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

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...

bauhaus-universität weimar

C

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...

jaguar



Jaguar

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

Jaguar - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Jaguar - Wikipedia -The Jaguar (/'dʒægju:er, 'dʒægju:or, 'dʒægju:or/ or /'dʒægwor/; Brazilian Portuguese: [ʒɐˈɡwaʁ], Spanish: [xaˈywar]), 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
Vikipedia
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 ...

jaguar



Jaguar

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

Jaguar - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Jaguar Wikipedia The Jaguar (/'dʒægju:er, 'dʒægju:or/ or /'dʒægwor/; Brazilian Portuguese: [ʒɛˈɡwaʁ], Spanish: [xaˈywar]), 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 ...

jaguar



Jaguar

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

Jaguar - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Jaguar Wikipedia The Jaguar (/'dʒægju:er, 'dʒægju:or/ or /'dʒægwor/; Brazilian Portuguese: [ʒɛˈɡwaʁ], Spanish: [xaˈywar]), 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
Vikipedia
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 ...

jaguar



Jaguar

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

Jaguar - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Jaguar Wikipedia The Jaguar (/'dʒægju:er, 'dʒægju:or/ or /'dʒægwor/; Brazilian Portuguese: [ʒɛˈɡwaʁ], Spanish: [xaˈywar]), 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
Vikipedia
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 ...



www.spgrestaurantsandbars.com Google+ page Markt 19 Weimar 04936 438020





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









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





Basic Idea



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.













excalibur mythical sword
Excalibur - Wikipedia, the free encyclopedia en wikipedia.org/wiki/excalibur - Excalibur or Galiburn is the begendary sword of King Arthur, somatimes attributed with megical powers are associated with the eight of source/grity of Groat Britain. Durendal - Excalibur (disambiguation) - Excalibur (film) - The Sword in the Stone
Excellbur - Simple English Wikipedia, the free encyclopedia atmpts withpaths.org/shifeExcellbur + Excellbur is a legendary several, in the mythology of Graft Erlain. It was owned by King Athur. The event and its name have become vary widespread in popular
Excellbur - Sword in the Stone - Crystalinks www.crystalinks.com/excellbur.html - Excellbure is the empirical examples of 610p.Artice, constitues attributed with magical poses or associated with the signified exversionly of Grast Brisin. Sometimes
Excellbur (1981) - IMDb www.imdb.com/dtle/M0082348/ ← ★★★★/ Vi Rading: 7.4/10-42,383 volae Excellbur - Undersoft times are abought to virid tife in stylish relating of King Uthur Pendagon is given the wystload swared Excellbur by the wizard Marin.
Excelibur - Myth Encyclopedia - mythology, story, king www.mythencyclopedia.com - Dr-R - In Arthurian legenda, Excelibur was flog Articr's magis eword. There are two accounts of how Arthur obtained Excelibur. According to one vanion, the sword
Legend of Excellbur - Timeless Myths www.timelessmyths.com/striturian/excellbur.html - Jamp to Knight with Two Swends - The following story stord Salin, can only found in two sevences: The Saline du Mathin (Balafin Continuation, Post-Vulgata, c. Binth of Asthur - Kingship and Emity Wass - Margaruna and the Quaeting Beast
Excellbur: the Sword in the Stone? - Missglen.net www.misglen.net/arthura/excellbur/rhm - Two ewords are presented in the Adhurin Logonda: Excellbur (date celled She sometimes is a negatical figure - an ancient Weich poddess of water, and
Britannia Articles: King Arthur's Sword, Excalibur www.britannia.com/httpl://warthur/excalibur.html * The Tealition: The Name "Excalibur" was their used for King Arthur's eword by the Antaend Odjins: Legendary Signers Houghout the World are associated with



excalibur mythical s	word	C.
Excalibur - Wikipedia, the free encycloped	ia	
en.wikipedia.org/wiki/Excalibur +		
Excallbur or Calibum is the legendary sword of King Art	hur, sometimes attribute	d with
magical powers or associated with the rightful soversignty	of Great Britain.	
Durendal - Excalibur (disambiguation) - Excalibur (film) - 1	he Sword in the Stone	
Excalibur - Simple English Wikipedia, the t	free encyclopedia	
simple vikinedia arakulki/Excelibur +	nee enegeropeara	
Excallburis a legendary sword, in the mythology of Gr	eat Britain. It was owned	6y
King Arthur. The sword and its name have become vary v	widespread in popular	
Excelling - Sword in the Stone - Causialian		
Excelled - Sword III the Storie - Crystellin	12	
Excelliburis the muthical sword of King Arbur, sometim	es attributed with made	al I
opaans or seepsisted with the sightful several or to f Grast	Britain, Sometimes	
Excalibur (1981) - IMDb		
www.imdb.com/lite/il0082348/ -		
******** Reling: 7.4/10 - 42,239 value		
Excallbur – bledieval times are brought to vivid life in sty		
Pendragon is given the mystical eword Excalibur by the		
Eventibur - Myth Encyclonedia - mythology	r stanr kina	
vow m/hencyclopedia.com - Dr.Fl +	te erenye rang	
In Arthurian legende, Excellibur was King Arthur's magic e	sword. There are two	
accounts of how Arthur obtained Excatibur. According to	one version, the sword .	
Length of Freedbarr, Theology Mailer		
Legend of Excalibur - Timeless Myths		
www.unelessmyths.cumournulaurexcalipur.nunu •	art Calin ann anly formd	See.
ban namana: the Suite do Marin Marin Continuation. Ros	succeanit, can only tourio	113
Birth of Arthur - Kingship and Early Wass - Morgawso and	the Questing Beast	
som o a com a confincts ma conta aano a mufarroo ma	the quantity trainer	
Excalibur: the Sword in the Stone? - Misso	tien.net	
www.misegien.nel/arihurian/excalibur.himi -		
Two swords are presented in the Arthurian Legends: Exce	allbur (also called She	5
somelimes is a mythical figure - an ancient Weish godde	as of water, and	
Pritagela Adialas: King Adiauta Curani Eu	antinue	
emainina Anticles: Ming Anthur's SWord, EX	canour	
www.usuanna.cunonssury/annun@ccalibur.nimi ▼ The Tradition: The Name "Excelling" was first used for 6	ing Artikurfe enward in the	
Ancient Origins: Legendary figures throughout the World :	are associated with	
and a subsequent and an and a start of the second s	and an	









44 Topics 468 Subtopics

4 680 Documents

Vocabulary Generation



Vocabulary Generation













Document Indexing – Explicit Semantic Analysis (ESA)



Document Indexing – Explicit Semantic Analysis (ESA)



Document Indexing – Relevance Constraint





Cluster Analysis – Constrained *k***-means**



Evaluation of Cluster Documents



Topic No.

Evaluation of Cluster Labels – Discriminative Power



Evaluation of Cluster Labels – Discriminative Power

Judgment	Constrained <i>k</i> -means	Descriptive <i>k</i> -means	<i>k</i> -means + chi-square
×	213	180	152
×	15	25	39
-	21	44	58
Σ	249	249	249
F-measure	0.9221	0.8392	0.7581
<i>p</i> -value	-	0.0049	0.0000

Evaluation of Cluster Labels – Descriptive Power



Voting	Constrained <i>k</i> -means	Descriptive <i>k</i> -means	<i>k-</i> means + chi-square
~	299	274	184
<i>p</i> -value	-	0.3525	0.0000

Summary

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

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

Thank you.

Literature

[Stein and Meyer zu Eißen, 2004] Benno Stein an Framework and

[Barker and Cornacchia, 2000]

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.

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



Number of Extracted Phrases

Document Indexing – Evaluation of Retrieval Models



Topic No.

Document Indexing – Result lists from Boolean to ESA model



Topic No.

Soft F-measure

Cluster		Documents			
	a_1	<i>a</i> ₂	b_1	b_2	
Α	1	1	0	0	
В	0	1	1	1	

			Truth	
		True	False	
Clustering	True	$tp = \frac{1}{2} + 1 = 1.5$	$fp = 0 + 0 + \frac{1}{2} + \frac{1}{2} = 1$	0.6 • 0.75
Clustering	False	$tp = \frac{1}{2} + 0 = 0.5$	$tn = 1 + 1 + \frac{1}{2} + \frac{1}{2} = 3$	 $F_1 = 2 \cdot \frac{1}{(1 \cdot 0.6) + 0.75} = 0.6$
		$a_1a_2 + b_1b_2$	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_2$	

Justification of Soft F-measure



Number of Cluster Documents

Cluster Analysis – Evaluation of Clustering Algorithms



Topic No.