Semantic Non-Negative Matrix Tri-Factorization for Document-word Co-Clustering

Melissa Ailem, Aghiles Salah and Mohamed Nadif

Outline

Outline

Context

Social media posts

amazon NETFLIX ★macy's tripadvisor

Reviews

Google PubMed

Search results

News stories

- Wide variety of textual content.
- Document clustering models are necessary.

Outline

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices.

Govaert & Nadif (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

Govaert & Nadif (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

Govaert & Nadif (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

	Term 4	Term 10	Term 3	Term 6	Term 12	Term 2	Term 9	Term 5	Term 11	Term 7	Term 1	Term 8
Doc 1	1	1	1	1	0	0	0	0	1	1	1	1
Doc 2	1	1	1	1	0	0	1	0	1	1	1	1
Doc 3	1	1	0	1	0	0	0	0	1	1	1	1
Doc 4	0	0	0	0	1	1	1	1	0	0	0	0
Doc 5	0	0	0	0	1	1	1	0	0	0	1	0
Doc 6	0	0	0	0	1	1	1	1	0	0	0	0
Doc 7	0	0	0	0	0	0	0	0	1	1	1	1
Doc 8	0	0	0	0	0	0	0	0	1	1	1	1
Doc 9	0	0	0	0	0	1	0	0	1	1	1	1

(c) Co-clustering

Govaert & Nadif (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

	Term 4	Term 10	Term 3	Term 6	Term 12	Term 2	Term 9	Term 5	Term 11	Term 7	Term 1	Term 8
Doc 1	1	1	1	1	0	0	0	0	1	1	1	1
Doc 2	1	1	1	1	0	0	1	0	1	1	1	1
Doc 3	1	1	0	1	0	0	0	0	1	1	1	1
Doc 4	0	0	0	0	1	1	1	1	0	0	0	0
Doc 5	0	0	0	0	1	1	1	0	0	0	1	0
Doc 6	0	0	0	0	1	1	1	1	0	0	0	0
Doc 7	0	0	0	0	0	0	0	0	1	1	1	1
Doc 8	0	0	0	0	0	0	0	0	1	1	1	1
Doc 9	0	0	0	0	0	1	0	0	1	1	1	1

(c) Co-clustering

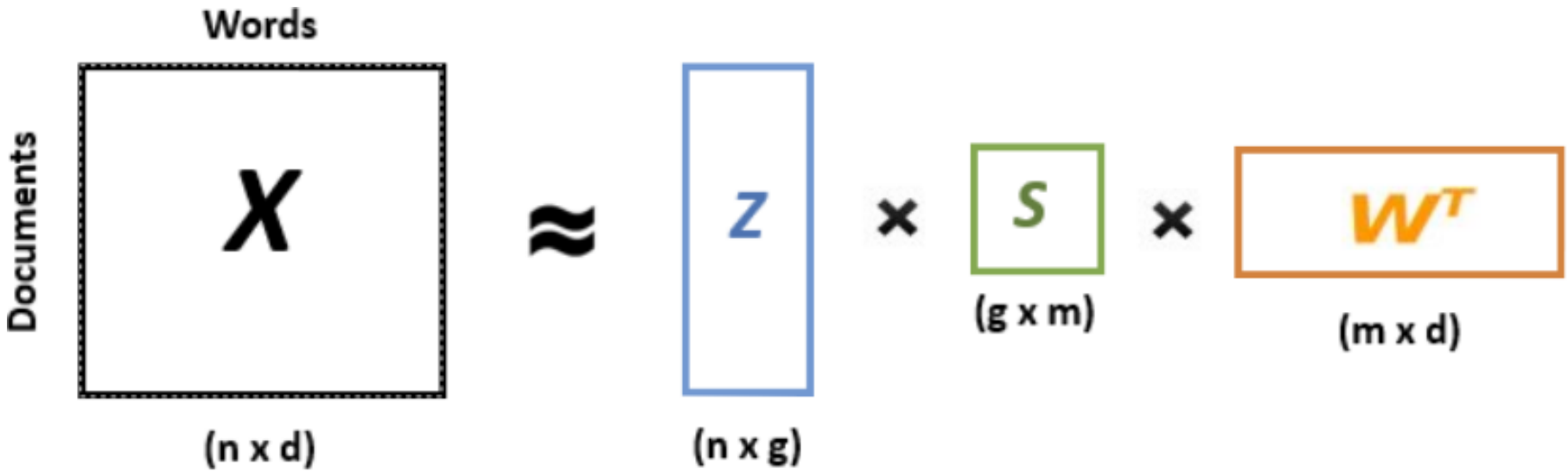
Why Co-clustering?

- Exploit the duality between object space and attribute space
- Cluster Characterization
- Technique for dimensionality reduction
- Reduce Computation time

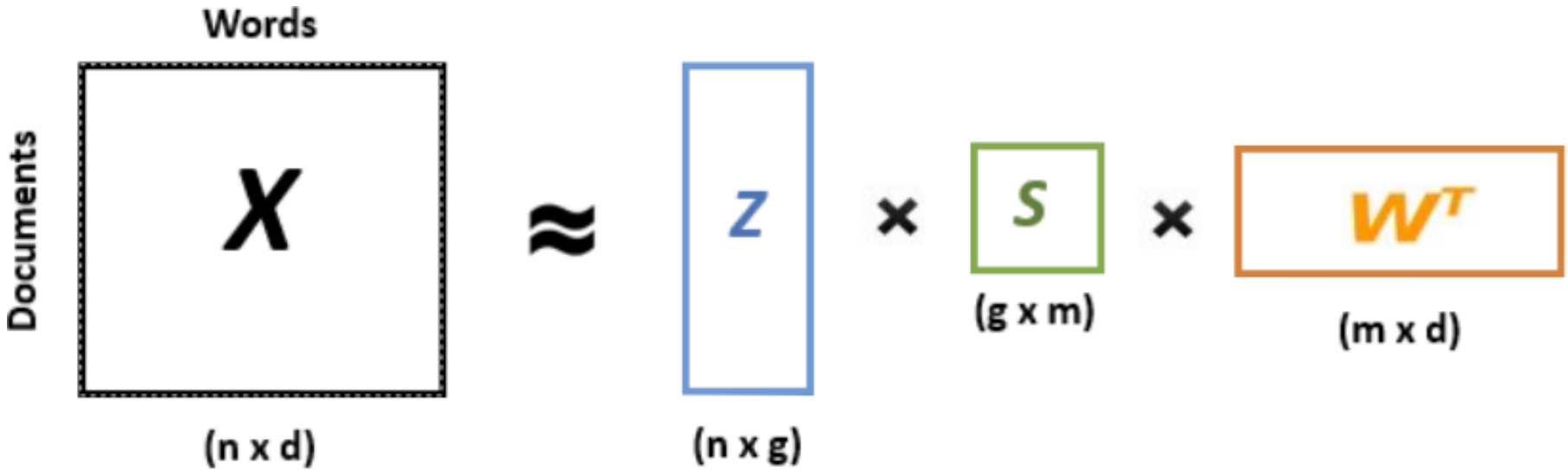
Govaert & Nadif (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

Outline

Non-Negative Matrix Tri-Factorization (NMTF)



Non-Negative Matrix Tri-Factorization (NMTF)



- Solve the following optimization problem:

$$F = \frac{1}{2}||\mathbf{X} - \mathbf{ZSW}^\top||^2, \quad s.t. \ \mathbf{Z} \geq 0, \ \mathbf{W} \geq 0, \ \mathbf{S} \geq 0. \tag{1}$$

Outline

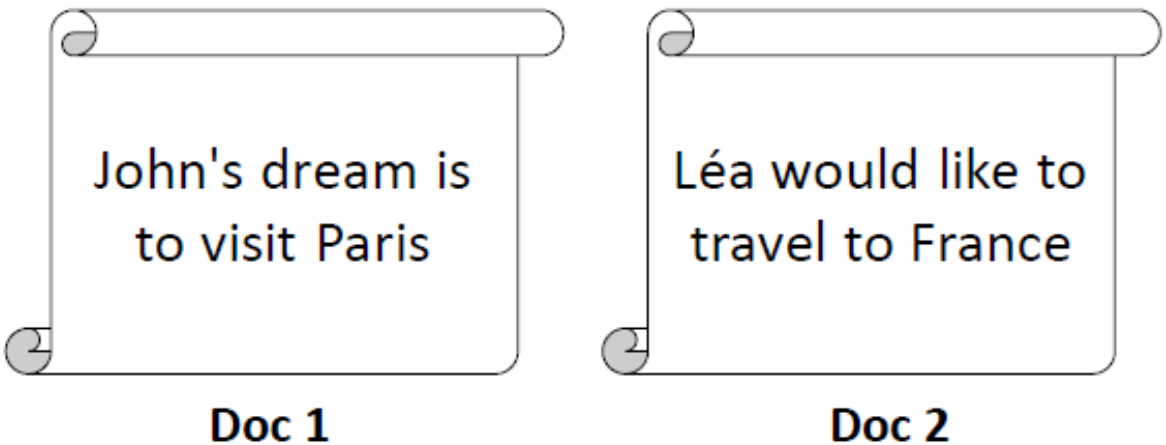
Contribution

- One important aspect when dealing with text data, is to preserve the semantic relationships between words.

Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Contribution

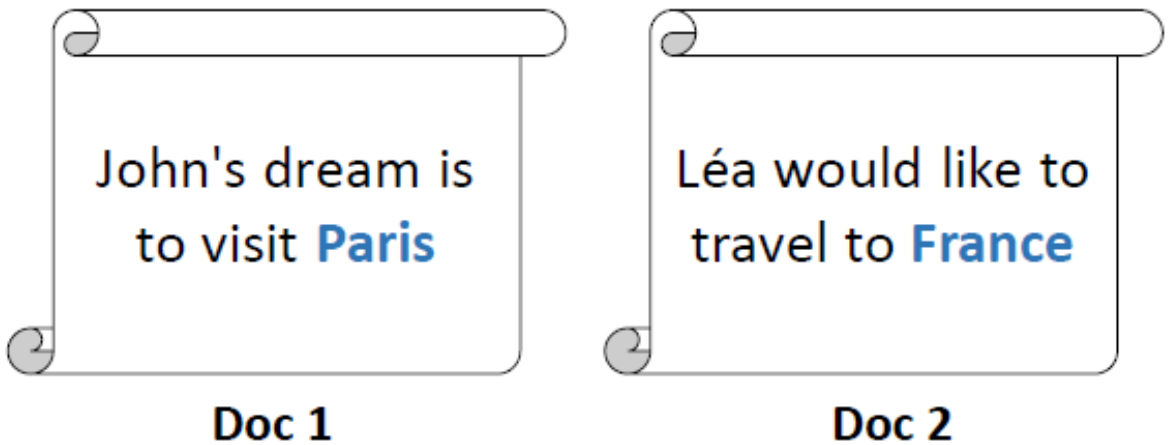
- One important aspect when dealing with text data, is to preserve the semantic relationships between words.



Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

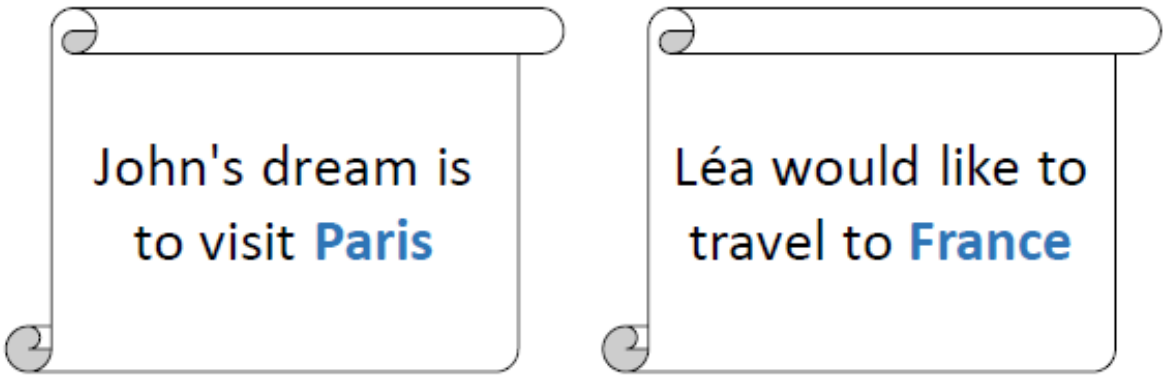
Contribution

- One important aspect when dealing with text data, is to preserve the semantic relationships between words.



Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Contribution

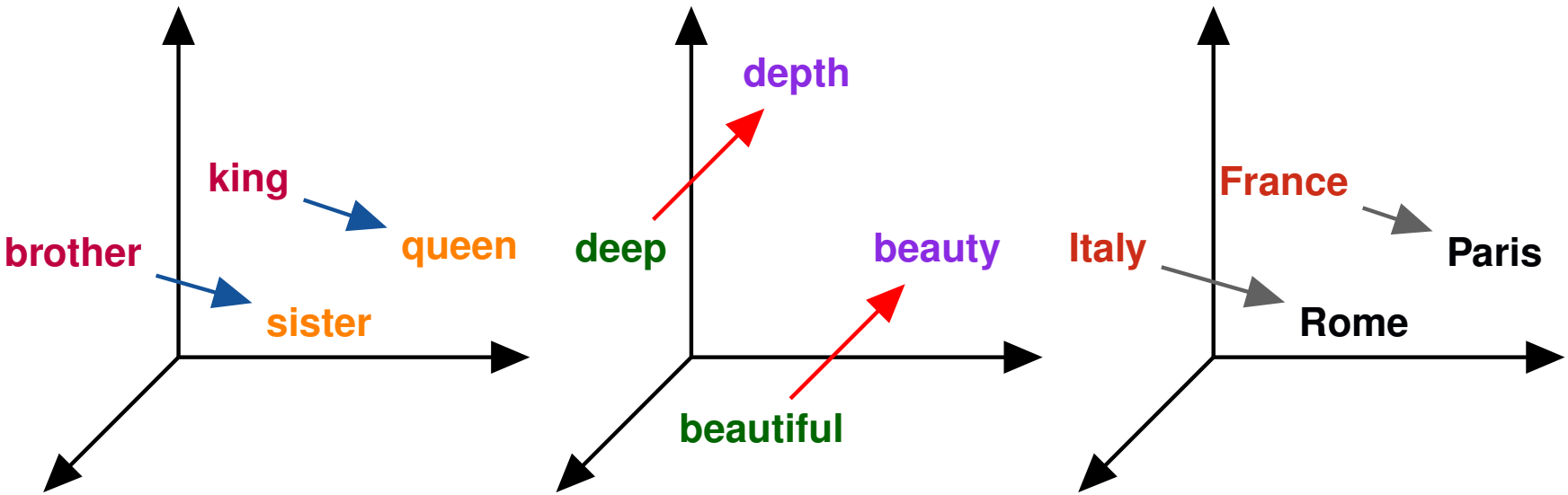


- Previous co-clustering methods, including NMTF, have overlooked this aspect.
- This may induce a significant loss of semantics.
- We propose a new NMTF model that leverage word embeddings so as to preserve more semantics (Ailem et al., 2017; Salah et al., 2017; Ailem et al., 2018).

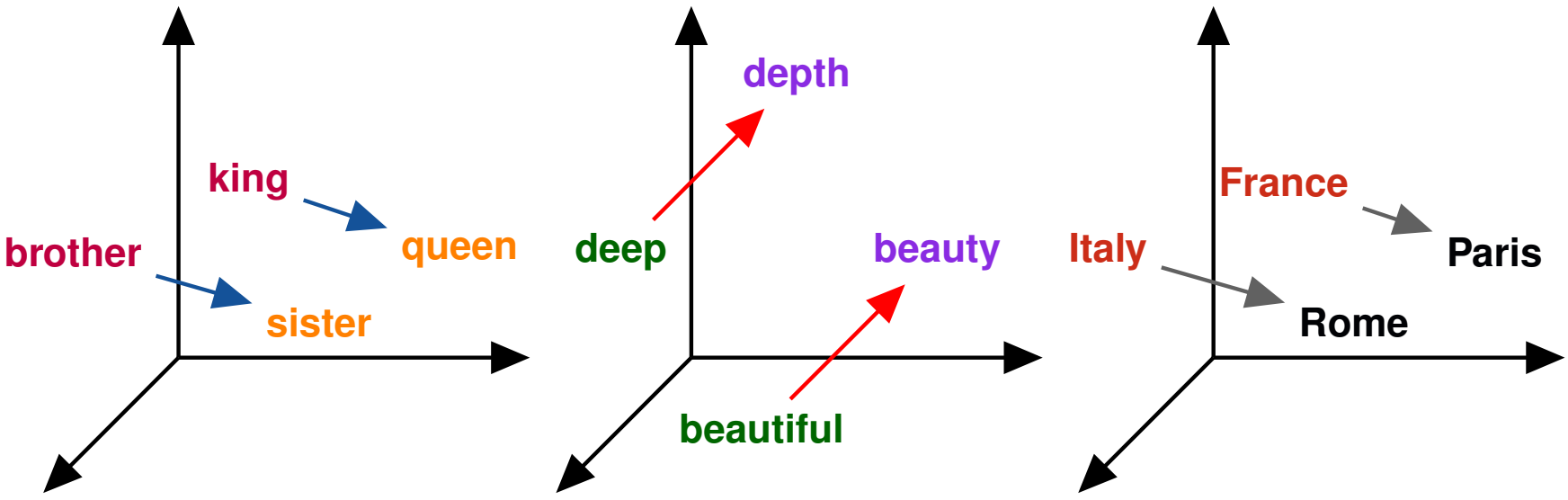
Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Outline

Word Embeddings



Word Embeddings



SKIP-GRAM with Negative Sampling (SGNS)

$$\sum_{w,c} y_{wc} n_{wc} \left[\log \sigma(\boldsymbol{v}_c^{\mathsf{T}} \boldsymbol{e}_w) + \sum_{i=1}^{N} \log \sigma(-\boldsymbol{v}_{c_i'}^{\mathsf{T}} \boldsymbol{e}_w) \right]. \tag{2}$$

Outline

Semantic NMTF (SNMTF)

Levy and Goldberg (2014) showed that SGNS is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) shifted by $\log(N)$

$$\text{PMI}(w_j, w_{j'}) = \log \frac{p(w_j, w_{j'})}{p(w_j)p(w_{j'})} \tag{3}$$

Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Semantic NMTF (SNMTF)

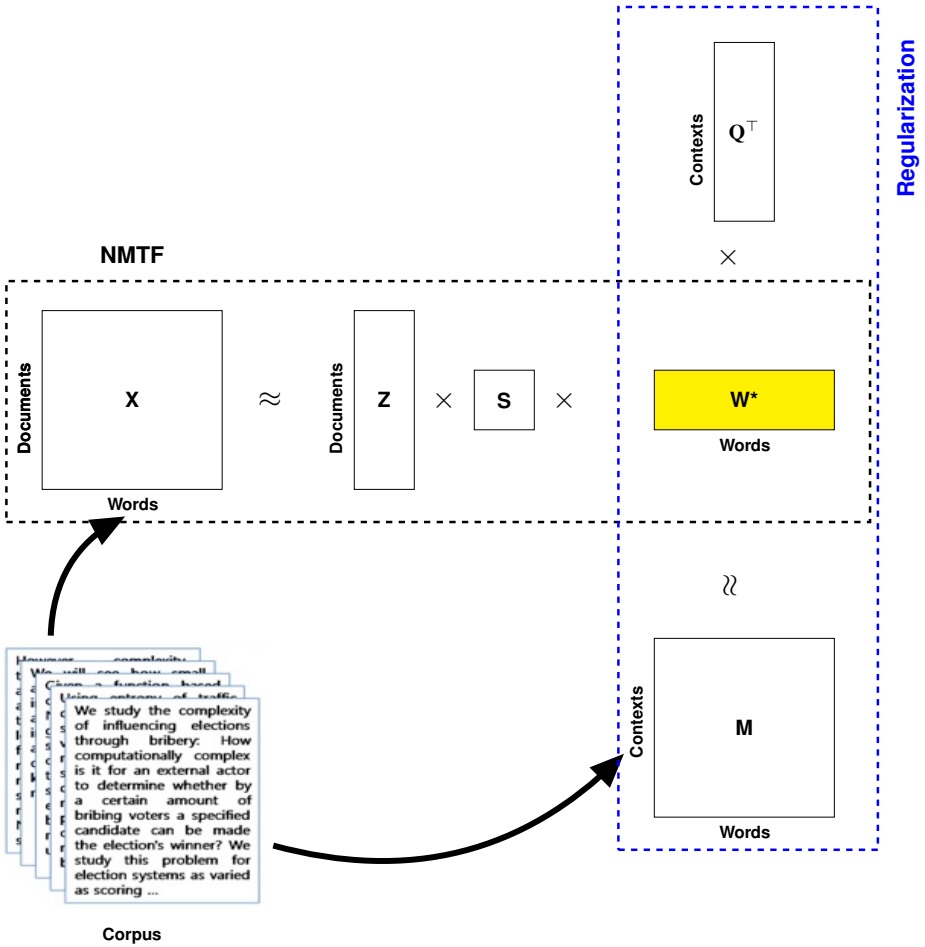
Levy and Goldberg (2014) showed that SGNS is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) shifted by $\log(N)$

$$\text{PMI}(w_j, w_{j'}) = \log \frac{p(w_j, w_{j'})}{p(w_j)p(w_{j'})} \tag{3}$$

- Objective : Regularize NMTF with Word embeddings

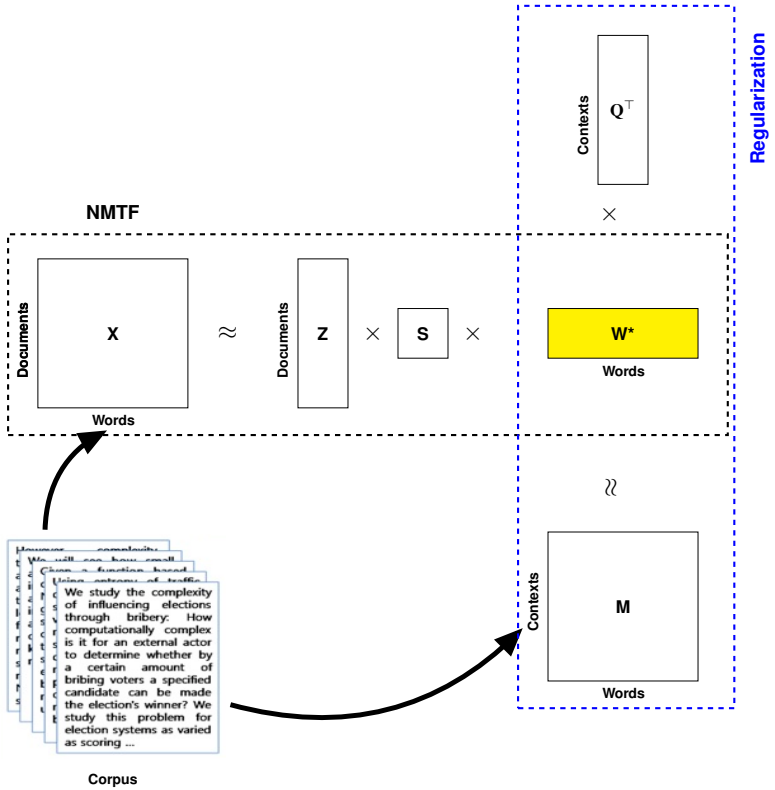
Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Semantic NMTF (SNMTF)



Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Semantic NMTF (SNMTF)



- The objective function of our model, SNMTF, is given by

$$F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) = \underbrace{\frac{1}{2}||\mathbf{X}-\mathbf{Z}\mathbf{S}\mathbf{W}^\top||^2}_{\text{NMTF}} + \underbrace{\frac{\lambda}{2}||\mathbf{M}-\mathbf{W}\mathbf{Q}^\top||^2}_{\text{Regularization term}}, \tag{4}$$

Inference

$$F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) = \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{X}\mathbf{X}^\top - 2\mathbf{X}\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top + \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top\right) + \frac{\lambda}{2}\,\mathrm{Tr}\left(\mathbf{M}\mathbf{M}^\top - 2\mathbf{M}\mathbf{Q}\mathbf{W}^\top + \mathbf{W}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^\top\right). \quad (5)$$

Inference

$$F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) = \frac{1}{2}\operatorname{Tr}\left(\mathbf{X}\mathbf{X}^\top - 2\mathbf{X}\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top + \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top\right) + \frac{\lambda}{2}\operatorname{Tr}\left(\mathbf{M}\mathbf{M}^\top - 2\mathbf{M}\mathbf{Q}\mathbf{W}^\top + \mathbf{W}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^\top\right). \quad (5)$$

- Enforce positivity constraints by introducing the Lagrange multipliers $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$:

$$L(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\gamma}) = F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) + \operatorname{Tr}(\boldsymbol{\alpha}\mathbf{Z}^\top) + \operatorname{Tr}(\boldsymbol{\beta}\mathbf{W}^\top) + \operatorname{Tr}(\boldsymbol{\mu}\mathbf{S}^\top) + \operatorname{Tr}(\boldsymbol{\gamma}\mathbf{Q}^\top).$$

Inference

$$F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) = \frac{1}{2}\operatorname{Tr}\left(\mathbf{X}\mathbf{X}^\top - 2\mathbf{X}\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top + \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top\mathbf{Z}^\top\right) + \frac{\lambda}{2}\operatorname{Tr}\left(\mathbf{M}\mathbf{M}^\top - 2\mathbf{M}\mathbf{Q}\mathbf{W}^\top + \mathbf{W}\mathbf{Q}^\top\mathbf{Q}\mathbf{W}^\top\right). \quad (5)$$

- Enforce positivity constraints by introducing the Lagrange multipliers $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$:

$$L(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\gamma}) = F(\mathbf{Z},\mathbf{W},\mathbf{S},\mathbf{Q}) + \operatorname{Tr}(\boldsymbol{\alpha}\mathbf{Z}^\top) + \operatorname{Tr}(\boldsymbol{\beta}\mathbf{W}^\top) + \operatorname{Tr}(\boldsymbol{\mu}\mathbf{S}^\top) + \operatorname{Tr}(\boldsymbol{\gamma}\mathbf{Q}^\top).$$

- Making use of the KKT conditions and solving resulting stationary equations, yields the following update rules

$$\mathbf{Z} \leftarrow \mathbf{Z} \odot \frac{\mathbf{X}\mathbf{W}\mathbf{S}^\top}{\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}\mathbf{S}^\top} \quad (7a) \qquad \mathbf{S} \leftarrow \mathbf{S} \odot \frac{\mathbf{Z}^\top\mathbf{X}\mathbf{W}}{\mathbf{Z}^\top\mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{W}} \quad (7c)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\mathbf{X}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{M}\mathbf{Q})}{\mathbf{W}(\mathbf{S}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{S} + \lambda\mathbf{Q}^\top\mathbf{Q})} \quad (7b) \qquad \mathbf{Q} \leftarrow \mathbf{Q} \odot \frac{\mathbf{M}^\top\mathbf{W}}{\mathbf{Q}\mathbf{W}^\top\mathbf{W}}. \quad (7d)$$

Outline

Document Clustering

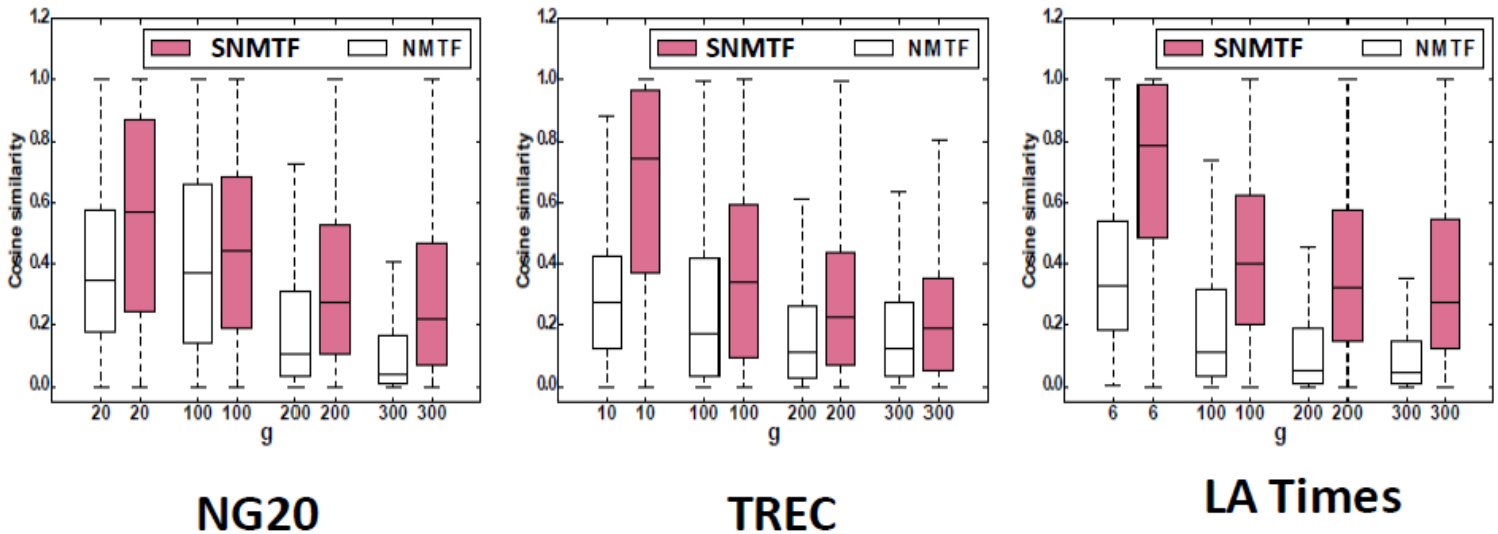
- Average NMI and ARI over different datasets.

Data	Metrics	NMTF	SNMTF
NG20	NMI	0.40±0.02	**0.63±0.01**
	ARI	0.23±0.02	**0.47±0.02**
TREC	NMI	0.59±0.02	**0.67±0.03**
	ARI	0.43±0.03	**0.53±0.05**
LA Times	NMI	0.42±0.02	**0.53±0.03**
	ARI	0.35±0.04	**0.50±0.06**

Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Word Clustering

- Distribution of pairwise cosine similarities between the top 30 words characterizing each document class, computed using the word factors obtained by `NMTF` and `SNMTF`



Melissa Ailem, Aghiles Salah and Mohamed Nadif (AAAI 2018). Word Co-occurrence Regularized Non-negative Matrix Tri-Factorization for Text Data Co-clustering.

Conclusion

- We propose SNMTF, a new co-clustering model that leverage word embeddings in NMTF, thus allowing to preserve the semantic relationships between words.
- SNMTF successfully preserves more semantics, which allows it to noticeably improve the performance of NMTF models in terms of co-clustering.

Conclusion

- We propose SNMTF, a new co-clustering model that leverage word embeddings in NMTF, thus allowing to preserve the semantic relationships between words.
- SNMTF successfully preserves more semantics, which allows it to noticeably improve the performance of NMTF models in terms of co-clustering.

Perspectives

- Extend the idea of of leveraging the word co-occurrences to capture the semantic relationships between words to other co-clustering models, including the different variants of NMTF.
- Investigate other type of contexts in which words co-occur, e.g., sentences.

References I

Ailem, M., Salah, A., and Nadif, M. (2017). Non-negative matrix factorization meets word embedding. *SIGIR'2017*.

Ailem, M., Salah, A., and Nadif, M. (2018). Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *NIPS*, pages 2177–2185.

Salah, A., Ailem, M., and Nadif, M. (2017). A way to boost semi-nmf for document clustering. *CIKM'2017*.