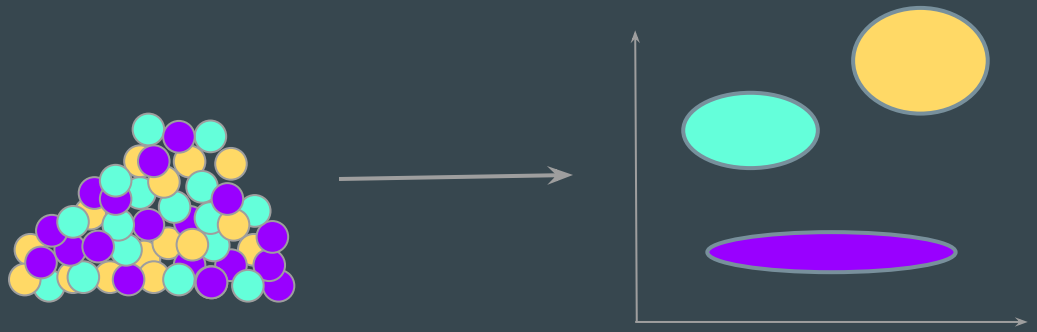
Effects of dimensionality reduction on clustering precision

...

Anna Sophia Stein • Advanced NLP with Python • 04.10.23

Motivation - What are clustering algorithms?

- Unsupervised learning algorithms
- Look for structure in data
- **Clusters** = data points that are similar in some way
- Based on proximity of data points (most of the time)



Motivation - What is dimensionality reduction?

Why not just use throw word embeddings into a clustering algorithm?

- Noise
  - **Noise** = unwanted or irrelevant information
  - Almost all data has some noise in it
- Too many dimensions
  - The more dimensions, the more potential for noise
  - The more dimensions, the less unique data points and unique information capture by them

Dimensionality reduction techniques can help reduce dimensions and noise!

Motivation - Dimensionality reduction techniques

- Principal component analysis (PCA) (Hotelling 1933)
  - Reshuffles your original variables into new variables (**principal components**)
  - Can correspond broadly to your original data but don't have to
  - Components that explain the most variance from original data become new variables
  - ➔ Some data loss

- T-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten, Hinton 2008)
  - Projects variables from original dimensions onto two or three dimensions
  - ➔ Some data loss

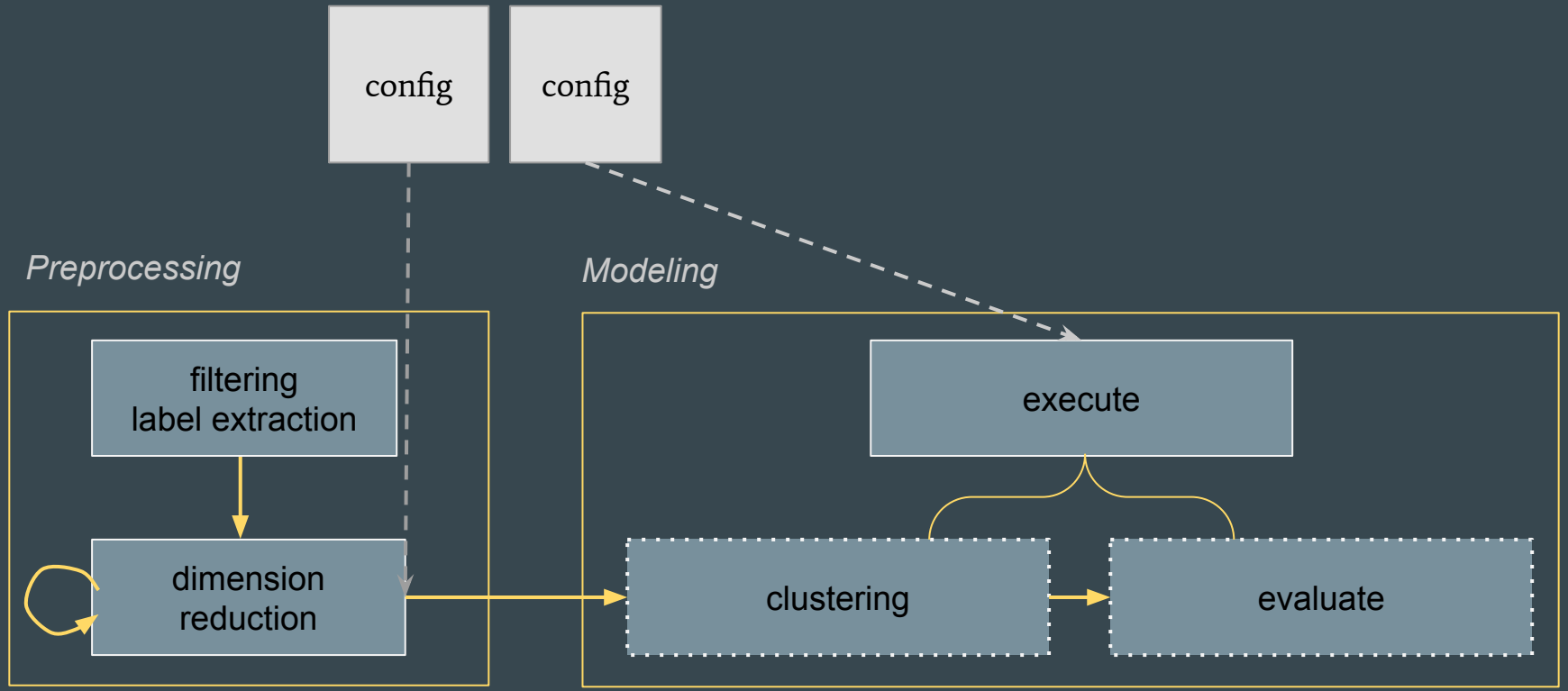
Research question

How do PCA and t-SNE influence the performance of the clustering algorithms k-means and DBSCAN?

Data

- Word embeddings from a trained FastText model (Bojanowski, et al. 2016)
- 14 cases for each noun
- Total of 603,286 forms---preprocessing----> 68,687 individual forms

Method - Pipeline



Applying the pipeline

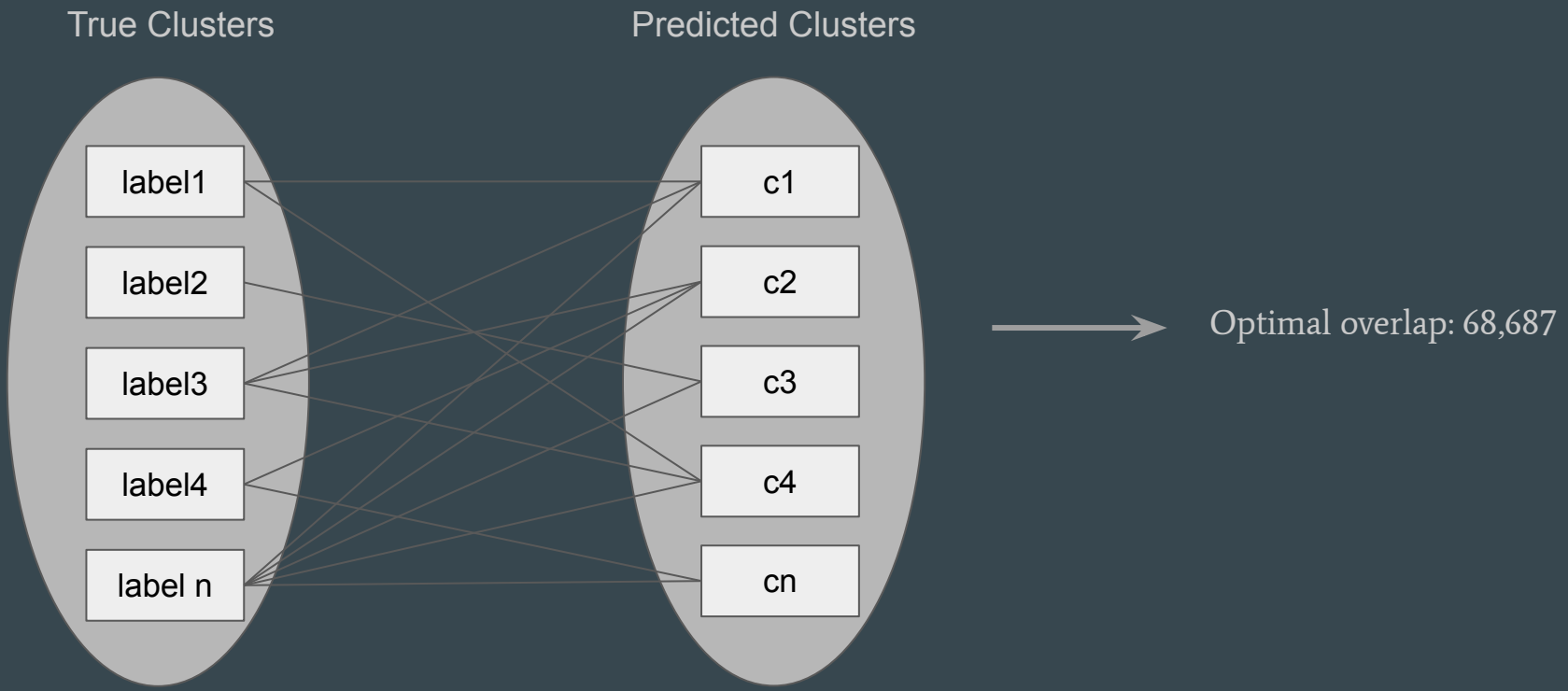
Input combinations:

- PCA with 215/300 components (95% explained variance)
- PCA with 277/300 components (99% explained variance)
- t-SNE with 2 dimensions
- t-SNE with 3 dimensions

Data sets:

PCA 215, t-SNE 2	PCA 277, t-SNE 2
PCA 215, t-SNE 3	PCA 277, t-SNE 3

Evaluation


True Clusters
Predicted Clusters
label1
label2
label3
label4
label n
c1
c2
c3
c4
cn
Optimal overlap: 68,687

Results

How do PCA and t-SNE influence the performance of the clustering algorithms k-means and DBSCAN?

PCA, t-SNE	k-means	DBSCAN
215, 2	-55,774	-55,326
215, 3	-52,272	-46,456
277, 2	-52,463	-50,844
277, 3	-48,264	-43,681

PCA, t-SNE	N clusters	Epsilon
215, 2	20	0.8
215, 3	19	1.8
277, 2	20	0.8
277, 3	20	1.8

Future Analyses

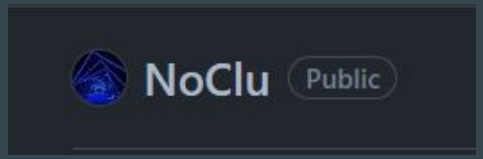
- More parameters!
- More clustering algorithms!
- More detailed analysis of predicted clusters

References

- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Thank you!

Steal my code here: github.com/ansost/NoClu



Minimum Cost Maximum Flow Algorithm with networkx