Centre for
Data Analytics

Insight

Stability Analysis For Topic Models

Dr. Derek Greene
Insight @ UCD

Motivation

- Key challenge in topic modeling: selecting an appropriate number of topics for a corpus.
  - Choosing too few topics will produce results that are overly broad.
  - Choosing too many will result in the"over-clustering" of a corpus into many small, highly-similar topics.
- In the literature, topic modeling results are often presented as lists of top-ranked terms. But how robust are these rankings?
- *Stability analysis* has been used elsewhere to measure ability of an algorithm to produce similar solutions on data originating from the same source (Levine & Domany, 2001).

Proposal: term-centric stability approach for selecting the number of topics in a corpus, based on agreement between term rankings.

Term Ranking Similarity

Initial Problem: Given a pair of ranked lists of terms, how can we measure the similarity between them?

Rank	Topic 1
1	film
2	music
3	awards
4	star
5	band
6	album
7	oscar
8	movie
9	cinema
10	song

Ranking R1

Rank	Topic 1
1	celebrity
2	music
3	awards
4	star
5	ceremony
6	band
7	movie
8	oscar
9	cinema
10	film

Ranking R2

- Simple approaches:
  - Measure correlation (e.g. Spearman).
  - Measure overlap between the two sets. $\frac{|R1 \cap R2|}{|R1 \cup R2|}$
- How do we deal with…
  - Indefiniteness (i.e. missing terms).
  - Positional information.

➡ We propose a "top-weighted" similarity measure that can also handle indefinite rankings.

Term Ranking Similarity

Average Jaccard (AJ) Similarity:
Calculate average of the Jaccard scores between every pair of subsets of *d* top-ranked terms in two ranked lists, for depths *d* ∈ *[1, t]*.

$$AJ(R_i, R_j) = \frac{1}{t}\sum_{d=1}^{t} \gamma_d(R_i, R_j)$$

$$\gamma_d(R_i, R_j) = \frac{|R_{i,d} \cap R_{j,d}|}{|R_{i,d} \cup R_{j,d}|}$$

Example - AJ Similarity for two ranked lists with *t=5* terms:

d	$R_{1,d}$	$R_{2,d}$	Jac_d	AJ
1	album	sport	0.000	0.000
2	album, music	sport, best	0.000	0.000
3	album, music, best	sport, best, win	0.200	0.067
4	album, music, best, award	sport, best, win, medal	0.143	0.086
5	album, music, best, award, win	sport, best, win, medal, award	0.429	0.154

➡ Differences at the top of the ranked lists have more influence than differences at the tail of the lists.

Topic Model Agreement

Next Problem: How to measure agreement between two topic models, each containing *k* ranked lists?

- **Proposed Strategy:**
  1. Build *k x k* Average Jaccard similarity matrix.
  2. Find optimal match between the rows and columns using Hungarian assignment method.
  3. Measure agreement as the average similarity between matched topics.

Ranking Set #1:

$R_{11} = \{\text{sport, win, award}\}$
$R_{12} = \{\text{bank, finance, money}\}$
$R_{13} = \{\text{music, album, band}\}$

Ranking Set #2:

$R_{21} = \{\text{finance, bank, economy}\}$
$R_{22} = \{\text{music, band, award}\}$
$R_{23} = \{\text{win, sport, money}\}$

	R_{21}	R_{22}	R_{23}
R_{11}	0.00	0.07	0.50
R_{12}	0.50	0.00	0.07
R_{13}	0.00	0.61	0.00

AJ Similarity Matrix

Optimal Match

$$\pi = (R_{11}, R_{23}), (R_{12}, R_{21}), (R_{13}, R_{23})$$

$$agree(\mathcal{S}_1, \mathcal{S}_2) = \frac{0.50+0.50+0.61}{3} = 0.54$$

Model Selection

Q. How can we use the agreement between pairs of topic models to choose the number of topics in a corpus?

- **Proposal:**
  - Generate topics on different samples of the corpus.
  - Measure term agreement between topics and a “reference set” of topics.
  - Higher agreement between terms ➢ A more stable topic model.

Rank	Topic 1	Topic 2
1	oil	win
2	bank	players
3	election	minister
4	policy	party
5	government	ireland
6	match	club
7	senate	year
8	democracy	election
9	firm	coalition
10	team	first

Run 1

Low agreement between top ranked terms

Low stability for k=2

Rank	Topic 1	Topic 2
1	cup	first
2	labour	sales
3	growth	year
4	team	minister
5	senate	firm
6	minister	match
7	ireland	coalition
8	players	team
9	year	election
10	economy	policy

Run 2

Model Selection

Q. How can we use the agreement between pairs of topic models to choose the number of topics in a corpus?

- **Proposal:**
  - Generate topics on different samples of the corpus.
  - Measure term agreement between topics and a "reference set" of topics.
  - Higher agreement between terms ➢ A more stable topic model.

Rank	Topic 1	Topic 2	Topic 3
1	growth	game	labour
2	company	ireland	election
3	market	win	vote
4	economy	cup	party
5	bank	goal	governmen
6	year	match	coalition
7	firm	team	minister
8	sales	first	policy
9	shares	club	democracy
10	oil	players	first

Run 1

High agreement between top ranked terms

High stability for $k=3$

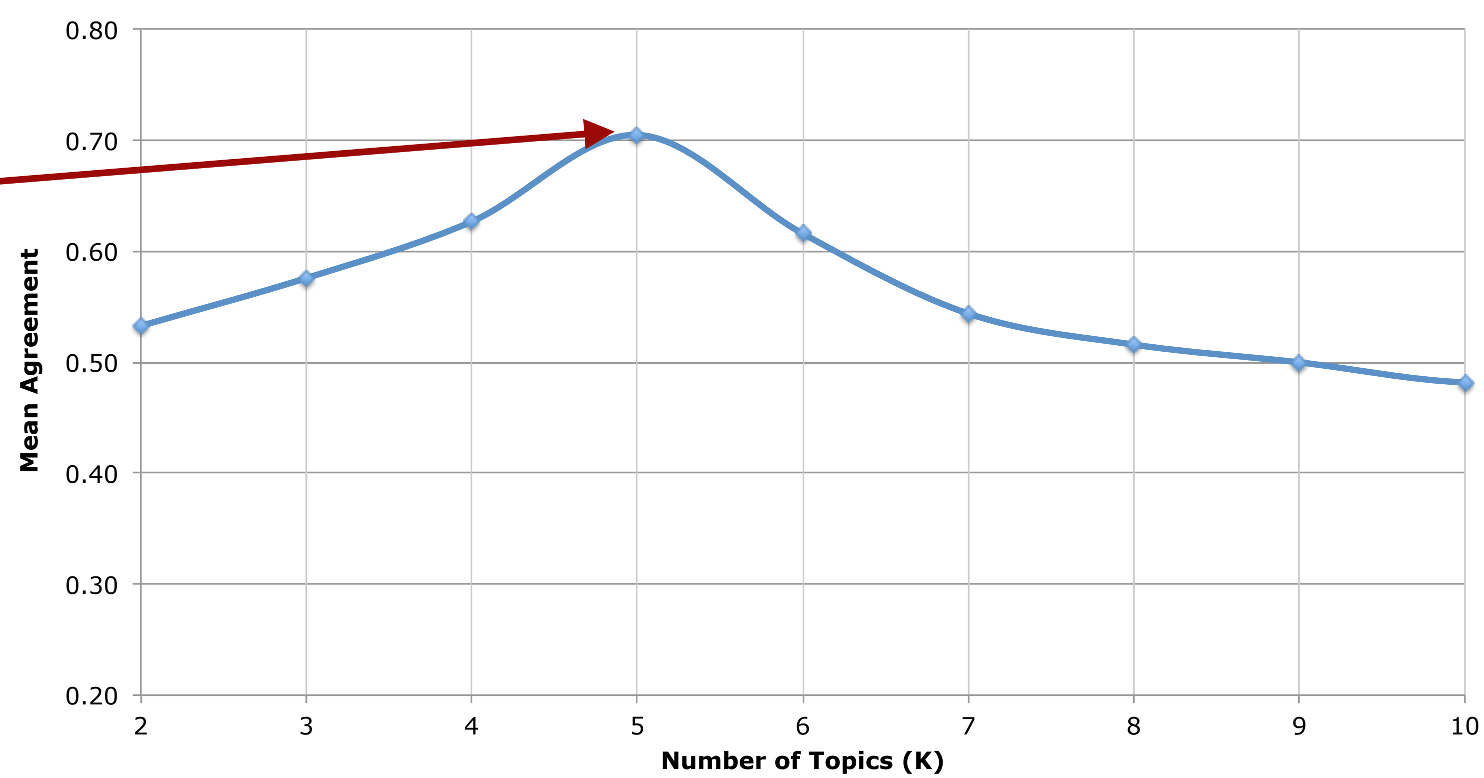
Rank	Topic 1	Topic 2	Topic 3
1	game	growth	labour
2	win	company	election
3	ireland	market	governmen
4	cup	economy	party
5	match	bank	vote
6	team	shares	policy
7	first	year	minister
8	players	firm	democracy
9	club	sales	senate
10	goal	oil	coalition

Run 2

Model Selection - Algorithm

1. Randomly generate τ samples of the data set, each containing $\beta \times n$ documents.
2. For each value of $k \in [k_{min}, k_{max}]$:
   1. Apply the topic modeling algorithm to the complete data set of n documents to generate k topics, and represent the output as the reference ranking set $\mathcal{S}_0$.
   2. For each sample $\mathbf{X}_i$:
      (a) Apply the topic modeling algorithm to $\mathbf{X}_i$ to generate k topics, and represent the output as the ranking set $\mathcal{S}_i$.
      (b) Calculate the agreement score $agree(\mathcal{S}_0, \mathcal{S}_i)$.
   3. Compute the mean agreement score for k over all τ samples
3. Select one or more values for k based upon the highest mean agreement scores.

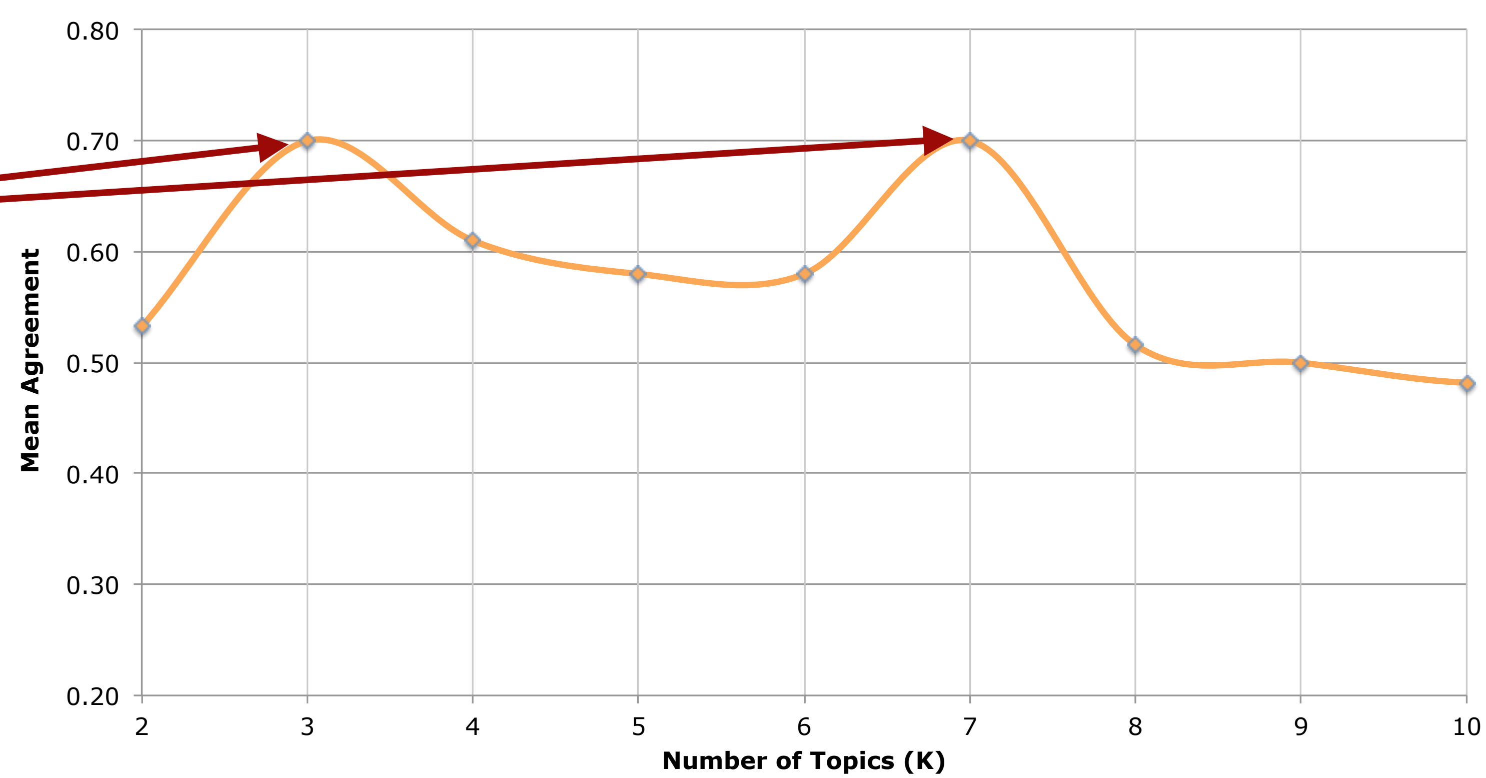
Single stability peak for *k=5*



Model Selection - Algorithm

1. Randomly generate τ samples of the data set, each containing $\beta \times n$ documents.
2. For each value of $k \in [k_{min}, k_{max}]$:
   1. Apply the topic modeling algorithm to the complete data set of n documents to generate k topics, and represent the output as the reference ranking set $\mathcal{S}_0$.
   2. For each sample $\mathbf{X}_i$:
      (a) Apply the topic modeling algorithm to $\mathbf{X}_i$ to generate k topics, and represent the output as the ranking set $\mathcal{S}_i$.
      (b) Calculate the agreement score $agree(\mathcal{S}_0, \mathcal{S}_i)$.
   3. Compute the mean agreement score for k over all τ samples
3. Select one or more values for k based upon the highest mean agreement scores.

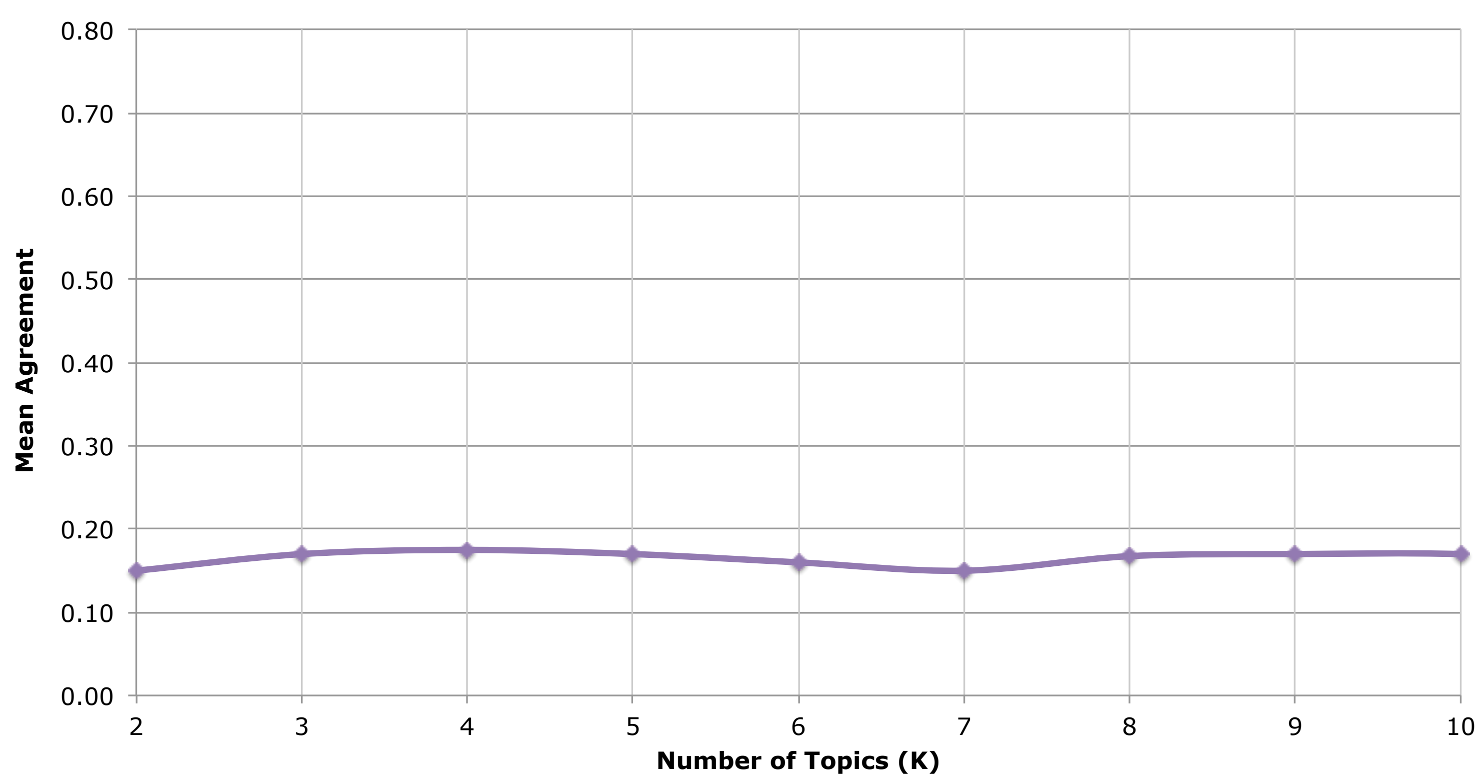
Two potentially good models



Model Selection - Algorithm

1. Randomly generate τ samples of the data set, each containing $\beta \times n$ documents.
2. For each value of $k \in [k_{min}, k_{max}]$:
   1. Apply the topic modeling algorithm to the complete data set of n documents to generate k topics, and represent the output as the reference ranking set $\mathcal{S}_0$.
   2. For each sample $\mathbf{X}_i$:
      (a) Apply the topic modeling algorithm to $\mathbf{X}_i$ to generate k topics, and represent the output as the ranking set $\mathcal{S}_i$.
      (b) Calculate the agreement score $agree(\mathcal{S}_0, \mathcal{S}_i)$.
   3. Compute the mean agreement score for k over all τ samples
3. Select one or more values for k based upon the highest mean agreement scores.

No coherent topics in the data?



Aside: NMF For Topic Models

- **Applying NMF to Text Data:**

  1. Construct vector space model for documents (after stop-word filtering), resulting in a document-term matrix **A**.
  2. Apply TF-IDF term weight normalisation to **A**.
  3. Normalize TF-IDF vectors to unit length.
  4. Apply Projected Gradient NMF to **A**.

- **NMF outputs two factors:**

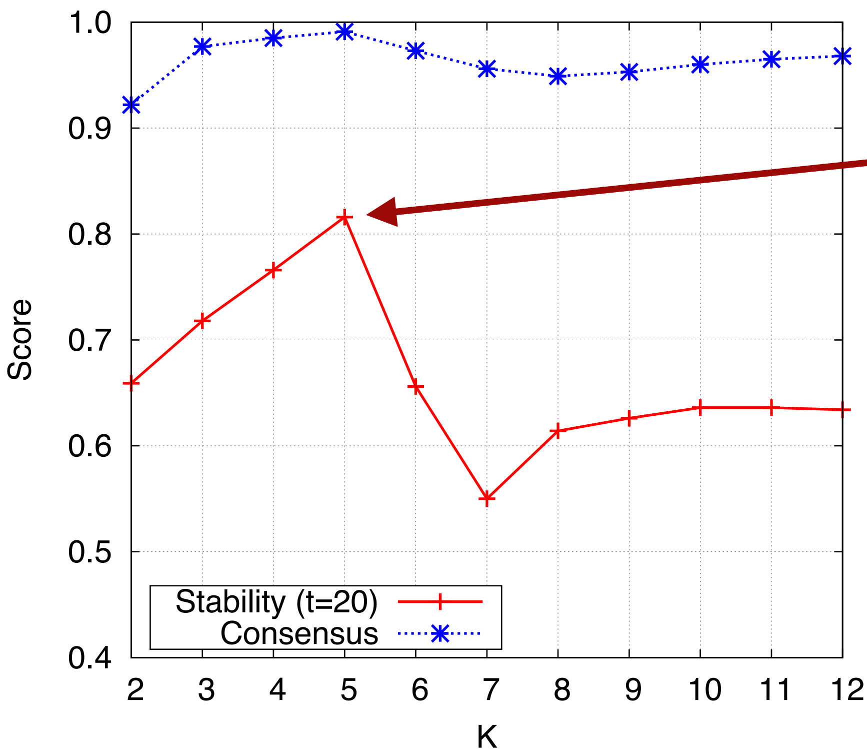
  1. *Basis matrix:* The topics in the data. Rank entries in columns to produce topic ranking sets.
  2. *Coefficient matrix*: The membership weights for documents relative to each topic.

Experimental Evaluation

- **Experimental Setup:**
  - Examine topic stability for $k \in [2, 12]$.
  - Reference ranking set produced using NNDSVD + NMF on the complete corpus.
  - Generated 100 test ranking sets using Random Initialisation + NMF, randomly sampling 80% of documents.
  - Measure agreement using top 20 terms.
- **Comparison:**
  - Apply popular existing approach for selecting rank for NMF based on the cophenetic correlation of a consensus matrix (Brunet et al, 2004).
  - Compare both results to ground truth labels for each corpus.

Experimental Results

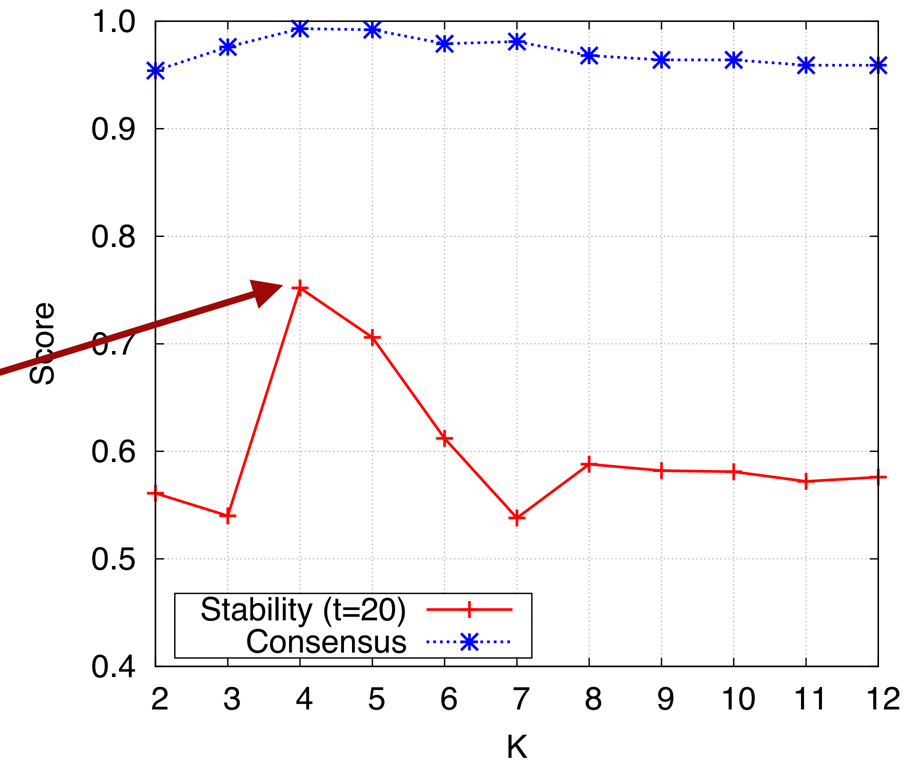
***bbc* corpus**



k=5 ground truth labels

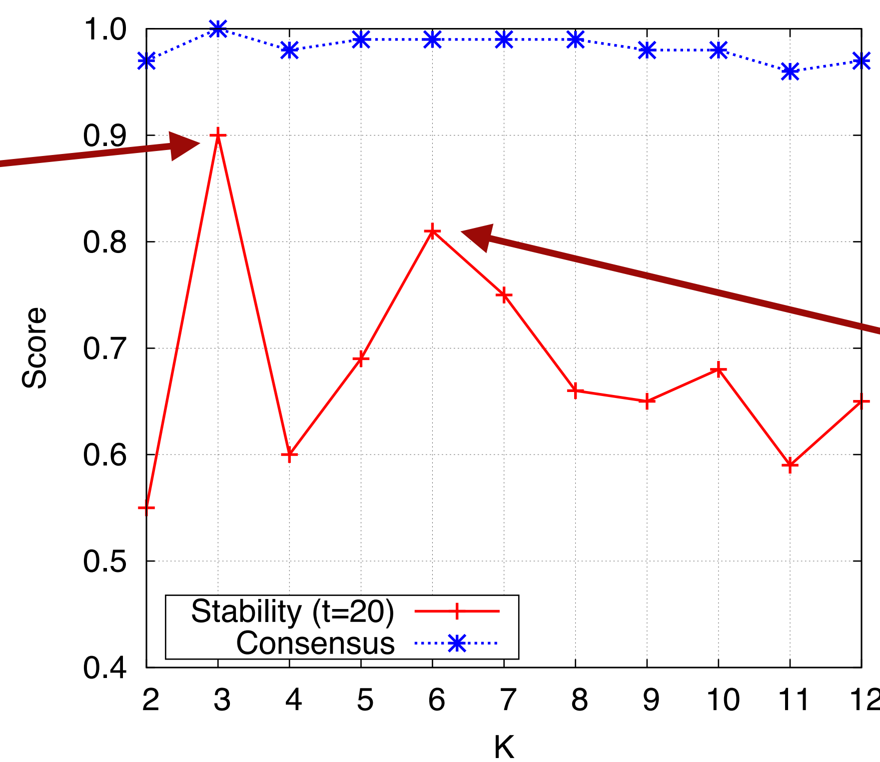
***bbcsport* corpus**

5 ground truth labels but "athletics" & "tennis" ofter merged



***guardian-2013* corpus**

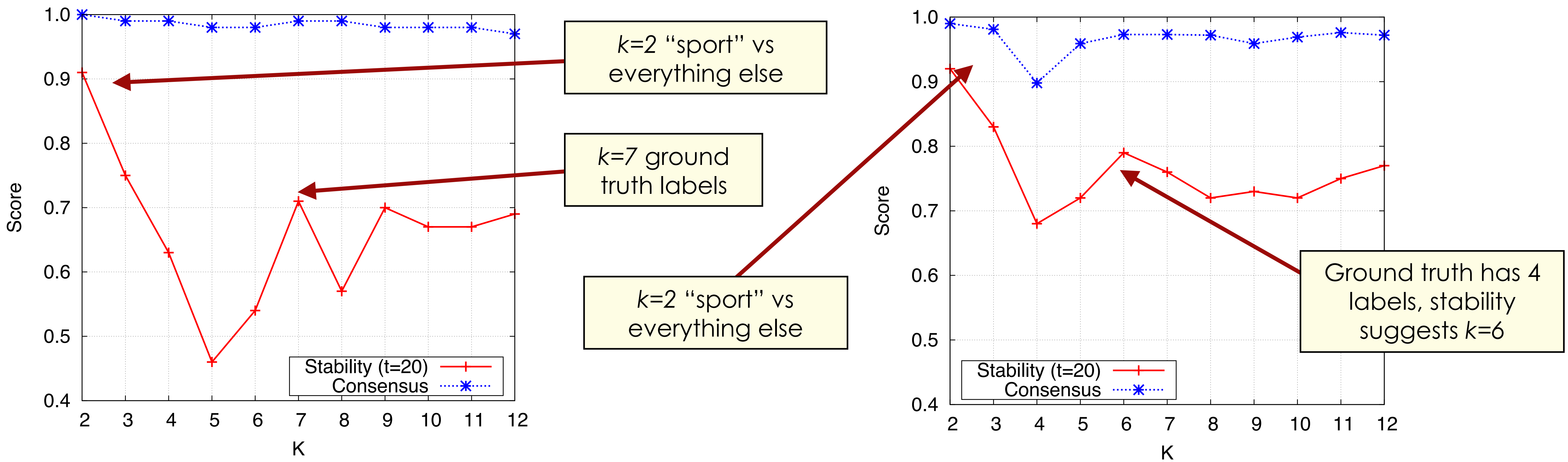
"Books", "Fashion" & "Music" merged into a culture topic at *k*=3



k=6 ground truth labels

Experimental Results

***irishtimes-2013* corpus** / ***nytimes-1999* corpus**



***irishtimes-1999* corpus (*k=2*)**

Rank	*Topic 1*	*Topic 2*
1	game	cent
2	against	government
3	team	court
4	ireland	health
5	players	ireland
6	time	minister
7	cup	people
8	back	tax
9	violates	dublin
10	win	irish

***nytimes-1999* corpus (*k=4*)**

Rank	*Topic 1*	*Topic 2*	*Topic 3*	*Topic 4*
1	game	company	yr	mets
2	knicks	stock	bills	yankees
3	team	market	bond	game
4	season	business	rate	inning
5	coach	companies	infl	valentine
6	points	shares	bds	season
7	play	stocks	bd	torre
8	league	york	month	baseball
9	players	investors	municipal	run
10	sprewell	bank	buyer	clemens

Ground truth does not always correspond well to the actual data!
Can arise when metadata is used as ground truth for ML experiments.

Summary

- Proposed new method for choosing number of topics using a term-centric stability analysis strategy.
- Using rankings rather than raw factor values or probabilities means we can generalise to any topic modeling approach that represents topics as term rankings.

- **Future work:**
  - Evaluate topic stability method with LDA.
  - Build ensemble of topic models to provide better term rankings and document clusters.
  - Apply term agreement measures in context of dynamic topic models.

Any Questions ?

http://arxiv.org/abs/1404.4606

https://github.com/derekgreene/topic-stability

References

- Greene, D., O’Callaghan, D. & Cunningham, P. How Many Topics? Stability Analysis for Topic Models. arXiv.org pre-print 1404.4606, April 2014.
- Levine, E. & Domany, E. Resampling method for unsupervised estimation of cluster validity. Neural Computation, 13. 2001
- Tibshirani, R., Walther, G., Botstein, D. & Coalition, P. Cluster validation by prediction strength. Tech. rep., Dept. Statistics, Stanford University. 2001
- Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proc. National Academy of Sciences 101(12) (2004).