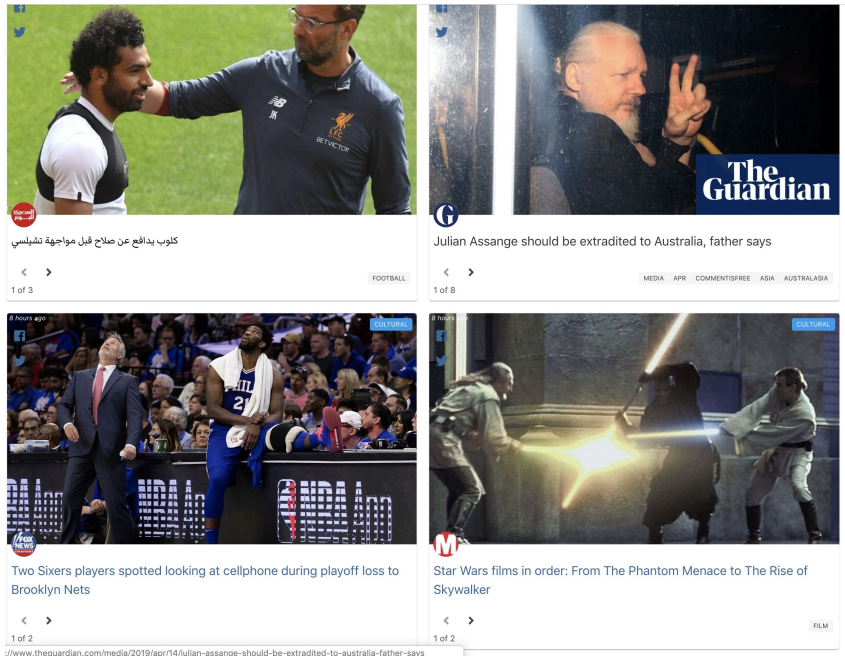


Dense vs. Sparse Representations for News Stream Clustering

Todor Staykovski Alberto Barrón-Cedeño* Giovanni Da San Martino Preslav Nakov

Sofia University
Sofia, Bulgaria

Qatar Computing Research Institute, HBKU
Doha, Qatar
* Università di Bologna, Italy

Overview

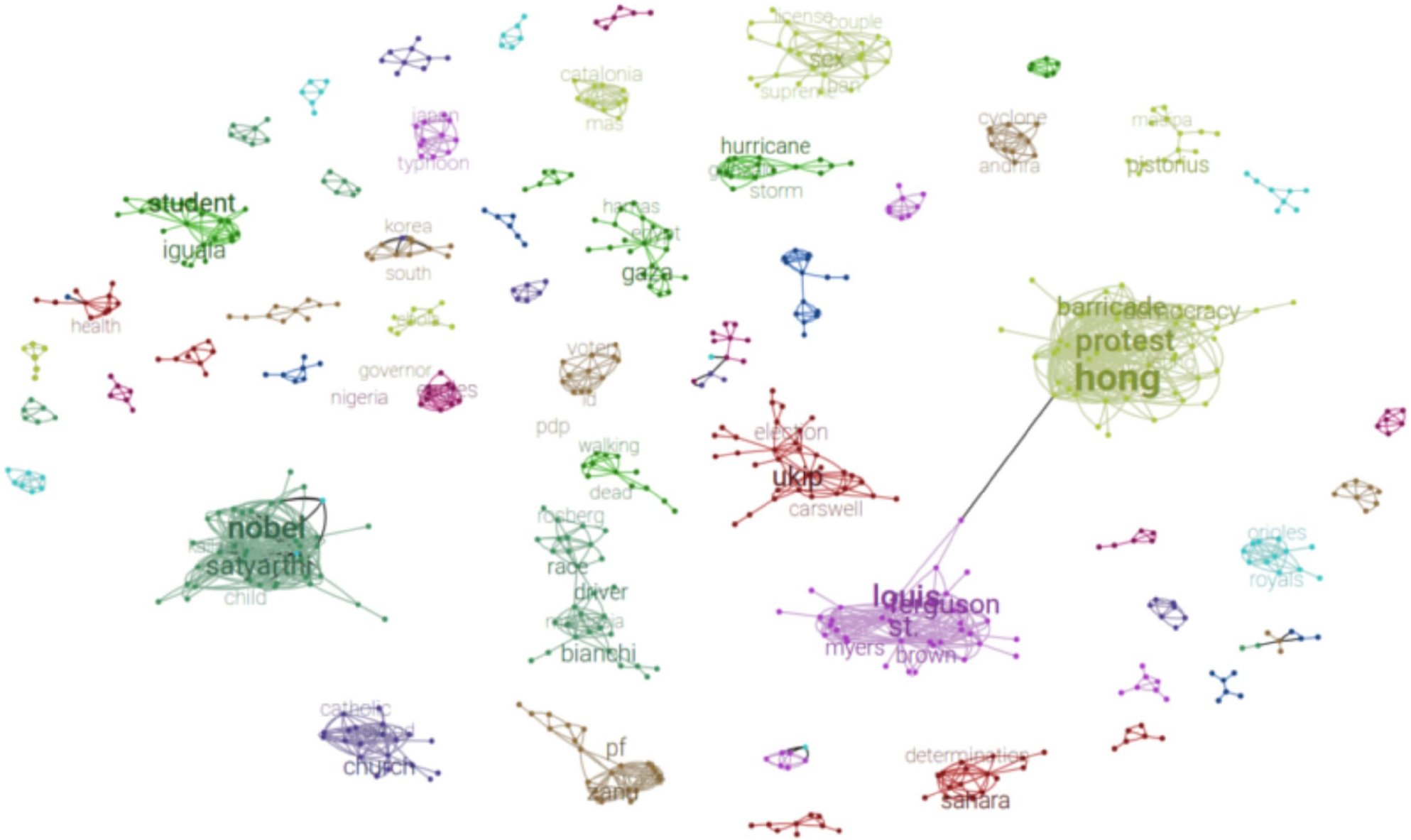
- News stream clustering
- Compare
  - dense vectors
  - sparse vectors
  - combinations
- State-of-the-art results
- B-cubed F1 for evaluation
- Integrated into a real system

Related Work

- **newsLens** [Laban & Hearst, 2017]
  - build clusters and storylines for news
  - used: keywords, graph, community detection
    - we use vectors instead
  - *We are inspired by their pipeline*

- **[Miranda et al., 2018]**
  - clustered a stream of news articles in English, Spanish and German
  - TF-IDF, timestamp
    - we use a local graph instead
  - *We use their dataset*

NewsLens Clustering Example



Our Model

1. **identify local topics**
   - sliding window over time with overlaps
   - using article vectors: TF-IDF, doc2vec, combination
   - graph edge if $sim(d_i, d_j) \geq T_1$
   - Louvain's community detection

2. **merge long-term topics**
   - using the mean of all news vectors belonging to the topic
   - merge topics if
     - *also, use machine* $sim(t_i, t_j) \geq T_2$ *and T2*

Representation

TF-IDF

Doc2Vec

- trained on Signal Media One-million news articles corpus:
  - 265,512 blog articles
  - 734,488 news articles
  - from 93k unique sources
  - over a period of one month

Combined Representations

Combination (unsupervised)

- using T1 and T2

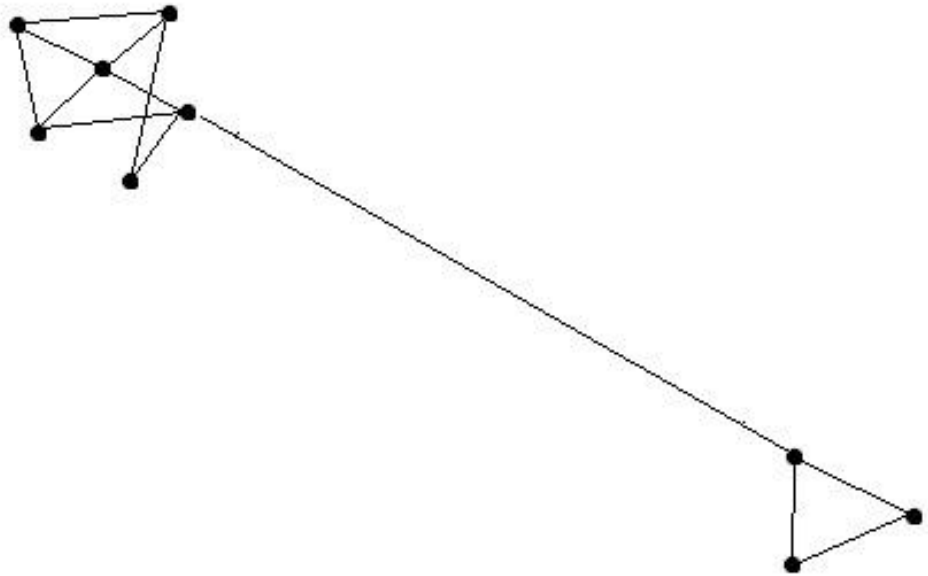
Combination (supervised)

- logistic regression with TF-IDF and doc2vec scores as input
- and whether a pair of articles are in the same cluster in the training data

Merging Long-Term Topics: Louvain's Community Detection

Topic 1

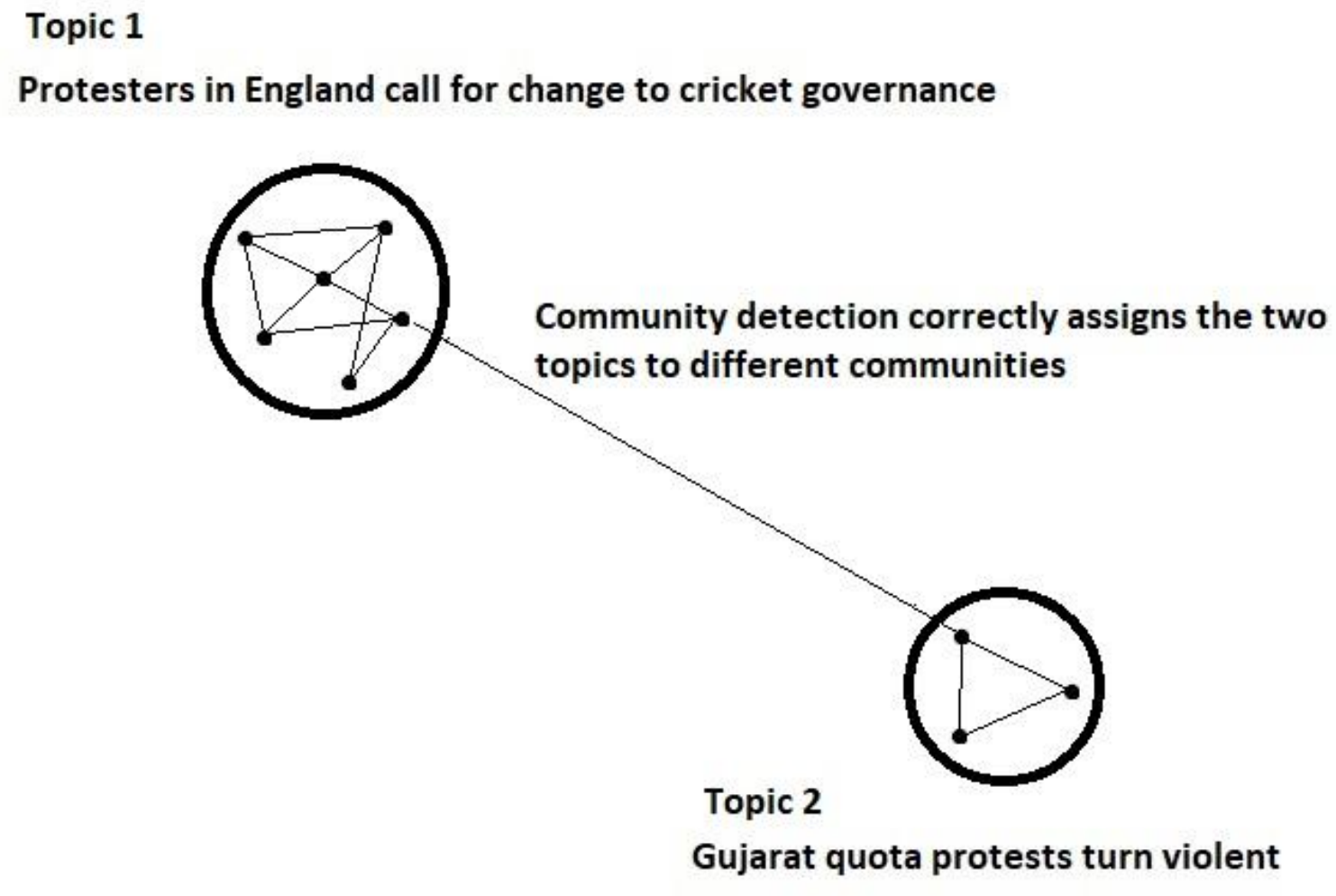
Protesters in England call for change to cricket governance



Topic 2

Gujarat quota protests turn violent

Merging Long-Term Topics: Louvain's Community Detection



Dataset

Partition	Docs	Tokens	Clusters	Cluster Size
Train	12,233	434±364	593	21±32
Test	8,726	521±495	222	39±88

Sebastinao Miranda, Arturs Znotins, Shay B. Cohen, Guntis Barzdins. *Multilingual clustering of streaming news.* EMNLP'18.

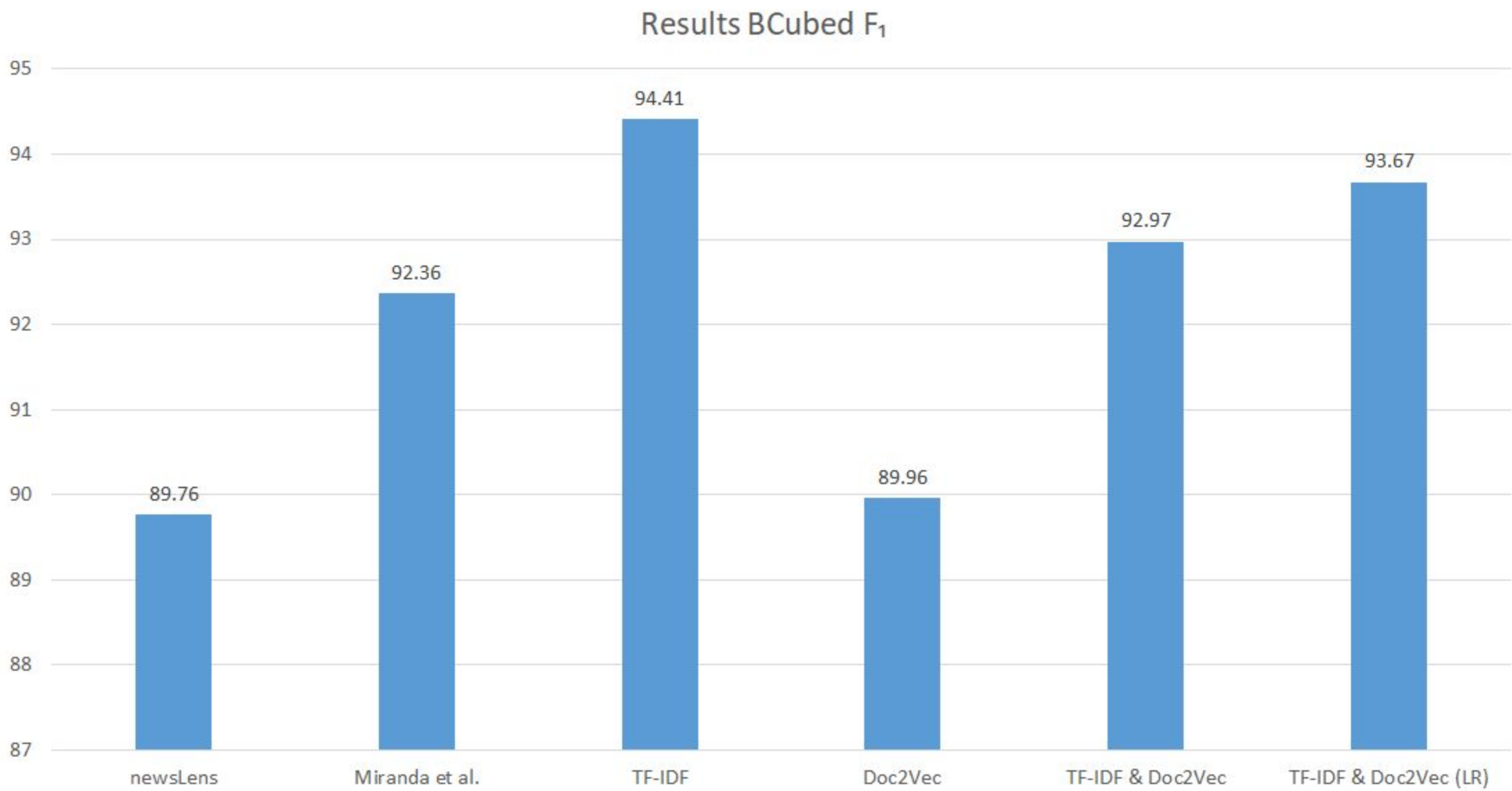
Train
- December 18, 2013 to February 2, 2014
- No time gaps

Test:
- November 2, 2014 to August 25, 2015
- Gaps as long as 3 months --> need to find long-term topics

Evaluation: BCubed measures

Favor clusterings that

- split a cluster that mixes two categories into two **pure clusters** (**cluster homogeneity**)
- unify two clusters that contain only items from the same category (**cluster completeness**)
- **add** an item of a different category **to an already noisy cluster** instead of a pure one
- **make small errors in a big cluster** rather than a large number of small errors in small clusters


Results BCubed F_1
95
94
93
92
91
90
89
88
87
89.76
92.36
94.41
89.96
92.97
93.67
newsLens
Miranda et al.
TF-IDF
Doc2Vec
TF-IDF & Doc2Vec
TF-IDF & Doc2Vec (LR)


Results Standard F_1
99
98
97
96
95
94
93
92
91
95.09
94.03
98.11
95.44
97.61
97.15
newsLens
Miranda et al.
TF-IDF
Doc2Vec
TF-IDF & Doc2Vec
TF-IDF & Doc2Vec (LR)

Evaluation Results

	BCubed						
model	F_1	P	R	F_1	P	R	clusters
baselines							
newsLens [LH17]	89.76	94.37	85.58	95.09	95.90	94.30	873
Miranda et al. [MZCB18]	92.36	94.57	90.25	94.03	98.14	90.25	326
unsupervised							
TF-IDF	**94.41**	95.16	93.66	**98.11**	97.60	98.63	484
doc2vec	89.96	93.00	87.12	95.44	95.55	95.34	785
TF-IDF & doc2vec	92.97	95.75	90.34	97.61	97.73	97.48	663
supervised (LR)							
TF-IDF	94.30	94.87	93.73	98.08	97.46	98.71	485
TF-IDF & doc2vec	93.67	92.71	94.65	97.15	95.39	98.98	431

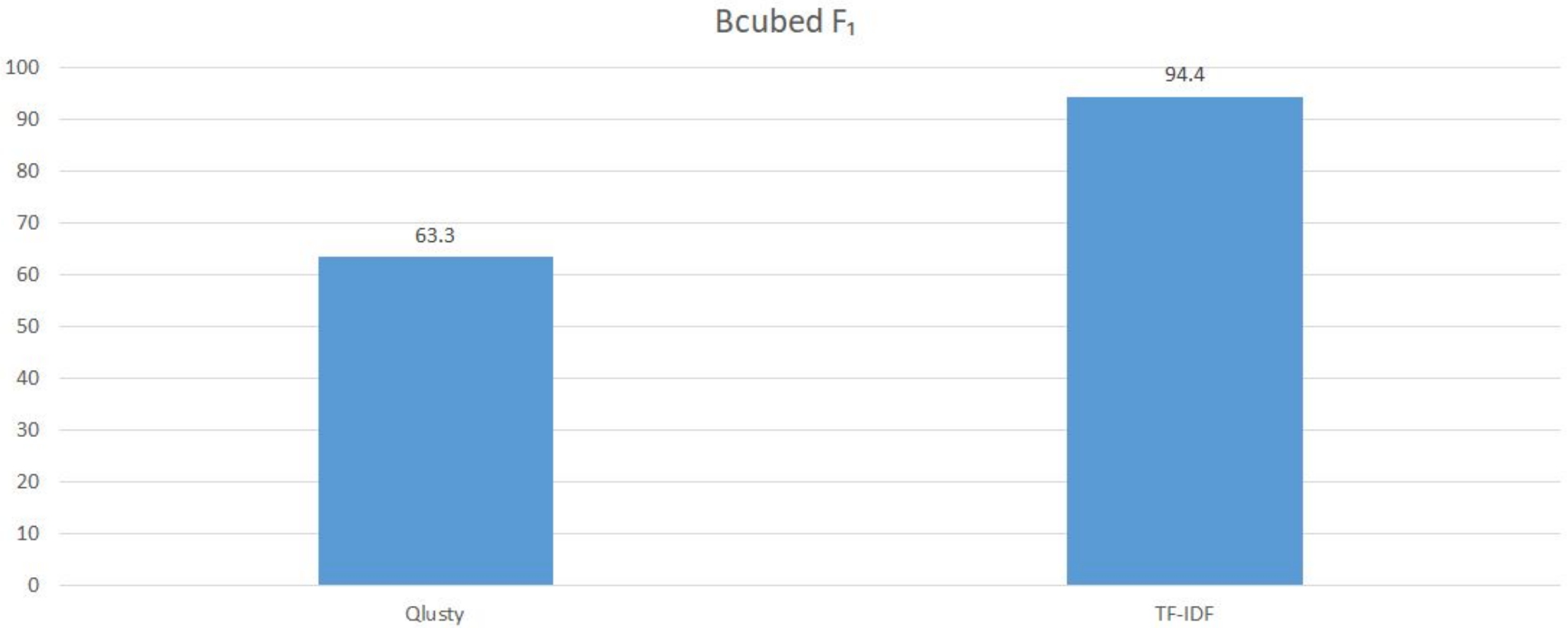
Expected: 222 clusters

Discussion:
Impact of Louvain's Community Detection

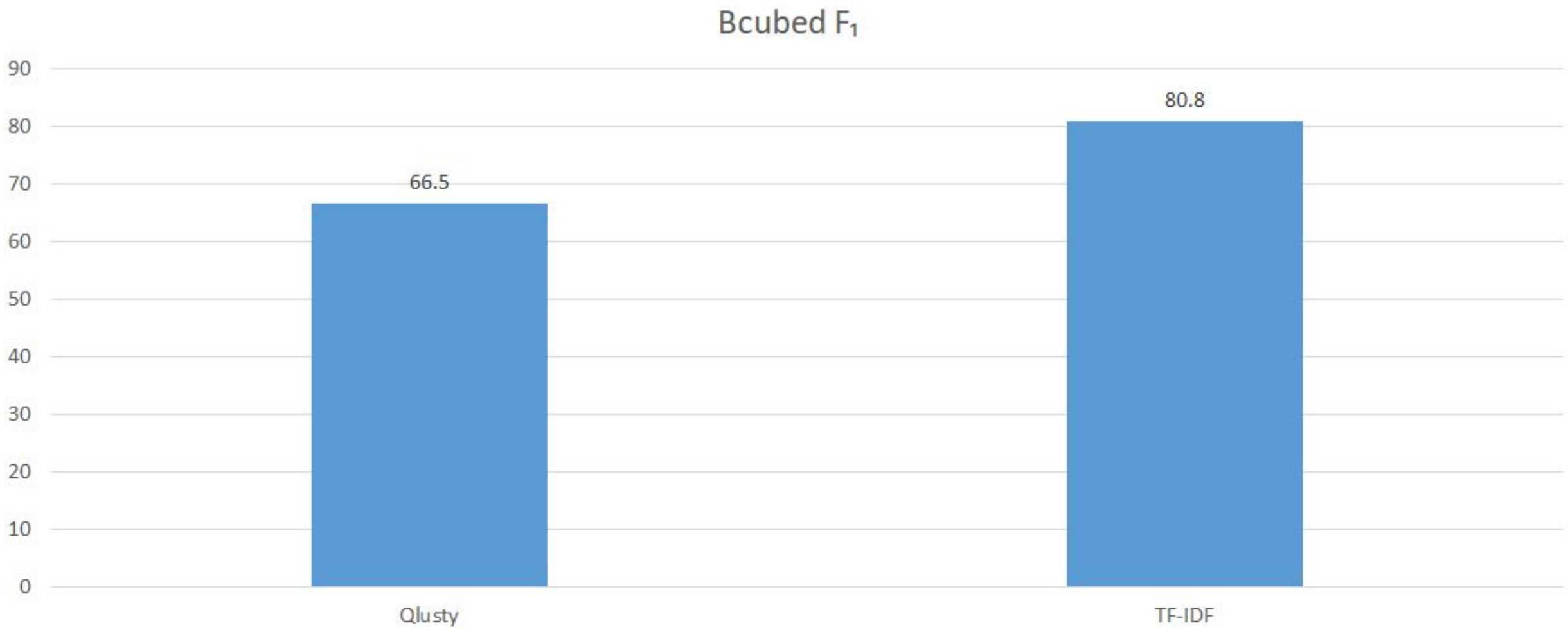
	BCubed						
	F_1	P	R	F_1	P	R	clusters
before	91.55	93.76	89.44	94.36	95.55	93.20	488
after	94.41	95.16	93.66	98.11	97.60	98.63	484

Discussion:

DBSCAN vs. Our Model: Miranda&al.-2018 corpus



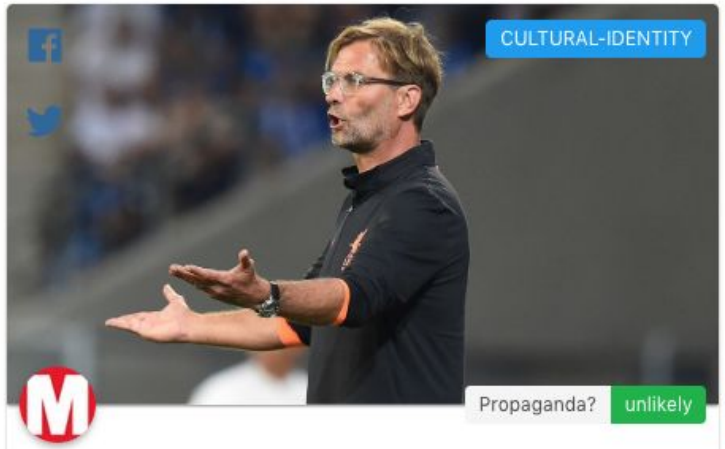
Discussion:
DBSCAN vs. Our Model: Meter Corpus



Our Clustering Algorithm (w/ TF-IDF) Powers Tanbih: https://www.tanbih.org/

Tutorial 2 / 6 : Read articles in a story

BUSINESS SPORTS POLITICS ARTS & ENT TECH & SCI



Joao Felix transfer: Man Utd and Chelsea 'to rival Liverpool for £60m wonderkid'

1 of 6 SPORT FOOTBALL TRANSFER NEWS

4 hours ago

Venezuelan FM Warns Foreign Journalists Against Working Without Accreditation

1 of 3 LATAM

4 hours ago

Each story on TANBIH will have multiple articles from multiple news outlets. If you are interested in a story, you can go through all articles of a story from these previous and next buttons

Conclusion and Future Work

Summary

- Compared dense vs. sparce vectors
- State-of-the-art results
- B-cubed F1 for evaluation
- Integrated into a real system

Future work

- Improve topic matching
- Cross-language extension

http://www.tanbih.org/