Document Clustering with Query Constraints

Bauhaus-Universität Weimar

Matthias Busse
matthias.busse@uni-weimar.de

Chair of Big Data Analytics
Prof. Dr. Matthias Hagen

Defense of Master's Thesis
March 10th, 2015

Motivation

Bauhaus-Universität Weimar: University
www.**uni-weimar**.de/en/**university**/start/ ▾ Bauhaus University, Weimar ▾
The online magazine BAUHAUS.JOURNAL ONLINE provides up-to-date information about what's happening at **Bauhaus-Universität Weimar**.
Architecture and Urbanism - Studies - Faculty of Media - Academic Programmes

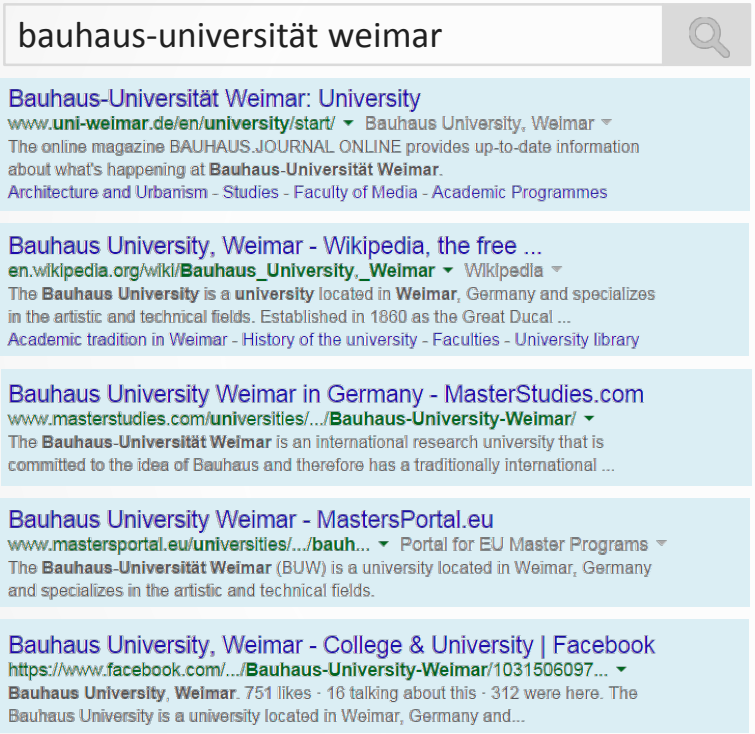
Bauhaus University, Weimar - Wikipedia, the free ...
en.wikipedia.org/wiki/**Bauhaus_University,_Weimar** ▾ Wikipedia ▾
The **Bauhaus University** is a **university** located in **Weimar**, Germany and specializes in the artistic and technical fields. Established in 1860 as the Great Ducal ...
Academic tradition in Weimar - History of the university - Faculties - University library

Bauhaus University Weimar in Germany - MasterStudies.com
www.masterstudies.com/**universities**/.../**Bauhaus-University-Weimar**/ ▾
The **Bauhaus-Universität Weimar** is an international research university that is committed to the idea of Bauhaus and therefore has a traditionally international ...

Bauhaus University Weimar - MastersPortal.eu
www.**mastersportal.eu**/**universities**/.../**bauh**... ▾ Portal for EU Master Programs ▾
The **Bauhaus-Universität Weimar** (BUW) is a university located in Weimar, Germany and specializes in the artistic and technical fields.

Bauhaus University, Weimar - College & University | Facebook
https://www.facebook.com/.../**Bauhaus-University-Weimar**/1031506097... ▾
Bauhaus University, **Weimar**. 751 likes · 16 talking about this · 312 were here. The Bauhaus University is a university located in Weimar, Germany and...

Motivation


bauhaus-universität weimar
Bauhaus-Universität Weimar: University
www.uni-weimar.de/en/university/start/ Bauhaus University, Weimar
The online magazine BAUHAUS.JOURNAL ONLINE provides up-to-date information about what's happening at Bauhaus-Universität Weimar.
Architecture and Urbanism - Studies - Faculty of Media - Academic Programmes
Bauhaus University, Weimar - Wikipedia, the free ...
en.wikipedia.org/wiki/Bauhaus_University,_Weimar Wikipedia
The Bauhaus University is a university located in Weimar, Germany and specializes in the artistic and technical fields. Established in 1860 as the Great Ducal ...
Academic tradition in Weimar - History of the university - Faculties - University library
Bauhaus University Weimar in Germany - MasterStudies.com
www.masterstudies.com/universities/.../Bauhaus-University-Weimar/
The Bauhaus-Universität Weimar is an international research university that is committed to the idea of Bauhaus and therefore has a traditionally international ...
Bauhaus University Weimar - MastersPortal.eu
www.mastersportal.eu/universities/.../bauh... Portal for EU Master Programs
The Bauhaus-Universität Weimar (BUW) is a university located in Weimar, Germany and specializes in the artistic and technical fields.
Bauhaus University, Weimar - College & University | Facebook
https://www.facebook.com/.../Bauhaus-University-Weimar/1031506097...
Bauhaus University, Weimar. 751 likes · 16 talking about this · 312 were here. The Bauhaus University is a university located in Weimar, Germany and...

Motivation

Jaguar
www.**jaguar**.com/ ▾ Jaguar Cars ▾
Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar** ▾ Wikipedia ▾
The **jaguar** (/ˈdʒægjuːər, ˈdʒægjʊər, ˈdʒægjuːɑr/ or /ˈdʒægwɑr/; Brazilian Portuguese: [ʒaˈgwaʁ], Spanish: [xaˈɣwar]), Panthera onca, is a big cat, ...

Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/**jaguar**/basic-facts ▾ Defenders of Wildlife ▾
The **jaguar** is the largest cat in the Americas. The **jaguar** has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾
Jaguar Cars is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by the Indian ...

Jacksonville Jaguars, Official Site of the Jacksonville Jaguars
www.**jaguars**.com/ ▾ Jacksonville Jaguars ▾
Take a look back at some of the best images from the Jacksonville **Jaguars** ...
Jaguars.com analyst Jeff Lageman breaks down the **Jaguars'** offensive line play in ...

Motivation

Jaguar
www.**jaguar**.com/ ▾ Jaguar Cars ▾
Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar** ▾ Wikipedia ▾
The **jaguar** (/ˈdʒægjuːər, ˈdʒægjʊər, ˈdʒægjuːɑr/ or /ˈdʒægwɑr/; Brazilian Portuguese: [ʒaˈgwaʁ], Spanish: [xaˈɣwar]), Panthera onca, is a big cat, ...

Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/**jaguar**/basic-facts ▾ Defenders of Wildlife ▾
The **jaguar** is the largest cat in the Americas. The **jaguar** has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾
Jaguar Cars is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by the Indian ...

Jacksonville Jaguars, Official Site of the Jacksonville Jaguars
www.**jaguars**.com/ ▾ Jacksonville Jaguars ▾
Take a look back at some of the best images from the Jacksonville **Jaguars** ...
Jaguars.com analyst Jeff Lageman breaks down the **Jaguars'** offensive line play in ...

Motivation

Jaguar
www.**jaguar**.com/ ▾ Jaguar Cars ▾
Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar** ▾ Wikipedia ▾
The **jaguar** (/ˈdʒægjuːər, ˈdʒægjʊər, ˈdʒægjuːɑr/ or /ˈdʒægwɑr/; Brazilian Portuguese: [ʒaˈgwaʁ], Spanish: [xaˈɣwar]), Panthera onca, is a big cat, ...

Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/**jaguar**/basic-facts ▾ Defenders of Wildlife ▾
The **jaguar** is the largest cat in the Americas. The **jaguar** has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...

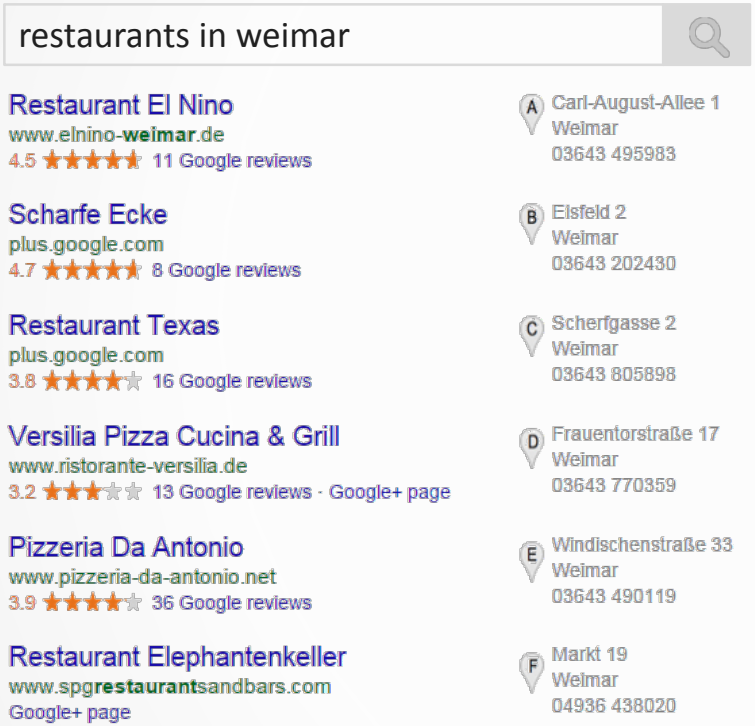
Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾
Jaguar Cars is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by the Indian ...

Jacksonville Jaguars, Official Site of the Jacksonville Jaguars
www.**jaguars**.com/ ▾ Jacksonville Jaguars ▾
Take a look back at some of the best images from the Jacksonville **Jaguars** ...
Jaguars.com analyst Jeff Lageman breaks down the **Jaguars'** offensive line play in ...

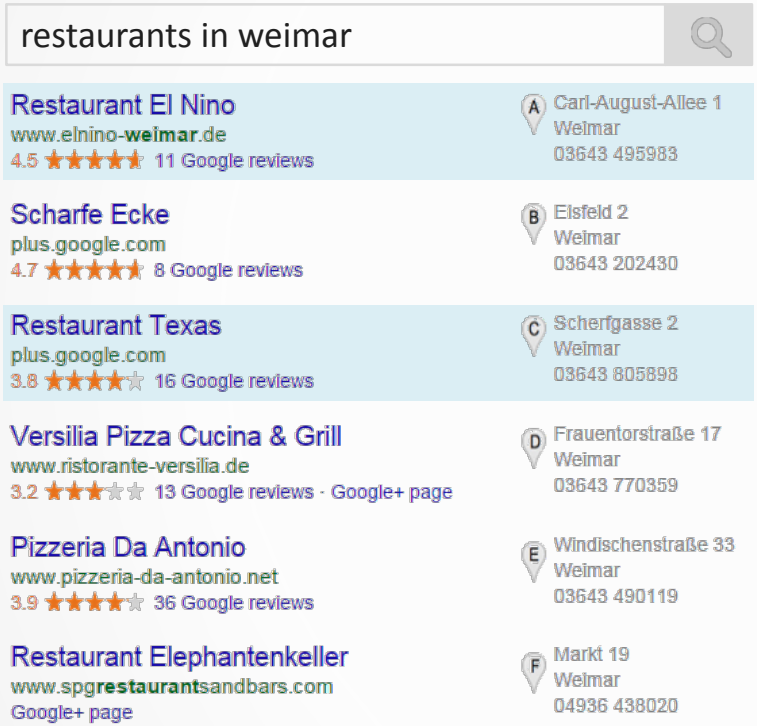
Motivation


jaguar
Jaguar
www.jaguar.com/ Jaguar Cars
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets and model-specific websites.
Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Jaguar Wikipedia
The jaguar (/ˈdʒægjuːər, ˈdʒægjʊər, ˈdʒægjuːɑr/ or /ˈdʒægwɑr/; Brazilian Portuguese: [ʒaˈgwaʁ], Spanish: [xaˈɣwar]), Panthera onca, is a big cat, ...
Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/jaguar/basic-facts Defenders of Wildlife
The jaguar is the largest cat in the Americas. The jaguar has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...
Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Jaguar_Cars Wikipedia
Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by the Indian ...
Jacksonville Jaguars, Official Site of the Jacksonville Jaguars
www.jaguars.com/ Jacksonville Jaguars
Take a look back at some of the best images from the Jacksonville Jaguars ...
Jaguars.com analyst Jeff Lageman breaks down the Jaguars' offensive line play in ...

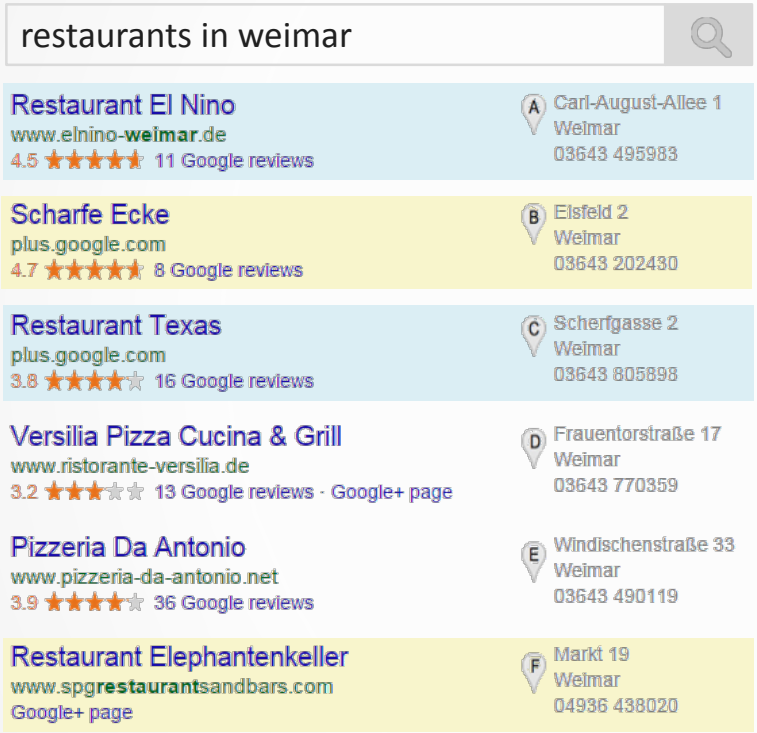
Motivation


restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1
Weimar
03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2
Weimar
03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2
Weimar
03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17
Weimar
03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33
Weimar
03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19
Weimar
04936 438020

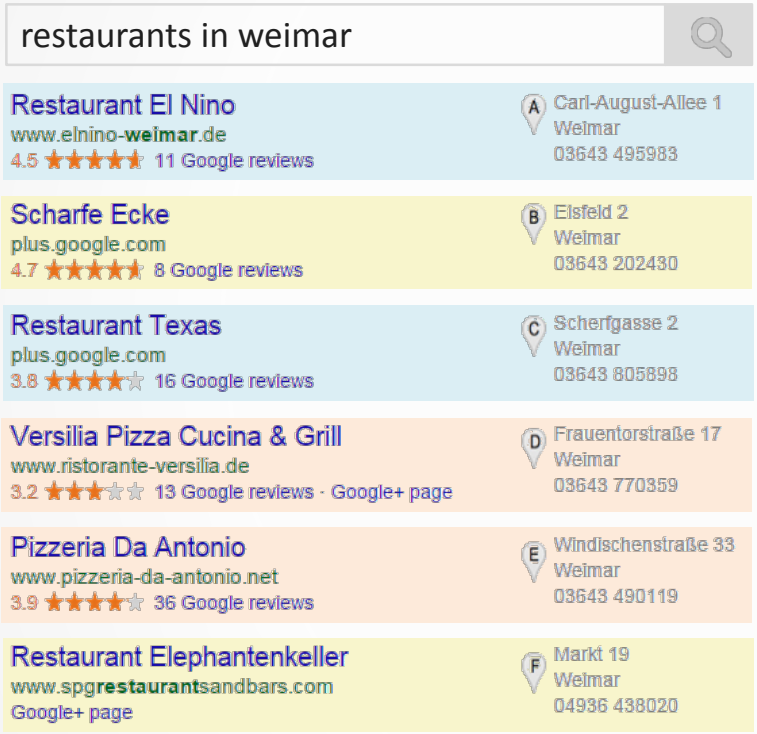
Motivation


restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1 Weimar 03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2 Weimar 03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2 Weimar 03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17 Weimar 03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33 Weimar 03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19 Weimar 04936 438020

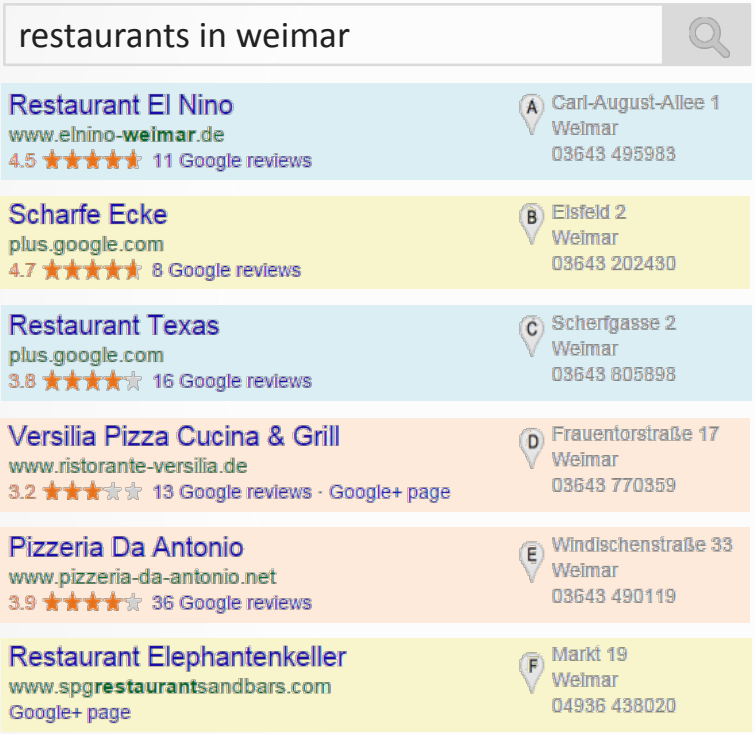
Motivation

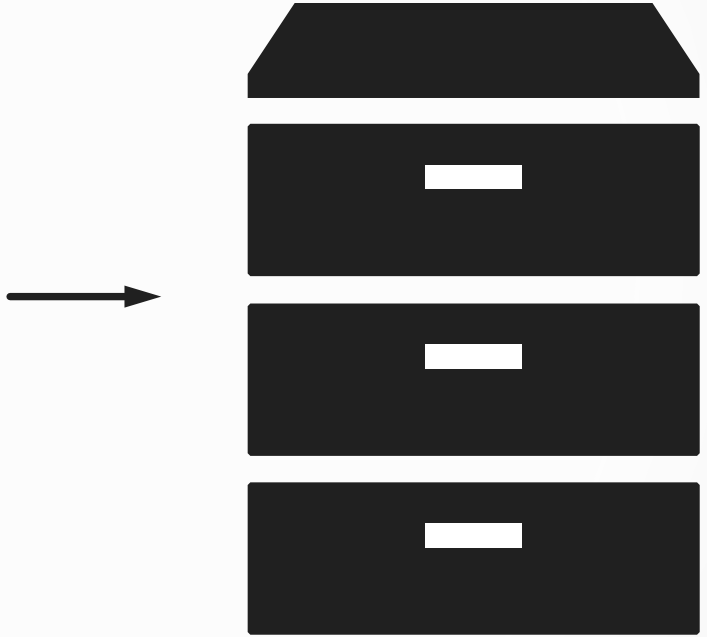

restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1 Weimar 03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2 Weimar 03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2 Weimar 03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17 Weimar 03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33 Weimar 03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19 Weimar 04936 438020

Motivation

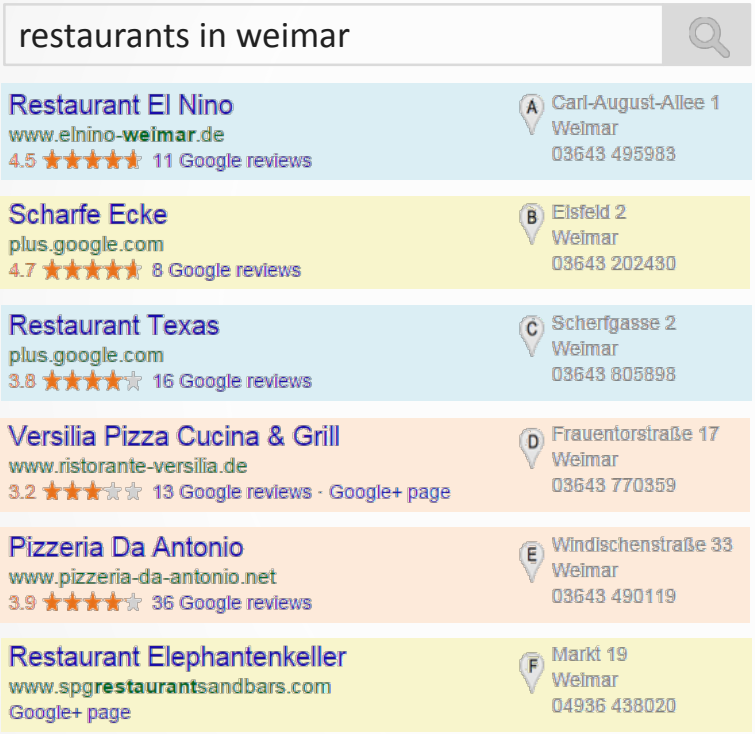

restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1 Weimar 03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2 Weimar 03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2 Weimar 03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17 Weimar 03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33 Weimar 03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19 Weimar 04936 438020

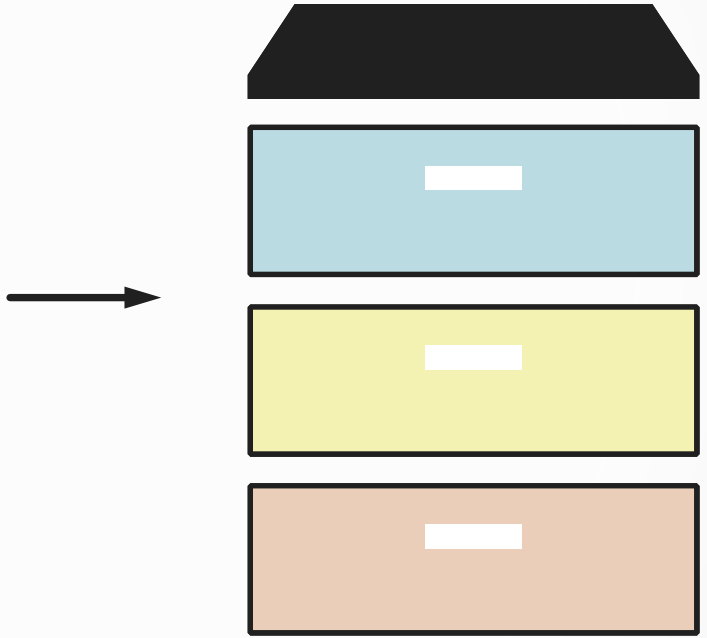
Motivation


restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1 Weimar 03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2 Weimar 03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2 Weimar 03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17 Weimar 03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33 Weimar 03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19 Weimar 04936 438020



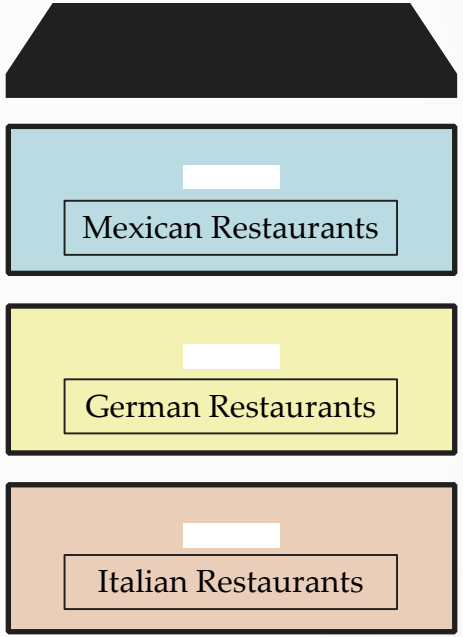
Motivation


restaurants in weimar
Restaurant El Nino
www.elnino-weimar.de
4.5 11 Google reviews
A Carl-August-Allee 1 Weimar 03643 495983
Scharfe Ecke
plus.google.com
4.7 8 Google reviews
B Elsfeld 2 Weimar 03643 202430
Restaurant Texas
plus.google.com
3.8 16 Google reviews
C Scherfgasse 2 Weimar 03643 805898
Versilia Pizza Cucina & Grill
www.ristorante-versilia.de
3.2 13 Google reviews · Google+ page
D Frauentorstraße 17 Weimar 03643 770359
Pizzeria Da Antonio
www.pizzeria-da-antonio.net
3.9 36 Google reviews
E Windischenstraße 33 Weimar 03643 490119
Restaurant Elephantenkeller
www.spgrestaurantsandbars.com
Google+ page
F Markt 19 Weimar 04936 438020

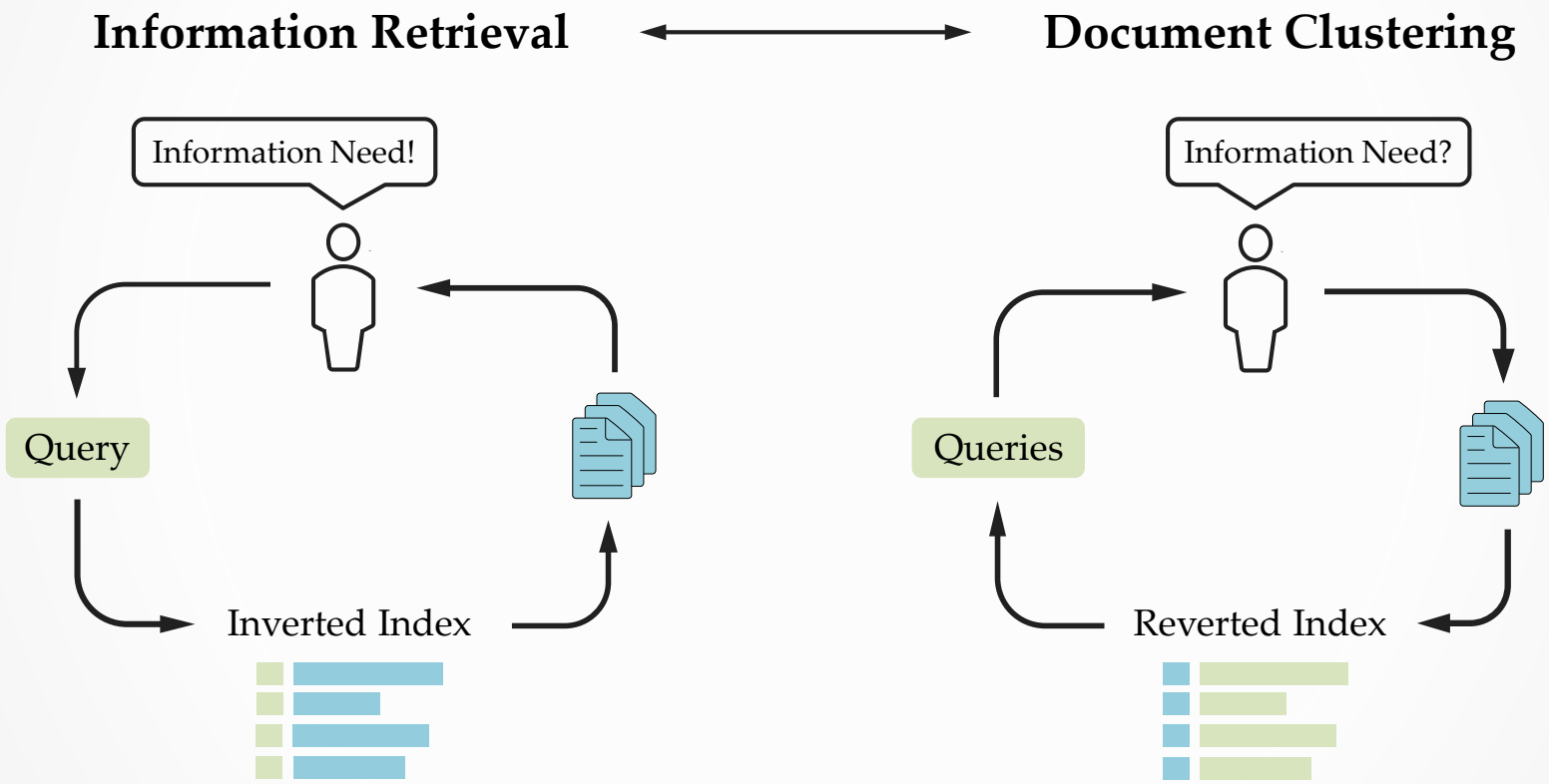


Goals

1. Group similar documents in the same cluster.
2. Label each cluster with a meaningful name.



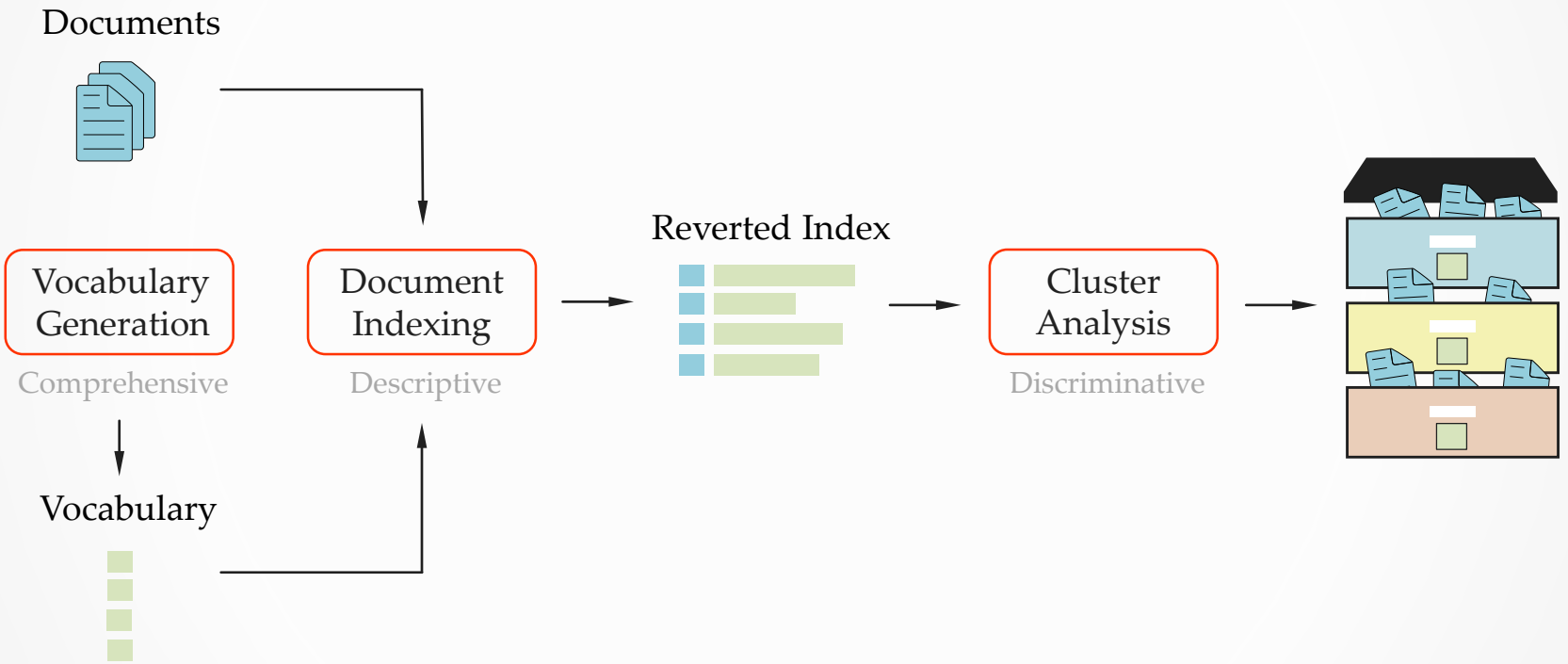
Basic Idea


Information Retrieval
Document Clustering
Information Need!
Query
Inverted Index
Information Need?
Queries
Reverted Index

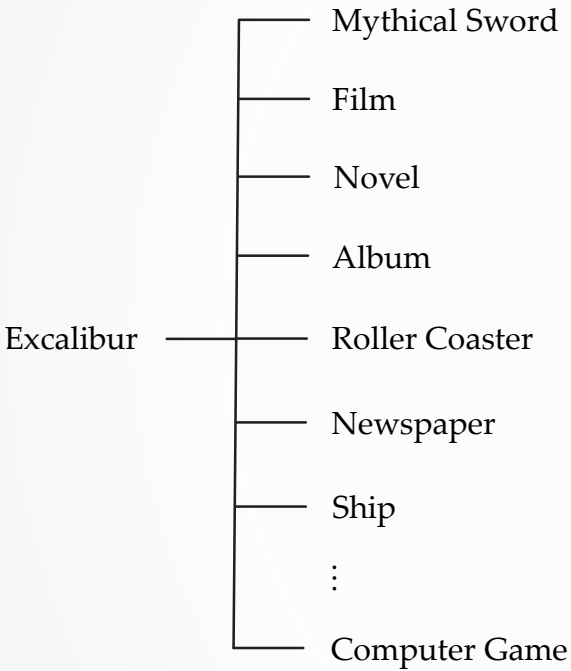
Characteristics of Cluster Labels [Stein and Meyer zu Eißen, 2004]

Comprehensive	The syntax of the label should be appropriate.
Descriptive	The label should speak for each document in a cluster.
Discriminative	The semantic overlap between two cluster labels should be minimal.

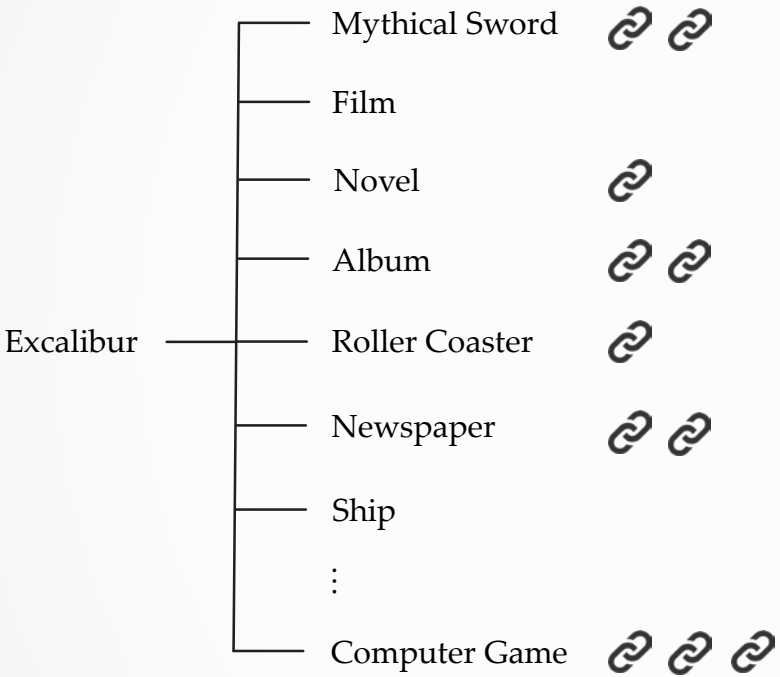
Processing Pipeline



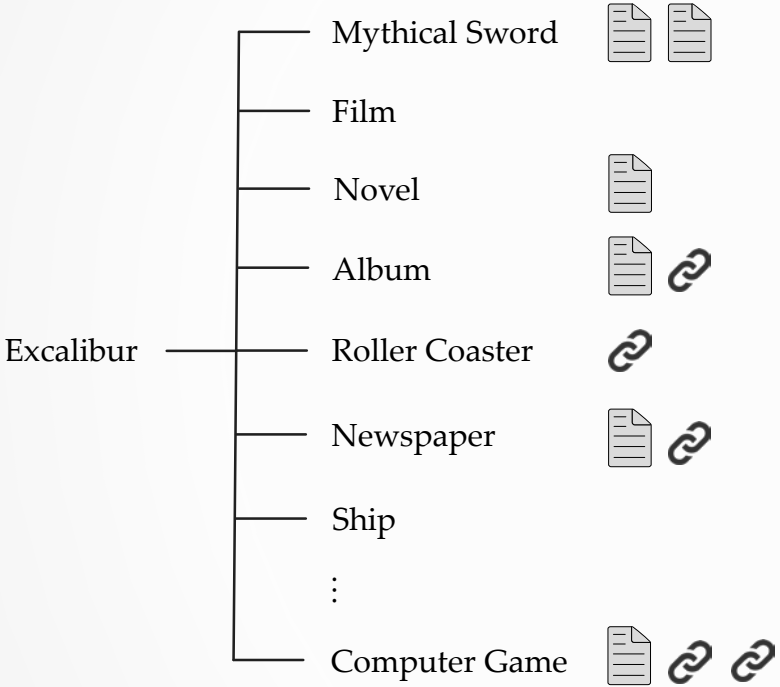
Ambient++ Data Set



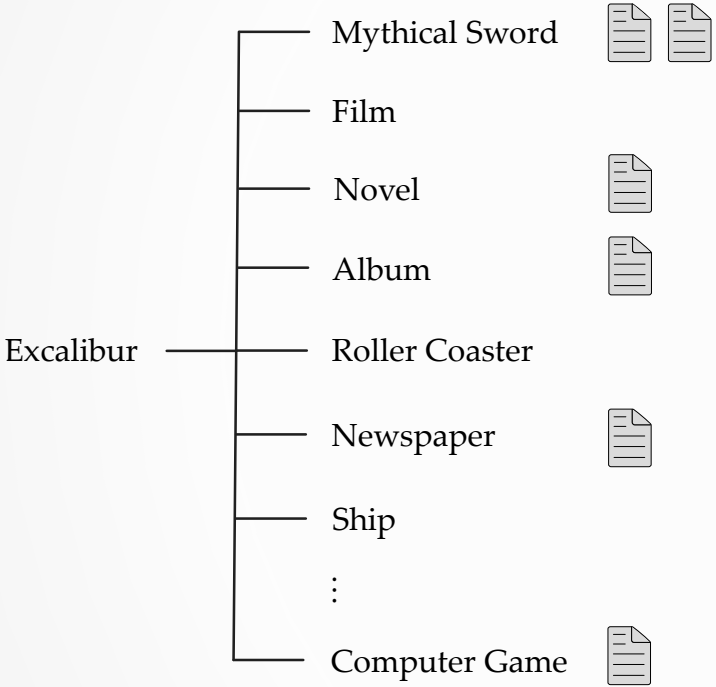
Ambient++ Data Set



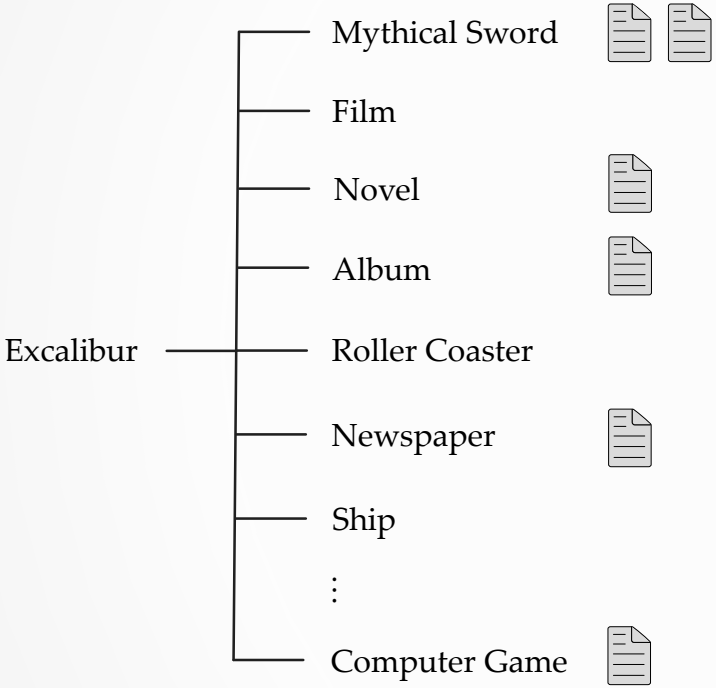
Ambient++ Data Set

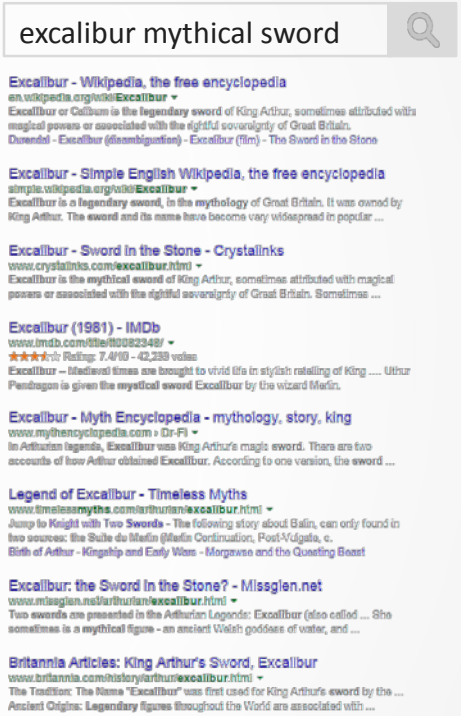


Ambient++ Data Set

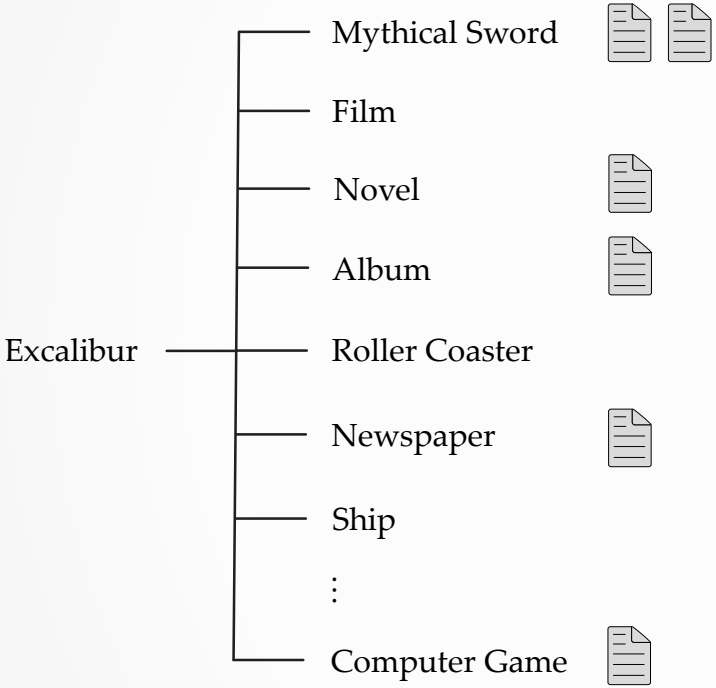


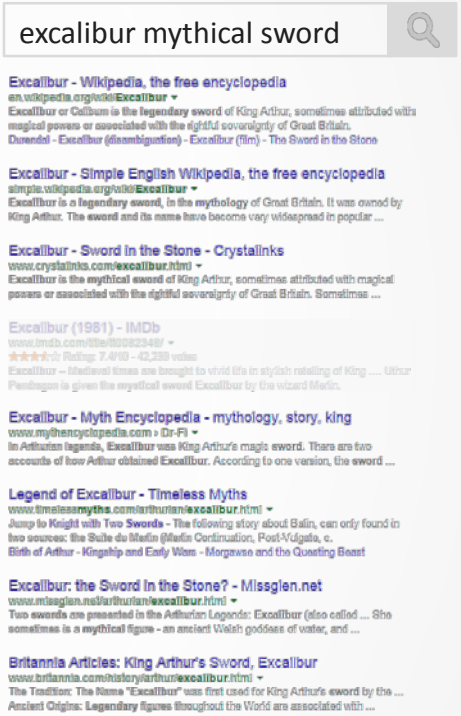
Ambient++ Data Set



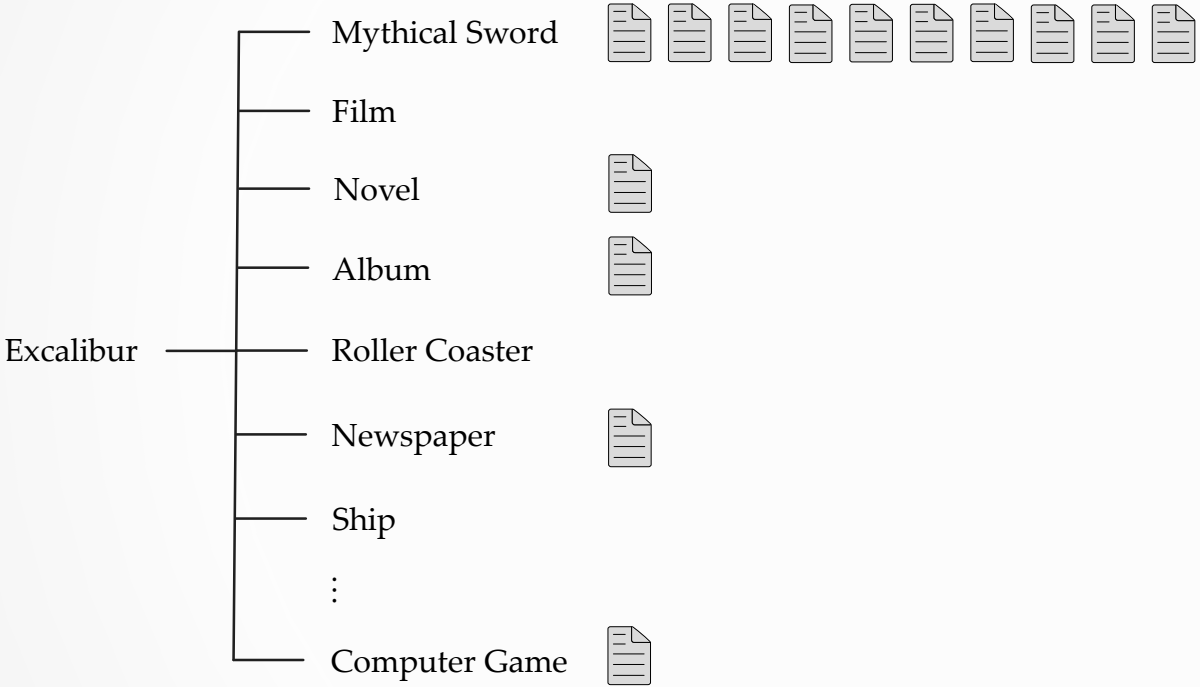


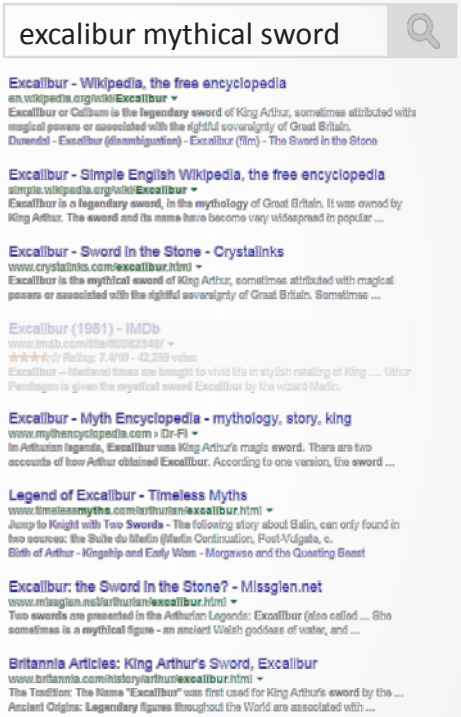
Ambient++ Data Set



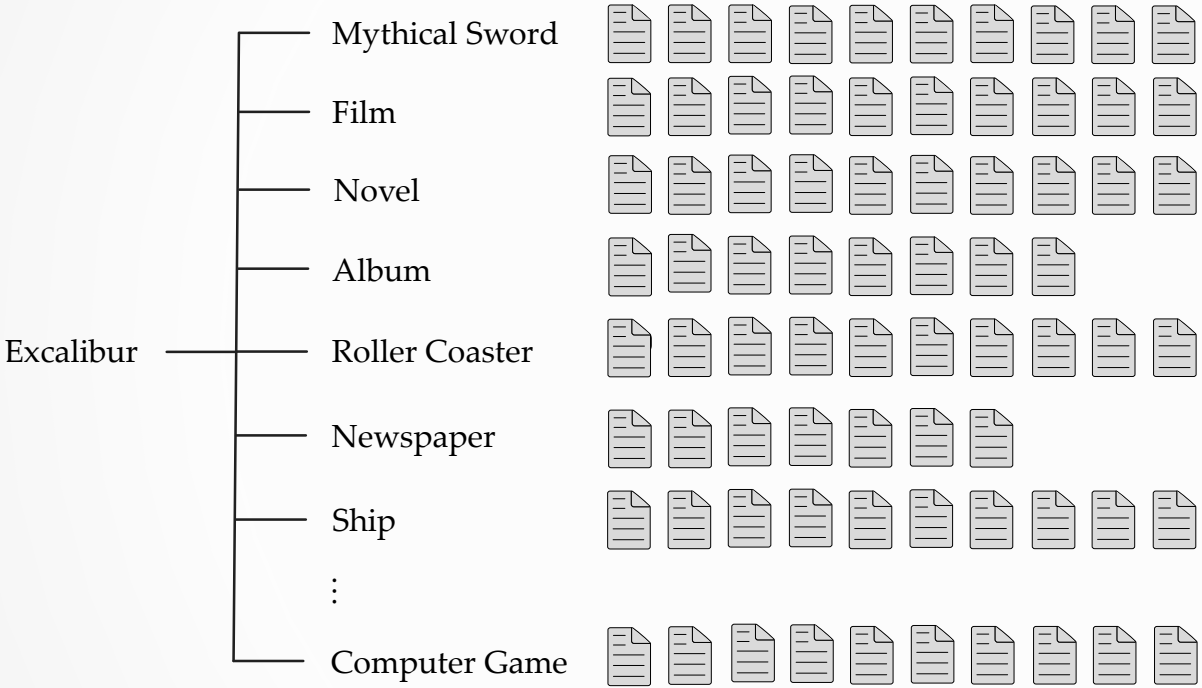


Ambient++ Data Set

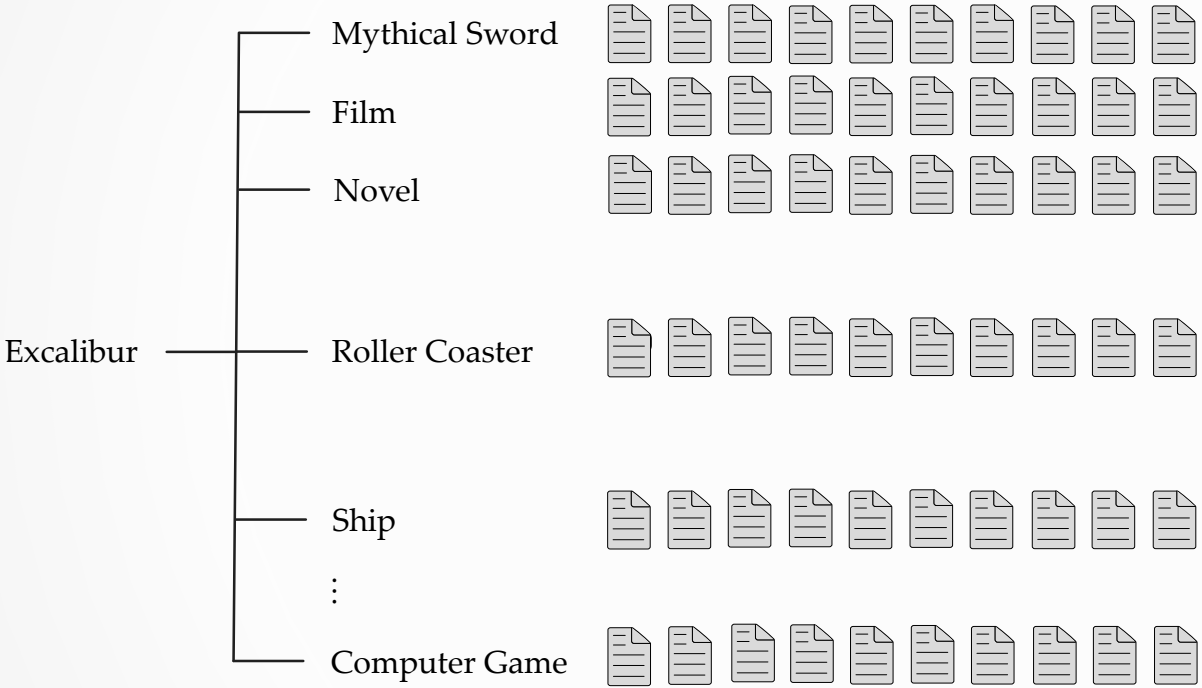




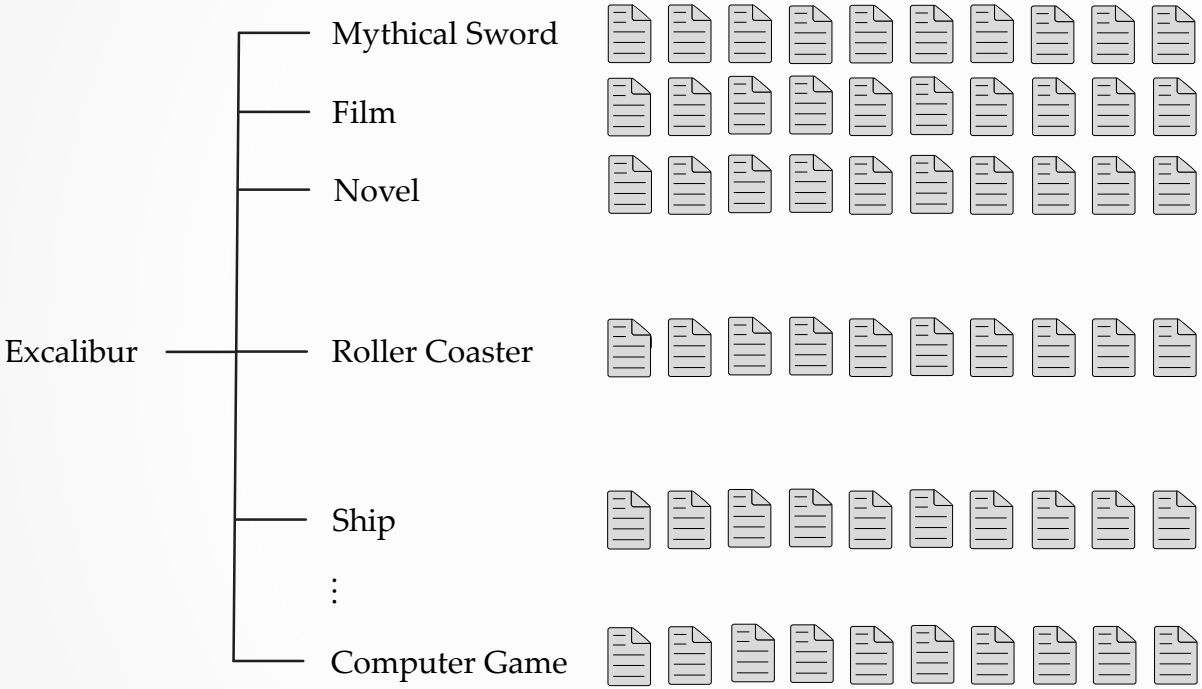
Ambient++ Data Set



Ambient++ Data Set



Ambient++ Data Set



44 Topics

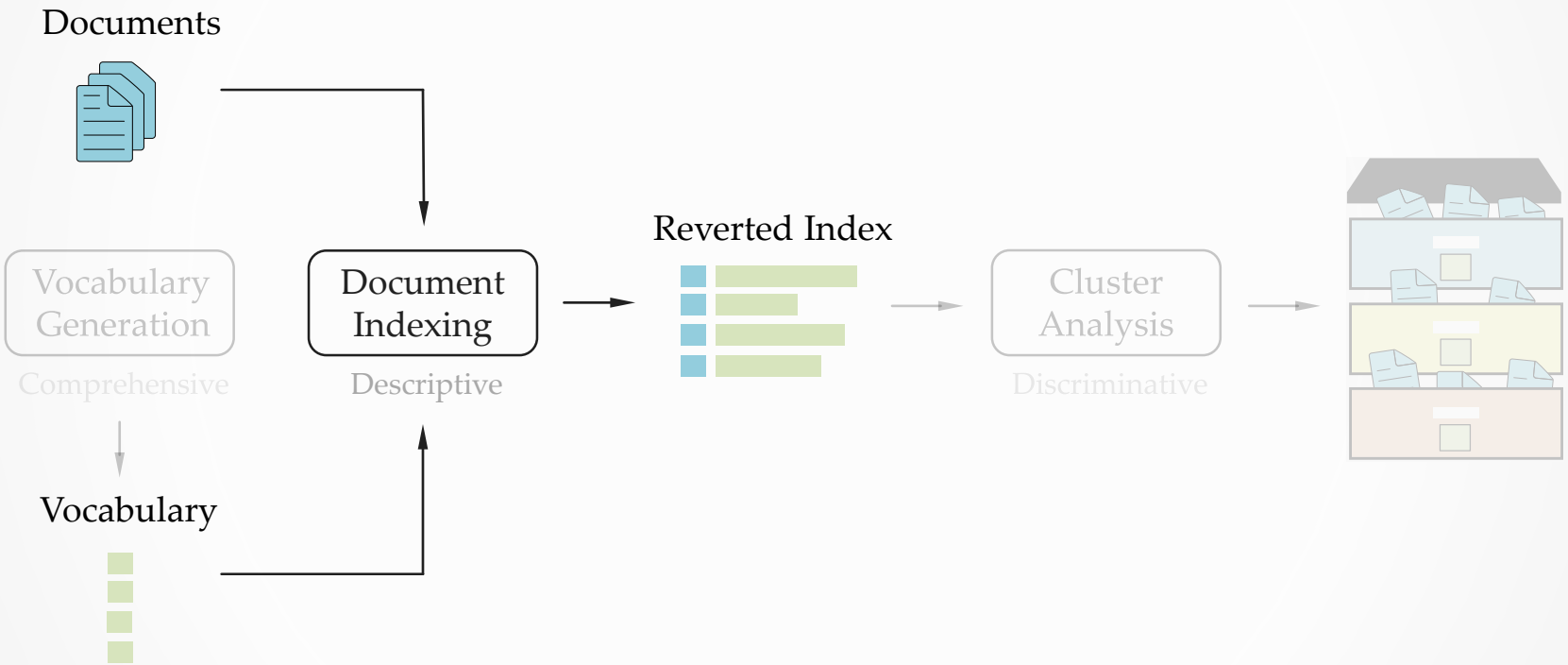
468 Subtopics

4 680 Documents

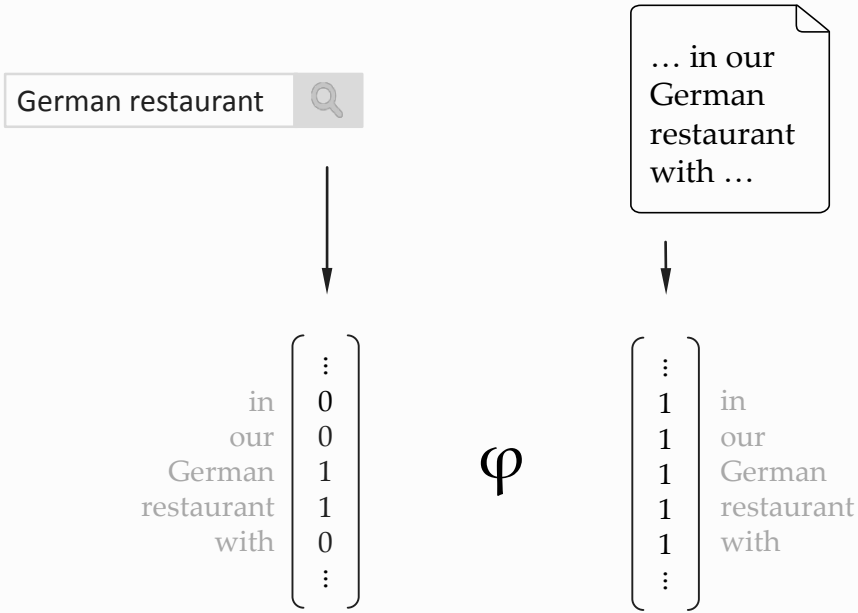
Vocabulary Generation

Vocabulary Generation

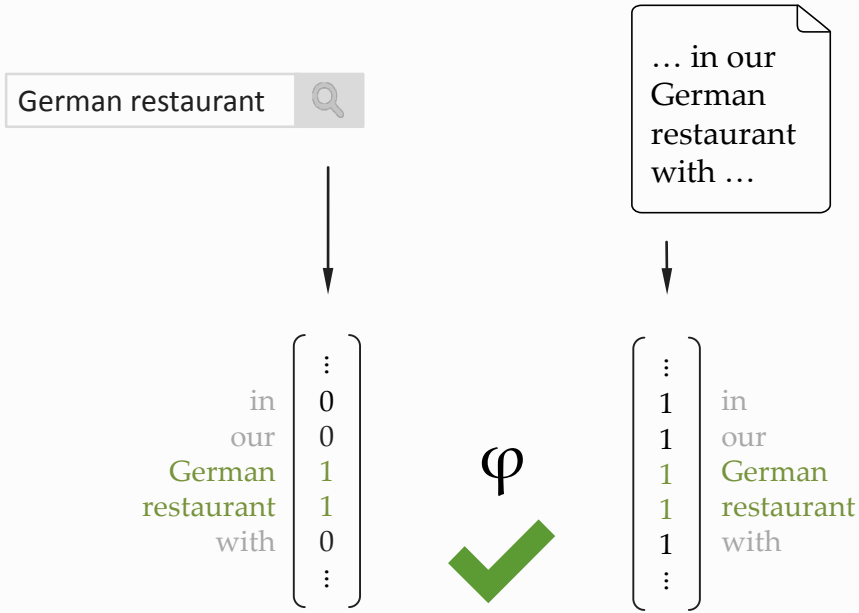
Document Indexing



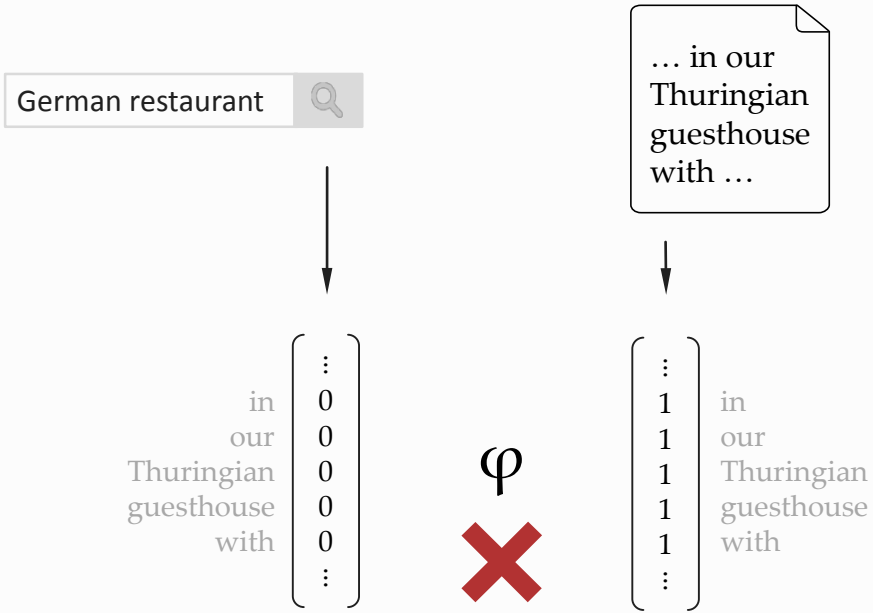
Document Indexing



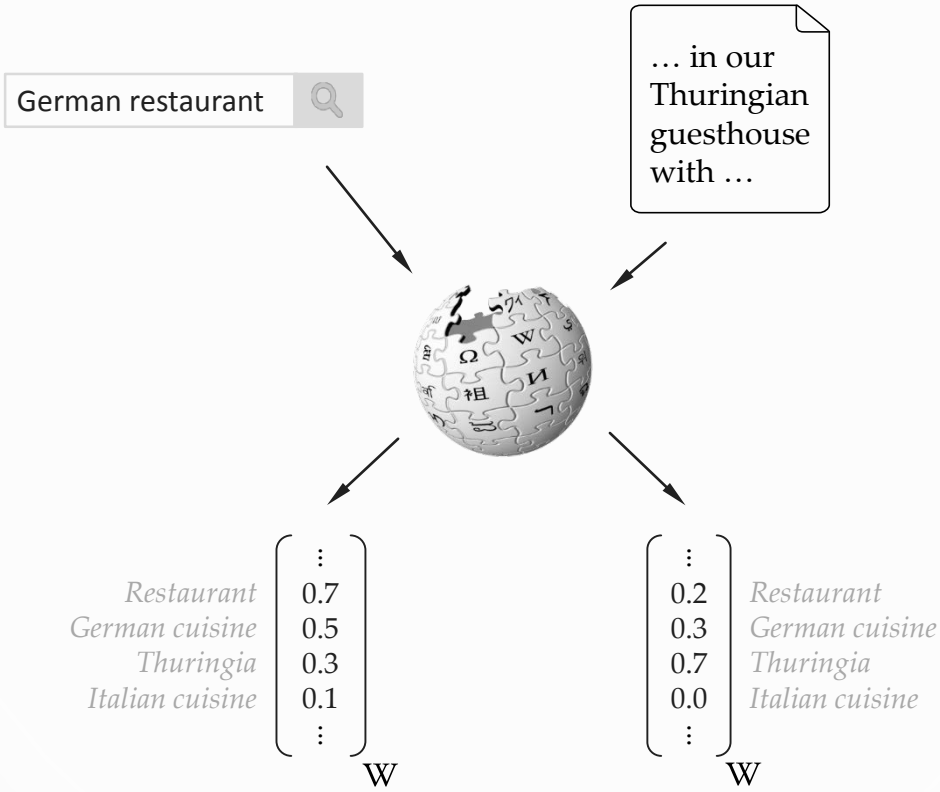
Document Indexing



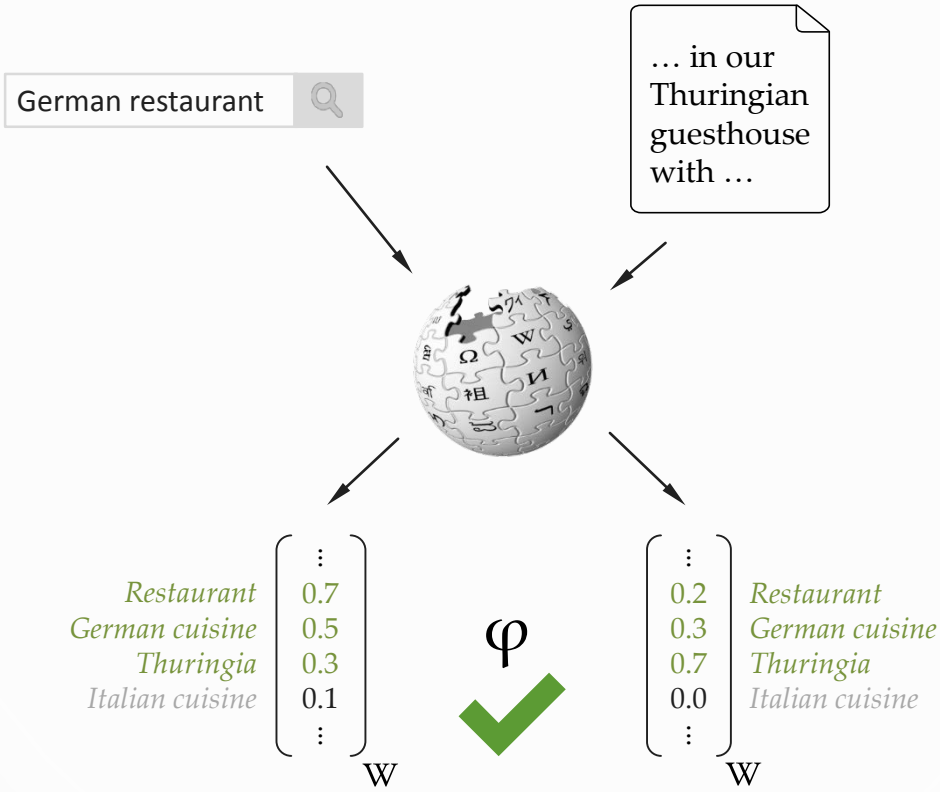
Document Indexing



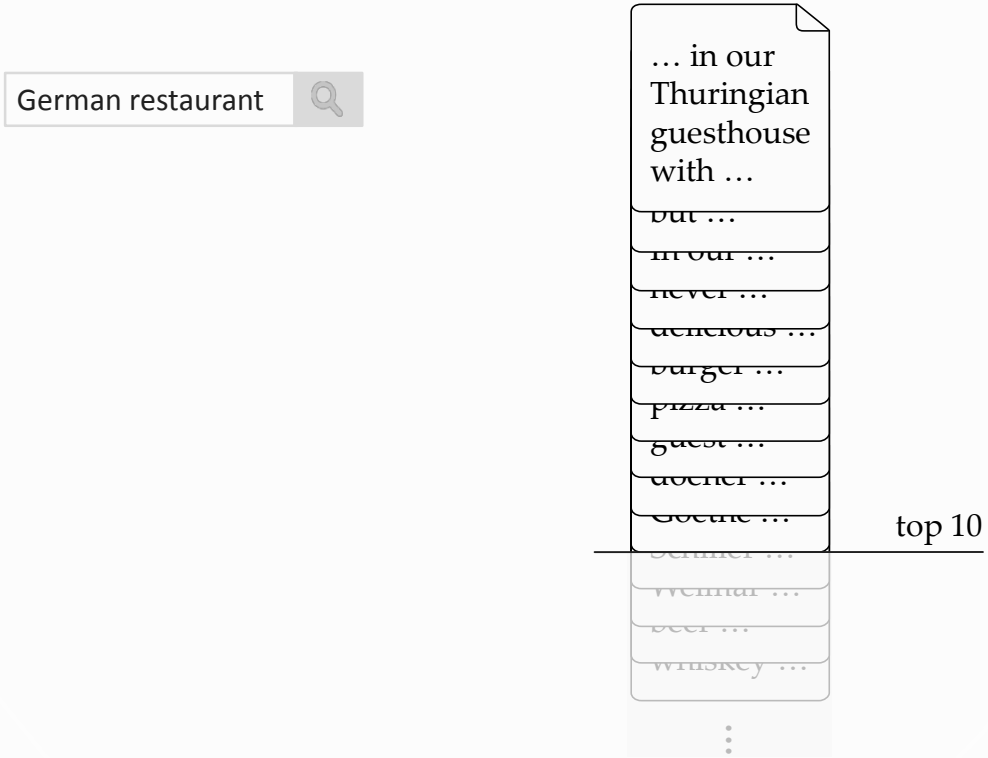
Document Indexing – Explicit Semantic Analysis (ESA)



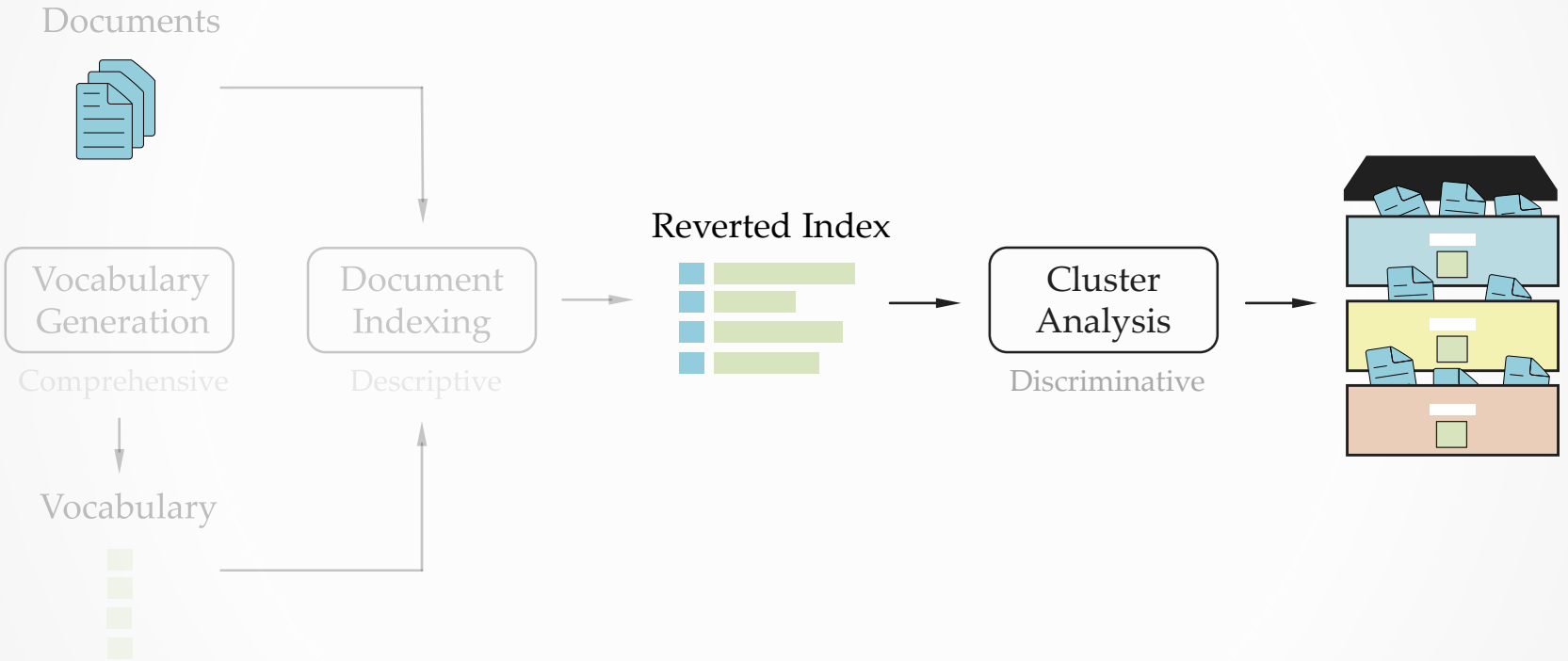
Document Indexing – Explicit Semantic Analysis (ESA)



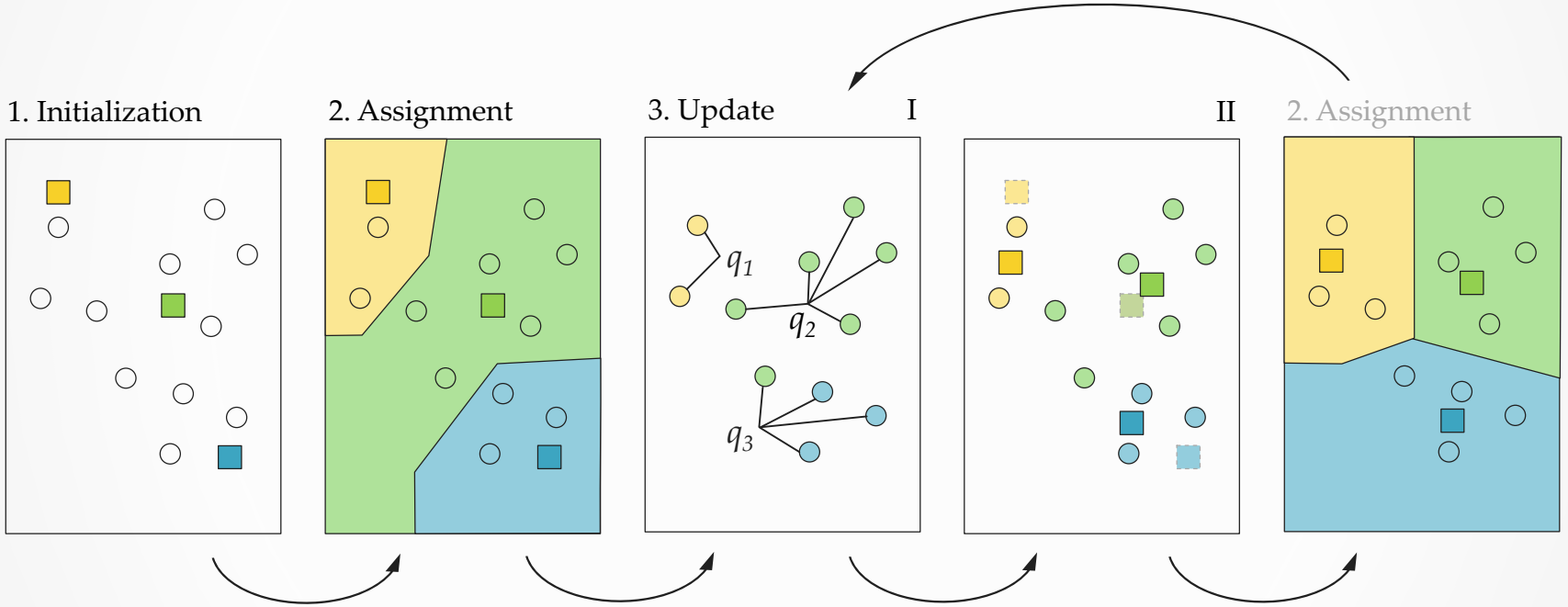
Document Indexing – Relevance Constraint



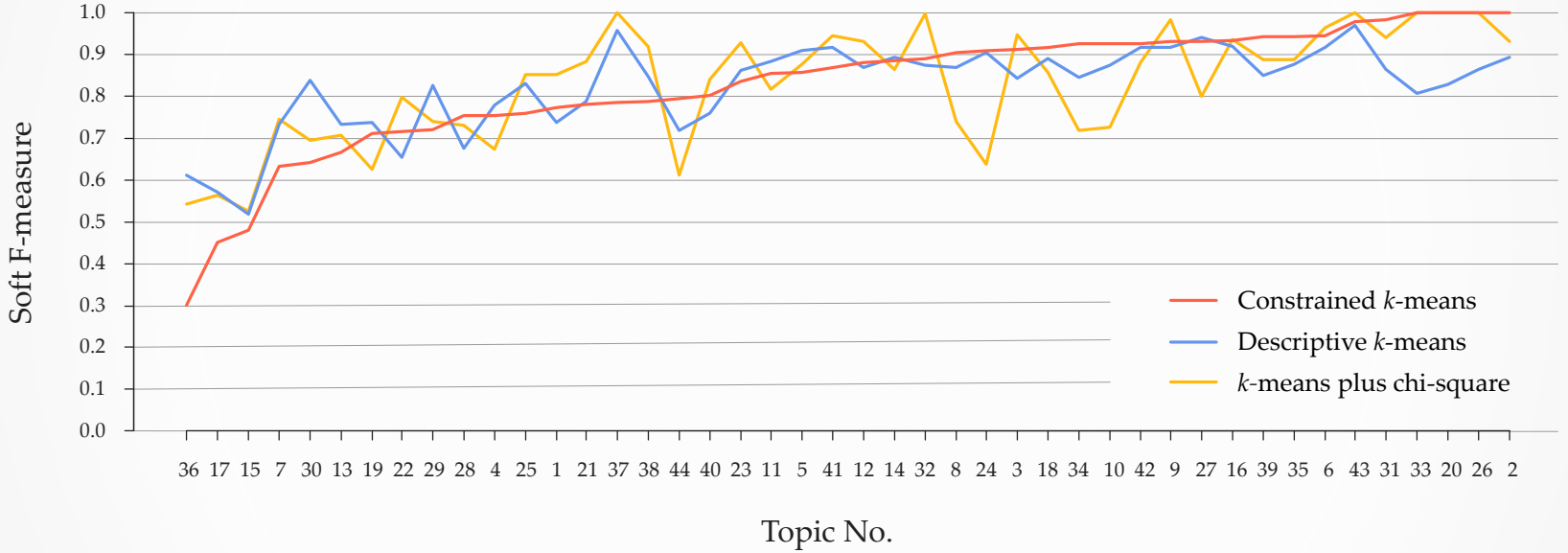
Cluster Analysis



Cluster Analysis – Constrained *k*-means



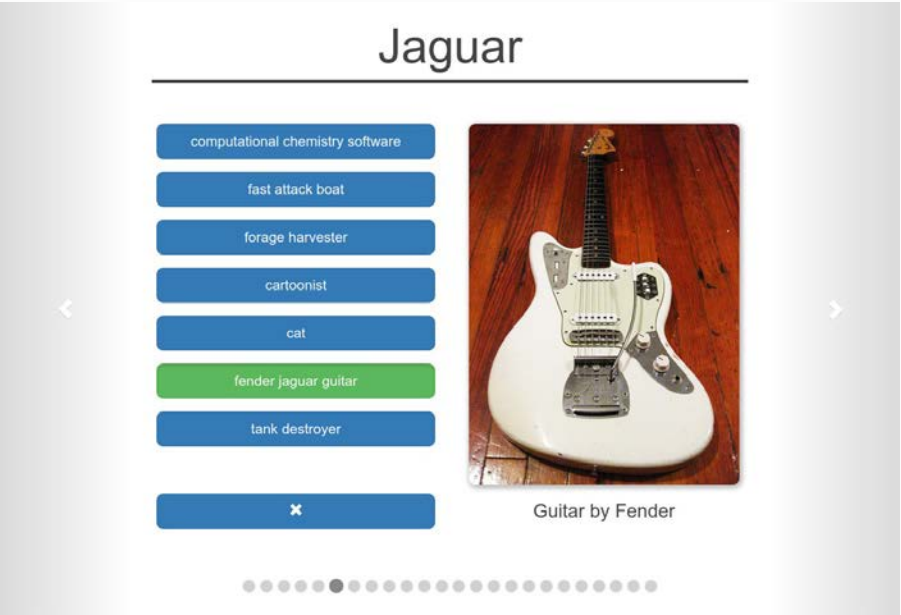
Evaluation of Cluster Documents



Evaluation of Cluster Labels – Discriminative Power

Topic

Cluster Labels

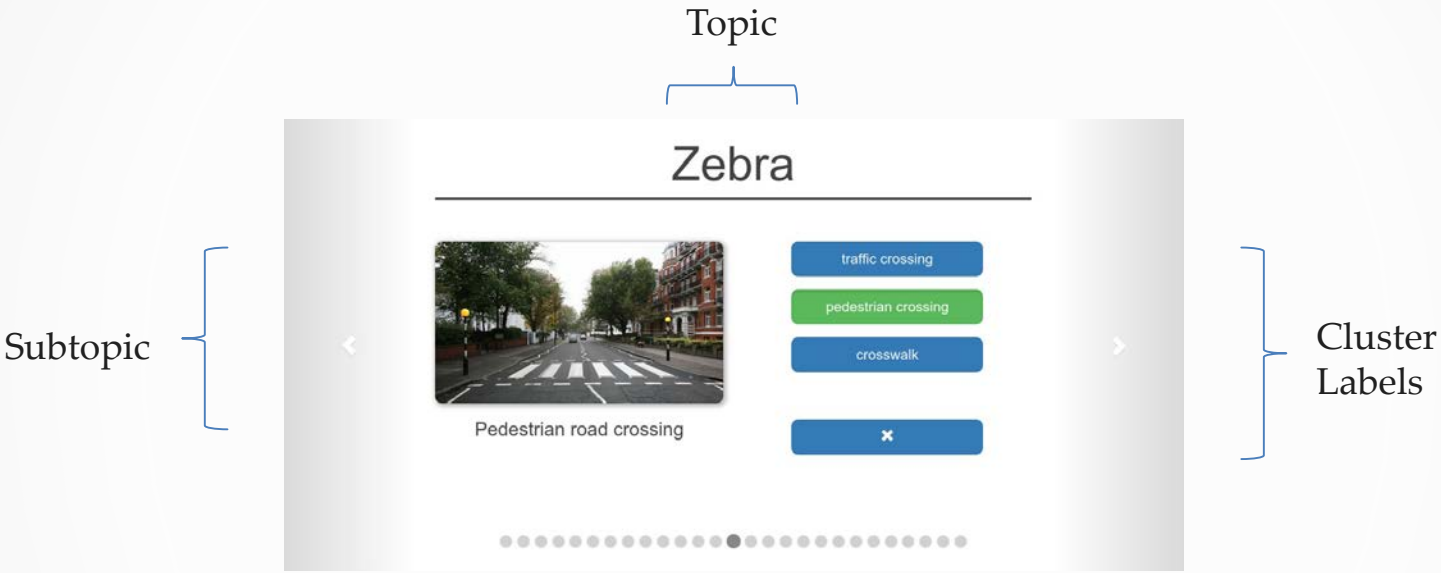


Subtopic

Evaluation of Cluster Labels – Discriminative Power

Judgment	Constrained *k*-means	Descriptive *k*-means	*k*-means + chi-square
✔	**213**	180	152
✘	**15**	25	39
–	**21**	44	58
∑	249	249	249
F-measure	**0.9221**	0.8392	0.7581
p-value	–	0.0049	0.0000

Evaluation of Cluster Labels – Descriptive Power



Voting	Constrained *k*-means	Descriptive *k*-means	*k*-means + chi-square
✔	**299**	274	184
p-value	–	0.3525	0.0000

Summary

- Novel perspective on document clustering
- Flexible processing pipeline
- Query-based evaluation measures
- User Study
- Ambient++ data set

Future work

- Predefined taxonomy as vocabulary
- Hierarchical clustering
- Other evaluation data sets

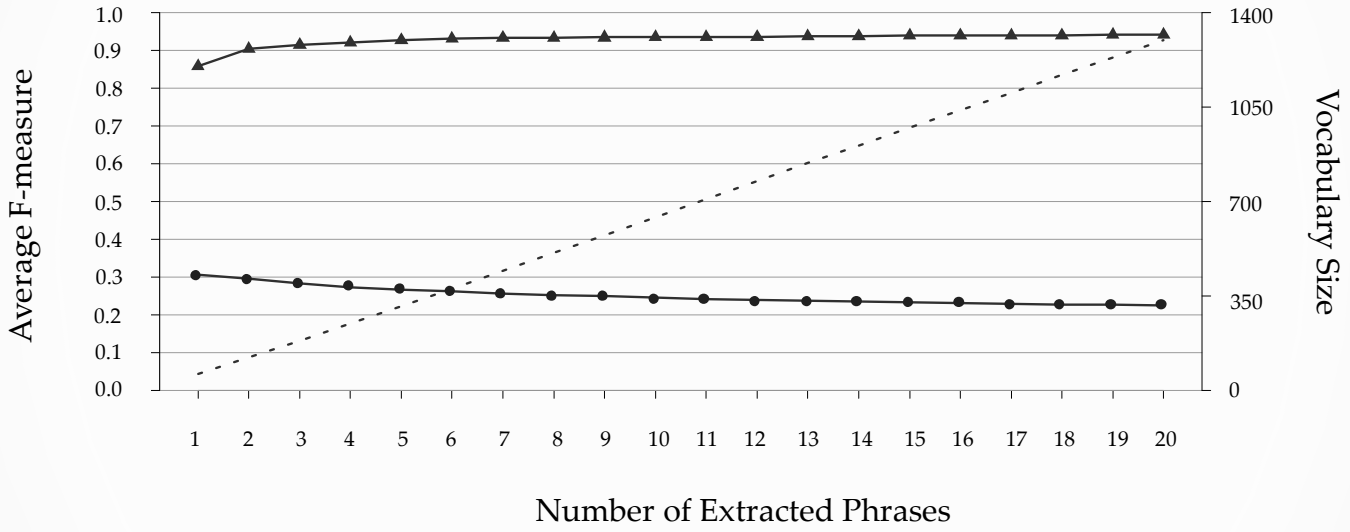
Thank you.

Literature

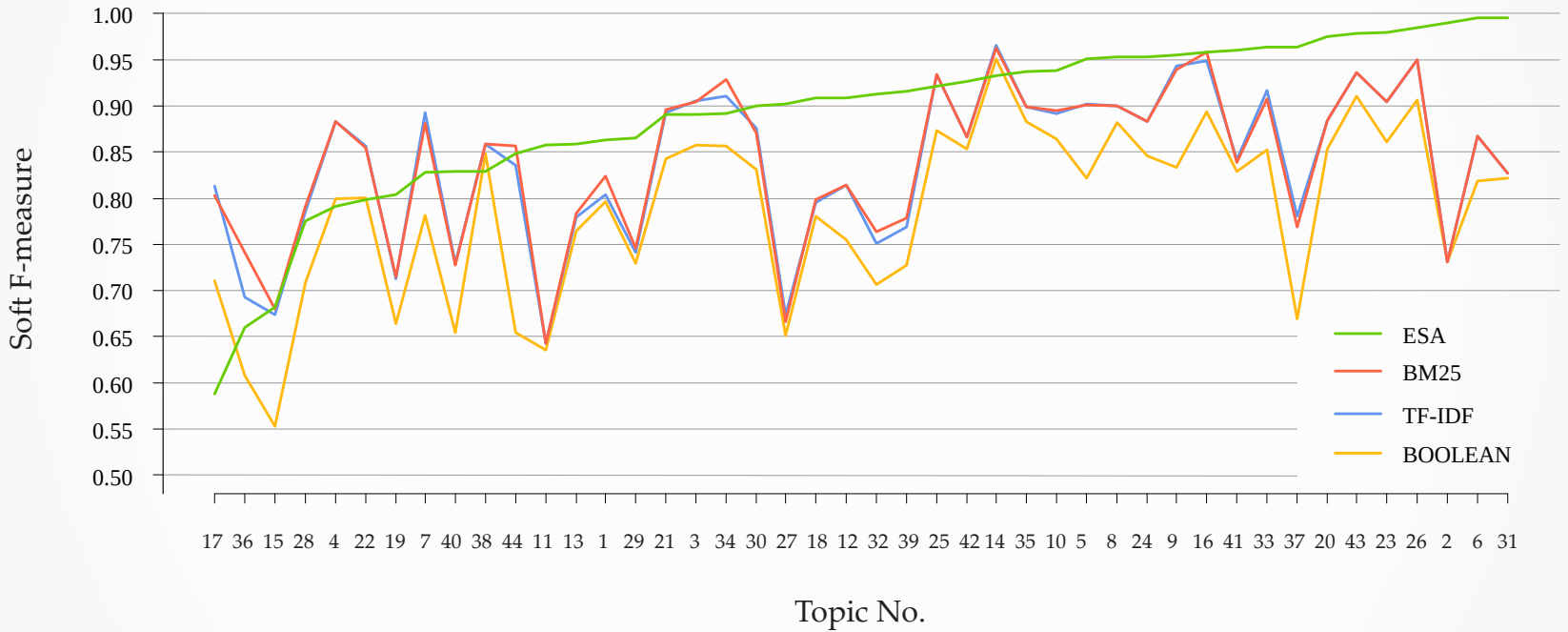
[Stein and Meyer zu Eißen, 2004] Benno Stein and Sven Meyer zu Eißen. Topic identification: Framework and application. In *Proceedings of the International Conference on Knowledge Management*, volume 400 of I-KNOW '14, pages 522–531, 2004.

[Barker and Cornacchia, 2000] Ken Barker and Nadia Cornacchia. Using noun phrase heads to extract document keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence* , AI '00, pages 40–52, London, UK, 2000. Springer.

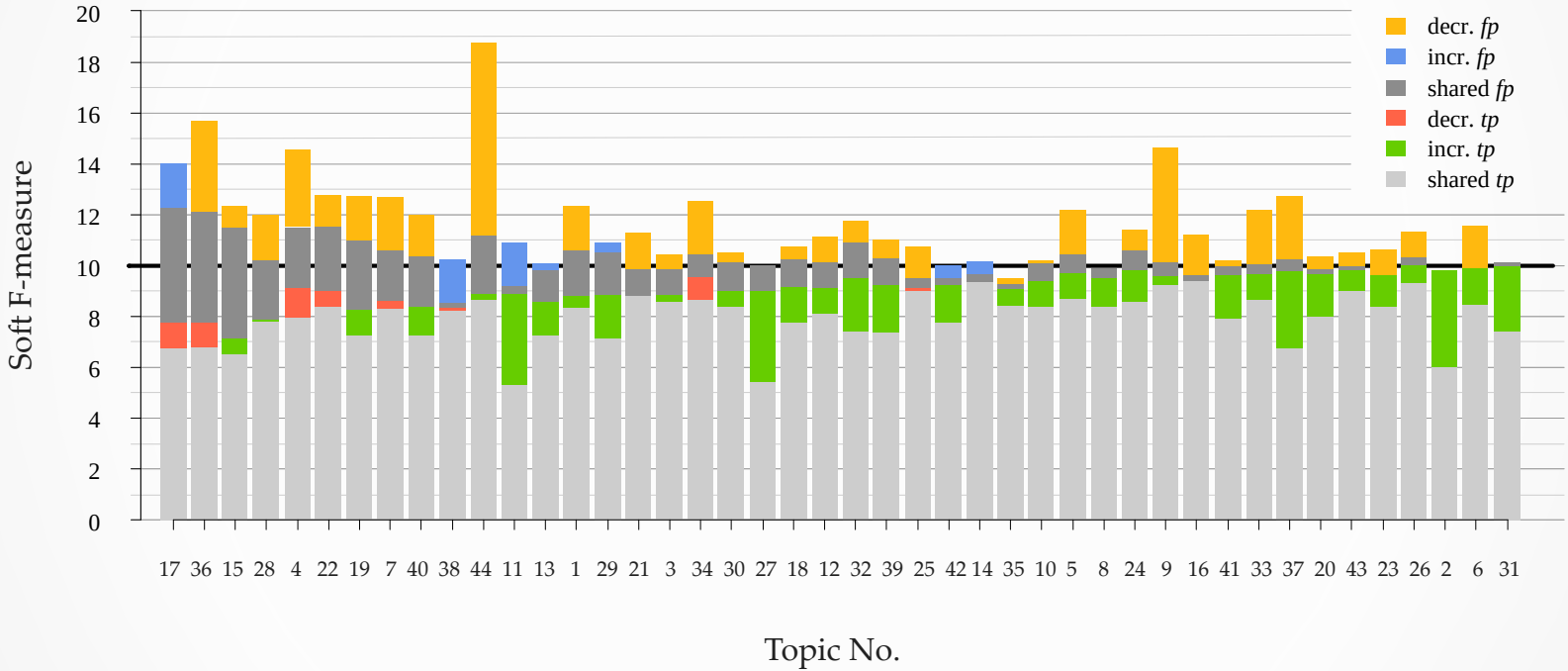
Vocabulary Generation – Number of Extracted Noun Phrases



Document Indexing – Evaluation of Retrieval Models



Document Indexing – Result lists from Boolean to ESA model



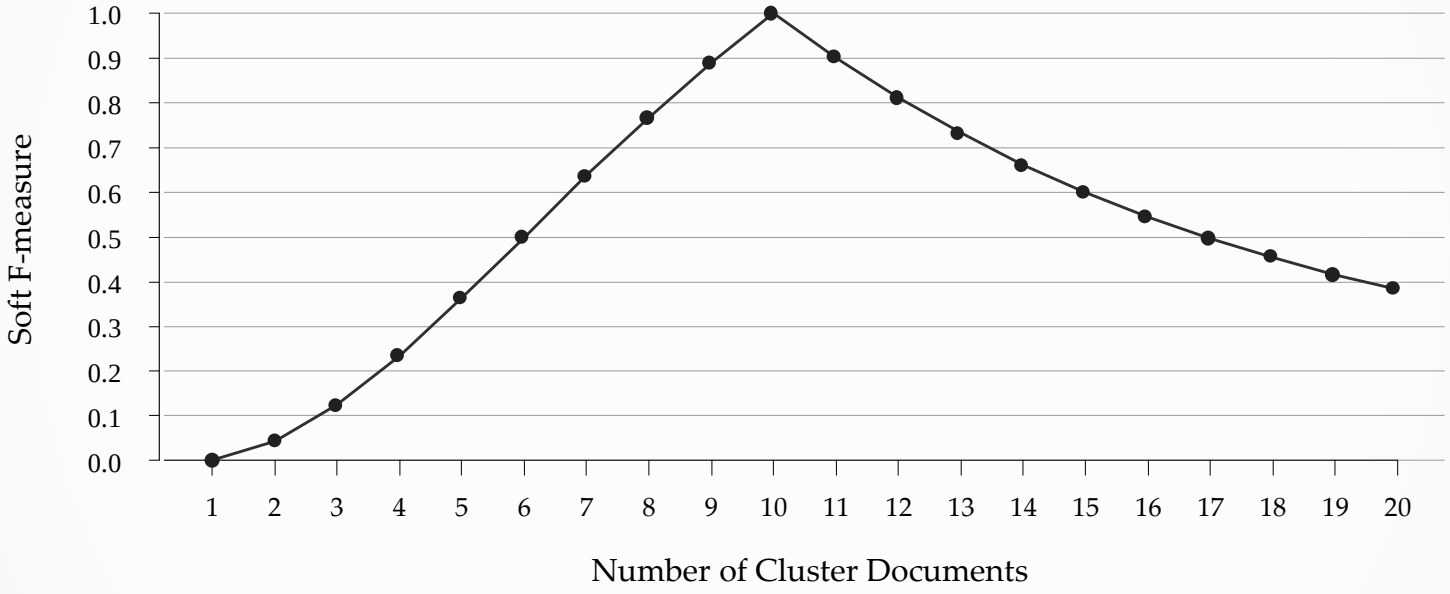
Soft F-measure

Cluster	Documents			
	a_1	a_2	b_1	b_2
A	1	1	0	0
B	0	1	1	1

		Truth	
		True	False
Clustering	True	$tp = \frac{1}{2} + 1 = 1.5$	$fp = 0 + 0 + \frac{1}{2} + \frac{1}{2} = 1$
	False	$tp = \frac{1}{2} + 0 = 0.5$	$tn = 1 + 1 + \frac{1}{2} + \frac{1}{2} = 3$
		$a_1a_2 + b_1b_2$	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_2$

$$\longrightarrow \quad F_1 = 2 \cdot \frac{0.6 \cdot 0.75}{(1 \cdot 0.6) + 0.75} = 0.\bar{6}$$

Justification of Soft F-measure



Cluster Analysis – Evaluation of Clustering Algorithms