Tandem Anchoring: a Multiword Anchor Approach for Interactive Topic Modeling

Jeffrey Lund[†], Connor Cook[†], Kevin Seppi[†], Jordan Boyd-Graber[‡]

[†] 

[‡] 

Offline Topic Modeling

- Topic modeling discovers latent topics in a corpus
- Topics are distributions over vocabulary of the corpus
- Most well known example is Latent Dirichlet Allocation*

*Blei et al., 2003

Example topics from Amazon

Topic	Top Words in Topic
cameras	lens camera canon 50mm lenses digital nikon
cables	cable cables hdmi quality price tv monster
routers	wireless linksys work product signal internet work
headphones	koss headphones sound good quality noise pair

Example topics produced from a collection of Amazon product reviews

Interactive Topic Modeling

Adding interaction to offline topic modeling enables new use cases:

- Add user input to influence model during inference
- Run models on new data in real time
- Interactive exploratory analysis

Hu et al., 2013

Runtime Considerations

Interaction adds new requirement: inference in under ten seconds*

Algorithm	#Doc	Runtime (seconds)
Fast ITM†	13284	24.8
Utopian‡	515	48.0

Reported run times for various interactive topic modeling algorithms.

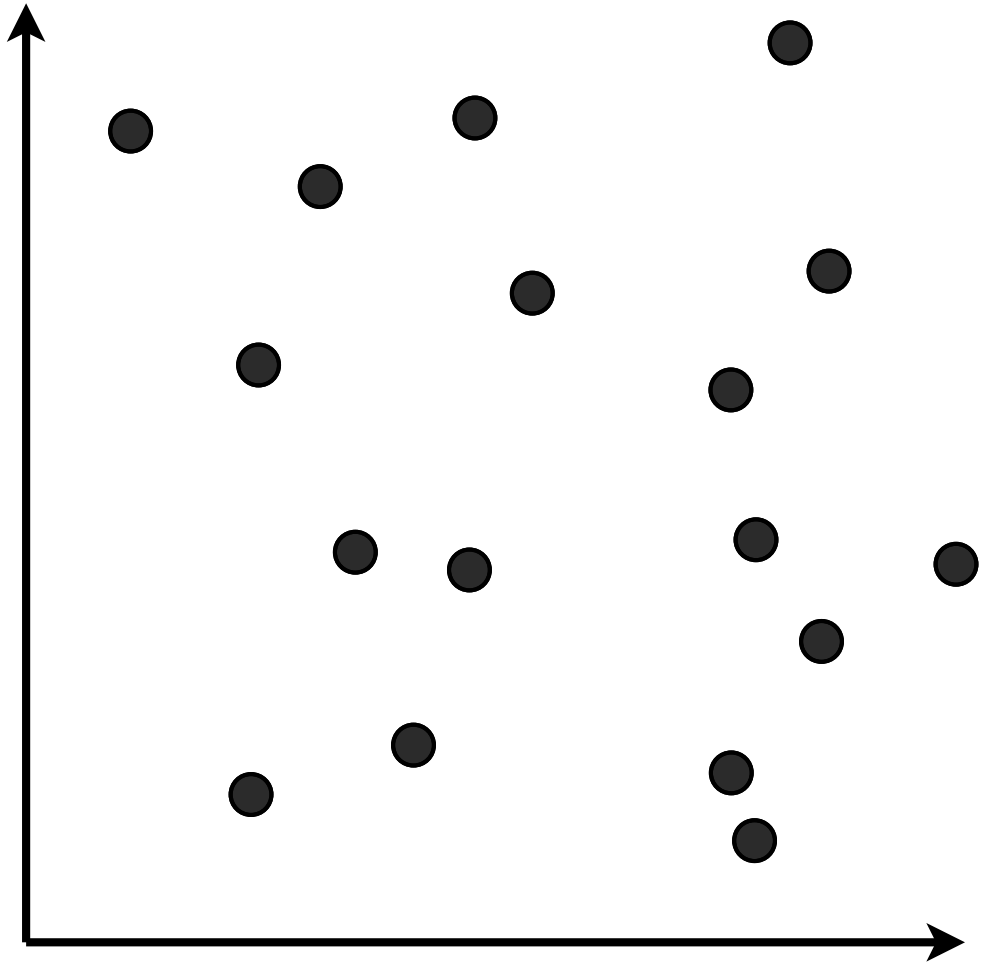
*Cook and Thomas, 2005
†Hu et al., 2013
‡Jaegul et al., 2013

Anchor Words Algorithm

- Views topic modeling as non-negative matrix factorization
- Assuming separability, provably recovers topics in polynomial time
- Fast and scalable

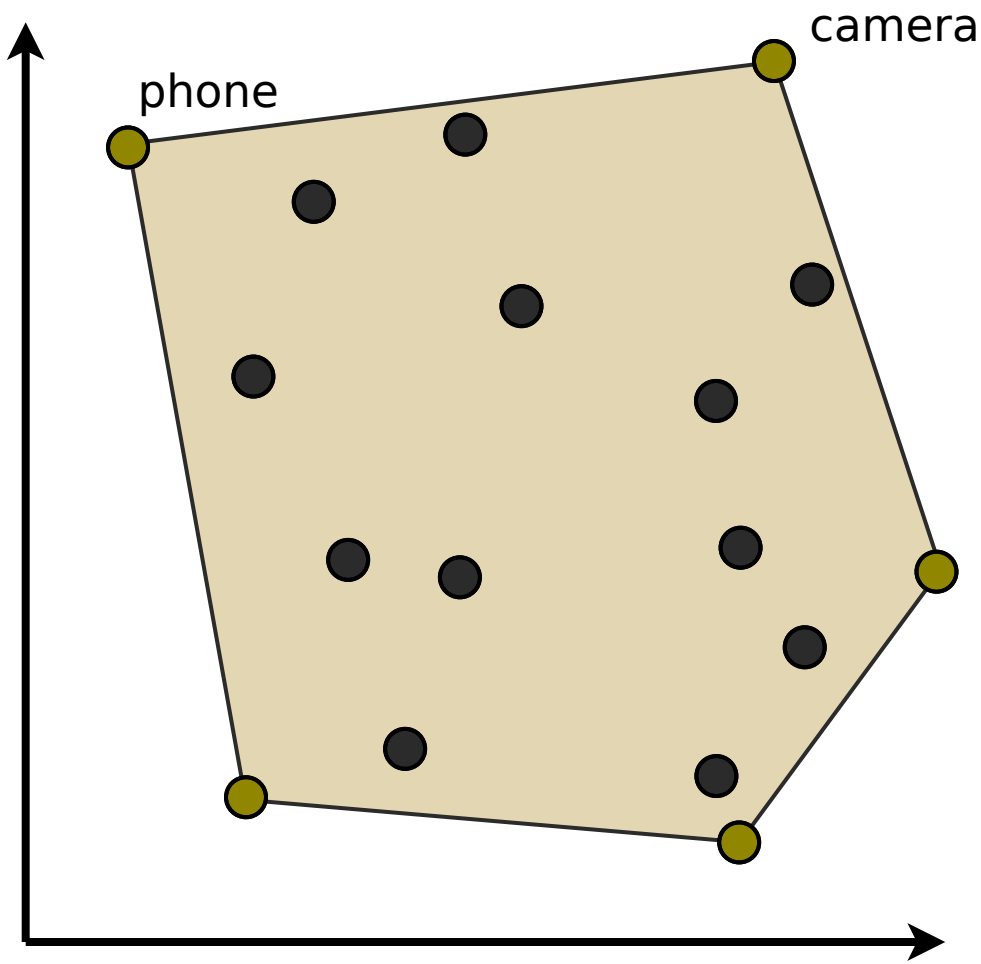
Arora et al., 2013

Anchor Words Algorithm



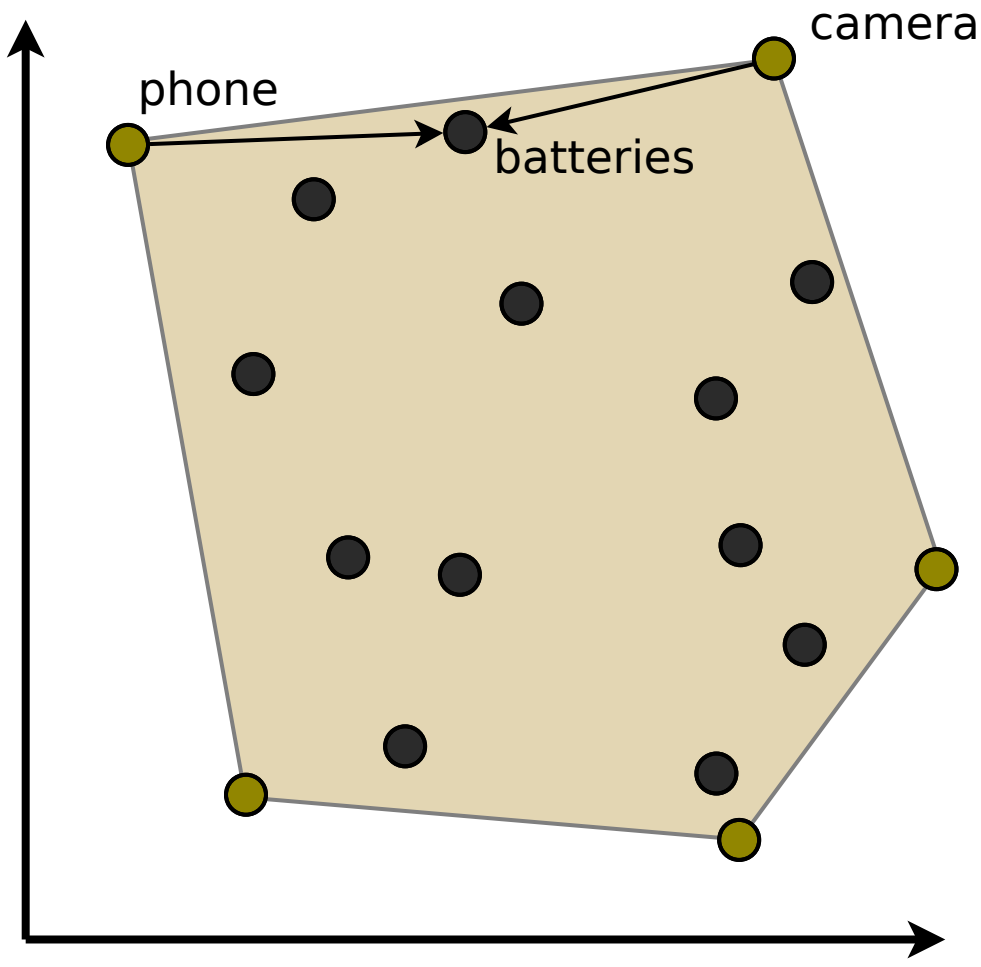
Represent words as vectors based on word cooccurrences

Anchor Words Algorithm



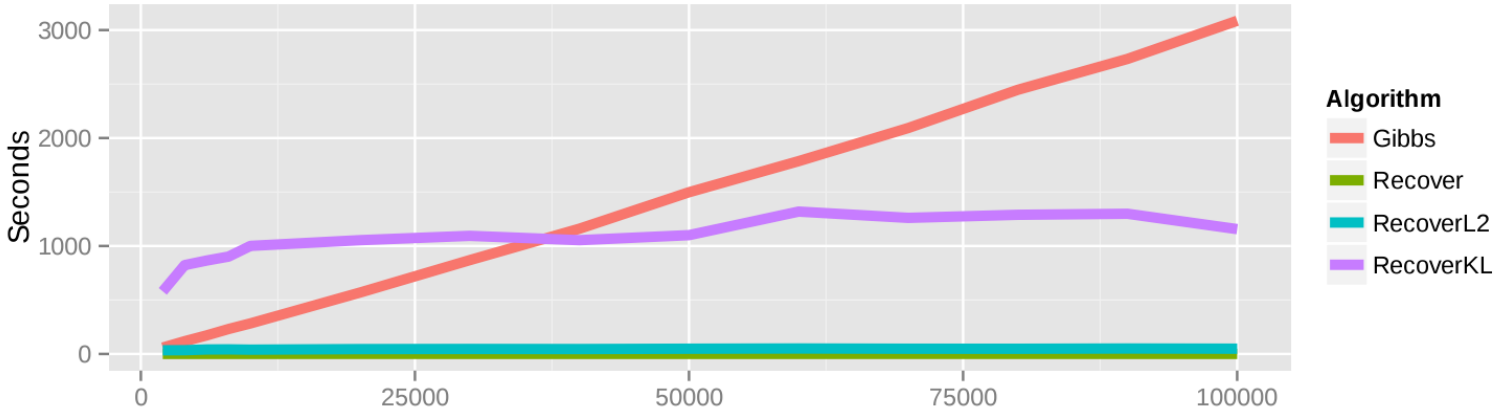
Select anchor words which uniquely identify topics using Gram-Schmidt

Anchor Words Algorithm



Represent each word as a linear combination of anchors

Anchor Words Scalability



(From Arora et al., 2013)

- Anchor words (Recover) scales with vocabulary size
- Model-based (Gibbs) scales with dataset size

Interactive Anchor Selection?

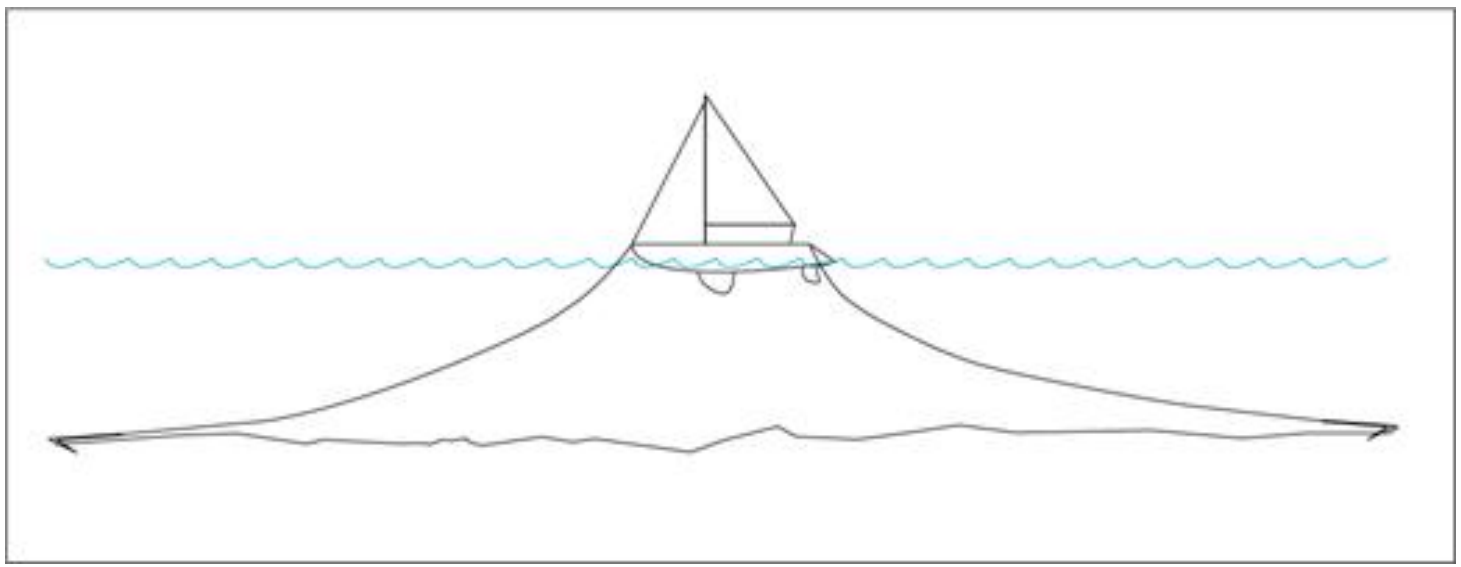
Idea: Let users specify anchors words

Anchor	Top Words in Topics
backpack	backpack camera lens bag room carry fit cameras equipment comfortable
camera	camera lens pictures canon digital lenses batteries filter mm photos
bag	bag camera diaper lens bags genie smell room diapers odor

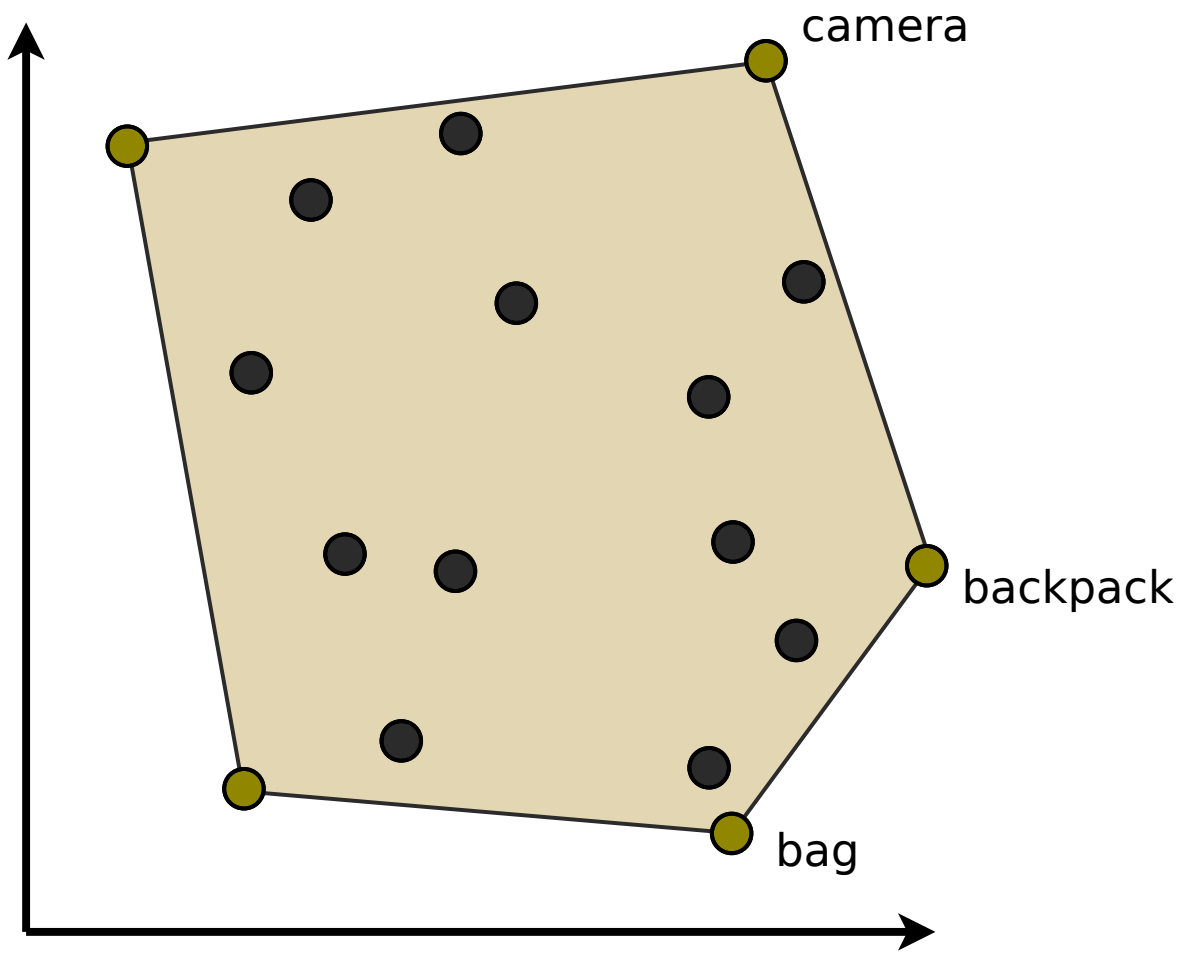
Attempts to construct a topic concerning camera bags in Amazon product reviews with single word anchors.

Tandem Anchors

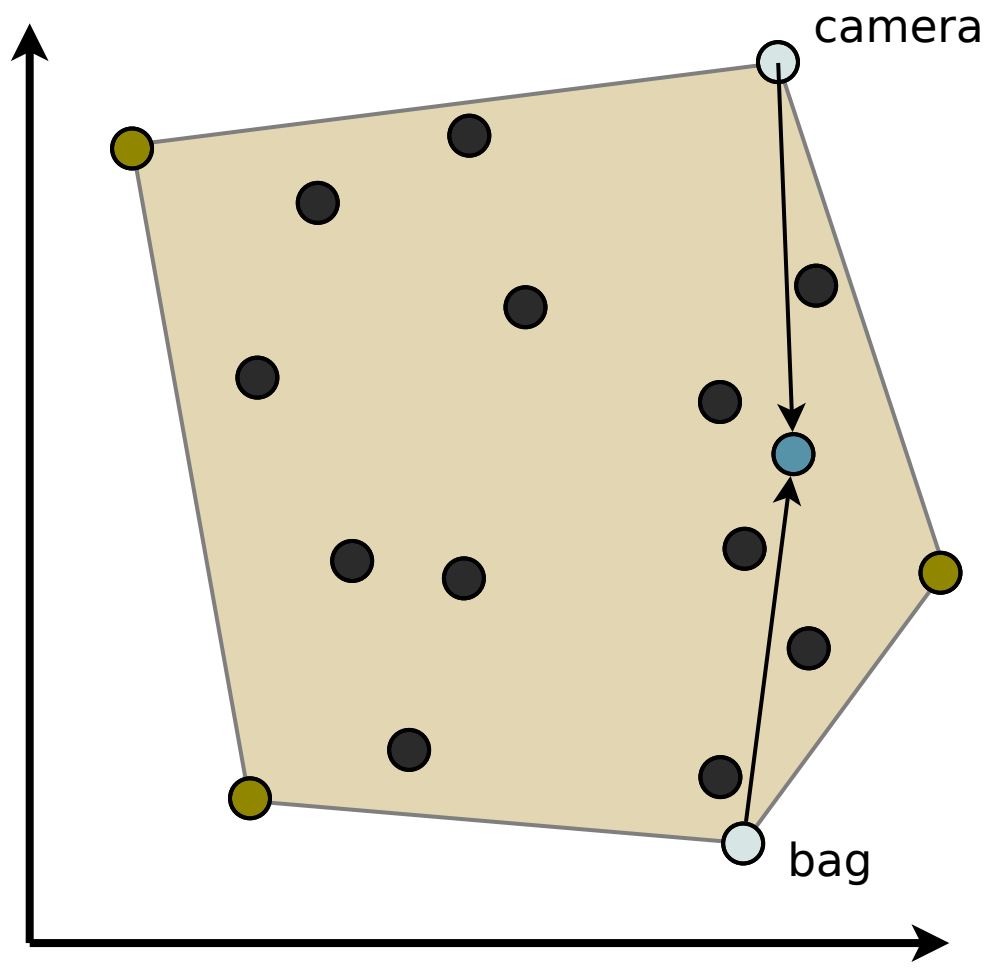
Solution: Form anchors from multiple words



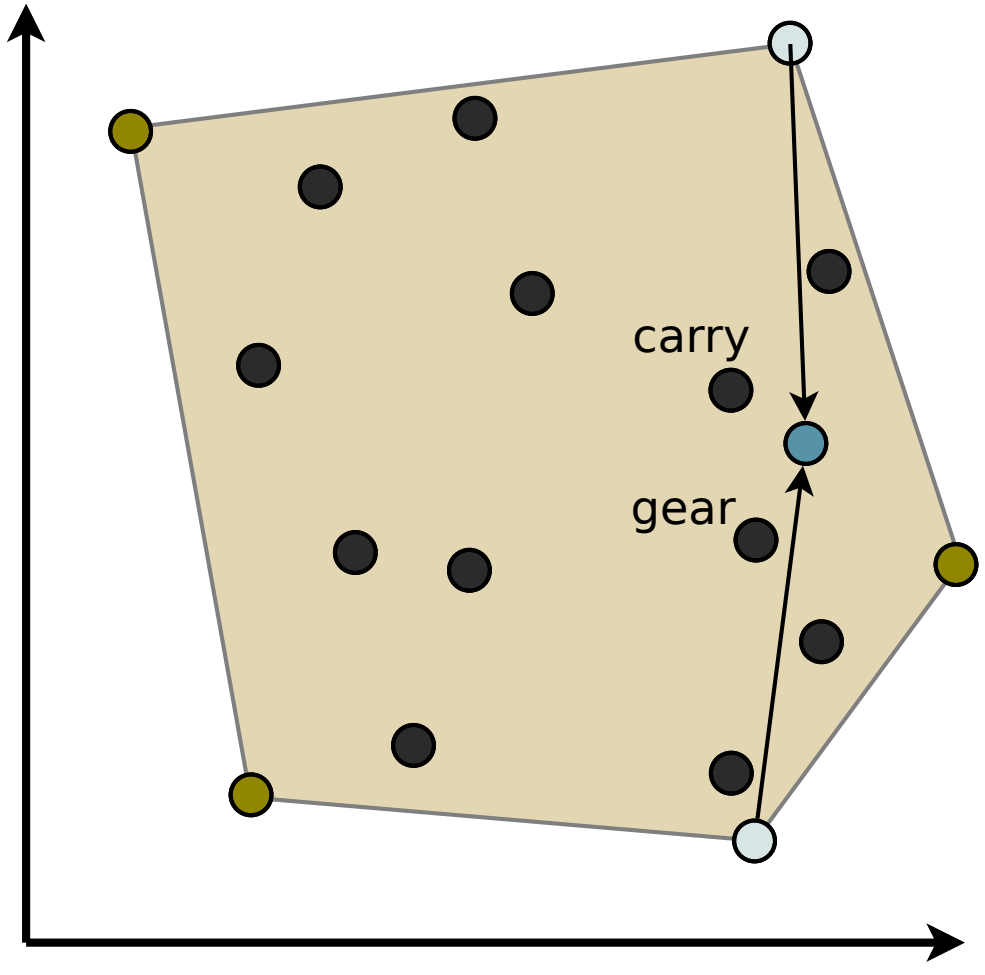
Tandem Anchors



Tandem Anchors



Tandem Anchors



Tandem Anchor Example

Anchor	Top Words in Topics
camera+bag	camera bag backpack carry digital equipment room fits space gear strap
backpack	backpack camera lens bag room carry fit cameras equipment comfortable
camera	camera lens pictures canon digital lenses batteries filter mm photos
bag	bag camera diaper lens bags genie smell room diapers odor

Tandem anchors allows users to correctly specify intuitive anchors

Tandem Anchors Runtime

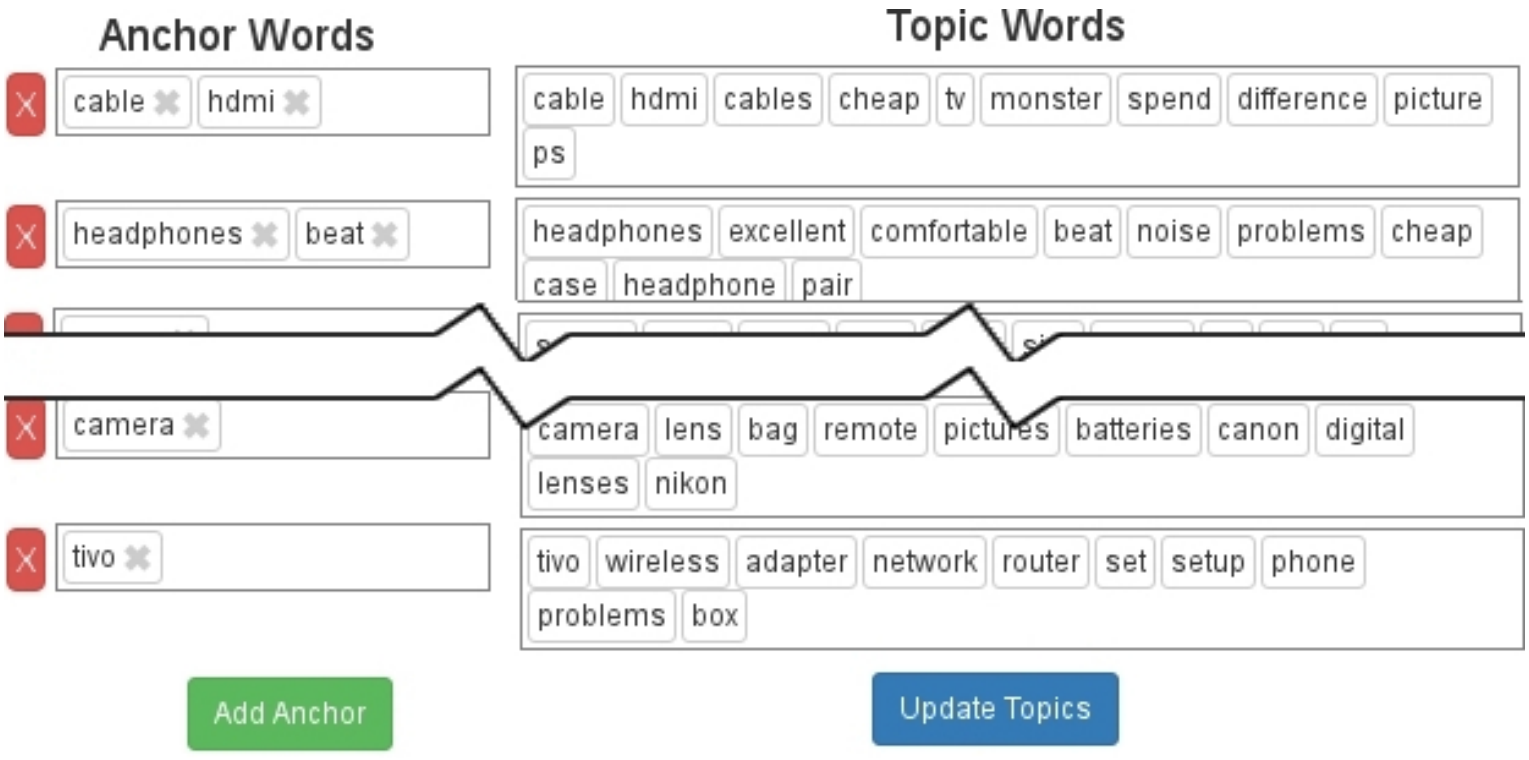
Algorithm	#Doc	Runtime (seconds)
Tandem Anchors	18846	2.13
Fast ITM	13284	24.8
Utopian	515	48.0

Only Tandem Anchors is fast enough to be interactive

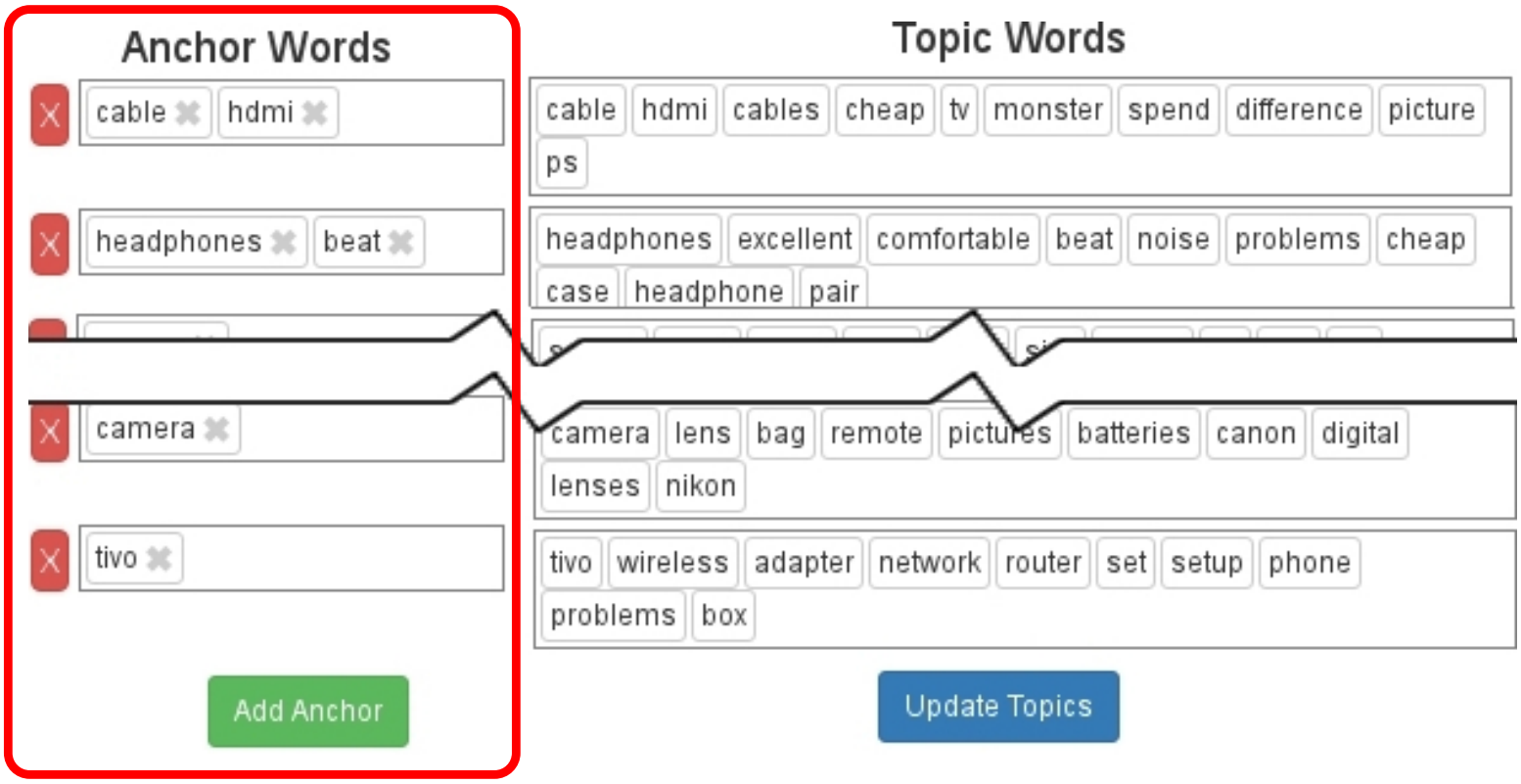
Interactive Topic Modeling User Study

- Used topics as features in classification task on 20 newsgroups
- Used topics produced by users using both tandem anchors and single-word anchors
- Compared to baseline topics produced using Gram-Schmidt anchors

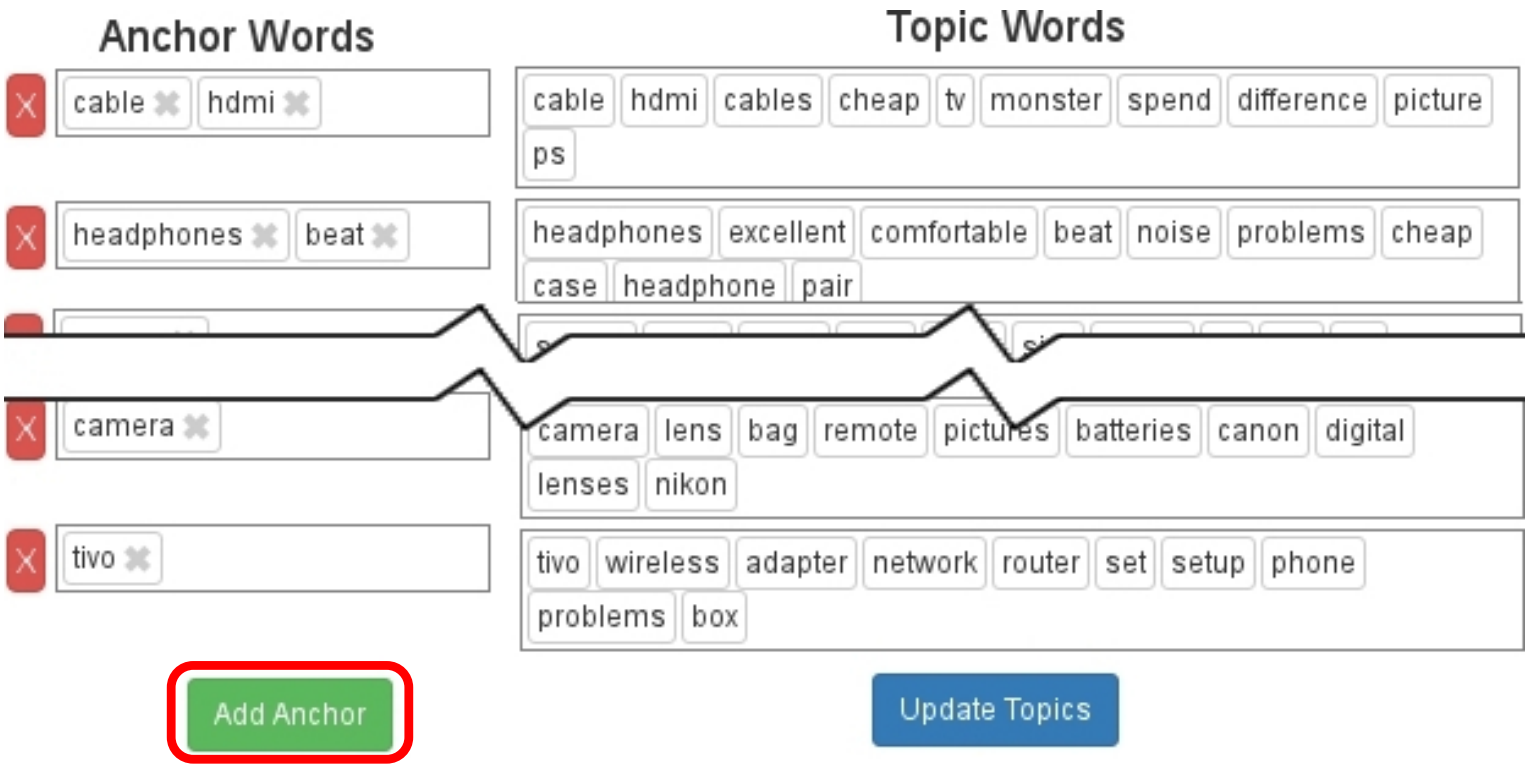
Interactive User Study Interface



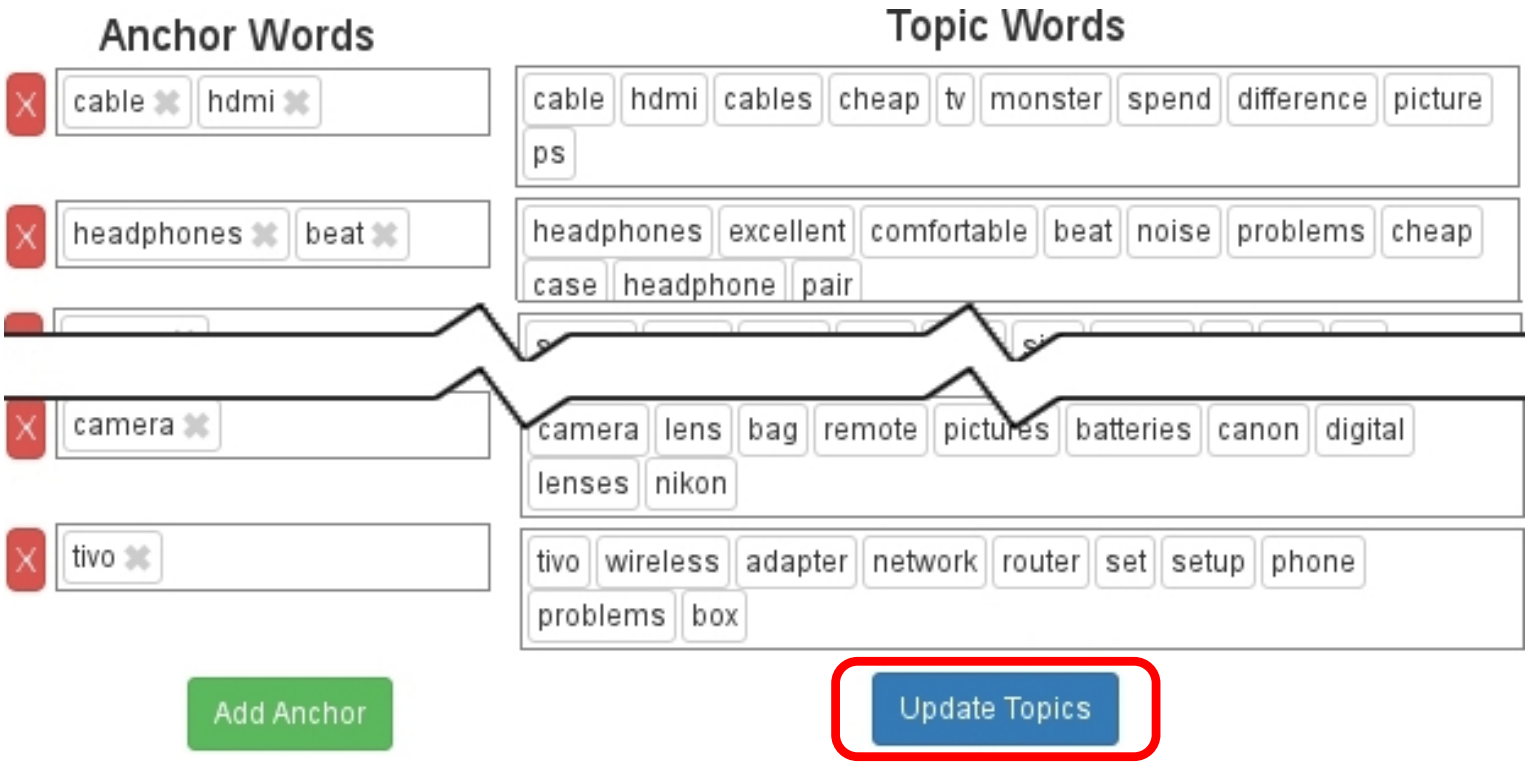
Interactive User Study Interface



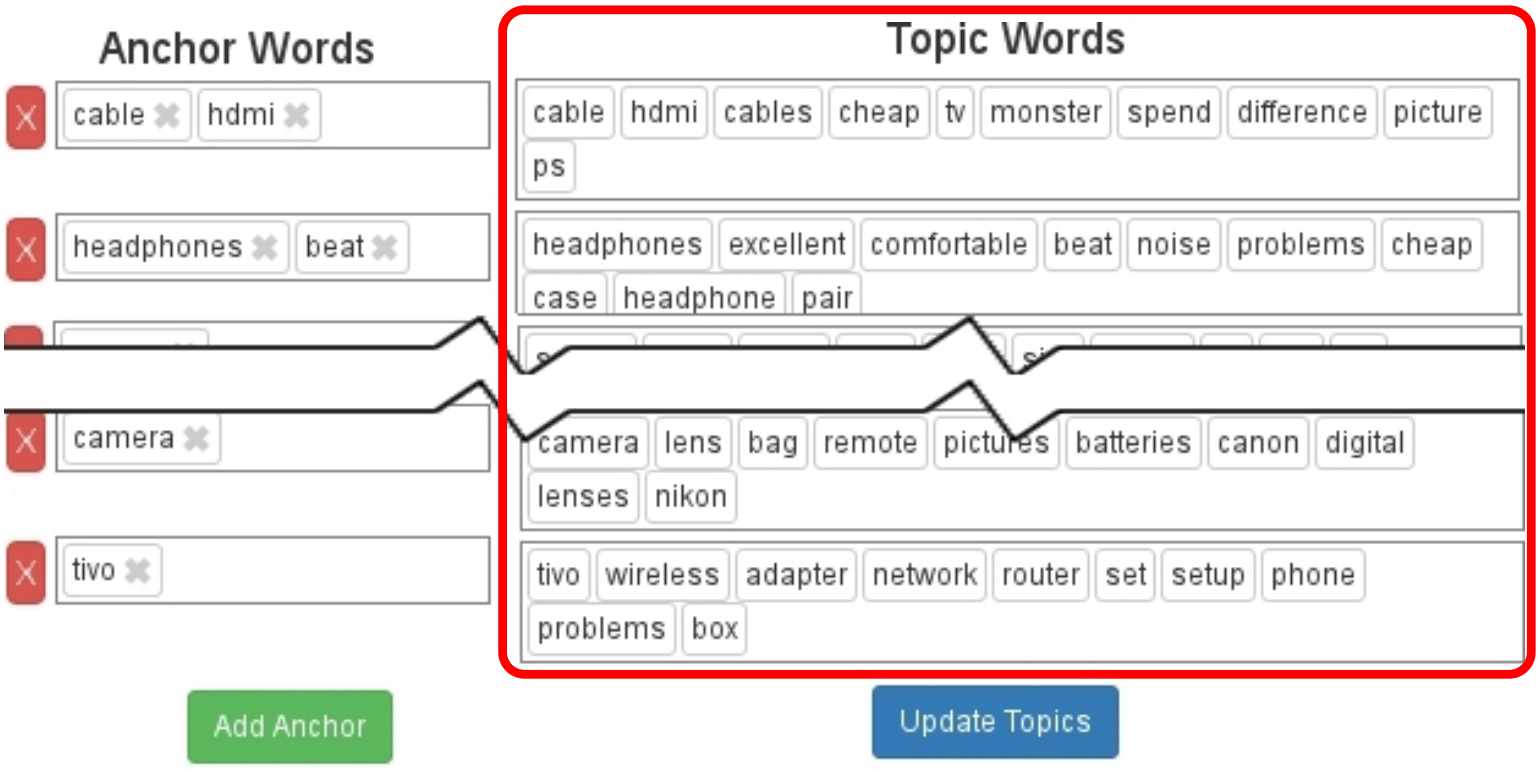
Interactive User Study Interface



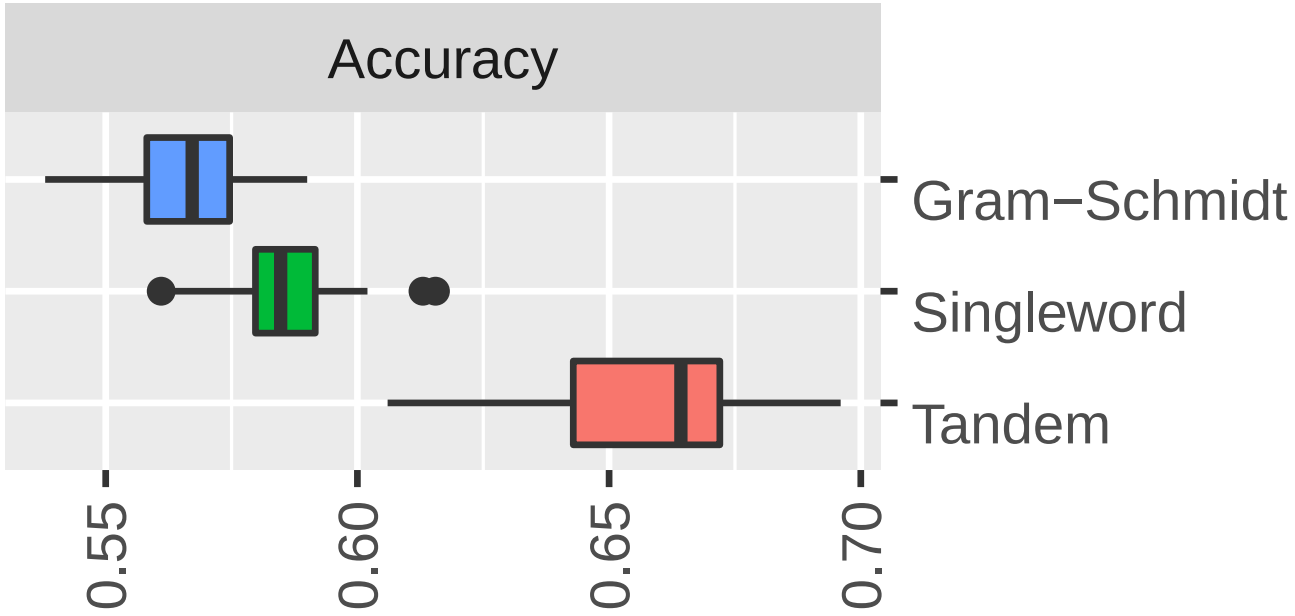
Interactive User Study Interface



Interactive User Study Interface

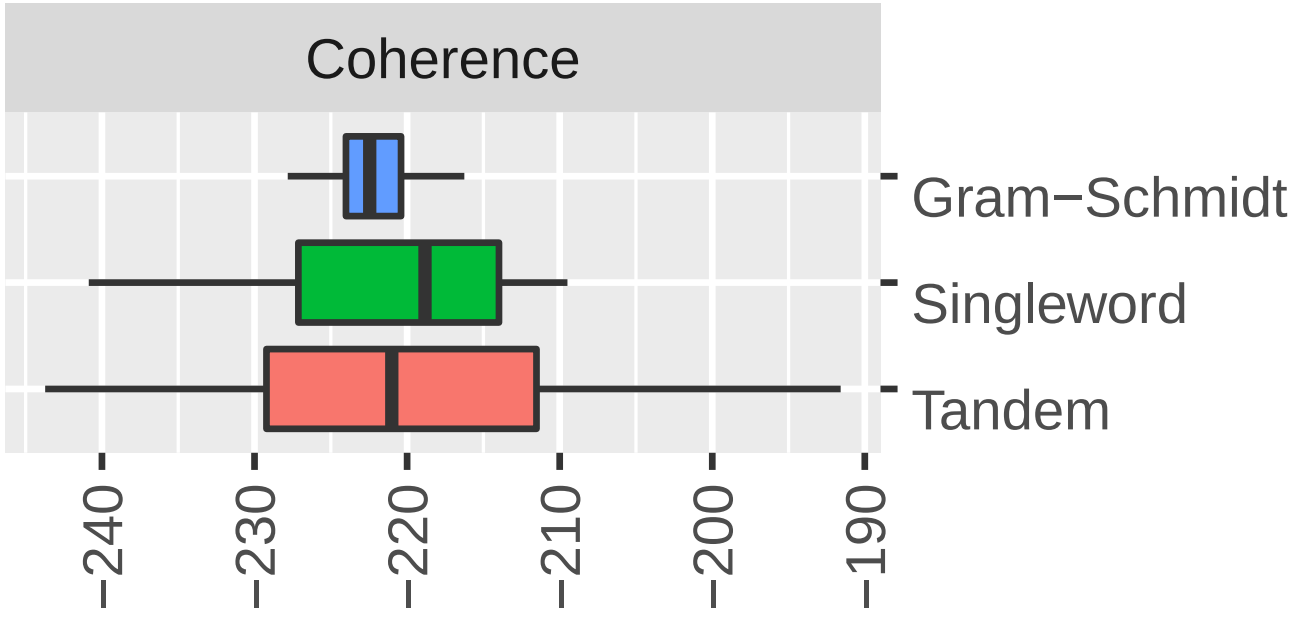


Interactive Topic Modeling Results



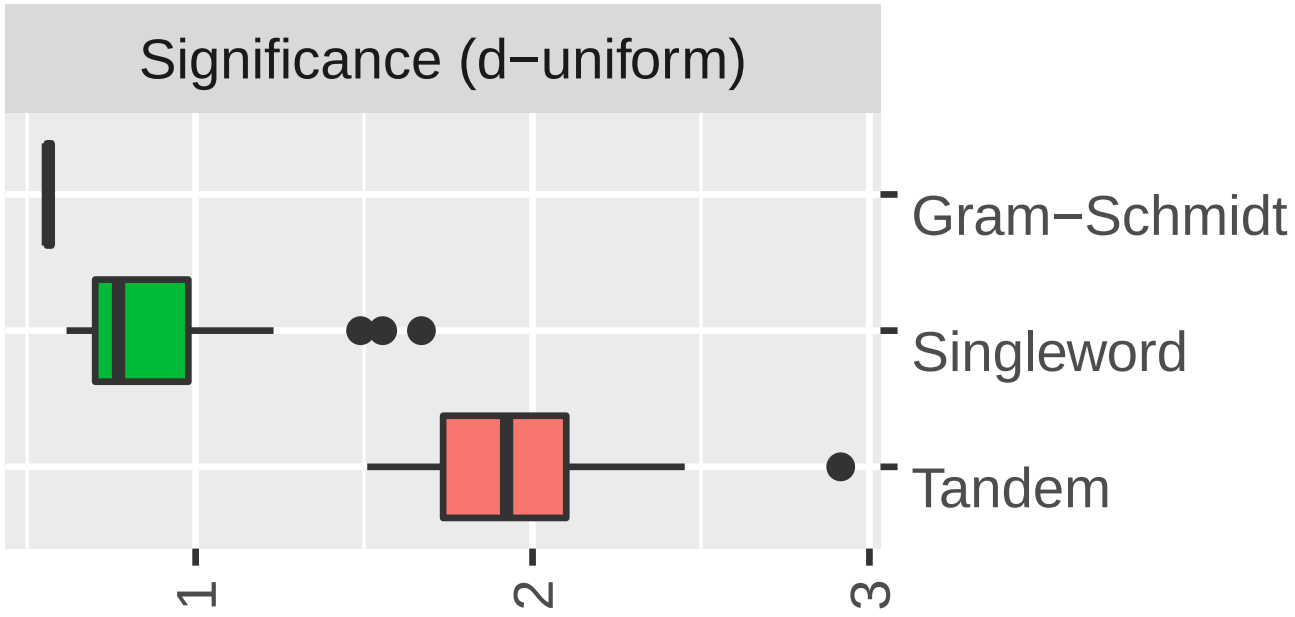
Topics from tandem anchors are superior features for classification

Interactive Topic Modeling Results



Differences in topic coherence are *not* statistically significant

Interactive Topic Modeling Results



- Topic significance measures distance of topic distributions from a background distributions*
- Topics from tandem anchors are more significant than single-word anchors

*AlSumait et al., 2009

Conclusion

Tandem anchors is:

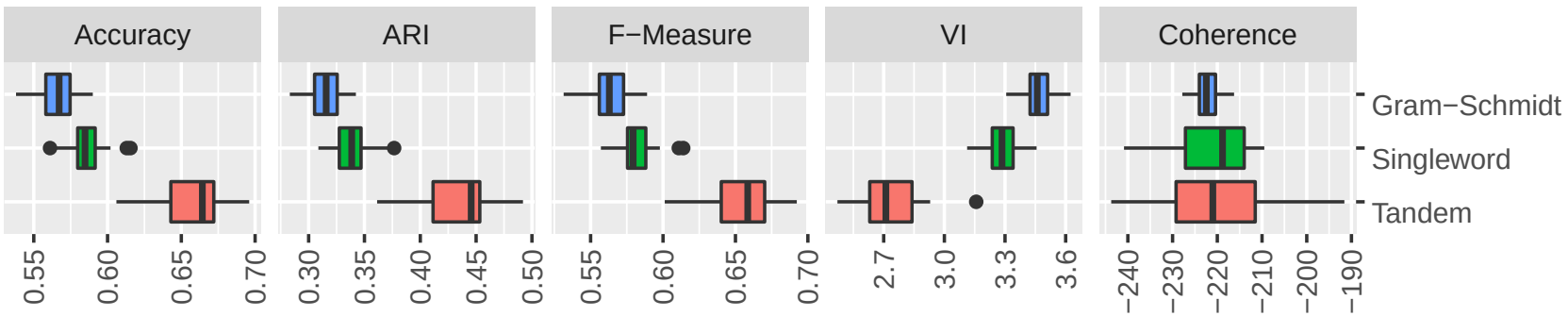
- More intuitive than single word anchors
- Produces superior topics compared to baseline
- Only interactive topic modeling algorithm that is actually interactive

Thank you!

Paper: http://cs.colorado.edu/~jbg/docs/2017_acl_multiword_anchors.pdf

Code: http://github.com/jefflund/ankura

Extra: Accuracy, Clustering and Coherence



Extra: Topic Significance