Evaluating Topic Model Dimensionality Reduction Performance

Nick Lines

JOHNS HOPKINS UNIVERSITY

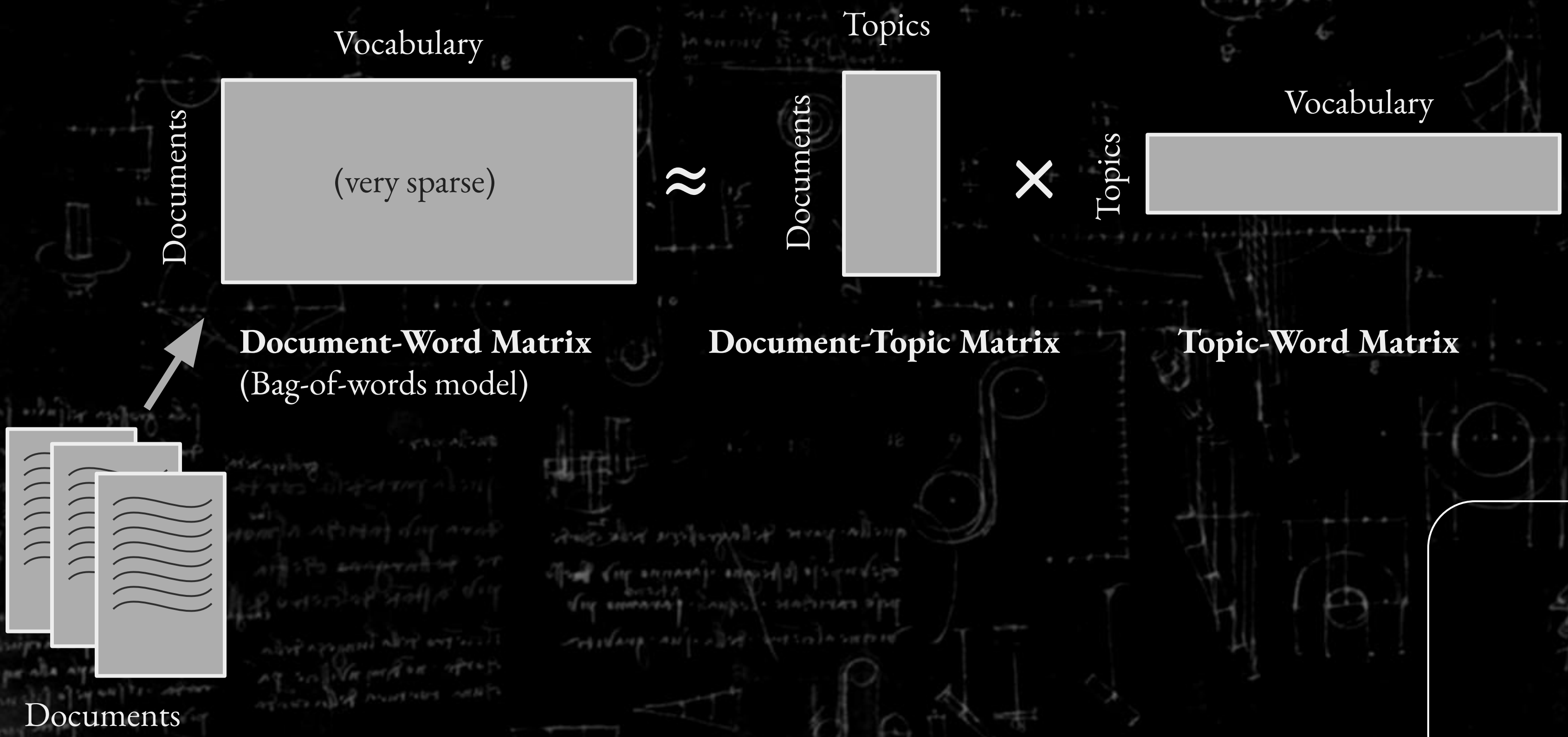
https://github.com/linesn/topic_model_dimensionality_reduction

Outline

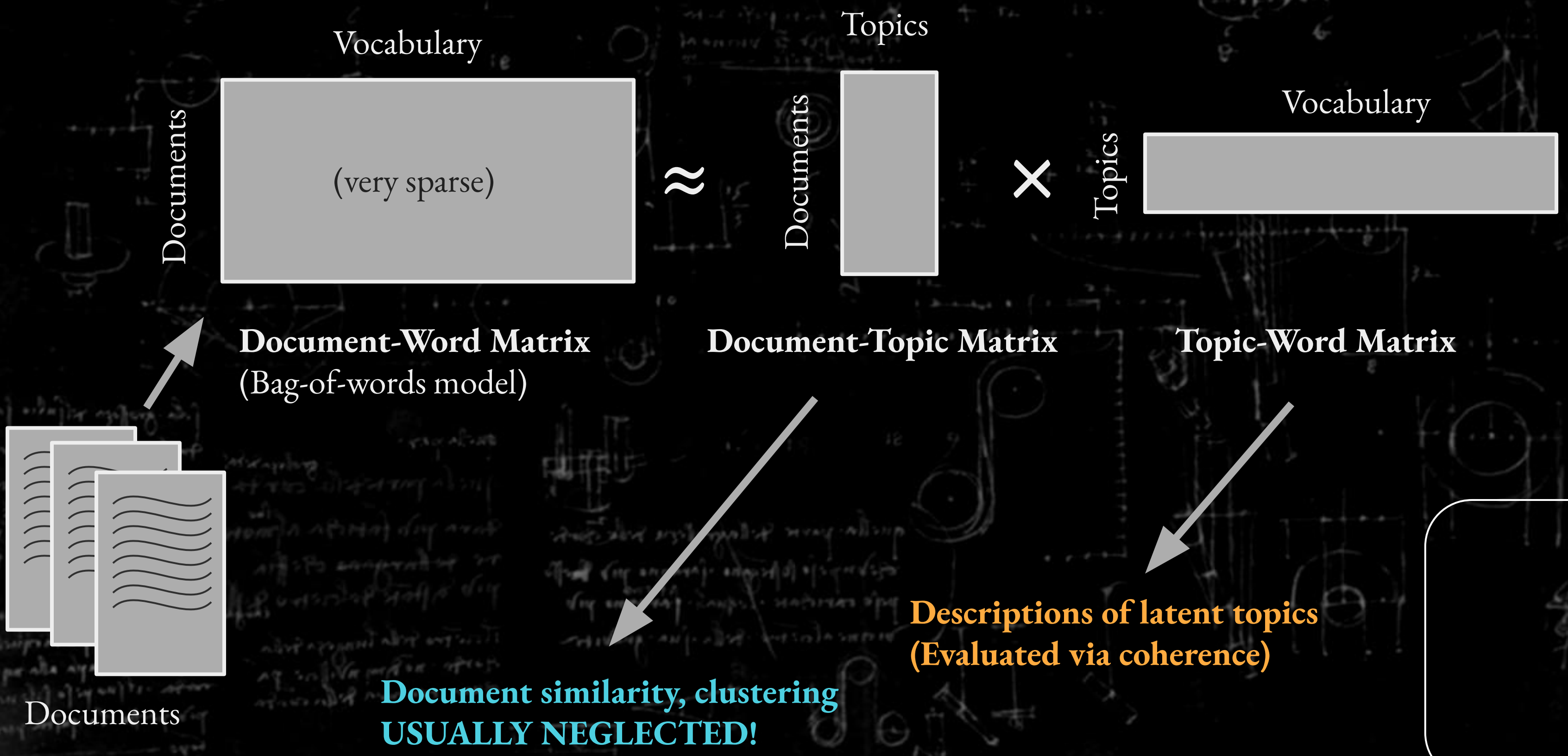
1. Topic Modeling
2. Evaluating Dimensionality Reduction
3. Experiment details
4. Results
5. Future work
6. Conclusion

1. Topic Modeling

Topic Modeling



Topic Modeling

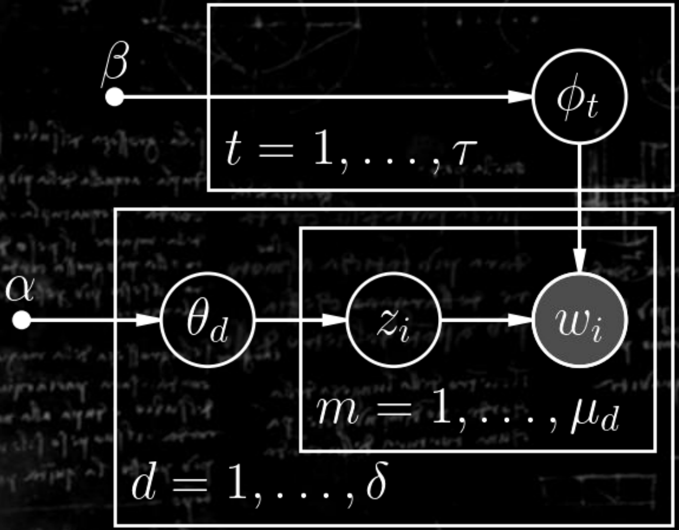


Recovering Latent Topics

Latent Dirichlet Allocation (LDA)

Pros: Probabilistic, very good at producing high-quality topics

Cons: Gibb’s sampling is non-parallelizable and slow



Non-negative Matrix Factorization (NMF)

Pros: Fast, easy to parallelize. Mathematically straightforward.

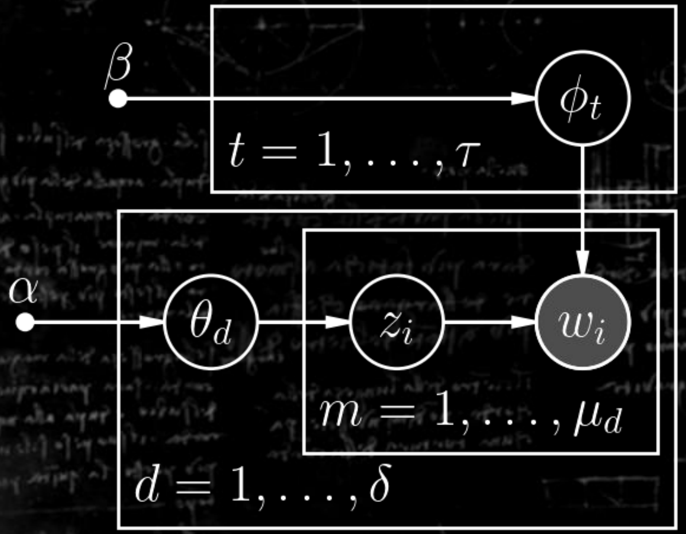
Cons: Inferior topics?

Recovering Latent Topics

Latent Dirichlet Allocation (LDA)

Pros: Probabilistic, very good at producing high-quality topics

Cons: Gibb's sampling is non-parallelizable and slow



Non-negative Matrix Factorization (NMF)

Pros: Fast, easy to parallelize. Mathematically straightforward.

Cons: Inferior topics?

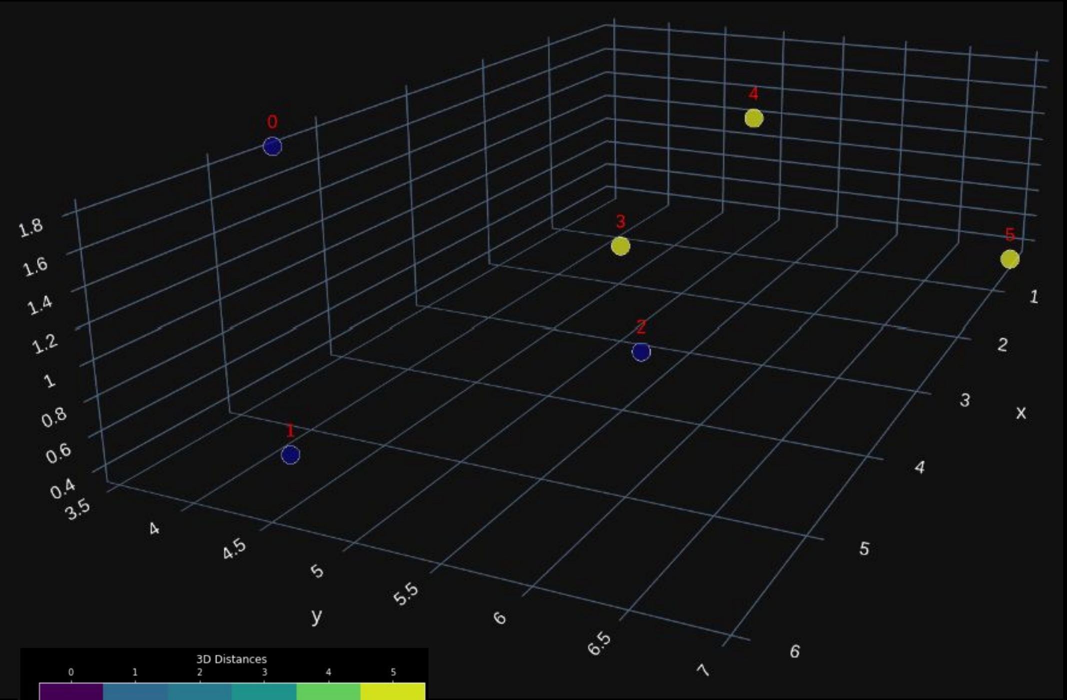
There is another…

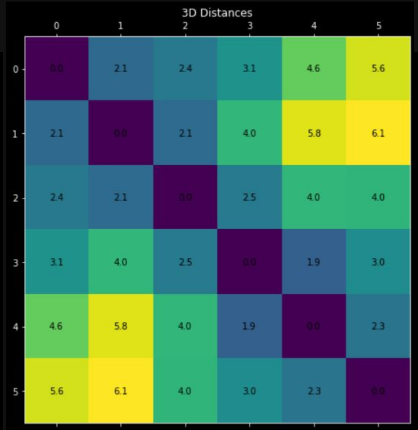
UNIFORM MANIFOLD
UMAP
APPROXIMATION & PROJECTION

And

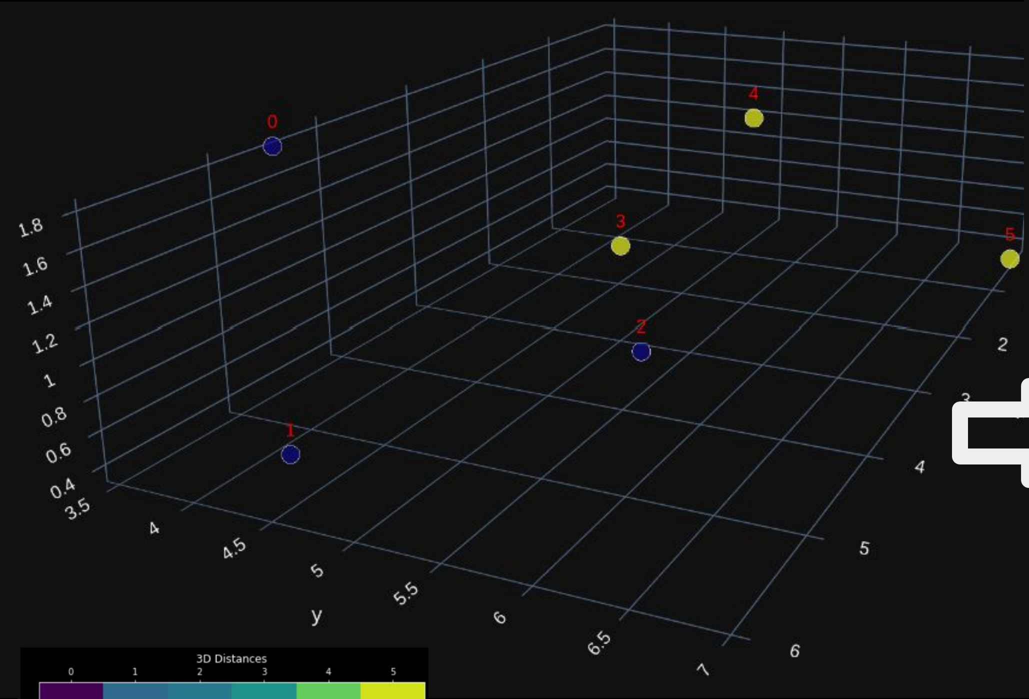
Top2Vec

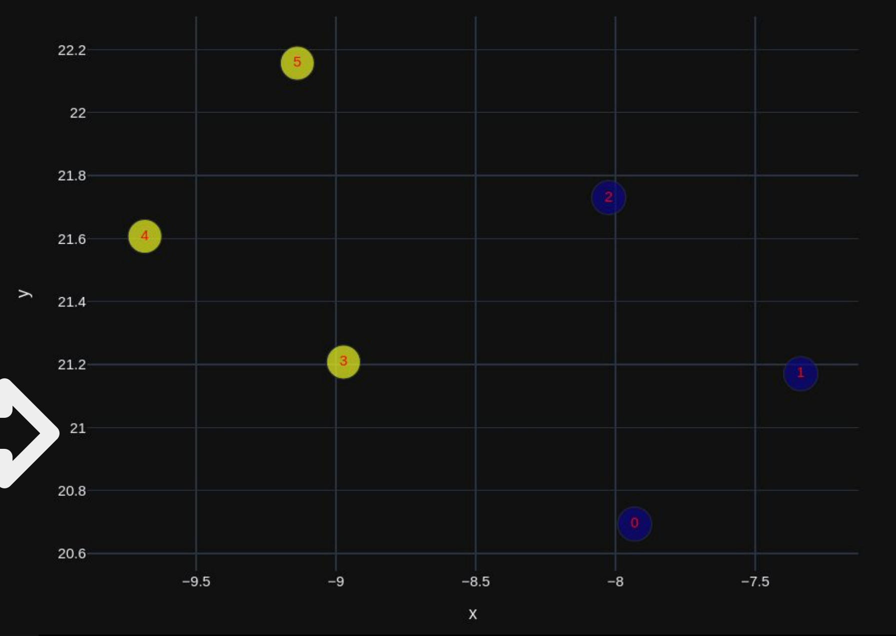
2. Evaluating Dimensionality Reduction





	0	1	2	3	4	5
0	0.0	2.1	2.4	3.1	4.6	5.6
1	2.1	0.0	2.1	4.0	5.8	6.1
2	2.4	2.1	0.0	2.5	4.0	4.0
3	3.1	4.0	2.5	0.0	1.9	3.0
4	4.6	5.8	4.0	1.9	0.0	2.3
5	5.6	6.1	4.0	3.0	2.3	0.0





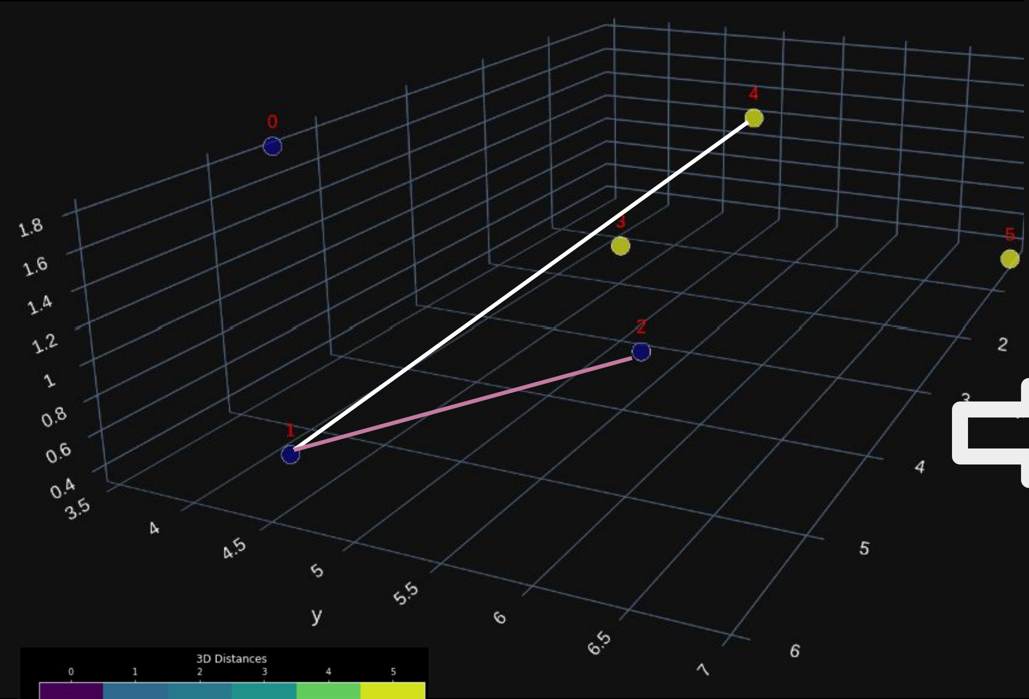
3D Distances

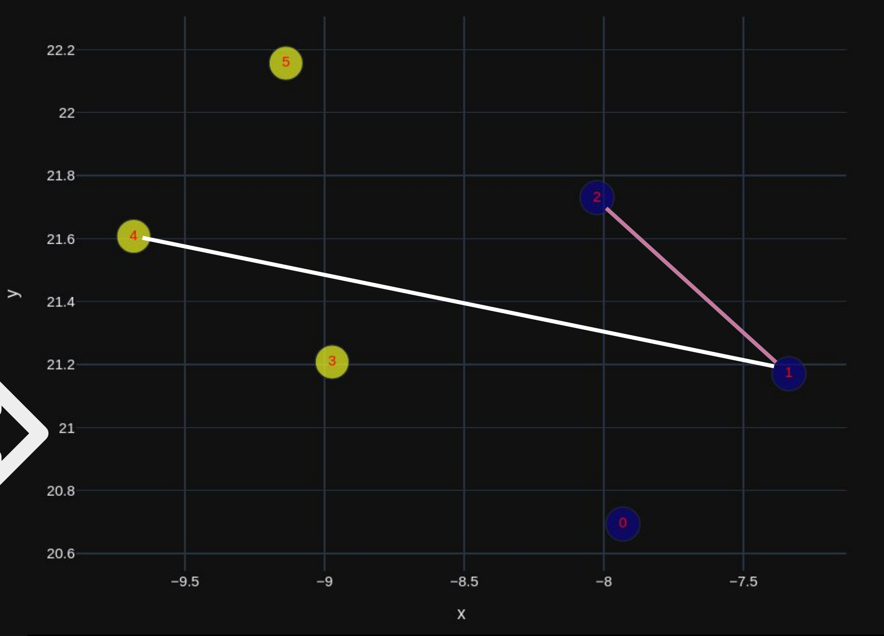
	0	1	2	3	4	5
0	0.0	2.1	2.4	3.1	4.6	5.6
1	2.1	0.0	2.1	4.0	5.8	6.1
2	2.4	2.1	0.0	2.5	4.0	4.0
3	3.1	4.0	2.5	0.0	1.9	3.0
4	4.6	5.8	4.0	1.9	0.0	2.3
5	5.6	6.1	4.0	3.0	2.3	0.0

Distances

2D Distances

	0	1	2	3	4	5
0	0.0	0.8	1.0	1.2	2.0	1.9
1	0.8	0.0	0.9	1.6	2.4	2.1
2	1.0	0.9	0.0	1.1	1.7	1.2
3	1.2	1.6	1.1	0.0	0.8	1.0
4	2.0	2.4	1.7	0.8	0.0	0.8
5	1.9	2.1	1.2	1.0	0.8	0.0





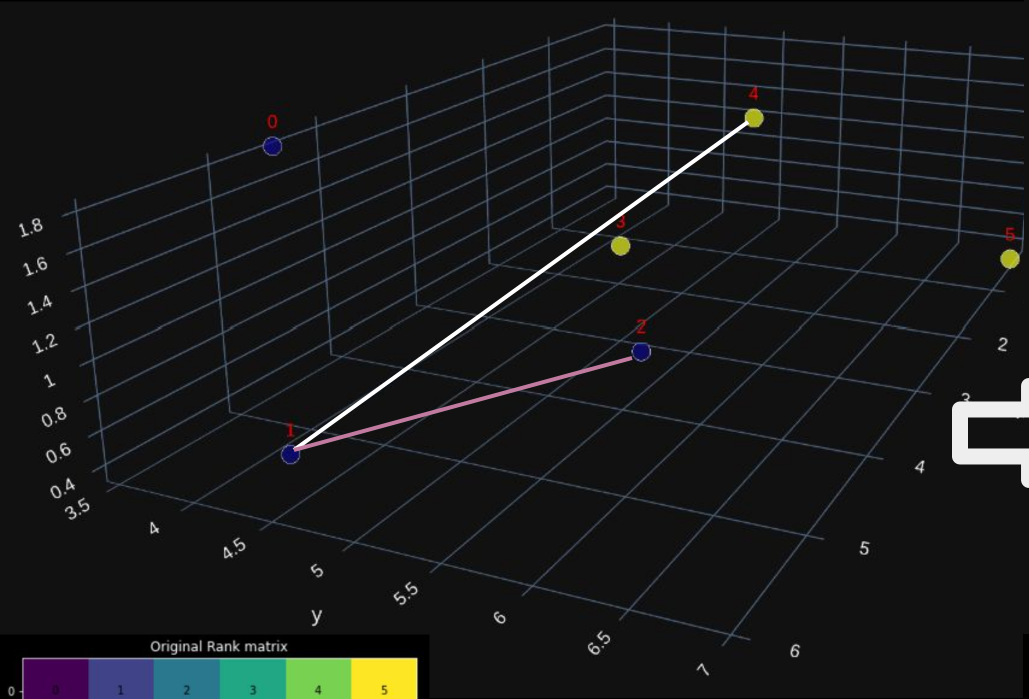
3D Distances

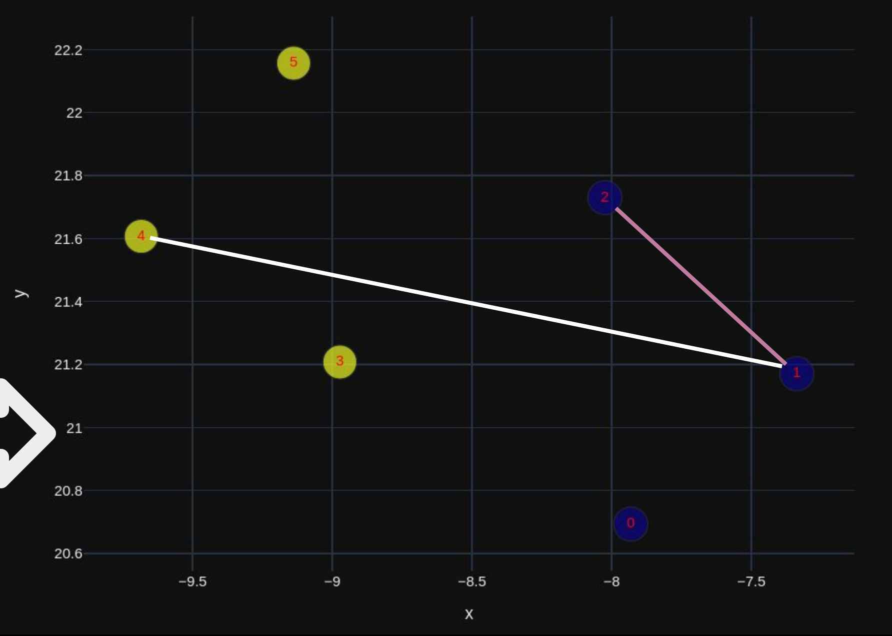
	0	1	2	3	4	5
0	0.0	2.1	2.4	3.1	4.6	5.6
1	2.1	0.0	2.1	4.0	5.8	6.1
2	2.4	2.1	0.0	2.5	4.0	4.0
3	3.1	4.0	2.5	0.0	1.9	3.0
4	4.6	5.8	4.0	1.9	0.0	2.3
5	5.6	6.1	4.0	3.0	2.3	0.0

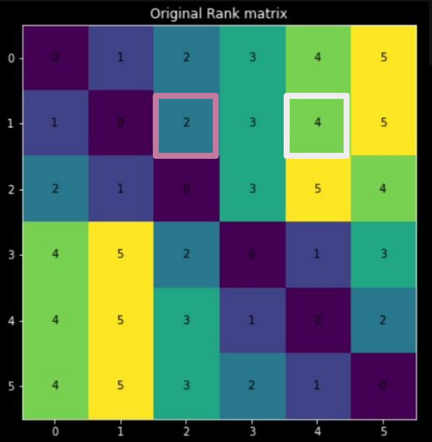
Distances

2D Distances

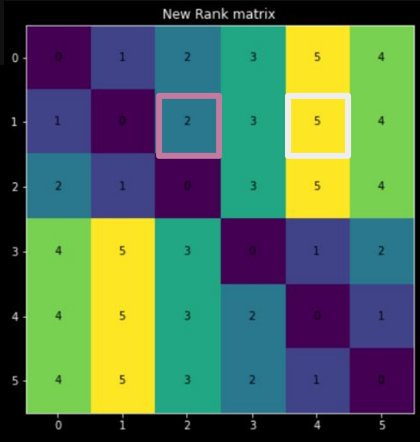
	0	1	2	3	4	5
0	0.0	0.8	1.0	1.2	2.0	1.9
1	0.8	0.0	0.9	1.6	2.4	2.1
2	1.0	0.9	0.0	1.1	1.7	1.2
3	1.2	1.6	1.1	0.0	0.8	1.0
4	2.0	2.4	1.7	0.8	0.0	0.8
5	1.9	2.1	1.2	1.0	0.8	0.0

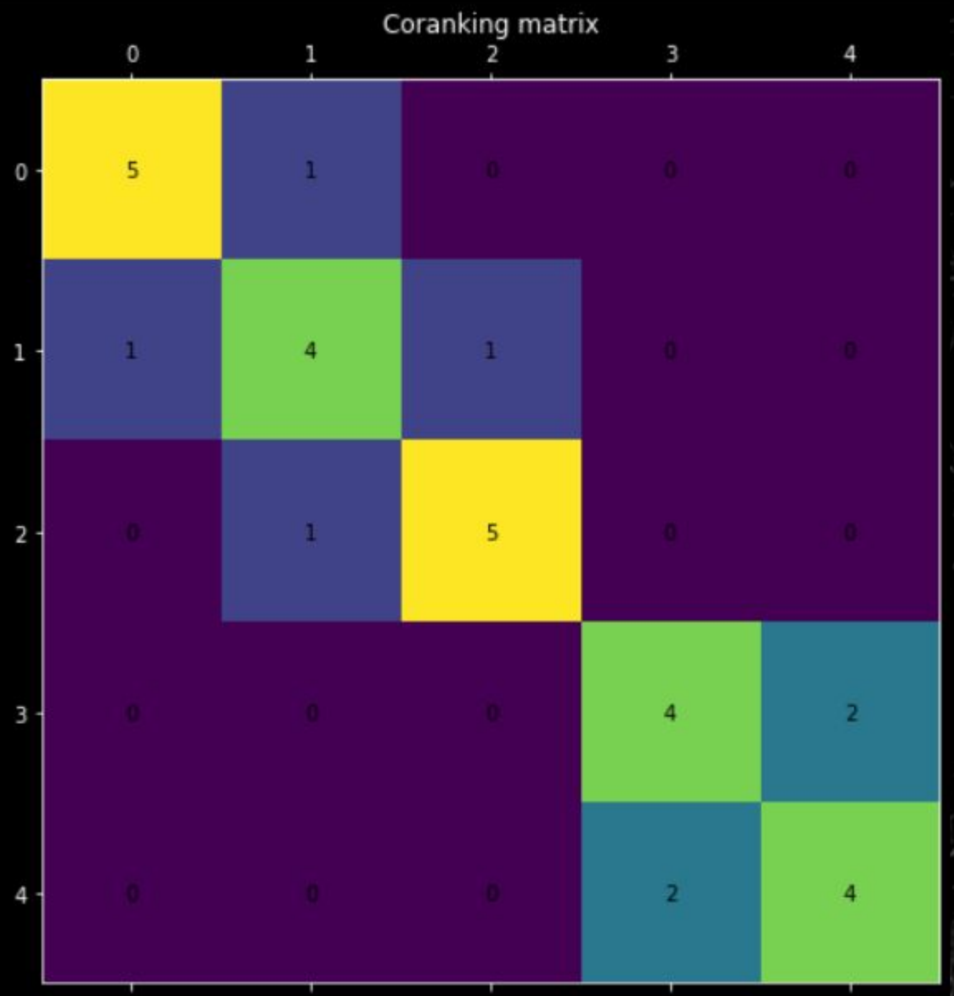

0
1
2
3
4
5
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
3.5
4
4.5
5
5.5
6
6.5
7
y
2
3
4
5
6


22.2
22
21.8
21.6
21.4
21.2
21
20.8
20.6
y
−9.5
−9
−8.5
−8
−7.5
x
0
1
2
3
4
5


Original Rank matrix
0 1 2 3 4 5
1 0 2 3 4 5
2 1 0 3 5 4
4 5 2 0 1 3
4 5 3 1 0 2
4 5 3 2 1 0

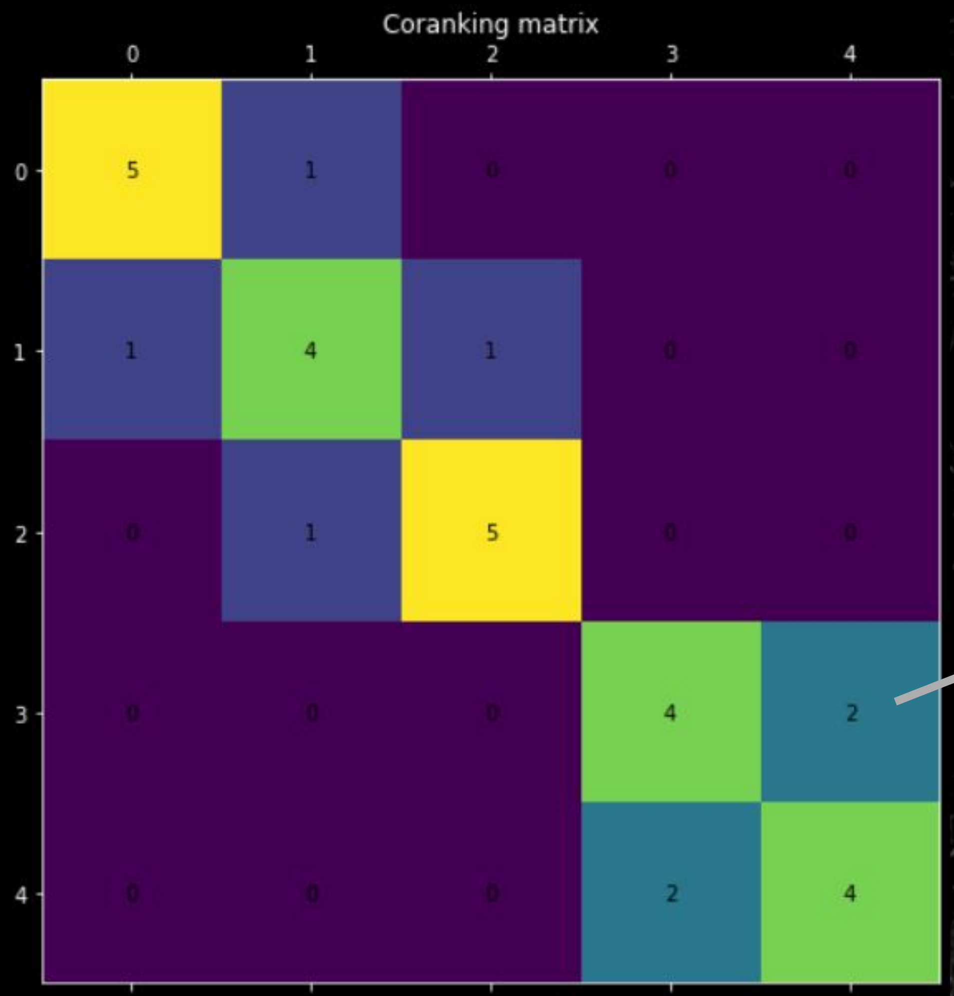
Ranking


New Rank matrix
0 1 2 3 5 4
1 0 2 3 5 4
2 1 0 3 5 4
4 5 3 0 1 2
4 5 3 2 0 1
4 5 3 2 1 0



The Coranking Matrix

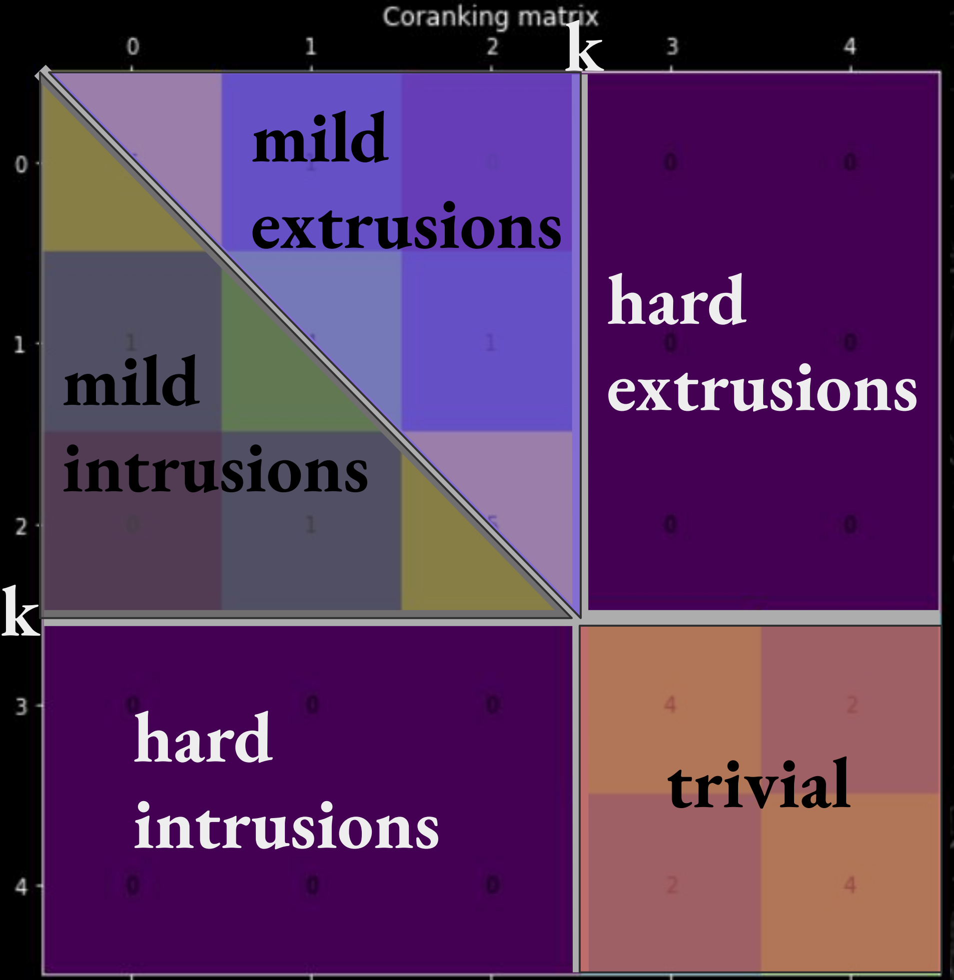
$$Q_{k,l} = \#\left\{(i,j) : R_{i,j} = k \text{ and } R'_{i,j} = l\right\}$$



The Coranking Matrix

$$Q_{k,l} = \#\left\{(i,j) : R_{i,j} = k \text{ and } R'_{i,j} = l\right\}$$

There are 2 times that a point had its 3rd closest neighbor re-ordered to be the 4th closest neighbor

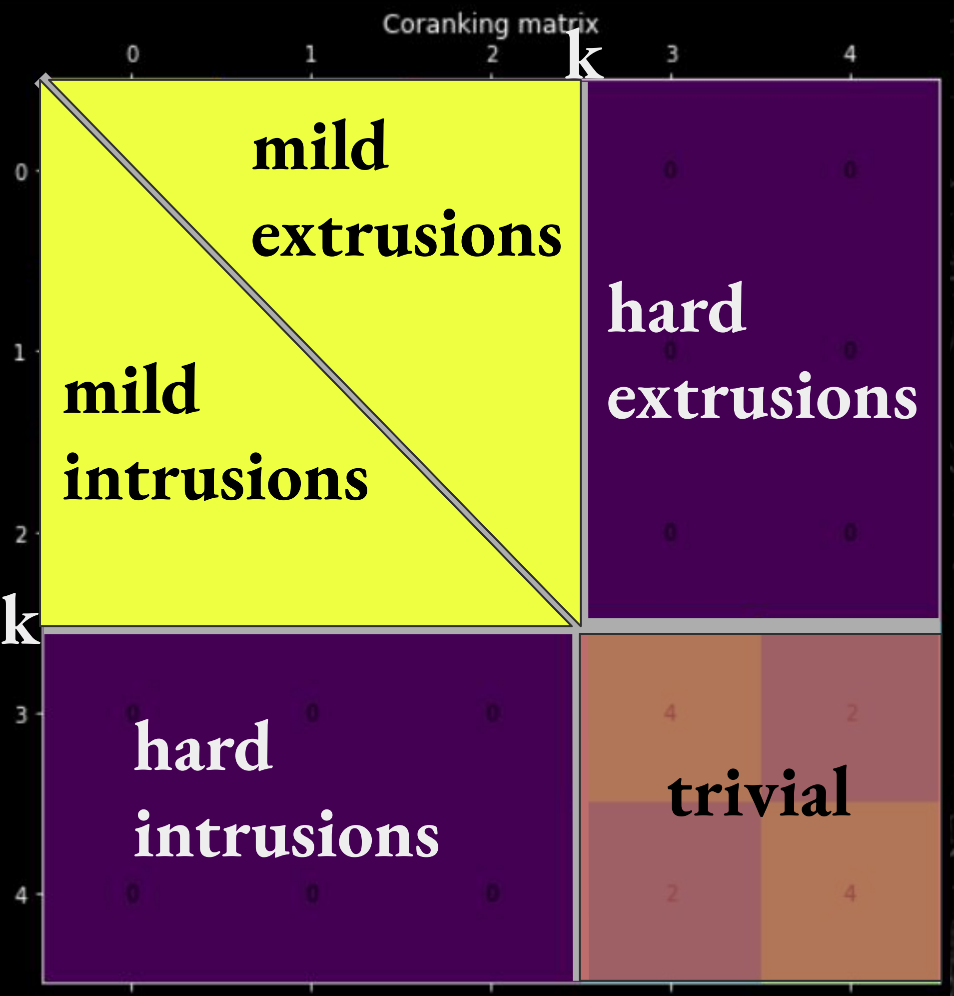


The Coranking Matrix

Extrusions: nearby points are pushed further away in rank (related to continuity)

Intrusions: far off points are pulled closer in rank (related to trustworthiness)

k: the neighborhood size.



The Coranking Matrix

$\mathbf{Q_{NN}(k)}$ sums up the highlighted region and divides by **k** for each **k.**

The Area Under this Curve gives a single real number measuring how well local structure is preserved.

3. Experiment Details

Research Question: How do vanilla LDA and NMF Compare in

- Local Dimensionality Reduction quality (AUC of $Q_{NN}(k)$)
- Topic quality (mean U_mass coherence)
- Algorithm run time (seconds)
- Normalized Reconstruction error

Hypothesis: NMF will perform better in the Local Dimensionality Reduction and run time categories, and LDA will get better topic quality and lower reconstruction error.

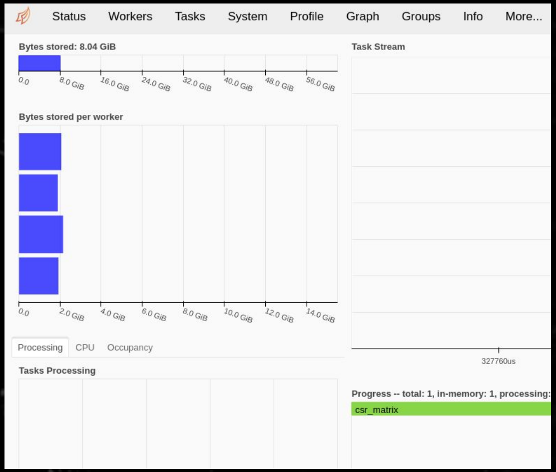
We expect BOTH to perform worse than UMAP in Local Dimensionality Reduction.

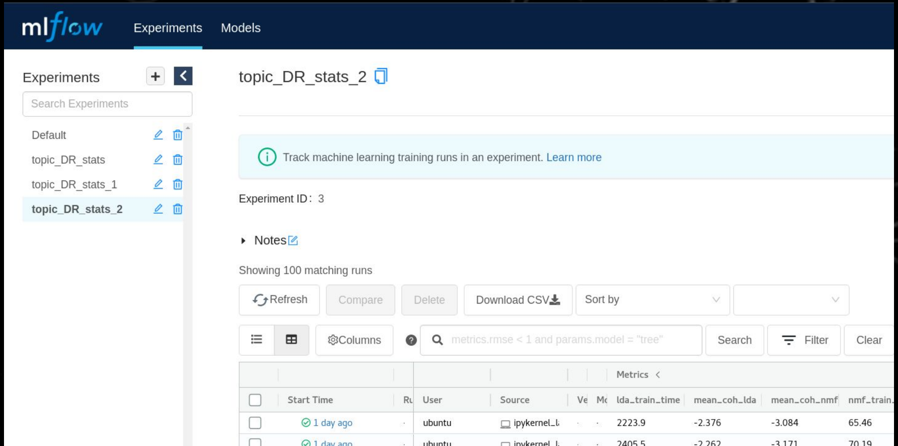
Data: English News articles from CNN Daily News

In our experiment we use the first 2000 articles. We'll perform English stemming and vocabulary reduction. Each algorithm will receive precisely the same input matrix.

Parameter space:

We'll build and test models for t=2,...,200 topics, using parallel job processing and model analysis with Dask and MLflow.



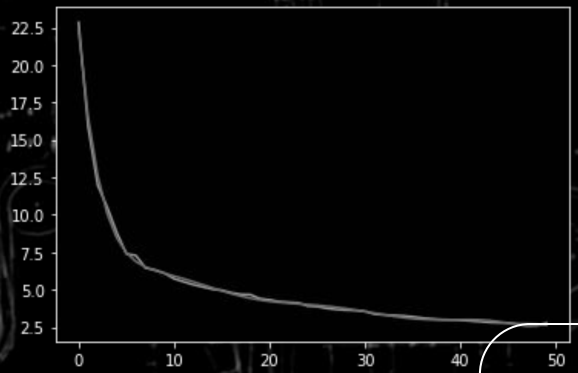


4. Experiment Results

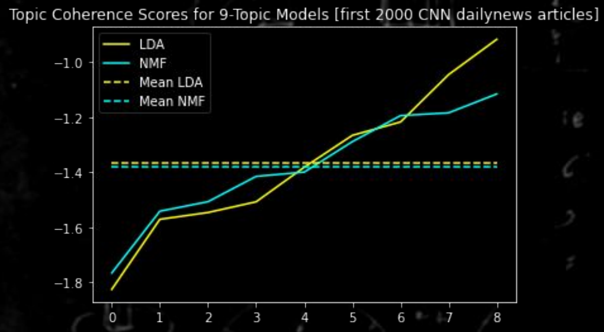
The “right topic number”

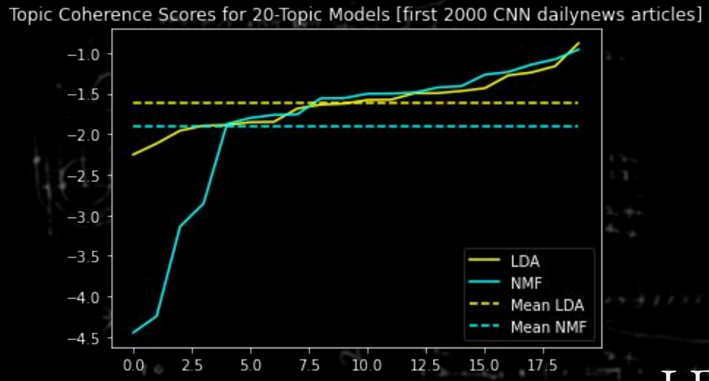
```
TOPIC MODELS:
   LDA
Topic 0: polic told offic court case investig charg accord
Topic 1: like just know think work make day want
Topic 2: attack govern militari forc al state secur countri
Topic 3: china countri world govern south chines north korea
Topic 4: use new million compani busi make like site
Topic 5: game team player play world win second match
Topic 6: health school student children food women care help
Topic 7: citi water area flight plane offici airport accord
Topic 8: presid obama state american republican elect democrat hous
   NMF
Topic 0: like just think work know make new day
Topic 1: govern attack militari forc al countri kill group
Topic 2: obama presid republican elect democrat hous state american
Topic 3: polic told court offic investig charg case famili
Topic 4: compani million new use china world like 000
Topic 5: flight citi plane water offici home area airport
Topic 6: game team world player play win second sport
Topic 7: state student school iran north korea unit nation
Topic 8: health children care medic hospit famili patient doctor
```

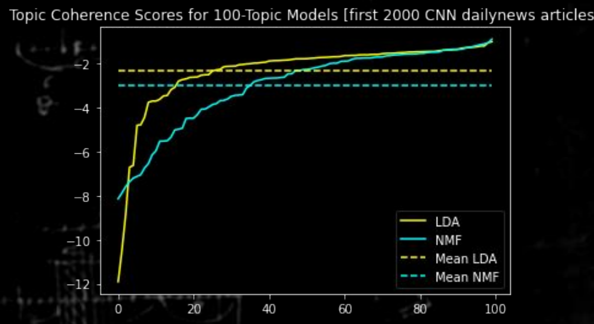
PCA Suggests that the best fitting number of topics parameter choice is 9

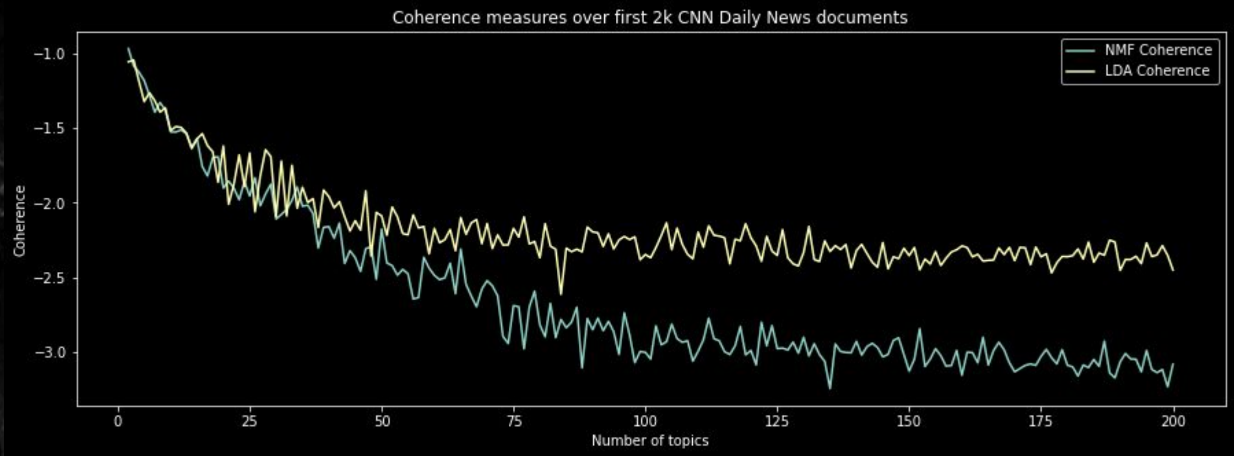


Coherence (topic quality)



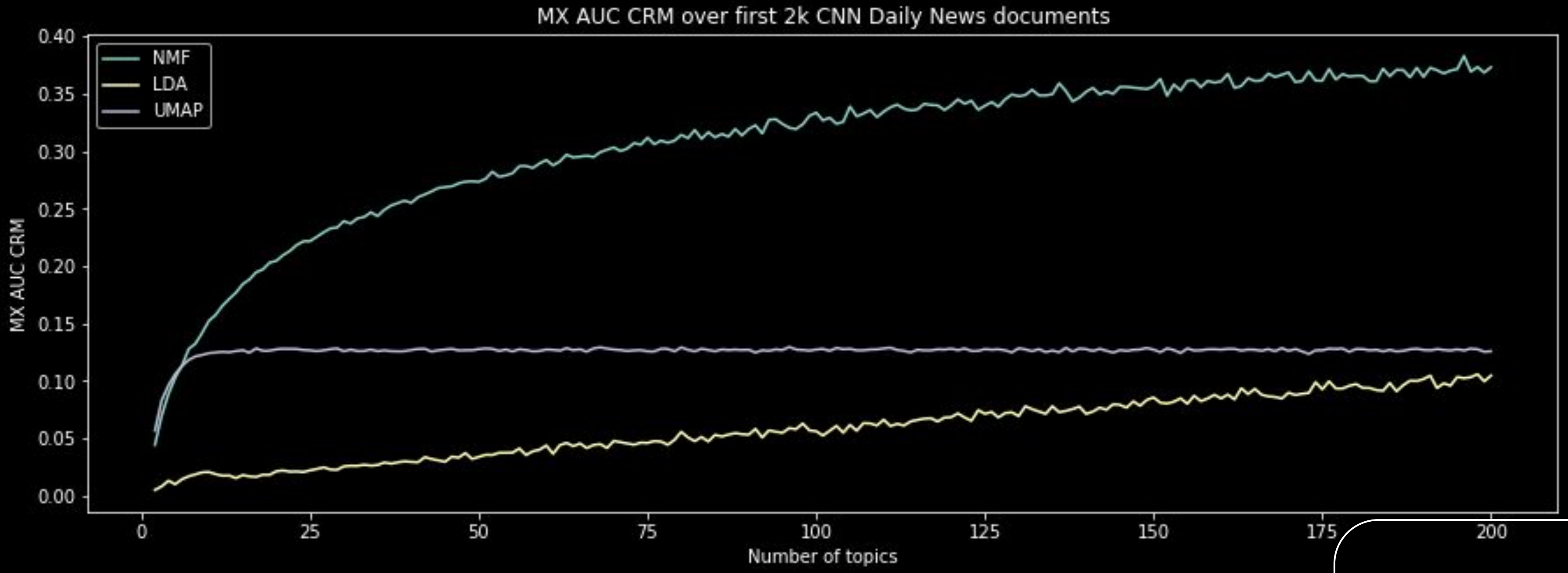




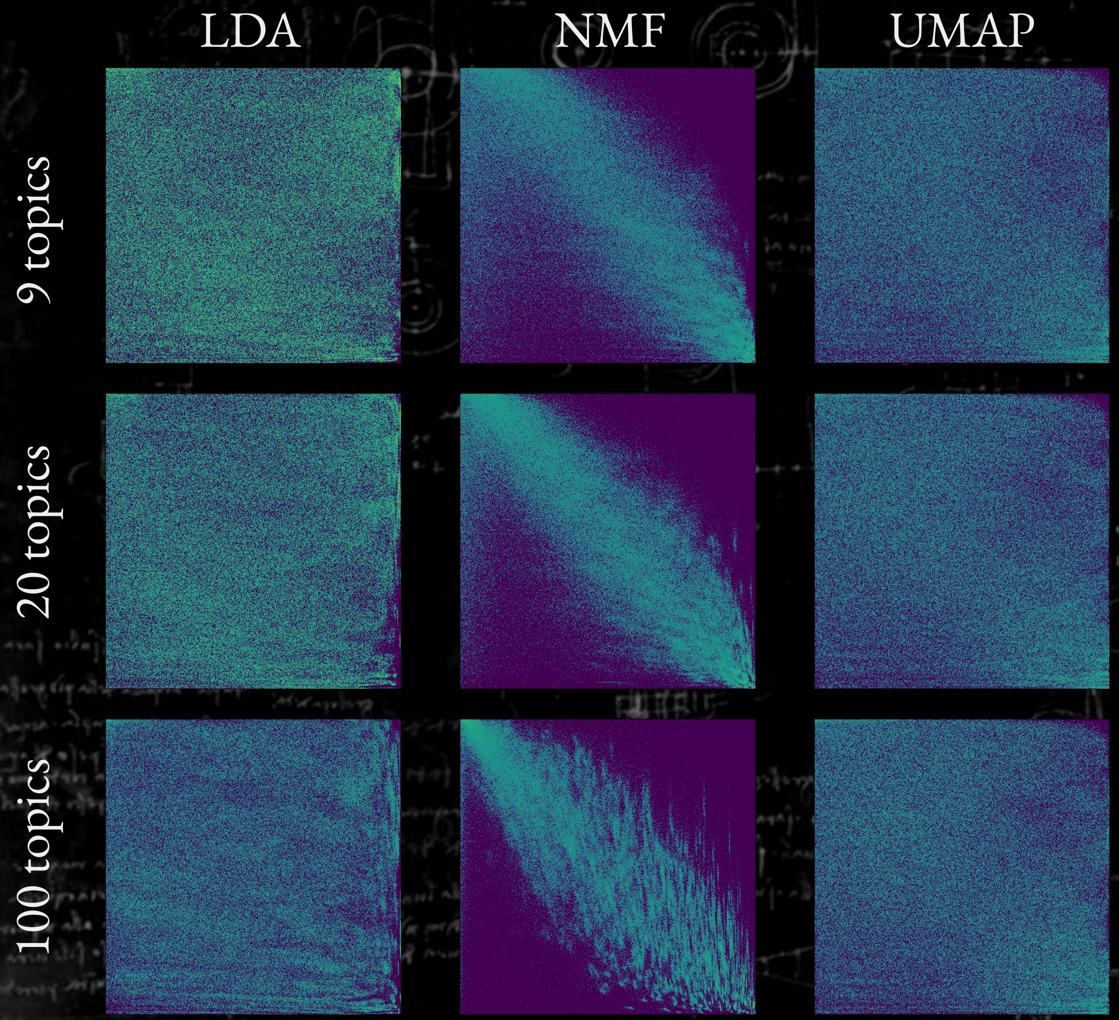


LDA doesn’t become a clear winner in terms of coherence until about 50 topics or more are used.

Dimensionality Reduction Quality



NMF performs slightly worse than UMAP at first, but soon becomes the clear winner. LDA's local structure preservation is little better than random (zero).

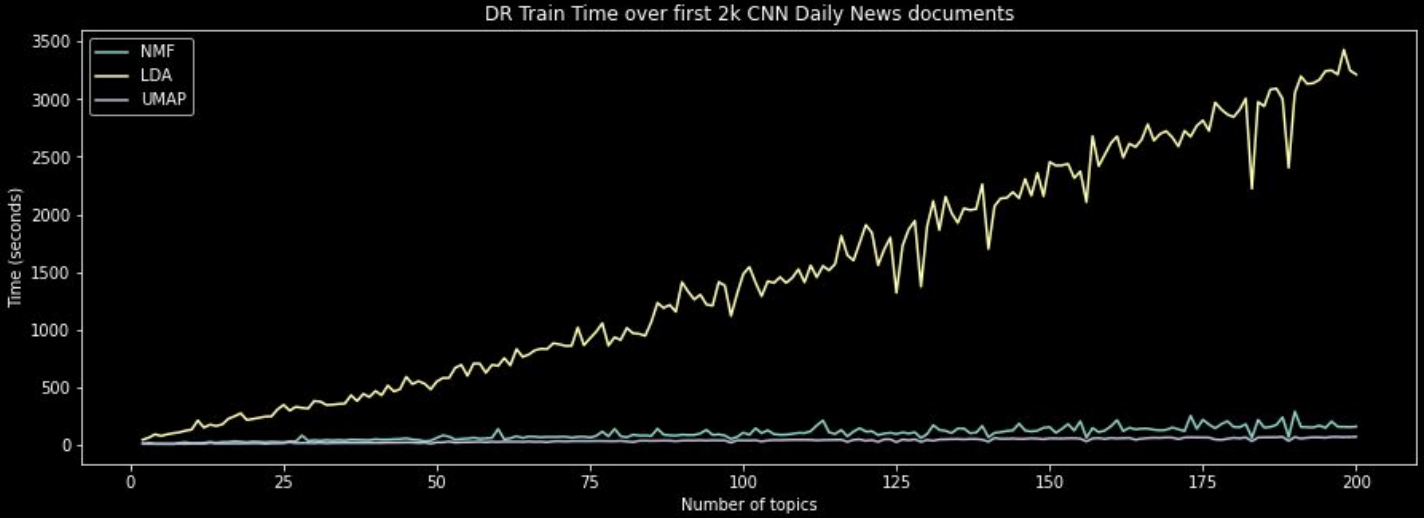


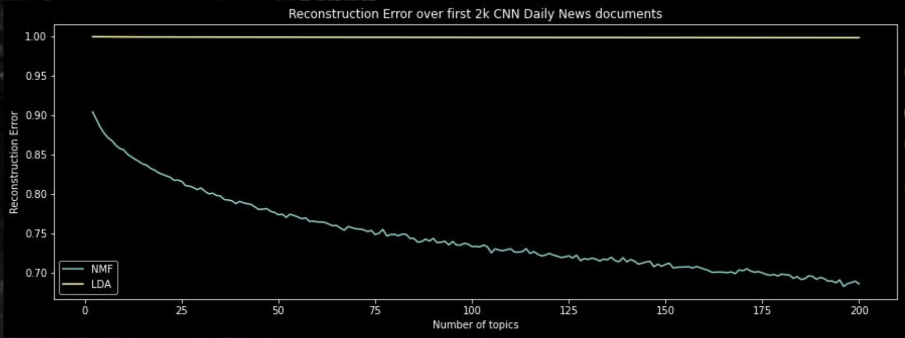
Dimensionality Reduction Quality

NMF does a great job of reducing extrusions, and in higher dimensions of reducing intrusions as well. UMAP really only focuses on hard extrusions/intrusions, and LDA is just a mess in general.

Run time and reconstruction error

LDA scales with the number of topics t, and eventually becomes 4 orders of magnitude slower than NMF!





By design, LDA will always reproduce the document-word matrix perfectly. NMF's reconstruction error gets worse in higher dimensions.

5. Future Work

Future work

- Repeat with alternative metrics for DR quality, e.g. Local Property Metric
- Repeat with alternative metrics for Topic quality, e.g. other coherence measures.
- Repeat in other languages and other datasets, e.g. Chinese news, English microblogs.

- Investigate why UMAP does not show up better in the metrics used here, and whether top2vec performs better with NMF or UMAP as the dimensionality reduction component.

6. Conclusion

Conclusions

- For small choices (<50) for the number of topics parameter, there's no reason not to use NMF over LDA.
- For higher topic dimensions, we need to be conscious of the trade-off between topic quality and dimensionality reduction quality LDA introduces.

Thanks for listening!

Credits and references:

- The UMAP logo is the property of the UMAP team.
- The speed-optimized AUC computation I used was written by Tim Sainburg. It's a nice bit of code!
- All images in this presentation, including the background, are original to me.
- Please see the details paper on GitHub for a full reference list of relevant papers.

Questions?

- Please reach out to me at nicholasalines@gmail.com or drop by the github project!