# Document Network Projection in Pretrained Word Embedding Space

Antoine Gourru, Julien Velcin, Julien Jacques, Adrien Guille

Université de Lyon, Lyon 2
ERIC EA3083

*Workshop on Representation Learning for Complex Data, Lyon, 2019*

## Data and Objective

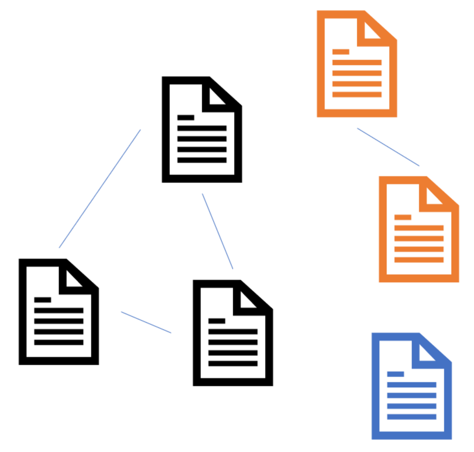

Figure 1: Document network

**We want to build continuous and dense representations of documents organised in network**

## Limitations of state-of-the art methods

- **Documents and Words do not *live* in the same space**
- **They do not consider external semantic information**
- **They rely on dense and complex neural networks: slow and hard to tune**

## General Model

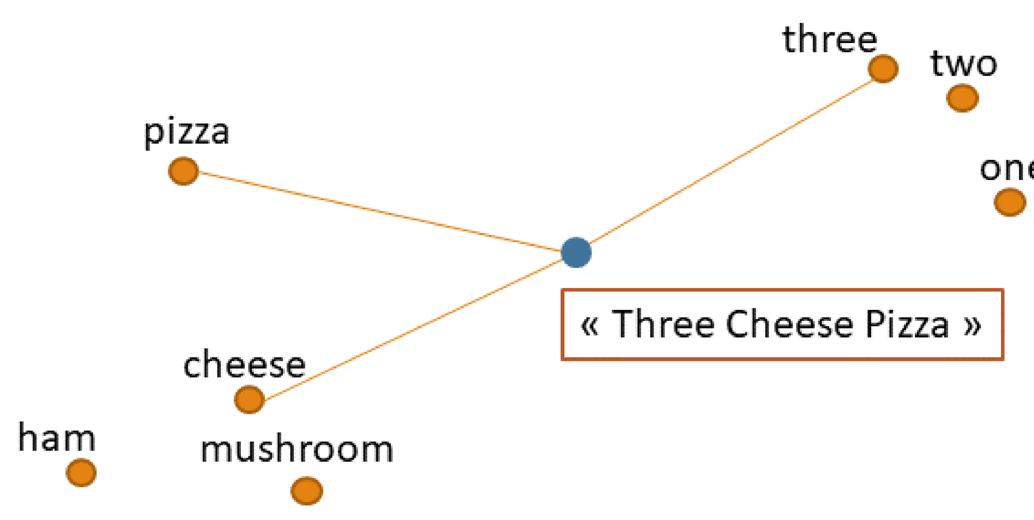


Figure 2: Word averaging method

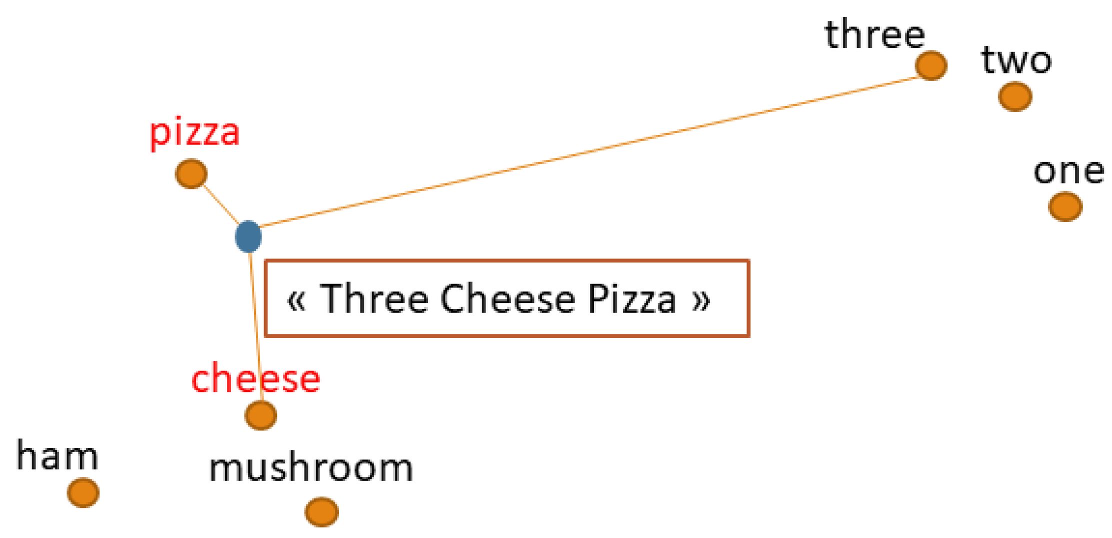


Figure 3: What smoothing should do

A vocabulary of size $v$, $n$ documents. With pretrained $U \in \mathbb{R}^{v \times r}$ (e.g. `fasttext.cc`) , let $d_i \in \mathbb{R}^r$ be the representation of document $i$. With $t_i$ a projection vector in $\mathbb{R}^v$

$$d_i = t_i U \tag{1}$$

How do we build a "relevant" T matrix ?
T = Term frequency ($Tf$)$\rightarrow$ Word Average method

## Contribution 1: RLE

Let $S \in \mathbb{R}^{n \times n}$ a pairwise proximity matrix (defined as $\frac{A+A^2}{2}$ with $A$ the transition matrix obtained from the document network)

$$T = (1-\lambda)Tf + \lambda(STf) \tag{2}$$

Cora and DBLP datasets: Ground truth given
We can evaluate how well documents embeddings allow to separate the classes using simple logistic regression

Table 1: Micro-F1 for classification on Cora and dblp for different train/test ratios.

| Dataset | Cora | | | | DBLP | | | |
|---|---|---|---|---|---|---|---|---|
| Train/Test (%) | 10 | 30 | 50 | Runtime(s) | 10 | 30 | 50 | Runtime(s) |
| DeepWalk | 69.85 | 75.70 | 79.39 | 142.78 | 52.12 | 53.26 | 53.41 | 1220.60 |
| LSA | 70.87 | 77.24 | 79.37 | 0.89 | 73.23 | 73.90 | 74.04 | 2.18 |
| DeepWalk + LSA | 70.28 | 79.48 | 83.63 | 86.88 | 77.65 | 78.18 | 78.29 | 708.74 |
| TADW [4] | 79.80 | 84.59 | 86.48 | 6.49 | 74.97 | 75.47 | 75.62 | 857.39 |
| Graph2Gauss [1] | 79.73 | 82.85 | 84.06 | 32.05 | 70.94 | 71.45 | 71.69 | 921.45 |
| STNE [2] | 76.40 | 82.84 | 85.46 | 510.33 | 72.19 | 72.70 | 72.82 | 12503.33 |
| CANE [3] | 83.27 | 85.47 | 86.31 | 1612.60 | 52.20 | 52.75 | 52.84 | 1622.16 |
| RLE | **83.93** | **86.70** | **87.77** | 0.28 | **78.76** | **80.08** | **80.51** | 0.93 |

## Contribution 2: DETM

T should explain both textual content of the document ($w$) and the edges ($c$) in the network
Using (1), probability of drawing an edge between two documents depends on the proximity of their representations
$t_i$ is drawn from a Dirichlet distribution with parameter $\alpha_k$ depending on document $i$ latent cluster, encoded by $z_i \in \{0,1\}^k$, with $k$ the number of cluster : $t_i | z_i \sim \prod_k \mathcal{D}ir(\alpha_k)^{z_{ik}}$

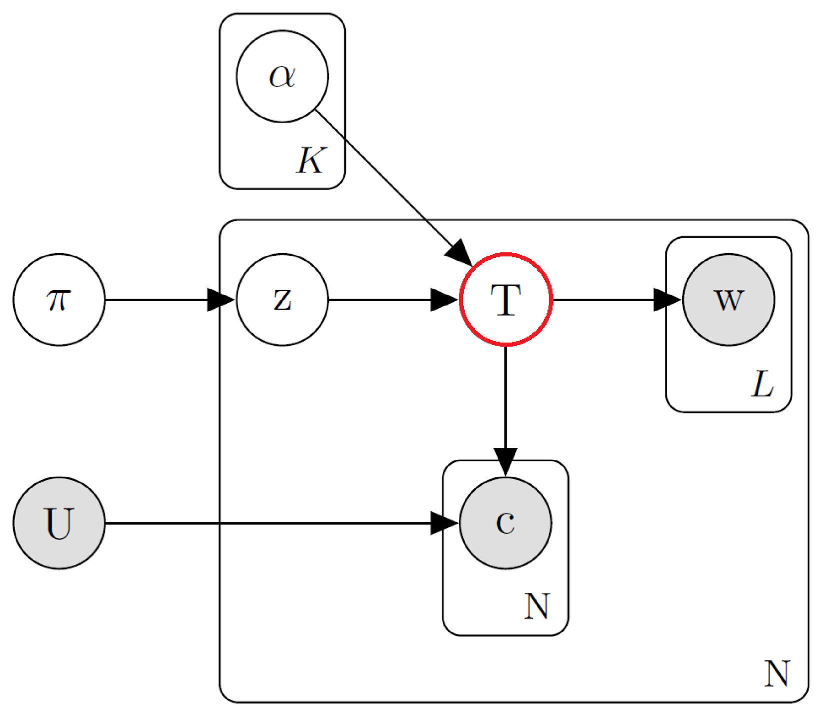


Figure 4: Graphical model representation of DETM. $\pi$ controls clusters' distribution, $w$ and $c$ are words and links counts for $i$ ($L$ is the size of $i$). Ongoing research

## Conclusion and future works

Projecting documents in word embedding space : provide good results
Adding topics: could provide interesting additional information on documents semantics $\rightarrow$ Optimisation problem

## References

[1] A. Bojchevski and S. Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *Proceeding of the International Conference on Learning Representations*, ICLR, 2018.

[2] J. Liu, Z. He, L. Wei, and Y. Huang. Content to node: Self-translation network embedding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1794–1802. ACM, 2018.

[3] C. Tu, H. Liu, Z. Liu, and M. Sun. Cane: Context-aware network embedding for relation modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1722–1731, 2017.

[4] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang. Network representation learning with rich text information. In *International Joint Conference on Artificial Intelligence*, 2015.