Text Similarity

Eva Gerlitz

January 31, 2018

Outline

1. Introduction
2. DBpedia
3. Training
    - Supervised
        - Preprocessing
        - Models
        - Results
    - Unsupervised
        - Preprocessing
        - Model
        - Results
4. Conclusion

Table of Contents

1. Introduction
2. DBpedia
3. Training
   - Supervised
     - Preprocessing
     - Models
     - Results
   - Unsupervised
     - Preprocessing
     - Model
     - Results
4. Conclusion

Introduction

- Take different classes of wikipedia articles (using DBpedia)
- Look at abstracts
- How well do different models learn to discriminate the classes?

Table of Contents

1. Introduction
2. DBpedia
3. Training
   - Supervised
     - Preprocessing
     - Models
     - Results
   - Unsupervised
     - Preprocessing
     - Model
     - Results
4. Conclusion

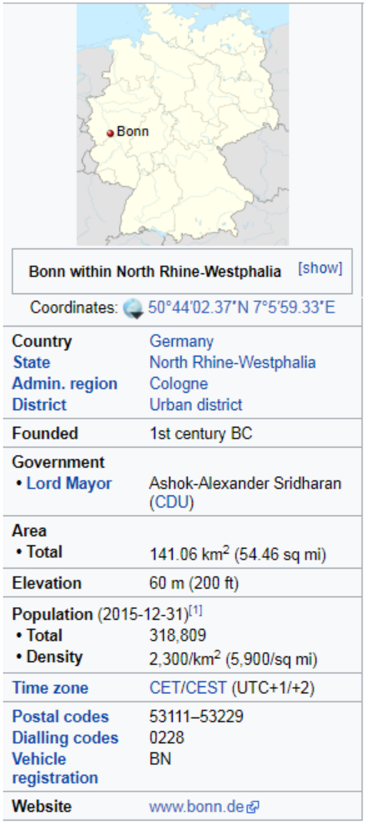
What is DBpedia?

Figure 1: DBpedia[1]

- Linked Data project
- Extracted, structured content from Wikipedia (Infoboxes)
- Possible to query properties and relationships

[1] Pic from https://en.wikipedia.org/wiki/DBpedia

DBpedia



Bonn within North Rhine-Westphalia [show]

Coordinates: 50°44′02.37″N 7°5′59.33″E

Country	Germany
State	North Rhine-Westphalia
Admin. region	Cologne
District	Urban district
Founded	1st century BC
Government	
• Lord Mayor	Ashok-Alexander Sridharan (CDU)
Area	
• Total	141.06 km^2 (54.46 sq mi)
Elevation	60 m (200 ft)
Population (2015-12-31)[1]	
• Total	318,809
• Density	2,300/km^2 (5,900/sq mi)
Time zone	CET/CEST (UTC+1/+2)
Postal codes	53111–53229
Dialling codes	0228
Vehicle registration	BN
Website	www.bonn.de

DBpedia - get data

- SPARQL queries [2]

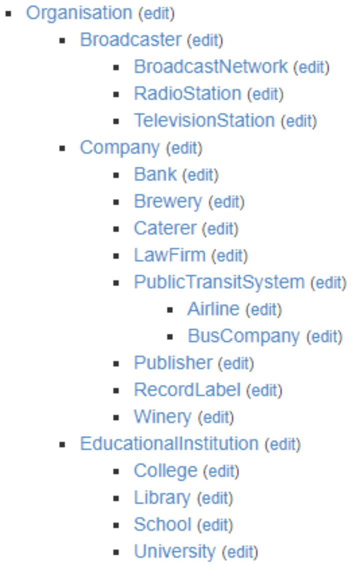
```
PREFIX dbpedia0: <http://dbpedia.org/ontology/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/resource/>

select distinct ?name ?abstract where {
    ?instance a dbpedia0:EducationalInstitution.
    ?instance foaf:name ?name.
    ?instance dbpedia0:abstract ?abstract.
    filter(langMatches(lang(?abstract),"en"))
    }
```

[2] http://dbpedia.org/snorql

DBpedia - Classes

- Use Subclasses that are still large enough



- Organisation (edit)
  - Broadcaster (edit)
    - BroadcastNetwork (edit)
    - RadioStation (edit)
    - TelevisionStation (edit)
  - Company (edit)
    - Bank (edit)
    - Brewery (edit)
    - Caterer (edit)
    - LawFirm (edit)
    - PublicTransitSystem (edit)
      - Airline (edit)
      - BusCompany (edit)
    - Publisher (edit)
    - RecordLabel (edit)
    - Winery (edit)
  - EducationalInstitution (edit)
    - College (edit)
    - Library (edit)
    - School (edit)
    - University (edit)

Figure 2: Classes in DBpedia[3]

- DBpedia Data not perfect (College!)

[3] http://mappings.dbpedia.org/server/ontology/classes/

Datasets

- **Actor** (10.000)
- **City** (10.000)
- **Celestial Body** (7699)
- **Educational Institution** (10.000)
- **Lake** (10.000)

- Athlete (10.000)
- Fictional Character (10.000)
- Musical Artist (10.000)

Table of Contents

1. Introduction
2. DBpedia
3. Training
   - Supervised
     - Preprocessing
     - Models
     - Results
   - Unsupervised
     - Preprocessing
     - Model
     - Results
4. Conclusion

Preprocessing - Hashing Vectorizer

- Turns abstracts into a sparse matrix
- No dictionary, but hashing function.

Preprocessing - Tfidf Vectorizer

- tf = term frequency, idf = inverse document frequency
- Term frequency = proportion of occurrences of a specific term to the total number of terms in a document
- Inverse document frequency = inverse of the proportion of documents that contain that word/phrase.

Model 1: Random Forest

- Consists of many decision trees, that are built during training
- When predicting a class, every tree gives a vote
- The class with most votes wins

Model 2: K-nearest neighbors

- Training examples in feature space, dimension = number of features
- Look at k neighbors → majority of classes will be predicted.

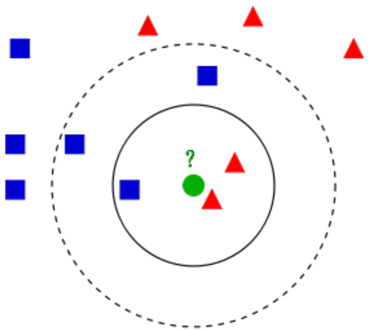


Figure 3: k-nearest neighbors[4]

[4] Pic from https://en.wikipedia.org/wiki/K-nearest$_neighbors_algorithm$

Model 3: Stochastic Gradient Descent

- Multiple binary classifiers, ”One versus all” scheme
- For classification: compute confidence score

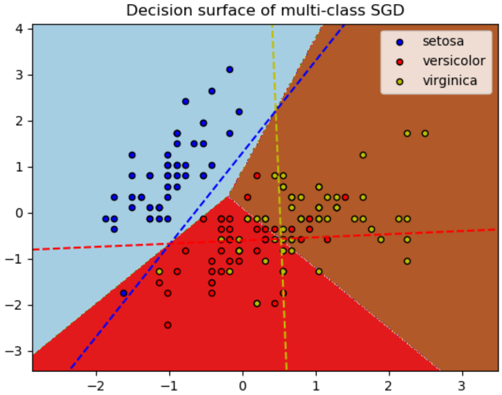


Figure 4: Multiple classes[5]

[5] Pic from http://scikit-learn.org/stable/modules/sgd.html

Results - 2 Classes

Classes	**Accuracy**
Actors and Fictional Characters	0.98
Actors and Athletes	0.989
Actors and Cities	0.997
Actor and MusicalArtist	0.954
City and Lake	0.986
City and Educational Institution	0.989
Actor and Celestial Body	1.00

Results - 5 Classes

Preprocessing → Models ↓	Hashing	tfidf
Random Forest	0.986	0.986
k-nearest Neighbors	0.993	0.977
Stochastic Gradient Descent	0.996	0.994

→ About 140 mistakes for hashing & random forest.

Results - What went wrong?

Classes	Actor	City	Cel. Body	Edu. Institution	Lake
Actor	-	5/4/8	4/2/1	16/14/4	1/2/2
City	*	-	1/0/0	34/19/17	57/26/6
Cel. Body	*	*	-	0/0/0	2/0/0
Edu. Institution	*	*	*	-	19/4/0
Lake	*	*	*	*	-

Table 1: How often were the class objects mistaken? - Hashing Vectorizer, Random Forest, K-nearest neighbors, Stochastic Gradient Descent

Results - What went wrong?

Classes	Actor	City	Cel. Body	Edu. Institution	Lake
Actor	-	5/4/8	4/2/1	16/14/4	1/2/2
City	*	-	1/0/0	34/19/17	57/26/6
Cel. Body	*	*	-	0/0/0	2/0/0
Edu. Institution	*	*	*	-	19/4/0
Lake	*	*	*	*	-

Table 2: How often were the class objects mistaken? - Hashing Vectorizer, Random Forest, K-nearest neighbors, Stochastic Gradient Descent

Results - What went wrong?

Classes	Actor	City	Cel. Body	Edu. Institution	Lake
Actor	-	5/4/8	4/2/1	16/14/4	1/2/2
City	*	-	1/0/0	34/19/17	57/26/6
Cel. Body	*	*	-	0/0/0	2/0/0
Edu. Institution	*	*	*	-	19/4/0
Lake	*	*	*	*	-

Table 3: How often were the class objects mistaken? - Hashing Vectorizer, Random Forest, K-nearest neighbors, Stochastic Gradient Descent

Results - What went wrong? - Example 1

- ”Pantnagar is a town and a university campus in Udham Singh Nagar district, Uttarakhand. Nainital, Kashipur, Rudrapur and Kiccha, Haldwani are the major cities surrounding Pantnagar.
  The town is famous for having the first agricultural university of India which was established on 17 November 1960. Previously the university was called the Uttar Pradesh Agricultural University or Pantnagar University. It was rechristened G. B. Pant University of Agriculture and Technology. keeping in view the contributions of Pt. Govind Ballabh Pant, the then Chief Minister of UP.
  In recent years, an integrated industrial estate has been established near the campus which houses companies such as Tata motor, Bajaj, Britannia, HP, HCL, Voltas, Schneider Electric, Nestle, Dabur, Vedanta Resources etc., as a part of SIDCUL industrial area developed by government owned State Industrial Development Corporation of Uttarakhand Limited.”

Results - What went wrong? - Example 1

- **”Pantnagar is a town** and a university campus in Udham Singh Nagar district, Uttarakhand. Nainital, Kashipur, Rudrapur and Kiccha, Haldwani are the major cities surrounding Pantnagar.
  The town is famous for having the first agricultural university of India which was established on 17 November 1960. Previously the university was called the Uttar Pradesh Agricultural University or Pantnagar University. It was rechristened G. B. Pant University of Agriculture and Technology. keeping in view the contributions of Pt. Govind Ballabh Pant, the then Chief Minister of UP.
  In recent years, an integrated industrial estate has been established near the campus which houses companies such as Tata motor, Bajaj, Britannia, HP, HCL, Voltas, Schneider Electric, Nestle, Dabur, Vedanta Resources etc., as a part of SIDCUL industrial area developed by government owned State Industrial Development Corporation of Uttarakhand Limited.”
- Belongs to class **City**

Results - What went wrong? - Example 1

- ”Pantnagar is a town and a university campus in Udham Singh Nagar district, Uttarakhand. Nainital, Kashipur, Rudrapur and Kiccha, Haldwani are the major cities surrounding Pantnagar.
  The town is famous for having the first agricultural university of India which was established on 17 November 1960. Previously the university was called the Uttar Pradesh Agricultural University or Pantnagar University. It was rechristened G. B. Pant University of Agriculture and Technology. keeping in view the contributions of Pt. Govind Ballabh Pant, the then Chief Minister of UP.
  In recent years, an integrated industrial estate has been established near the campus which houses companies such as Tata motor, Bajaj, Britannia, HP, HCL, Voltas, Schneider Electric, Nestle, Dabur, Vedanta Resources etc., as a part of SIDCUL industrial area developed by government owned State Industrial Development Corporation of Uttarakhand Limited.”
- Belongs to class **City**
- Labeled as: **Educational Institution**
- Problem: Overlapping of differnet classes

Results - What went wrong? - Example 2

- ”Rangsit is a city in Pathum Thani Province, Thailand. Rangsit is the home of Rangsit University.”

Results - What went wrong? - Example 2

- ”Rangsit is a city in Pathum Thani Province, Thailand. Rangsit is the home of Rangsit University.”
- Belongs to class **City**

Results - What went wrong? - Example 2

- "Rangsit is a city in Pathum Thani Province, Thailand. Rangsit is the home of Rangsit University."
- Belongs to class **City**
- Labeled as: **Educational Institution**
- Problem: Short tetxts

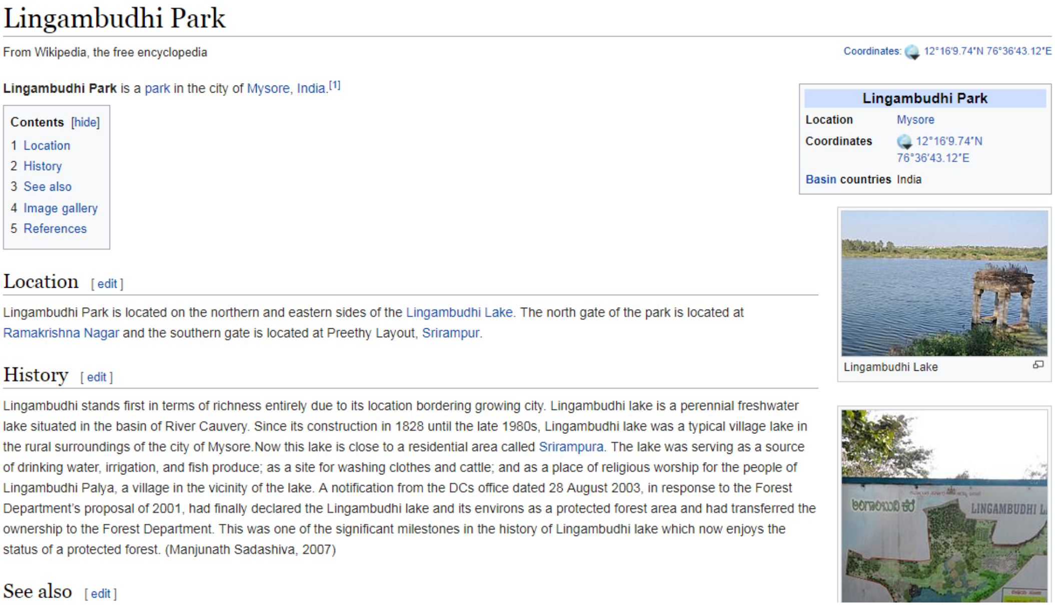
Results - What went wrong? - Example 3

- ”Lingambudhi Park is a park in the city of Mysore, India.”

Results - What went wrong? - Example 3

- ”Lingambudhi Park is a park in the city of Mysore, India.”
- Belongs to class **Lake**
- Labeled as: **City**

Results - What went wrong? - Example 3



Lingambudhi Park

From Wikipedia, the free encyclopedia

Coordinates: 12°16'9.74"N 76°36'43.12"E

Lingambudhi Park is a park in the city of Mysore, India.[1]

Lingambudhi Park	
Location	Mysore
Coordinates	12°16'9.74"N 76°36'43.12"E
Basin countries	India

Contents [hide]

1 Location
2 History
3 See also
4 Image gallery
5 References

Location [edit]

Lingambudhi Park is located on the northern and eastern sides of the Lingambudhi Lake. The north gate of the park is located at Ramakrishna Nagar and the southern gate is located at Preethy Layout, Srirampur.

Lingambudhi Lake

History [edit]

Lingambudhi stands first in terms of richness entirely due to its location bordering growing city. Lingambudhi lake is a perennial freshwater lake situated in the basin of River Cauvery. Since its construction in 1828 until the late 1980s, Lingambudhi lake was a typical village lake in the rural surroundings of the city of Mysore.Now this lake is close to a residential area called Srirampura. The lake was serving as a source of drinking water, irrigation, and fish produce; as a site for washing clothes and cattle; and as a place of religious worship for the people of Lingambudhi Palya, a village in the vicinity of the lake. A notification from the DCs office dated 28 August 2003, in response to the Forest Department's proposal of 2001, had finally declared the Lingambudhi lake and its environs as a protected forest area and had transferred the ownership to the Forest Department. This was one of the significant milestones in the history of Lingambudhi lake which now enjoys the status of a protected forest. (Manjunath Sadashiva, 2007)

See also [edit]

- Problem: Wrong class in DBpedia

Topic Modeling

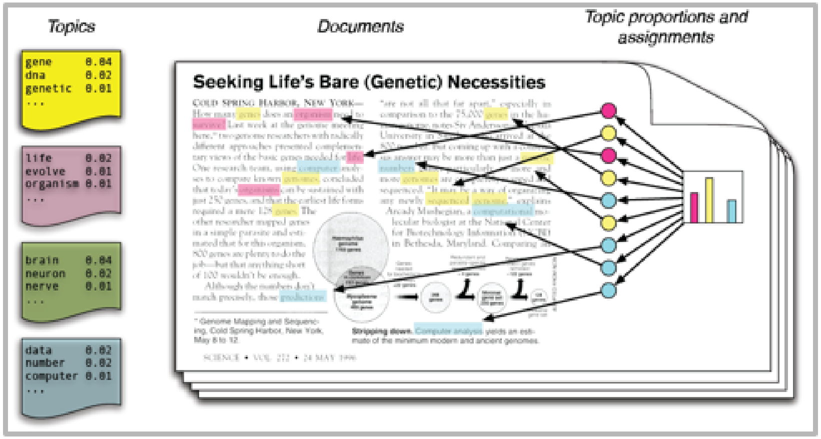


Figure 5: Topic Modeling[6]

[6] Pic from https://www.analyticsvidhya.com/wp-content/uploads/2016/08/Modeling1.png

Preprocessing

- 1. Remove punctuation
- 2. Remove Stop words
- 3. Lemmatize all words
- 4. Build dictionary
- 5. Build a matrix: For every abstract: Tupel (word id, word frequency)

LDA - Latent Dirichlet Allocation

- Statistical model
- What topics could have created this text?
- Which words are the most probable for the topics?

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student
- lake, river, located, water, area
- star, asteroid, galaxy, year, constellation
- loch, also, ontario, film, john
- chinese, hong, china, kong, quechua

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student → **Educational Institution**
- lake, river, located, water, area
- star, asteroid, galaxy, year, constellation
- loch, also, ontario, film, john
- chinese, hong, china, kong, quechua

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student → Educational Institution
- lake, river, located, water, area → **Lake**
- star, asteroid, galaxy, year, constellation
- loch, also, ontario, film, john
- chinese, hong, china, kong, quechua

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student → Educational Institution
- lake, river, located, water, area → Lake
- star, asteroid, galaxy, year, constellation → **Celestial Body**
- loch, also, ontario, film, john
- chinese, hong, china, kong, quechua

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student → Educational Institution
- lake, river, located, water, area → Lake
- star, asteroid, galaxy, year, constellation → Celestial Body
- loch, also, ontario, film, john → **Actor?**
- chinese, hong, china, kong, quechua

Topic Modeling - 1. Results

5 topis, 5 words:

- school, high, university, college, student → Educational Institution
- lake, river, located, water, area → Lake
- star, asteroid, galaxy, year, constellation → Celestial Body
- loch, also, ontario, film, john → Actor?
- chinese, hong, china, kong, quechua → **City??**

Preprocessing

- Remove all parts of names as well
- → stopwords = stopwords + ”also” + all names

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student
- river, located, area, reservoir, water
- asteroid, year, galaxy, constellation, approximately
- chinese, film, actor, actress, known
- province, county, norway, lough, district

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student → **Educational Institution**
- river, located, area, reservoir, water
- asteroid, year, galaxy, constellation, approximately
- chinese, film, actor, actress, known
- province, county, norway, lough, district

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student → Educational Institution
- river, located, area, reservoir, water → **Lake**
- asteroid, year, galaxy, constellation, approximately
- chinese, film, actor, actress, known
- province, county, norway, lough, district

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student → Educational Institution
- river, located, area, reservoir, water → Lake
- asteroid, year, galaxy, constellation, approximately → **Celestial Body**
- chinese, film, actor, actress, known
- province, county, norway, lough, district

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student → Educational Institution
- river, located, area, reservoir, water → Lake
- asteroid, year, galaxy, constellation, approximately → Celestial Body
- chinese, film, actor, actress, known → **Actor**
- province, county, norway, lough, district

Topic Modeling - 2. Results

5 topics, 5 words:

- school, high, university, college, student → Educational Institution
- river, located, area, reservoir, water → Lake
- asteroid, year, galaxy, constellation, approximately → Celestial Body
- chinese, film, actor, actress, known → Actor
- province, county, norway, lough, district → **City**

Table of Contents

1. Introduction
2. DBpedia
3. Training
   - Supervised
     - Preprocessing
     - Models
     - Results
   - Unsupervised
     - Preprocessing
     - Model
     - Results
4. Conclusion

Conclusion

- Supervised and unsupervised learning for wikipedia abstracts
- 5 classes → How well can they be discriminated?
- Supervised learning: Problems with overlapping classes, short texts, false classes