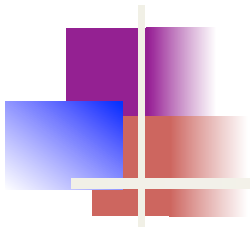


Search Results Clustering without External Resources

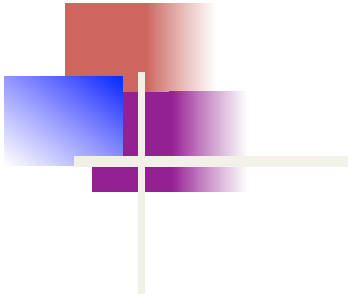
Chris Staff, Joel Azzopardi, Colin Layfield, and Daniel Mercieca

University of Malta

Malta

Search Engine Results

- Search Engine Results usually presented as a list
- For short and/or ambiguous queries, results are often *diversified* to increase chances of user finding relevant results



Jaguar: Luxury Cars & Sports Cars | Jaguar USA

www.**jaguar**usa.com/ ▾ Jaguar Cars ▾

The official home of **Jaguar** USA. Our luxury cars feature innovative designs along with legendary performance to deliver one of the top sports cars in the ...

Jaguar - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/**Jaguar** ▾ Wikipedia ▾

The **jaguar** (/ˈdʒægjuːər, ˈdʒægjʊər, ˈdʒægjuːɑr/ or /ˈdʒægwɑr/; Brazilian Portuguese: [ʒɐˈgwaʁ], Spanish: [xaˈɣwar]), Panthera onca, is a big cat, ...

Jaguar Cars - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾

Jaguar Cars (/ˈdʒægjuːər/ JAG-ew-ər) is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, ...

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 ...

The Jaguar Freeway

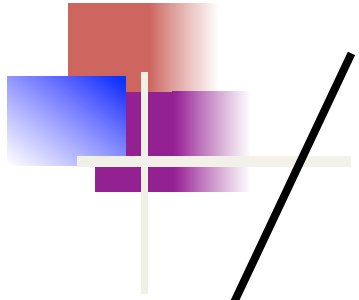
www.smithsonianmag.com/.../the-**jaguar**-freeway-73586097/ - Smithsonian

Half a mile in we see them: two Brazilian biologists and a veterinarian are kneeling in a semicircle, their headlamps spotlighting a tranquilized **jaguar**. It's a young male, about 4 years old: He's not fully ...

Pippa Bartolotti: 'Yes I drive a Jaguar – but why ...

www.independent.co.uk/.../pippa-bartolotti-yes-i-drive-... - The Independent

It may seem unlikely, but a businesswoman who drives a **Jaguar** is one of the four candidates to succeed Caroline Lucas MP as leader of the Green Party.



Jaguar: Luxury Cars & Sports Cars | Jaguar USA
www.jaguarusa.com/ ▾ Jaguar Cars ▾
The official home of **Jaguar** USA. Our luxury cars feature innovative designs along with legendary performance to deliver one of the top sports cars in the ...

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

Jaguar Cars - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾
Jaguar Cars (/ˈdʒægjuːər/ JAG-ew-ər) is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, ...

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 ...

The Jaguar Freeway
www.smithsonianmag.com/.../the-**jaguar**-freeway-73586097/ - Smithsonian
Half a mile in we see them: two Brazilian biologists and a veterinarian are kneeling in a semicircle, their headlamps spotlighting a tranquilized **jaguar**. It's a young male, about 4 years old: He's not fully ...

Pippa Bartolotti: 'Yes I drive a Jaguar – but why ...
www.independent.co.uk/.../pippa-bartolotti-yes-i-drive-... - The Independent
It may seem unlikely, but a businesswoman who drives a **Jaguar** is one of the four candidates to succeed Caroline Lucas MP as leader of the Green Party.

The Problem

- Subjectively relevant results are interspersed among non-relevant results
- May be OK if user looking for just one relevant document;
- Not so great if user aiming for *coverage*, or has difficulty finding the correct search terms

Why can queries be ambiguous?

- The same query term(s) may have multiple meanings – not only due to having different parts-of-speech, but also due to having different *senses* of the word in that specific part-of-speech.

Search Results Clustering

- Partition a results list into a number of clusters so that result snippets related to the same query *sense* are clustered together.

SemEval-2013 Task 11

- Evaluating Word Sense Induction & Disambiguation within an End-User Application:
  - Search Results Clustering (SRC)
- Navigli & Vannella organised the task; prepared a test dataset & gold standard; released an automatic evaluator (Navigli and Vannella, 2013)

SemEval-2013 Dataset

- Dataset available from: https://www.cs.york.ac.uk/semeval-2013/task11/
- Task: Given an ambiguous term (or phrase), and a list of result snippets, create a cluster for each *word sense*

SemEval-2013 Dataset

- 100 ambiguous queries (topics) with 64 results each
  - Topics that have a Wikipedia Disambiguation page
  - Topics submitted as queries to Google search engine

Search Results List Example

Topic/Query 1: polaroid

1.1 http://www.polaroid.com/ Polaroid | Home | 74.208.163.206 Create and share like never before at Polaroid.com. Find instant film and cameras reinvented for the digital age. Plus, digital cameras, digital camcorders, LCD ...

1.2 http://www.polaroid.com/products products | www.polaroid.com Come check out a listing of Polaroid products, by category.

1.3 http://en.wikipedia.org/wiki/Polaroid_Corporation Polaroid Corporation - Wikipedia, the free encyclopedia Polaroid Corporation is an American-based international consumer electronics and eyewear company, originally founded in 1937 by Edwin H. Land. It is most ...

1.4 http://en.wikipedia.org/wiki/Polaroid Polaroid - Wikipedia, the free encyclopedia Jump to: navigation, search. Polaroid may refer to: Polaroid Corporation, a multinational consumer electronics and eyewear company, and former instant camera ...

…

SemEval-2013 Gold Standard

- Each results list is organised into non-overlapping sub-topics (clusters) using Amazon Mechanical Turk to form a Gold Standard

Topics and Sub-topics Example

Topic/Query 1: polaroid

Subtopics:

1.1 Polaroid Corporation, a multinational consumer electronics and eyewear company, and former instant camera and film maker

1.2 Instant film photographs are sometimes known as “Polaroids”, after the company that invented and sold them originally

1.3 Instant camera (or Land camera), sometimes known as a Polaroid camera after the company that invented the concept

1.4 Polaroid Eyewear, eyewear based around glare-reducing polarized lenses made from the aforementioned Polaroid polarizer

1.5 Polaroid (polarizer), a type of synthetic plastic sheet which is used to polarize light, developed by Polaroid Corporation

1.6 Other

1.7 Polaroid (album) Polaroid (album), a bootleg album by American band Phantom Planet

SemEval-2013 Dataset Evaluation

- Navigli and Vannella, 2013 describe and provide automatic WSI-Evaluator that measures F1, Rand Index (RI), Adjusted Rand Index (ARI), and Jaccard Index (JI)
  - Measures of generated cluster quality compared to Gold Standard

WS Induction vs. Disambiguation

- Word Sense *Disambiguation*:
  - Can use a *sense inventory* to solve the problem
- Word Sense *Induction*:
  - No *sense inventory* used
- We do induction

No-K-Means

- Inspired by K-Means (MacQueen, 1967)…
- … but we don't need to know the number of clusters …

No-K-Means

- … without pre-processing external textual resources …
  - UniMelb (HDP-Clusters) performs Hierarchical Dirichlet Processing on English Wikipedia Dump (EWD) (Lau *et al.*, 2013);
  - UKP-WSI derives term co-occurrence statistics from EWD and ukWaC (Zorn and Gurevych, 2013);
  - In one config. Duluth performs LSA on part of the GigaWord collection (Pederson, 2013)

No-K-Means

- … without using external resources
  - UKP-WSI (Zorn and Gurevych, 2013) uses the full-text of documents

No-K-Means

- … without using external resources
  - SenseSearcher/SnS (Kozłowski and Rybiński, 2014) perform Part-of-Speech tagging and Named Entity Recognition

No-K-Means

- We perform unsupervised clustering using the result snippets only, without knowing in advance the number of clusters or query senses present in the results.

No-K-Means

- Two variants: withLSA (Latent Semantic Analysis, Deerwester, *et al.*, 1990) and noLSA
  - For withLSA, the background corpus is the results snippets for the query only
  - Duluth *also* experiment with just the results snippets, but achieve worse results than us.
  - Everyone else who use LSA (or LDA, or HDP) use larger background corpora.

Brief outline of No-K-Means

- Input: Query terms, snippets in results list (RL)
- Output: Result clusters
- Remove stop words; stem terms; remove query term stems from RL
- Create Term-by-Document (Snippet) matrix using term frequency as the term weight
- If withLSA, perform Singular Value Decomposition on matrix

Brief outline of No-K-Means (contd)

- For each snippet: If no existing clusters, create cluster with snippet representation as centroid, else…
- … use Cosine Similarity Measure to find most similar existing cluster; if maximum similarity is below some threshold, create a new cluster for the snippet
- Update cluster centroid (simple average of term weights)

Brief outline of No-K-Means (contd)

- Finally, merge singleton clusters into one 'Others' cluster.

Parameters

- *simThres* – the similarity threshold for a snippet to be added to a cluster
- (*k* – number of dimensions in LSA)

Parameters

- 0.01 <= *simThres* <= 0.9 (in steps of 0.01 to 0.1 and then in steps of 0.1 to 0.9)
- (5 <= *k* <= 60 in steps of 5)

Generalized Dunn's Index

- For each <query, results list> pair generate 18 cluster configurations (noLSA)
- GDI (Bezdek and Pal, 1998) generates a validity index for a cluster configuration.
- Ideally, we want clusters that have centroids that are far from each other, where each cluster has members that are quite similar to the centroid.

GDI_Fixed

- In GDI_Fixed we generate clusters for each query using the same value of *simThres*.
  - We run though all combinations of *simThres*. We use GDI to identify which value of *simThres* yields the most stable clusters on average.
- All clusters for all queries have been generated using the same *simThres*.

GDI_Varied

- For each query, we vary *simThres* using GDI to identify which *simThres* yields the most stable clusters.
  - The *simThres* for each query is independent.

Results:

- Correct results in Errata accompanying paper on the Workshop website.
- However, claims made in the paper are still valid.

Results: Compared to Prior Work

TABLE I. BEST PERFORMERS ON THE SEMEVAL-2013 TASK 11 DATASET

	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
hdp-clusters-lemma	68.30	65.22	21.31	33.02	6.63	11.07
hdp-clusters-nolemma	68.03	64.86	21.49	33.75	6.54	11.68
SnS	**70.16**	**65.84**	**22.19**	**34.26**	8.82	8.46
SemEval singletons	100.00	60.09	0.00	0.00	64.00	1.00
SemEval all-in-one	54.42	39.90	0.00	39.90	1.00	64.00
Gold Standard					**7.69**	**11.56**

TABLE I. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

QT	*SimThres*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	**71.78**	**68.30**	**26.19**	**35.13**	**8.00**	9.49
	GDI_Fixed	**70.19**	**67.77**	**26.40**	**36.26**	**7.14**	9.84
	0.01	64.72	62.06	19.34	**36.82**	5.01	14.72
✓	GDI_Varied	56.69	43.43	3.55	38.49	2.47	27.74
✓	GDI_Fixed	59.77	51.08	8.54	36.33	3.47	22.19
✓	0.01	54.56	40.19	0.13	39.91	1.10	60.80

Results: Keeping Query Terms

TABLE I. BEST PERFORMERS ON THE SEMEVAL-2013 TASK 11 DATASET

	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
hdp-clusters-lemma	68.30	65.22	21.31	33.02	6.63	11.07
hdp-clusters-nolemma	68.03	64.86	21.49	33.75	6.54	11.68
SnS	**70.16**	**65.84**	**22.19**	**34.26**	8.82	8.46
SemEval singletons	100.00	60.09	0.00	0.00	64.00	1.00
SemEval all-in-one	54.42	39.90	0.00	39.90	1.00	64.00
Gold Standard					**7.69**	**11.56**

TABLE I. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

QT	*SimThres*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	**71.78**	**68.30**	**26.19**	**35.13**	**8.00**	9.49
	GDI_Fixed	**70.19**	**67.77**	**26.40**	**36.26**	**7.14**	9.84
	0.01	64.72	62.06	19.34	**36.82**	5.01	14.72
✓	GDI_Varied	56.69	43.43	3.55	38.49	2.47	27.74
✓	GDI_Fixed	59.77	51.08	8.54	36.33	3.47	22.19
✓	0.01	54.56	40.19	0.13	39.91	1.10	60.80

Results: Omitting Query Terms 1

TABLE I. BEST PERFORMERS ON THE SEMEVAL-2013 TASK 11 DATASET

	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
hdp-clusters-lemma	68.30	65.22	21.31	33.02	6.63	11.07
hdp-clusters-nolemma	68.03	64.86	21.49	33.75	6.54	11.68
SnS	**70.16**	**65.84**	**22.19**	**34.26**	8.82	8.46
SemEval singletons	100.00	60.09	0.00	0.00	64.00	1.00
SemEval all-in-one	54.42	39.90	0.00	39.90	1.00	64.00
Gold Standard					**7.69**	**11.56**

TABLE I. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

QT	*SimThres*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	**71.78**	**68.30**	**26.19**	**35.13**	**8.00**	9.49
	GDI_Fixed	**70.19**	**67.77**	**26.40**	**36.26**	**7.14**	9.84
	0.01	64.72	62.06	19.34	**36.82**	5.01	14.72
✓	GDI_Varied	56.69	43.43	3.55	38.49	2.47	27.74
✓	GDI_Fixed	59.77	51.08	8.54	36.33	3.47	22.19
✓	0.01	54.56	40.19	0.13	39.91	1.10	60.80

Results: Omitting Query Terms 2

TABLE I. BEST PERFORMERS ON THE SEMEVAL-2013 TASK 11 DATASET

	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
hdp-clusters-lemma	68.30	65.22	21.31	33.02	6.63	11.07
hdp-clusters-nolemma	68.03	64.86	21.49	33.75	6.54	11.68
SnS	**70.16**	**65.84**	**22.19**	**34.26**	8.82	8.46
SemEval singletons	100.00	60.09	0.00	0.00	64.00	1.00
SemEval all-in-one	54.42	39.90	0.00	39.90	1.00	64.00
Gold Standard					**7.69**	**11.56**

TABLE I. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

QT	*SimThres*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	**71.78**	**68.30**	**26.19**	**35.13**	**8.00**	9.49
	GDI_Fixed	**70.19**	**67.77**	**26.40**	**36.26**	**7.14**	9.84
	0.01	64.72	62.06	19.34	**36.82**	5.01	14.72
✓	GDI_Varied	56.69	43.43	3.55	38.49	2.47	27.74
✓	GDI_Fixed	59.77	51.08	8.54	36.33	3.47	22.19
✓	0.01	54.56	40.19	0.13	39.91	1.10	60.80

Results: Omitting Query Terms 3

TABLE II. OUR RESULTS WITH NO-K-MEANS - WITHLSA, WITHOUT AND WITH QUERY TERMS (QT).

QT	*Mode*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	64.33	60.63	19.71	38.57	3.90	21.07
	GDI_Fixed	67.92	62.74	18.02	27.44	5.69	11.41
✓	GDI_Varied	54.91	40.68	0.50	39.73	1.27	55.57
✓	GDI_Fixed	64.81	54.61	8.81	31.13	4.76	14.12

Results: noLSA Beats withLSA

TABLE I. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

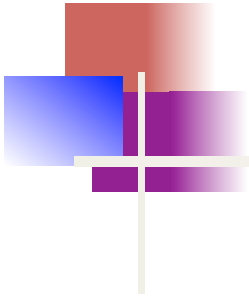
QT	*SimThres*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	**71.78**	**68.30**	**26.19**	**35.13**	**8.00**	9.49
	GDI_Fixed	**70.19**	**67.77**	**26.40**	**36.26**	**7.14**	9.84

TABLE II. OUR RESULTS WITH NO-K-MEANS - WITHLSA, WITHOUT AND WITH QUERY TERMS (QT).

QT	*Mode*	F1	RI	ARI	JI	Ave. # clusters	Ave. clus. size
	GDI_Varied	64.33	60.63	19.71	38.57	3.90	21.07
	GDI_Fixed	67.92	62.74	18.02	27.44	5.69	11.41

Conclusions

- Can cluster accurately, *as long as query terms are removed*.
- No-K-Means is unsupervised; does not need to pre-process huge text repositories; no need for snippet mark-up; no prior knowledge of K.
- Already outperform SOTA, but can we obtain better results if we start with different initial cluster centroids?



Thank you!

Questions?