Similarity in Wikipedia Articles

Badenes, Carlos (*cbadenes*)
Garijo, Daniel (*dgarijo*)
Priyatna, Freddy (*fpriyatna*)
{*}@fi.upm.es

EDBT Summer School 2015

Problem

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!

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

Languages
Acèh
Адыгэбзэ
Afrikaans
Alemannisch

Madrid

From Wikipedia, the free encyclopedia

Coordinates: 40°23′N 3°43′W

This article is about the capital city of Spain. For the autonomous community, see Community of Madrid. For other uses, see Madrid (disambiguation).

Madrid (/məˈdrɪd/, Spanish: [maˈðrið], locally: [maˈðriθ, -ˈðri]) is a south-western European city and the capital and largest municipality of Spain. The population of the city is almost 3.2 million[4] and that of the Madrid metropolitan area, around 6 million. It is the third-largest city in the European Union, after London and Berlin, and its metropolitan area is the third-largest in the European Union after Paris and London.[5][6][7][8] The city spans a total of 604.3 km² (233.3 sq mi).[9]

The city is located on the Manzanares River in the centre of both the country and the Community of Madrid (which comprises the city of Madrid, its conurbation and extended suburbs and villages); this community is bordered by the autonomous communities of Castile and León and Castile-La Mancha. As the capital city of Spain, seat of government, and residence of the Spanish monarch, Madrid is also the political, economic and cultural centre of Spain.[10] The current mayor is Manuela Carmena from Ahora Madrid.

The Madrid urban agglomeration has the third-largest GDP[11] in the European Union and its influences in politics, education, entertainment, environment, media, fashion, science, culture, and the arts all contribute to its status as one of the world's major global cities.[12][13] Due to its economic output, high standard of living, and market size, Madrid is considered the major financial centre of Southern Europe[14][15] and the Iberian Peninsula; it hosts the head offices of the vast majority of the major Spanish companies, such as Telefónica, Iberia and Repsol. Madrid is the 17th most livable city in the world according to Monocle magazine, in its 2014 index.[16][17]

Madrid houses the headquarters of the World Tourism Organization (WTO), belonging to the United Nations Organization (UN), the SEGIB, the Organization of Ibero-American States (OEI), and the Public Interest Oversight Board (PIOB). It also hosts major international regulators of Spanish: the Standing Committee of the Association of Spanish Language Academies, headquarters of the Royal Spanish Academy (RAE), the Cervantes Institute and the Foundation of Urgent Spanish (Fundéu BBVA). Madrid organizes fairs such as FITUR,[18] ARCO,[19] SIMO TCI[20] and the Cibeles Madrid Fashion Week.[21]

Madrid

From upper left: view of business districts of AZCA and CTBA, Gran Vía street and Metropolis Building, the Palace of Communication, view of Royal Palace and Almudena Cathedral.

Flag

Wikipedia Article:

- text
- links
- categories

Similarity between Wikipedia Articles

Hypothesis

Create account Log in

Article Talk Read Edit View history Search

Wikipedia
The Free Encyclopedia

Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!

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

Languages
Acèh
Адыгэбзэ
Afrikaans
Alemannisch

Madrid

From Wikipedia, the free encyclopedia

Coordinates: 40°23′N 3°43′W

This article is about the capital city of Spain. For the autonomous community, see Community of Madrid. For other uses, see Madrid (disambiguation).

Madrid (/məˈdrɪd/, Spanish: [maˈðrið], locally: [maˈðriθ, -ˈðri]) is a south-western European city and the capital and largest municipality of Spain. The population of the city is almost 3.2 million[4] and that of the Madrid metropolitan area, around 6 million. It is the third-largest city in the European Union, after London and Berlin, and its metropolitan area is the third-largest in the European Union after Paris and London.[5][6][7][8] The city spans a total of 604.3 km² (233.3 sq mi).[9]

The city is located on the Manzanares River in the centre of both the country and the Community of Madrid (which comprises the city of Madrid, its conurbation and extended suburbs and villages); this community is bordered by the autonomous communities of Castile and León and Castile-La Mancha. As the capital city of Spain, seat of government, and residence of the Spanish monarch, Madrid is also the political, economic and cultural centre of Spain.[10] The current mayor is Manuela Carmena from Ahora Madrid.

The Madrid urban agglomeration has the third-largest GDP[11] in the European Union and its influences in politics, education, entertainment, environment, media, fashion, science, culture, and the arts all contribute to its status as one of the world's major global cities.[12][13] Due to its economic output, high standard of living, and market size, Madrid is considered the major financial centre of Southern Europe[14][15] and the Iberian Peninsula; it hosts the head offices of the vast majority of the major Spanish companies, such as Telefónica, Iberia and Repsol. Madrid is the 17th most livable city in the world according to Monocle magazine, in its 2014 index.[16][17]

Madrid houses the headquarters of the World Tourism Organization (WTO), belonging to the United Nations Organization (UN), the SEGIB, the Organization of Ibero-American States (OEI), and the Public Interest Oversight Board (PIOB). It also hosts major international regulators of Spanish: the Standing Committee of the Association of Spanish Language Academies, headquarters of the Royal Spanish Academy (RAE), the Cervantes Institute and the Foundation of Urgent Spanish (Fundéu BBVA). Madrid organizes fairs such as FITUR,[18] ARCO,[19] SIMO TCI[20] and the Cibeles Madrid Fashion Week.[21]

Madrid

From upper left: view of business districts of AZCA and CTBA, Gran Vía street and Metropolis Building, the Palace of Communication, view of Royal Palace and Almudena Cathedral.

Flag

Wikipedia Article:

- **text** $\alpha \cdot simText$ +
- **links** $\beta \cdot simLinks$ +
- **categories** $\gamma \cdot simCtg$

$$simWA(R1,R2) = \alpha \cdot simTxt(R1,R2) + \beta \cdot simLinks(R1,R2) + \gamma \cdot simCtg(R1,R2)$$

where $\alpha+\beta+\gamma=1$

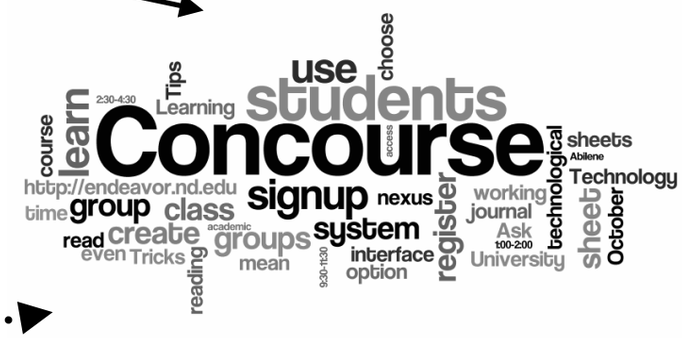
Similarity based on Text

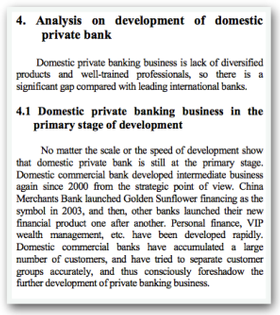
Latent
Dirichlet
Allocation

TOPIC_1

TOPIC_2

TOPIC_n





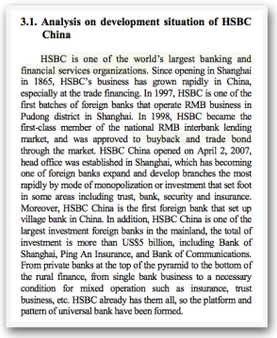
4. Analysis on development of domestic private bank

Domestic private banking business is lack of diversified products and well-trained professionals, so there is a significant gap compared with leading international banks.

4.1 Domestic private banking business in the primary stage of development

No matter the scale or the speed of development show that domestic private bank is still at the primary stage. Domestic commercial bank developed intermediate business again since 2000 from the strategic point of view. China Merchants Bank launched Golden Sunflower financing as the symbol in 2003, and then, other banks launched their new financial product one after another. Personal finance, VIP wealth management, etc. have been developed rapidly. Domestic commercial banks have accumulated a large number of customers, and have tried to separate customer groups accurately, and thus consciously foreshadow the further development of private banking business.

Ri



3.1. Analysis on development situation of HSBC China

HSBC is one of the world's largest banking and financial services organizations. Since opening in Shanghai in 1865, HSBC's business has grown rapidly in China, especially at the trade financing. In 1997, HSBC is one of the first batches of foreign banks that operate RMB business in Pudong district in Shanghai. In 1998, HSBC became the first-class member of the national RMB interbank lending market, and was approved to buyback and trade bond through the market. HSBC China opened on April 2, 2007, head office was established in Shanghai, which has becoming one of foreign banks expand and develop branches the most rapidly by mode of monopolization or investment that set foot in some areas including trust, bank, security and insurance. Moreover, HSBC China is the first foreign bank that set up village bank in China. In addition, HSBC China is one of the largest investment foreign banks in the mainland, the total of investment is more than US$5 billion, including Bank of Shanghai, Ping An Insurance, and Bank of Communications. From private banks at the top of the pyramid to the bottom of the rural finance, from single bank business to a necessary condition for mixed operation such as insurance, trust business, etc. HSBC already has them all, so the platform and pattern of universal bank have been formed.

Rj

$$sim_{cont}(R_i, R_j) = 10^{-JSD(p,q)}$$

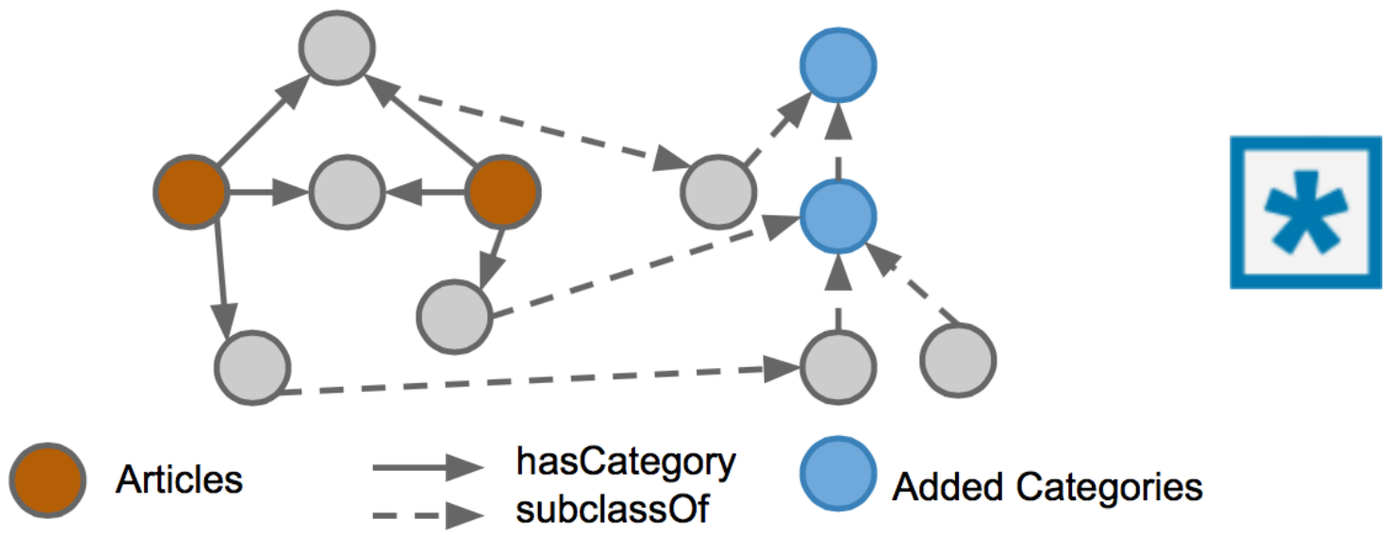
p = [0.5, 0.3,.., 0.7]

q = [0.2, 0.4,.., 0.9]

Similarity based on Categories

Articles with multiple common **categories** are likely to be similar

$$Sim(A,B) = cat(A) \cap cat(B) / ((cat(A) \cup cat(B))/2)$$

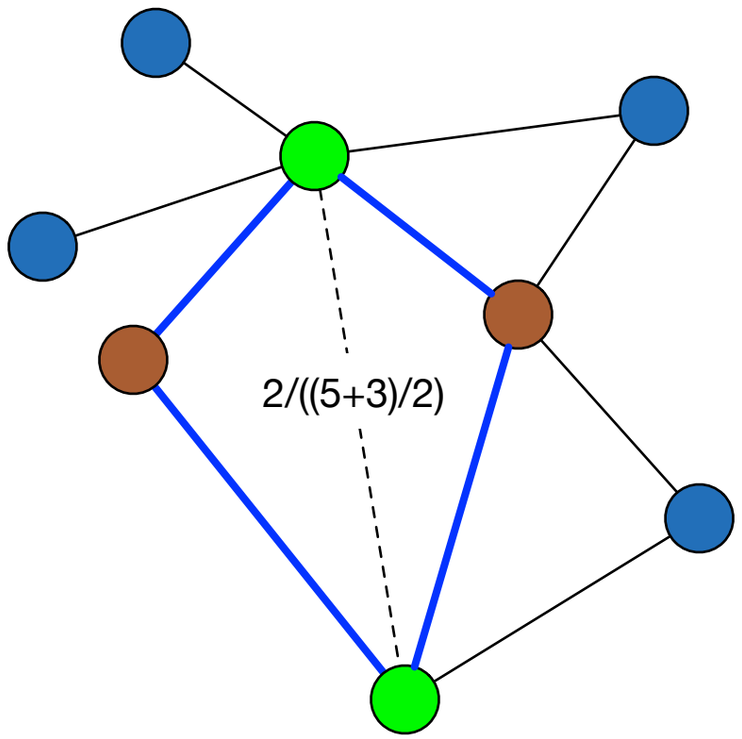


Noise filtering is necessary (e.g., “All articles lacking in-text citations”).
See https://github.com/cbadenes/siminwikart-challenge4/blob/master/category/wikipedia_bad_categories.txt

Similarity based on Links

Articles with multiple common **links** are likely to be similar

Sim(A,B) = links(A) ∩ links(B) / ((links(A) U links(B)) / 2)



Proof of Concept

(simLinks) α = 0.2
(simCtg) β = 0.2
(simTxt) γ = 0.6

Fernando Alonso

simTxt = 0.065
simLinks = 0.0
simCtg=*[1]*0.095
*[3]*0.161

simTxt = 0.059
simLinks = 0.019
simCtg=*[1]*0.117
*[3]*0.181

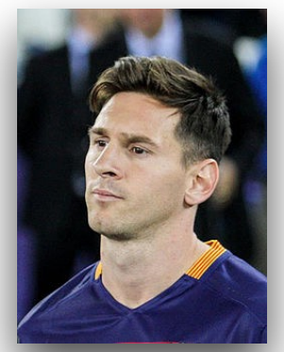
[1]0.062
[3]0.075

[1]0.068
[3]0.069

simTxt = 0.052
simLinks = 0.019
simCtg=*[1]*0.166
*[3]*0.172

[1]0.019
[3]0.023

Lionel Messi



[1]0.666
[3]0.683

[1]0.043
[3]0.072

Iker Casillas

simTxt = 0.980
simLinks = 0.175
simCtg=*[1]*0.217
*[3]*0.302

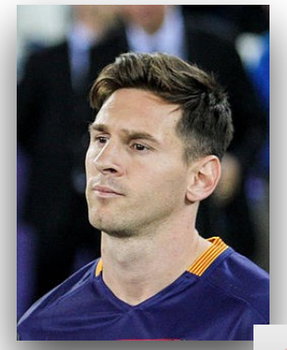
simTxt = 0.060
simLinks = 0.008
simCtg=*[1]*0.030
*[3]*0.172

Princess Akiko

simTxt = 0.069
simLinks = 0.004
simCtg=*[1]*0.080
*[3]*0.134

[1]0.058
[3]0.069

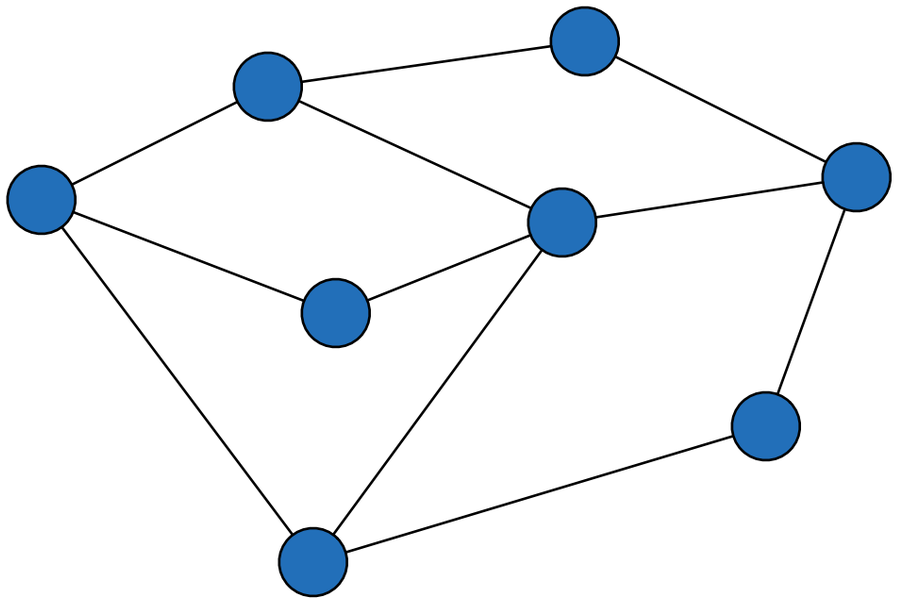
Lionel Messi



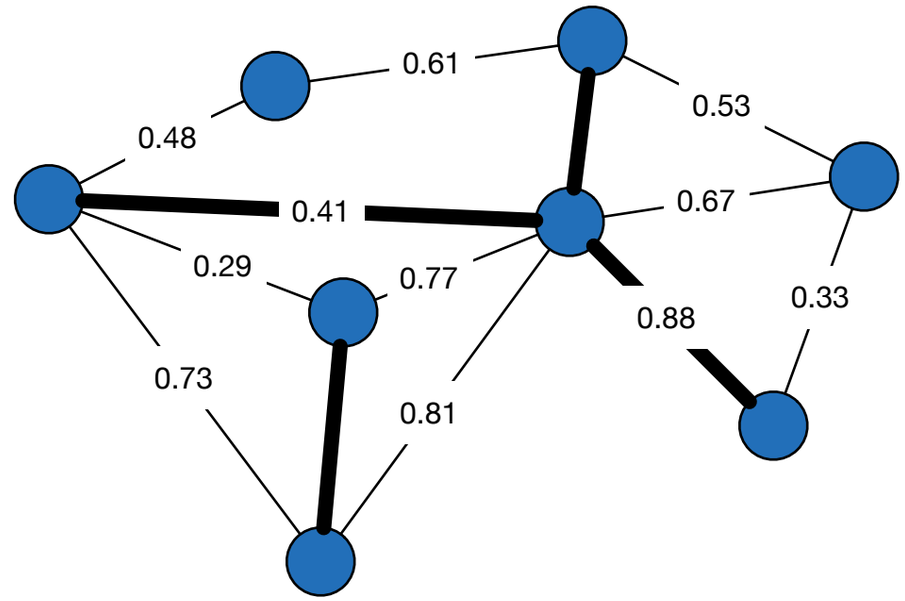

```
simTxt = 0.060     ->  <common words>
simLinks = 0.008   -> (England,Buenos_Aires,Chile,Madrid,Argentina)
simCtg=[1]0.030    ->  living_person
```

Princess Akiko

Graph based on Links



Graph based on Similarities



Problem

Create account Log in

Article Talk | Read Edit View history | Search

Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!

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

Languages
Acèh
Адыгэбзэ
Afrikaans
Alemannisch

Madrid

From Wikipedia, the free encyclopedia

Coordinates: 40°23′N 3°43′W

This article is about the capital city of Spain. For the autonomous community, see Community of Madrid. For other uses, see Madrid (disambiguation).

Madrid (/məˈdrɪd/, Spanish: [maˈðrið], locally: [maˈðriθ, -ˈðri]) is a south-western European city and the capital and largest municipality of Spain. The population of the city is almost 3.2 million[4] and that of the Madrid metropolitan area, around 6 million. It is the third-largest city in the European Union, after London and Berlin, and its metropolitan area is the third-largest in the European Union after Paris and London.[5][6][7][8] The city spans a total of 604.3 km² (233.3 sq mi).[9]

The city is located on the Manzanares River in the centre of both the country and the Community of Madrid (which comprises the city of Madrid, its conurbation and extended suburbs and villages); this community is bordered by the autonomous communities of Castile and León and Castile-La Mancha. As the capital city of Spain, seat of government, and residence of the Spanish monarch, Madrid is also the political, economic and cultural centre of Spain.[10] The current mayor is Manuela Carmena from Ahora Madrid.

The Madrid urban agglomeration has the third-largest GDP[11] in the European Union and its influences in politics, education, entertainment, environment, media, fashion, science, culture, and the arts all contribute to its status as one of the world's major global cities.[12][13] Due to its economic output, high standard of living, and market size, Madrid is considered the major financial centre of Southern Europe[14][15] and the Iberian Peninsula; it hosts the head offices of the vast majority of the major Spanish companies, such as Telefónica, Iberia and Repsol. Madrid is the 17th most livable city in the world according to Monocle magazine, in its 2014 index.[16][17]

Madrid houses the headquarters of the World Tourism Organization (WTO), belonging to the United Nations Organization (UN), the SEGIB, the Organization of Ibero-American States (OEI), and the Public Interest Oversight Board (PIOB). It also hosts major international regulators of Spanish: the Standing Committee of the Association of Spanish Language Academies, headquarters of the Royal Spanish Academy (RAE), the Cervantes Institute and the Foundation of Urgent Spanish (Fundéu BBVA). Madrid organizes fairs such as FITUR,[18] ARCO,[19] SIMO TCI[20] and the Cibeles Madrid Fashion Week.[21]

Madrid

From upper left: view of business districts of AZCA and CTBA, Gran Vía street and Metropolis Building, the Palace of Communication, view of Royal Palace and Almudena Cathedral.

Flag

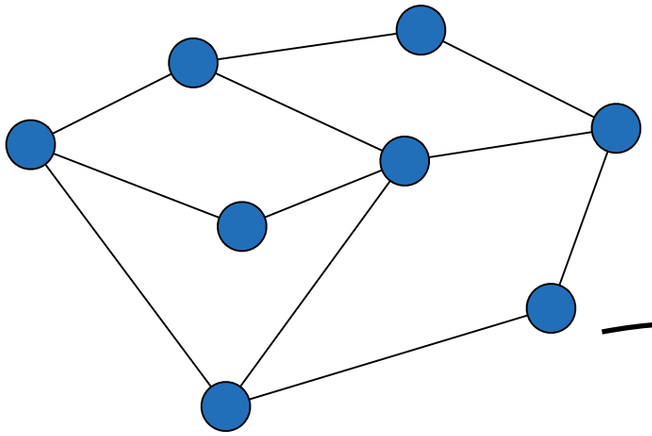
Wikipedia Article:

- text
- links
- categories

Wikipedia links reliability (missing links)

Similarities between categories (as topics) can define relations between articles

Graph based on Links

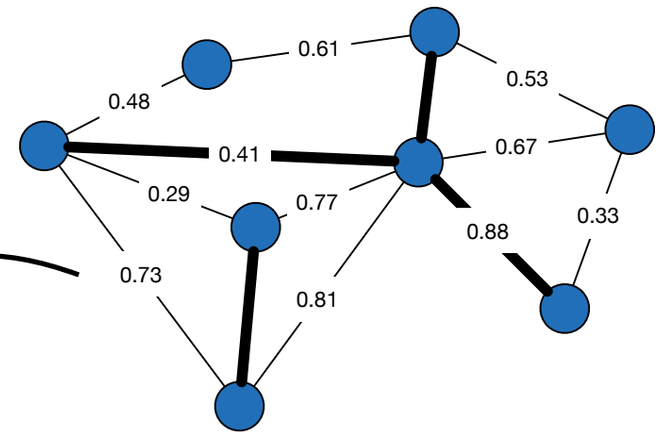


+

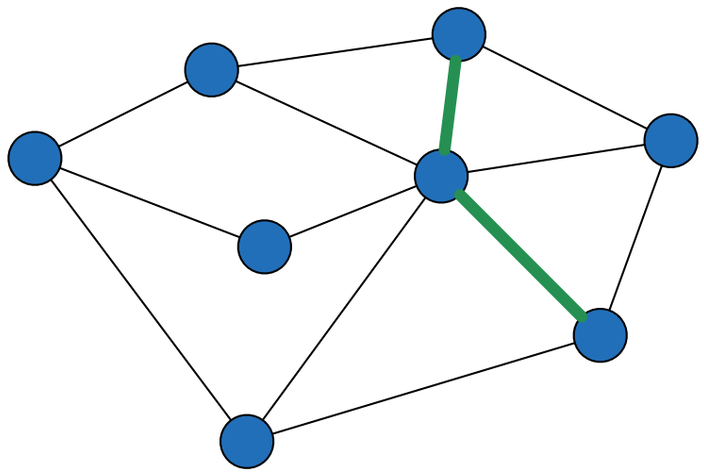
Latent Dirichlet Allocation Topic Model

+

Graph based on Similarities



Subgraph Pattern Matching



Code

https://github.com/cbadenes/siminwikart-challenge4

*Sparksee