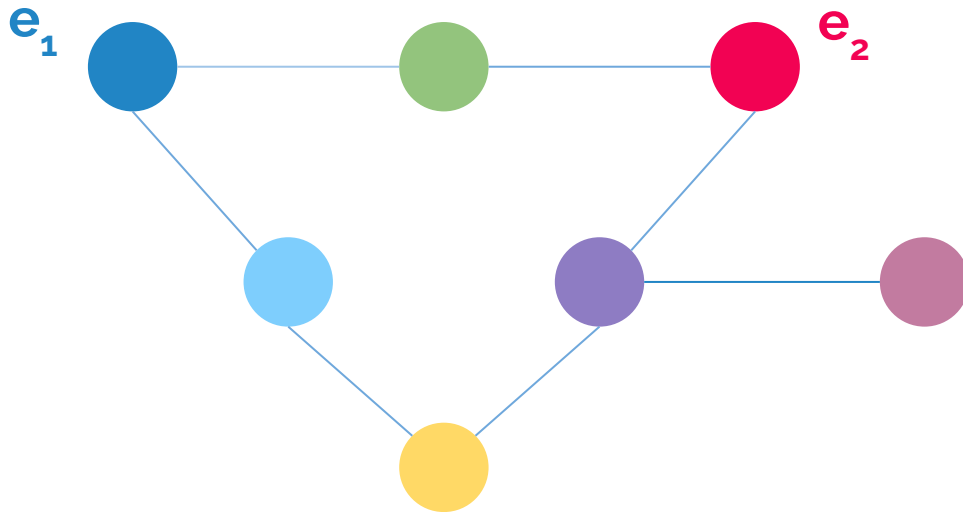
$p^1(e_2)$



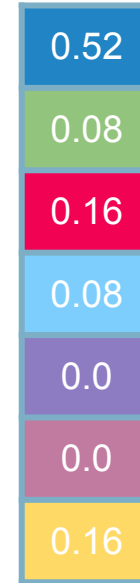
$$\text{Relatedness}^1(e_1, e_2) = 0.16$$

Entity Relatedness

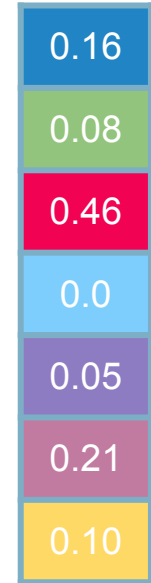
CoSimRank (Rothe, ACL '14)



$p^2(\mathbf{e}_1)$



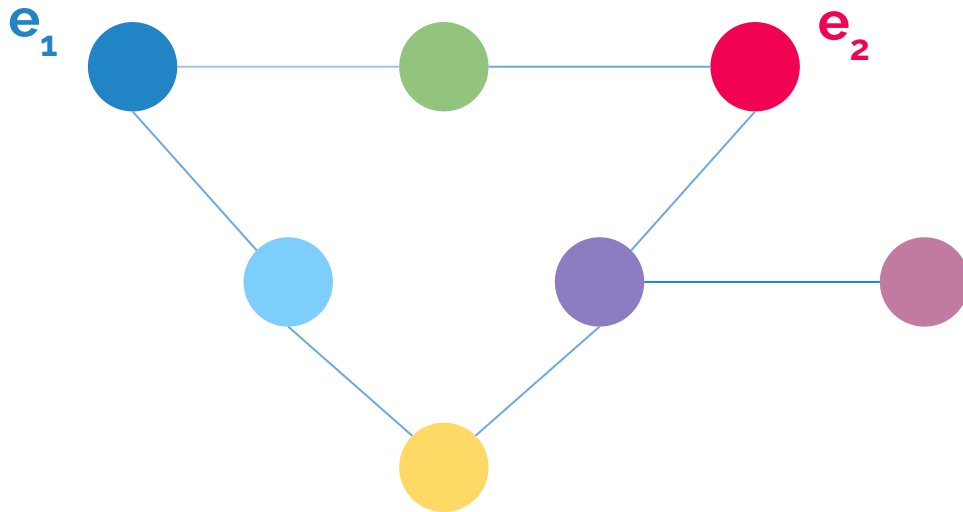
$p^2(\mathbf{e}_2)$



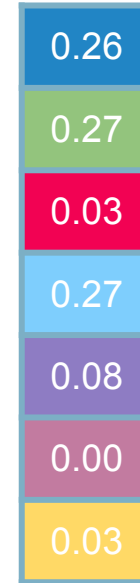
$$\text{Relatedness}^2(\mathbf{e}_1, \mathbf{e}_2) = 0.33$$

Entity Relatedness

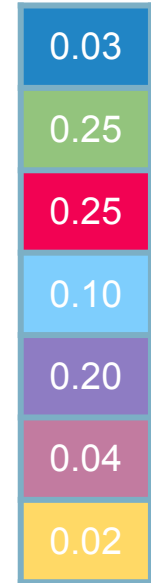
CoSimRank (Rothe, ACL '14)



$p^3(\mathbf{e}_1)$



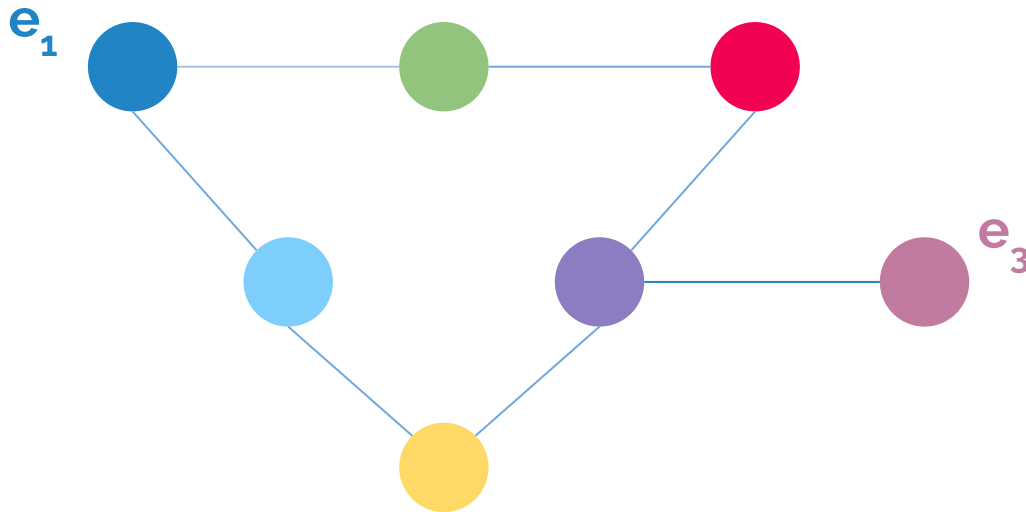
$p^3(\mathbf{e}_2)$



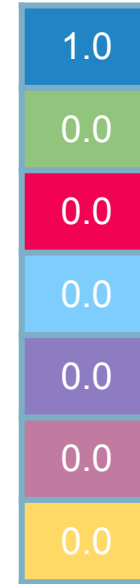
$$\text{Relatedness}^3(\mathbf{e}_1, \mathbf{e}_2) = 0.47$$

Entity Relatedness

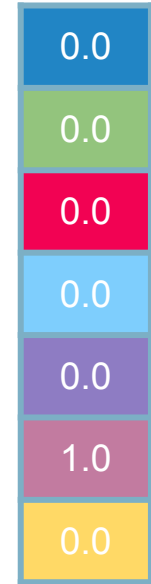
CoSimRank (Rothe, ACL '14)



$p^0(\mathbf{e}_1)$



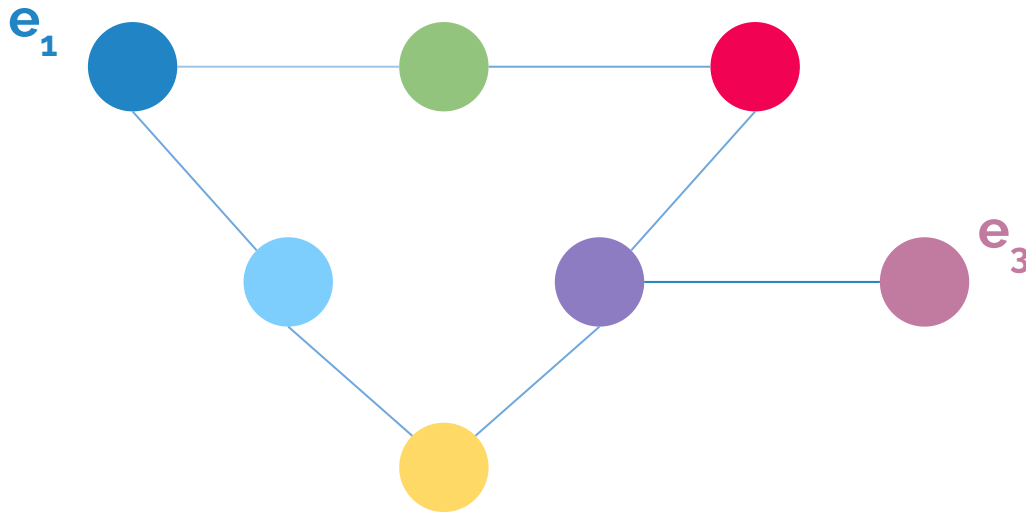
$p^0(\mathbf{e}_3)$



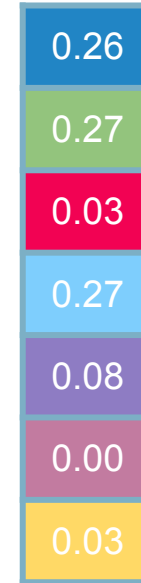
$$\text{Relatedness}^0(\mathbf{e}_1, \mathbf{e}_3) = 0.0$$

Entity Relatedness

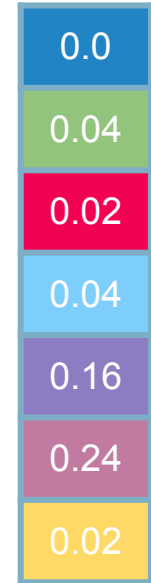
CoSimRank (Rothe, ACL '14)



$p^3(e_1)$



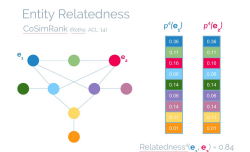
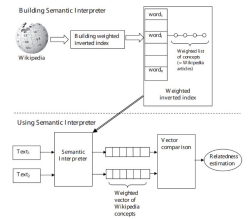
$p^3(e_2)$



$$\text{Relatedness}^3(e_1, e_3) = 0.13$$

Entity Relatedness

- ▷ Two entities are related whether...
 - ...they are **described** by related texts (**Corpus-based**)
 - Example: **ESA** (Gabrilovich, IJCAI '07)
 - Concepts grounded in **human** cognition
 - Opposite to *latent* concepts
 - ...they are **referenced** by related entities (**Graph-based**)
 - Example: **CoSimRank** (Rothe, ACL '14)
- ▷ Need of a **fair** and **meaningful** comparison

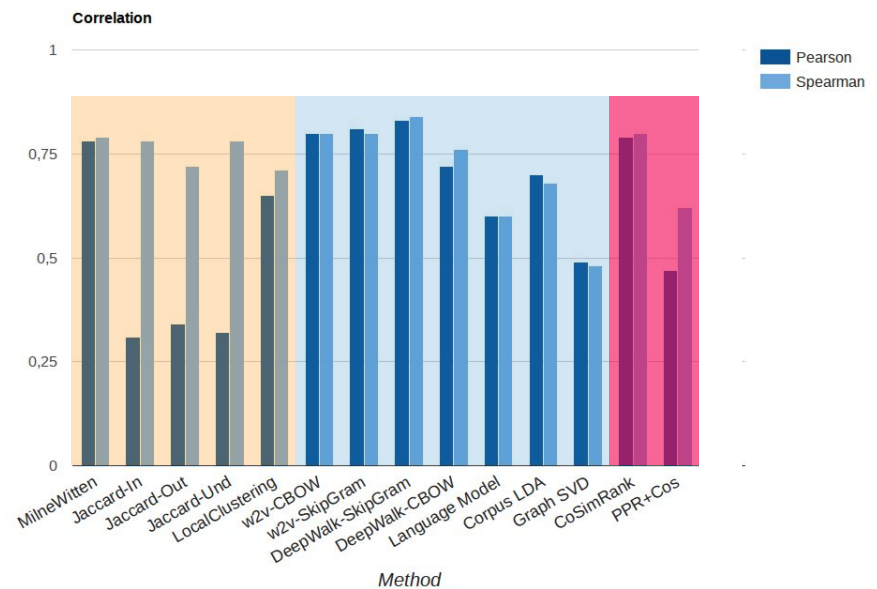


Entity Relatedness

Preliminary Results: The Relatedness Framework

- ▷ Design algorithms based on
 - Set Operations (Milne & Witten, Jaccard, ...)
 - Embeddings (Word2Vec, LDA, ...)
 - Random Walk (CoSimRank, PPR+Cos)

- ▷ Preliminary results
 - Analyse entity pairs
 - Deploy corpus-based algorithms
 - A new algorithm: LLP



Entity Relatedness

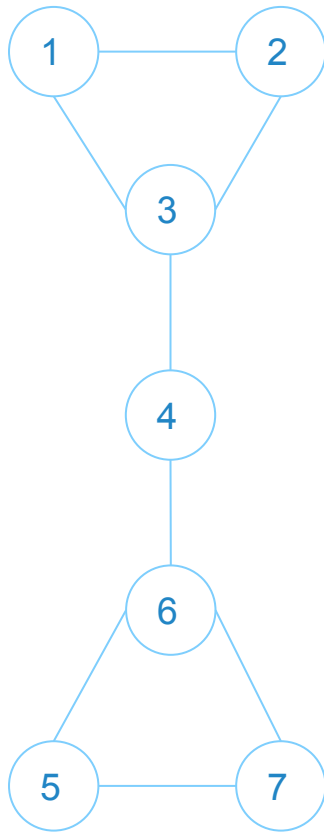
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

- ▷ **Standard Label Propagation** (Newman, Phys. Rev. '04)
 - Clustering algorithm (node labeling)
 - **Pro: Scales** on very large graphs
 - **Cons:** Can generate **few big clusters**

- ▷ **Layered** Label Propagation
- ▷ Standard Label Propagation with a resolution parameter γ
 - Graph compression
 - Algorithm (in brief)
 1. *Randomly initialize each node with a **label (cluster)***
 2. *Update label according to a specific **rule***
 - **Maximize nonlocal discount** (Ronhovde, Phys. Rev. '10)

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

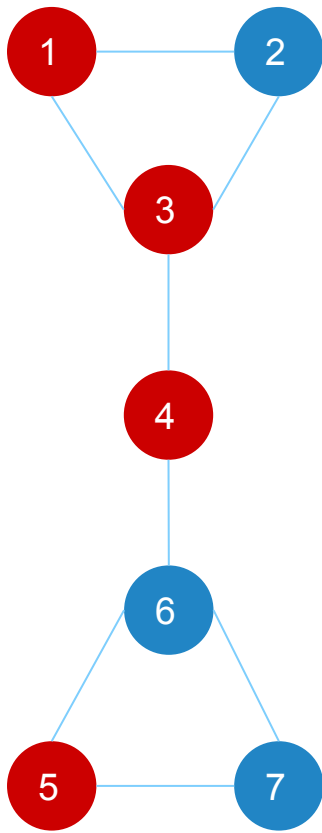


Round: 1
Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

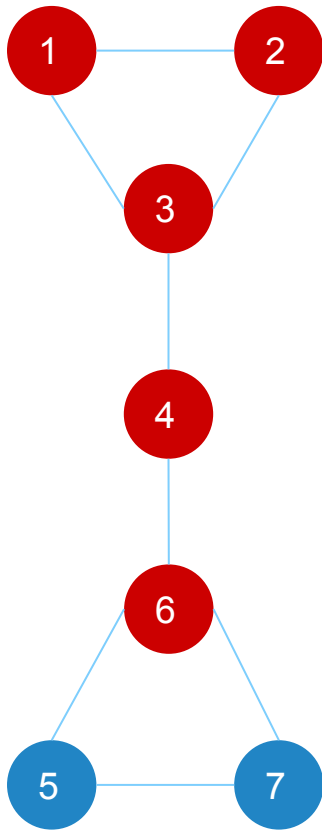


Round: 1
Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

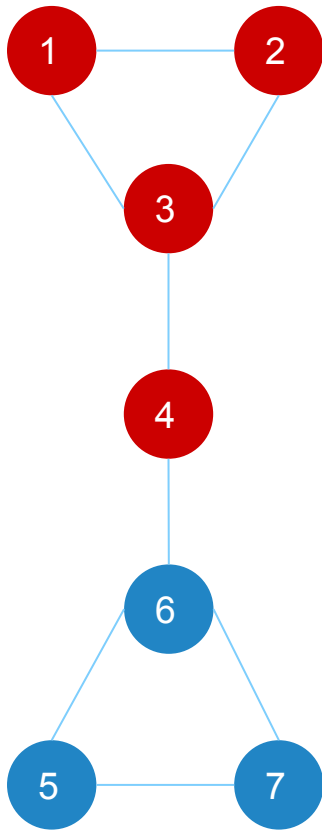


Round: 1
Step: Updating (1)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

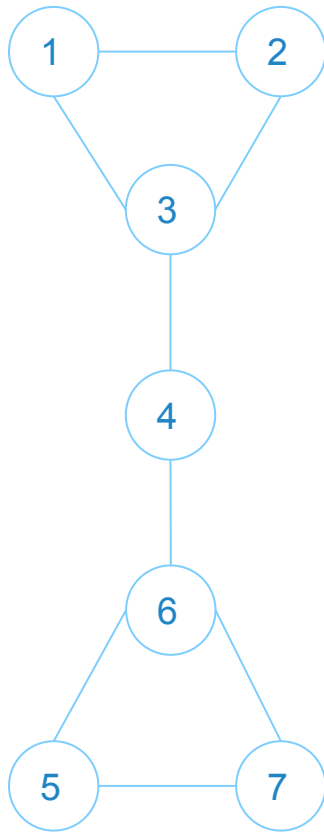


Round: 1
Step: Updating (2)

Node	Labels
1	Red
2	Red
3	Red
4	Red
5	Blue
6	Blue
7	Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

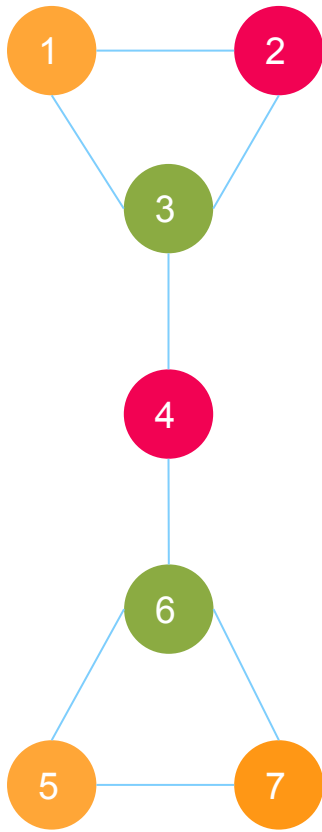


Round: 2
Step: Initialization

Node	Labels
1	Red
2	Red
3	Red
4	Red
5	Blue
6	Blue
7	Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

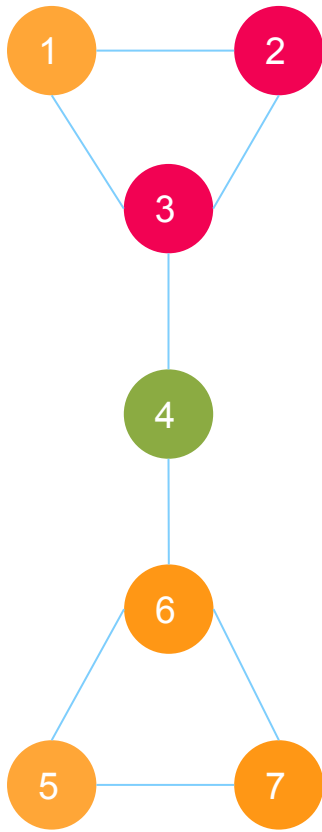


Round: 2
Step: Initialization

Node	Labels
1	Red
2	Red
3	Red
4	Red
5	Blue
6	Blue
7	Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

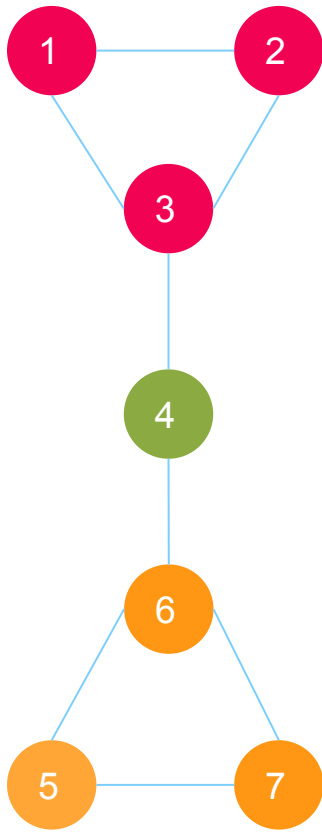


Round: 2
Step: Updating (1)

Node	Labels
1	Red
2	Red
3	Red
4	Red
5	Blue
6	Blue
7	Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

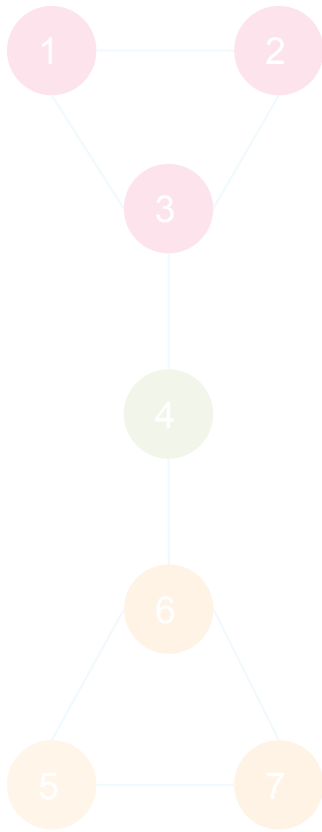


Round: 2
Step: Updating (2)

Node	Labels
1	Red, Pink
2	Red, Pink
3	Red, Pink
4	Red, Green
5	Blue, Orange
6	Blue, Orange
7	Blue, Orange

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

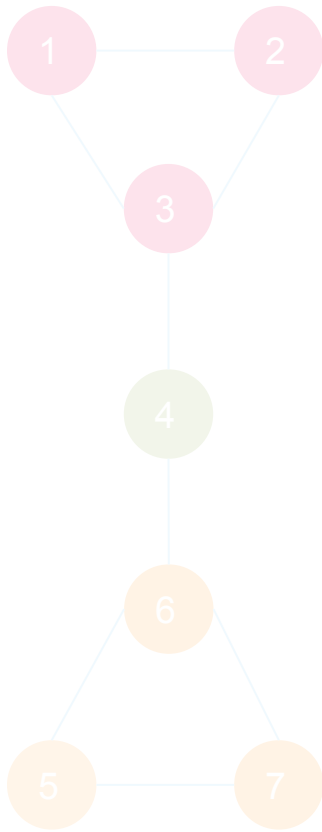


Round: 2
Step: Updating (2)

Node	Labels		
1	Red	Pink	Purple
2	Red	Pink	Brown
3	Red	Pink	Dark Brown
4	Red	Green	Light Blue
5	Blue	Orange	Light Blue
6	Blue	Orange	Light Blue
7	Blue	Orange	Light Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



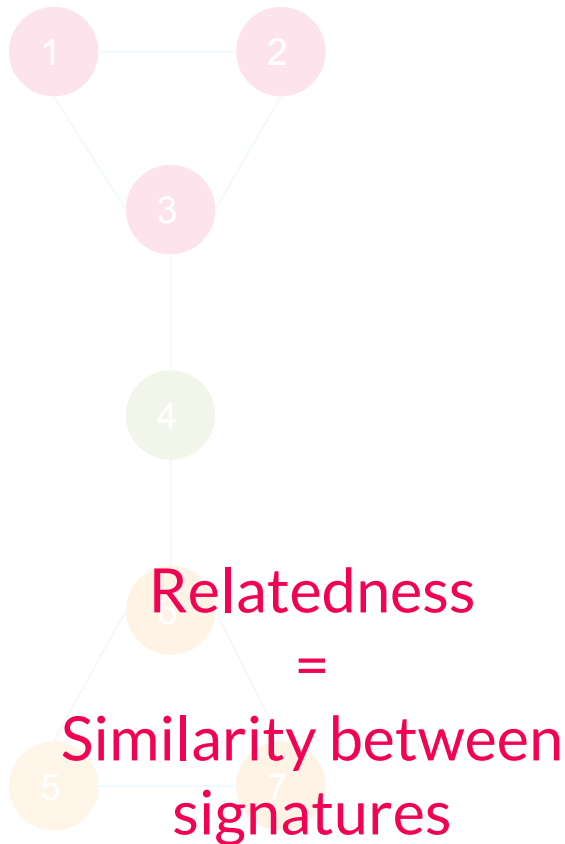
Round: 2

Step: Updating (2)

Node	Labels			
1	Red	Magenta	Purple	Teal
2	Red	Magenta	Brown	Teal
3	Red	Magenta	Dark Brown	Teal
4	Red	Green	Light Blue	Grey
5	Blue	Orange	Light Blue	Yellow
6	Blue	Orange	Light Blue	Purple
7	Blue	Orange	Light Blue	Blue

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



Round: 2
Step: Updating (2)

Node	Labels			
1	Red	Pink	Purple	Teal
2	Red	Pink	Brown	Teal
3	Red	Pink	Dark Brown	Teal
4	Red	Green	Light Blue	Grey
5	Blue	Orange	Light Blue	Yellow
6	Blue	Orange	Light Blue	Purple
7	Blue	Orange	Light Blue	Blue

Signatures



Document Aboutness



Document Aboutness



Aboutness = Succinct representation of a
document's subject matter (Hutchins, 1977)

Goal

- ▷ **Weight** information (e.g. entities, words, ...) within a document

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from **Senators Barack Obama** and **Hillary Rodham Clinton** anytime soon, but stay tuned.

At least that is the word from their associates. **Mr. Obama**, **Democrat of Illinois**, is not likely to say whether he intends to seek the party's presidential nomination until after **President Bush's State of the Union** address on Jan. 23. As he walked out of the **Capitol** on a recent afternoon, **Mr. Obama** only smiled when asked about his timing. Then, he rushed to change the subject.








Initially, **Mr. Obama** said he intended to announce his decision after returning from a holiday vacation in **Hawaii**, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

Still, **Mr. Obama** has been busy telephoning crucial **Democrats** in **Iowa**, **New Hampshire** and other states. There is, of course, only one reason for him to be making such inquiries.

Last week on **Capitol Hill**, **Mr. Obama** bumped into **Ethel Kennedy**, who has been a big admirer. When asked about him, she said, "He can't run soon enough."

Mrs. Clinton, meanwhile, plans to announce her decision in the next several weeks, her advisers say. According to several **Democrats** who have spoken to her, as well as advisers, **Mrs. Clinton** has given every indication that she is running, short of saying so, and no signals that she is not.

She is making phone calls to **Democratic** officials, labor leaders and supporters in early nominating states. And she continues to talk to possible consultants and donors, yet she has not made any travel plans to kick off a campaign. **JEFF ZELENY** and **PATRICK HEALY**

-  FACEBOOK
-  TWITTER
-  GOOGLE+
-  EMAIL
-  SHARE
-  PRINT
-  REPRINTS

Aboutness

Entity	Weight
Barack_Obama	0.85
Hillary_Clinton	0.8
Hawaii	0.3
George_W._Bush	0.2
...	...

Document Aboutness



Aboutness = Succinct representation of a
document's subject matter (Hutchins, 1977)


Goal

- ▷ **Weight** information (e.g. entities, words, ...) within a document
- ▷ Wide range of practical applications:
 - a. Recommendation
 - b. Categorization
 - c. Exploratory search
 - d. Web Ranking
 - e. ...

Document Aboutness

Keyphrase Extraction

Entity Saliency

Aboutness	Words Proper nouns Sentences ...	Entities 
Candidate Extraction	Dictionary POS tags ...	Entity Annotation
Subject Matter Identification	Ranking/Classification	
Issues	Interpretation Overgeneration Infrequency Redundancy	Dependency on KG ?



Keyphrase Extraction



“

*“[...] errors could be addressed using background **knowledge**.”*

(Hasen, ACL ‘14)

*“[...] features more directly linked to **Wikipedia** [...] can provide more focused background information.”*

(Dunietz, EACL ‘14)



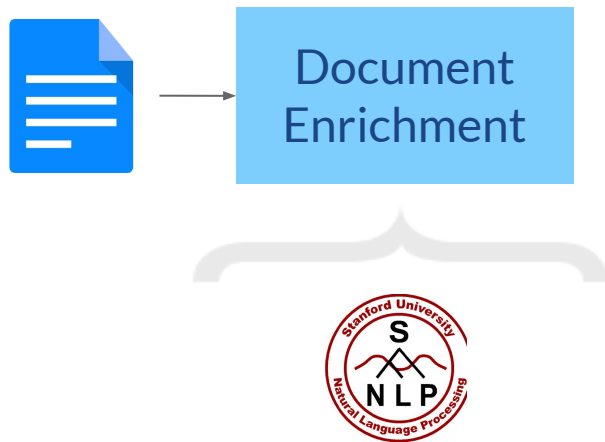
Entity Salience



Document Aboutness

Our Proposal

- ▷ **Entity Salience** Approach



- ▷ Pos tagging
- ▷ Mention detection
- ▷ Dependency parsing
- ▷ Co-reference resolver.

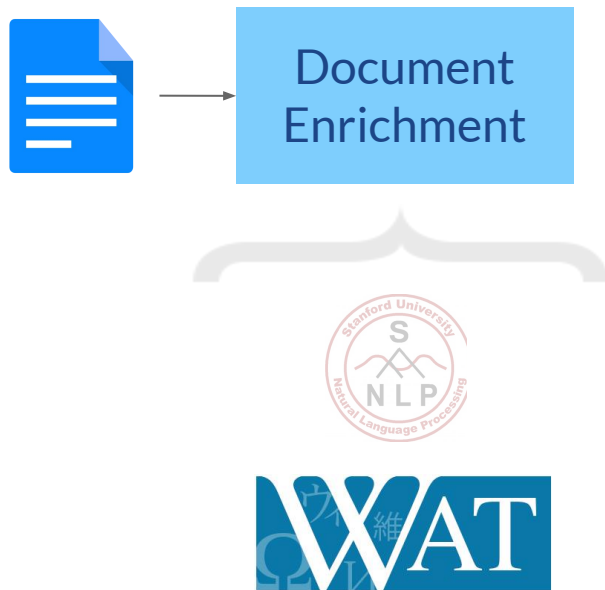


Obama loves hamburgers

Document Aboutness

Our Proposal

- ▷ **Entity Salience** Approach



- ▷ Entity annotation
- ▷ Graph of entities
- ▷ Relatedness

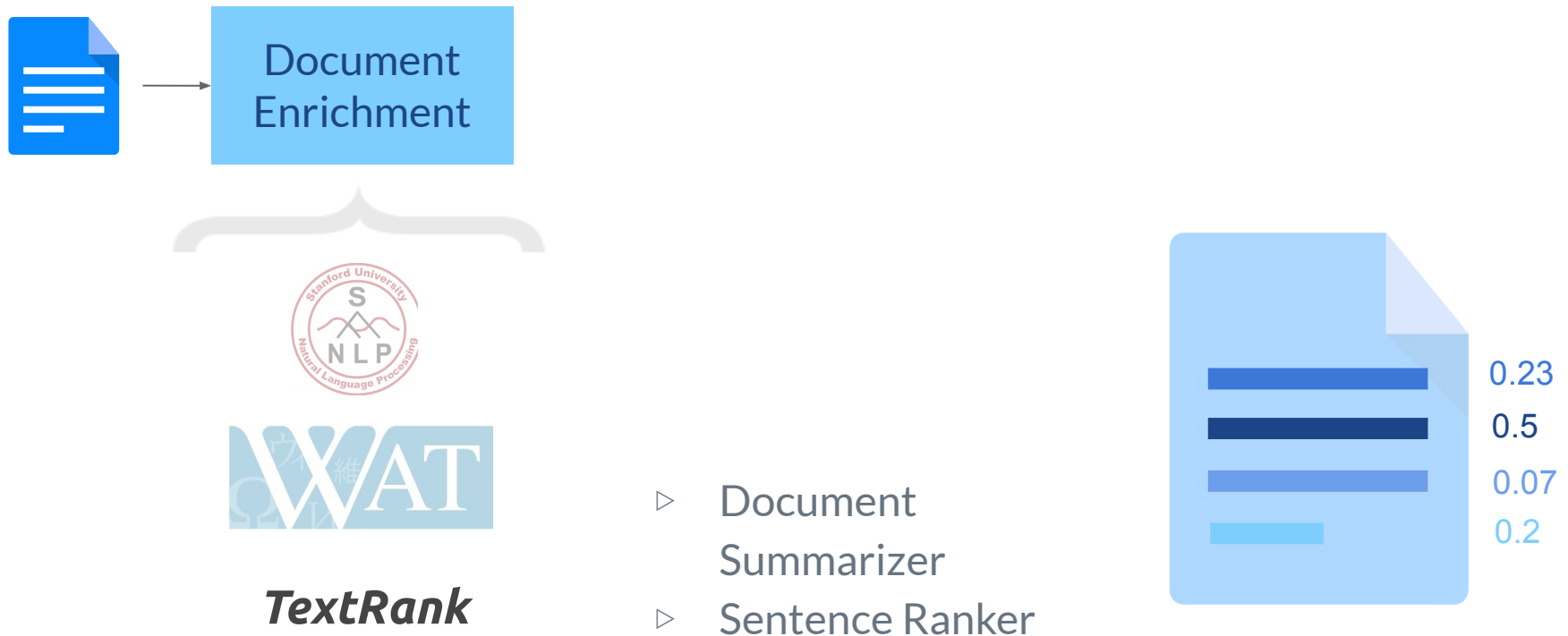
Obama loves hamburgers



Document Aboutness

Our Proposal

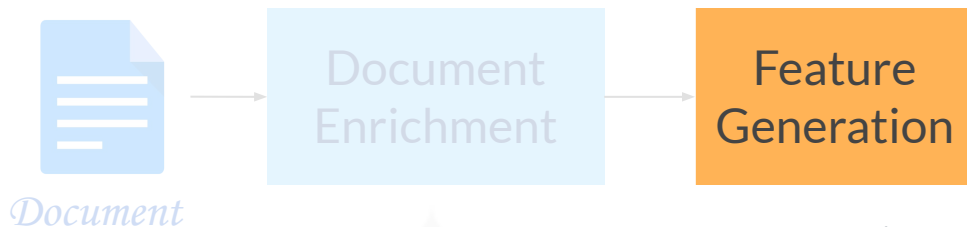
- ▷ **Entity Salience** Approach



Document Aboutness

Our Proposal

- ▷ **Entity Salience** Approach



- ▷ Entity → Feature Vector
- ▷ Classical **syntactic** features
 - Frequency
 - Position
 - ...
- ▷ New **syntactic** and **semantic** features:
 - Position-based
 - Dependency-based
 - Relatedness-based
 - Centrality-based
 - ...

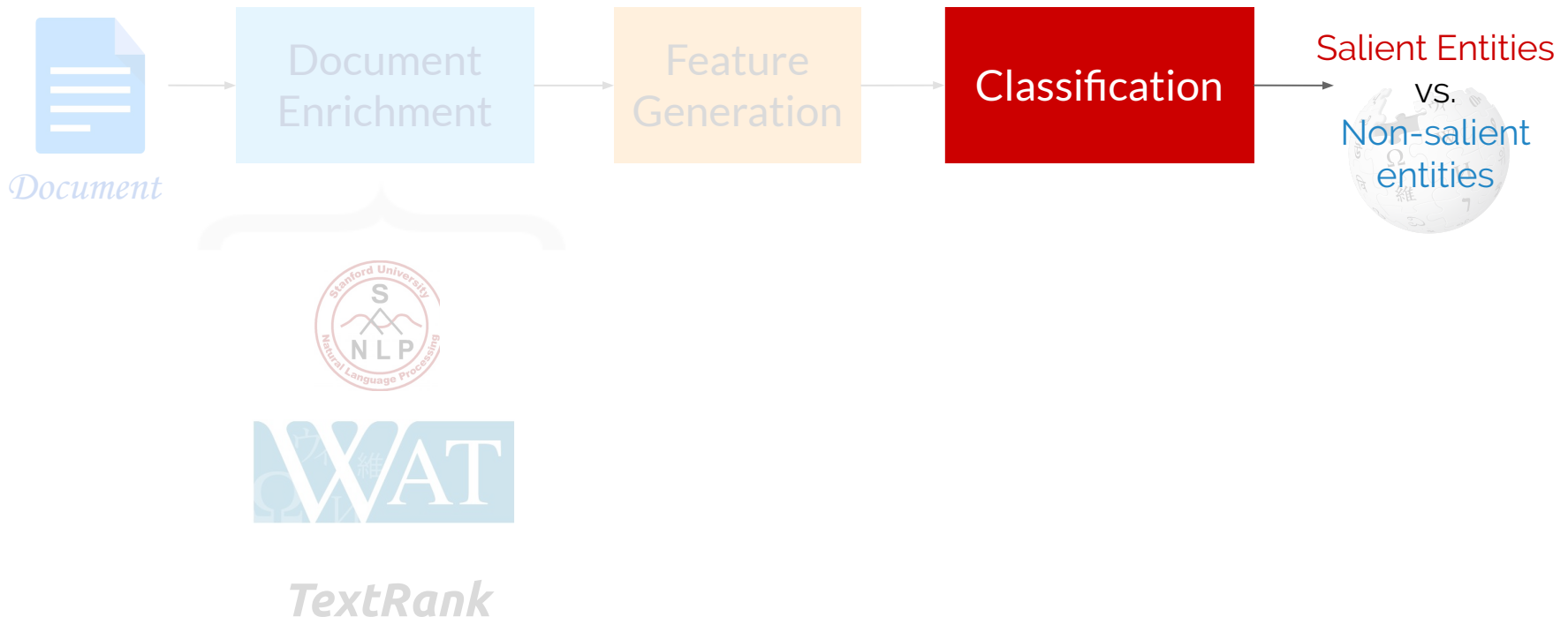


TextRank

Document Aboutness

Our Proposal

- ▷ **Entity Salience** Approach



Document Aboutness

Our Proposal: Main Contributions

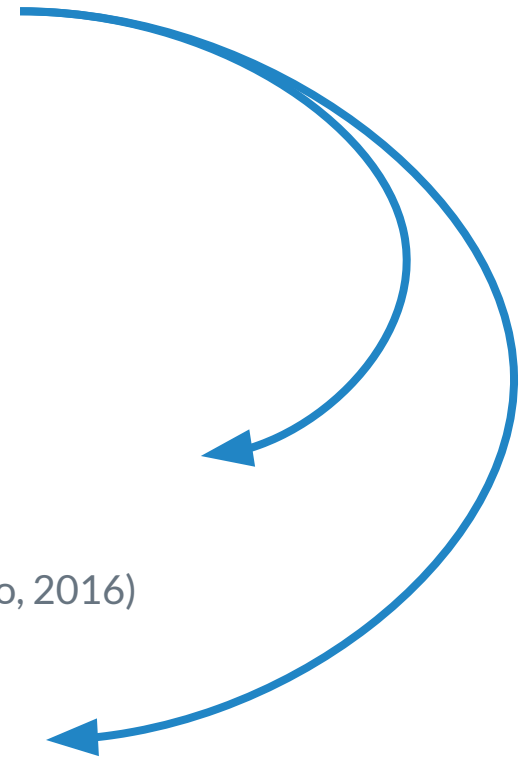
- ▷ Fully documented system
- ▷ Public available via Web-API
- ▷ **Improvement** of state-of-the-art (**Cmu-Google, F1: 61.5**)
 - New York Times' dataset (110,000 news, 1.3M entities)
 - *62.6 micro-F1 (+1.1%) and 59.5 macro-F1 (+2.5%)*
 - More robust when entities are not biased at the beginning (+9%)
- ▷ Deep Feature and Error Analysis

3.

Future Work

Future Work

- ▷ Conclude **Entity Relatedness**
 - Finalize experiments
 - Related vs Non-related
 - Scalability
- ▷ Improve our **Entity Saliency System**
 - Deep Learning (i.e. w2v)
 - Abstractive Summarization
 - Create and test new datasets
 - Plug the new TagMe-Wat 2.0 (Piccinno, 2016)
- ▷ **Entity Annotation Improvement**



Thanks!

Any questions?