Анализ медиатекстов на основе методов понижения размерности

Sergei Sidorov and Dmitriy Melnichuk

Saratov State University

Липецк 6 октября 2023 г.

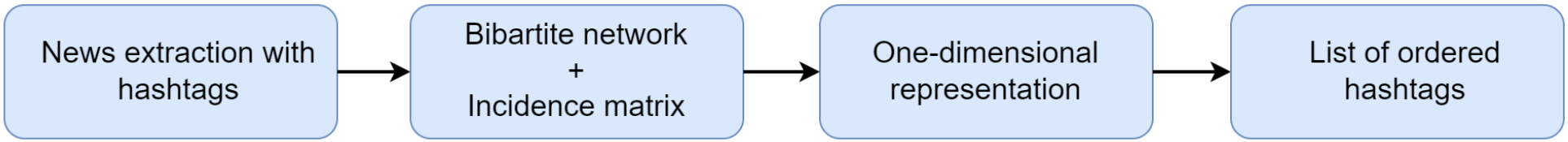
Supported by the Russian Science Foundation, 22-18-00153

Index

Outline

- We propose a new methodology for analyzing topics of interest in media space.
- This approach makes it easy to visualize and interpret the results, and it can also be used to examine changes in the structure of media space over time.

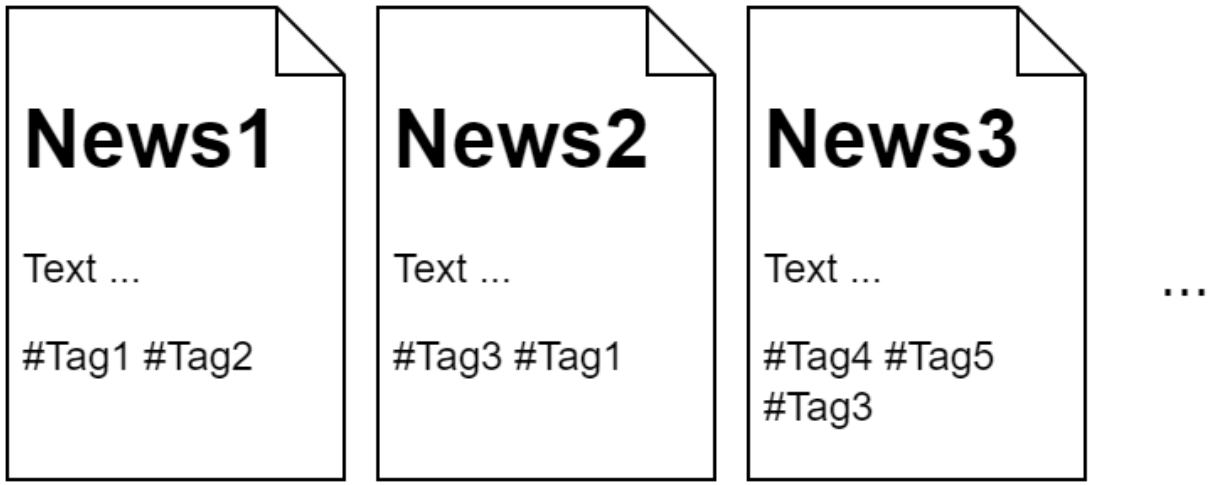
Introduction



- Hashtags (or keywords) extraction;
- Create a bipartite network of hashtags and publications represented in the form of an incident matrix and a square matrix of hashtag co-mentions;
- Build a projection to one-dimensional representation, solving the optimization problem of minimizing the distance of the path around all the vertices of the graph;
- Get a ranked looped list of hashtags.

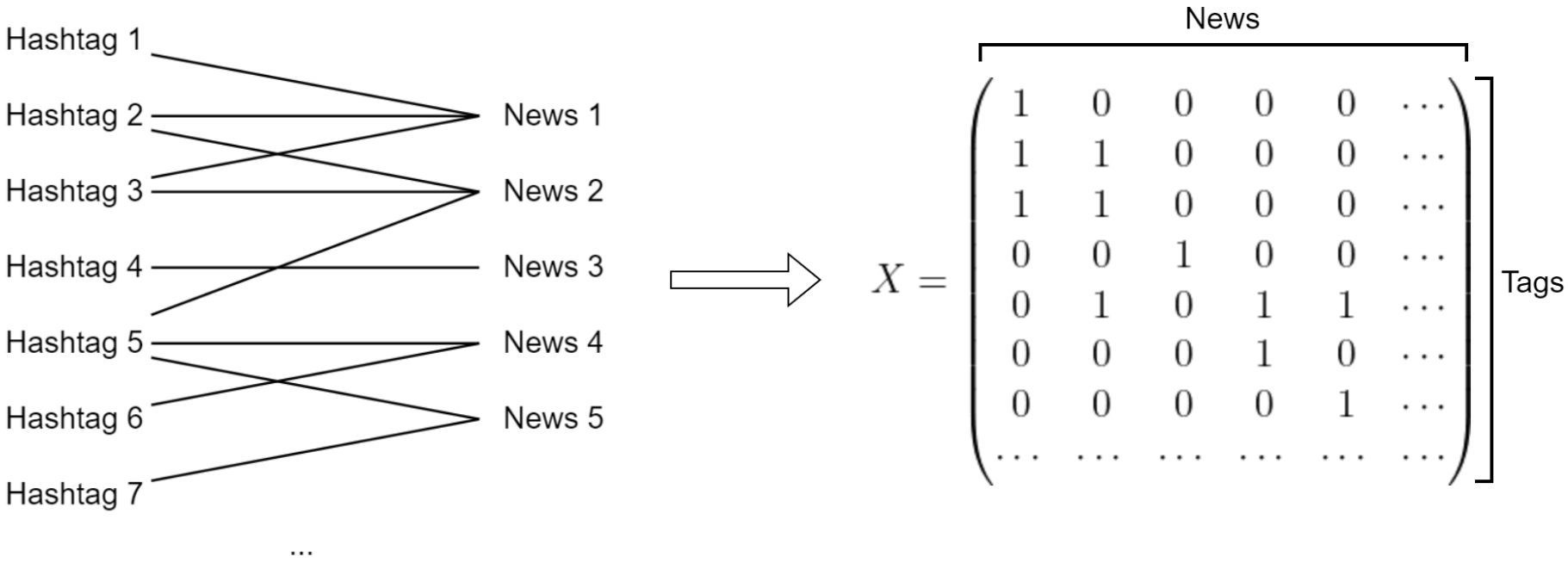
917 articles, 236 tags, from GDELT news base, between 01/01/2022 and 01/01/2023.

Notation and Problem Definition



- $P = \{p_j\}_{j=1}^{m}$: Set of all posts or publications in the corpus of texts (m posts).
- $H = \{h_i\}_{i=1}^{n}$: Set of hashtags used in the corpus of texts (n hashtags).
- $\Phi([n])$: Set of all permutations of indices in set $[n]$.

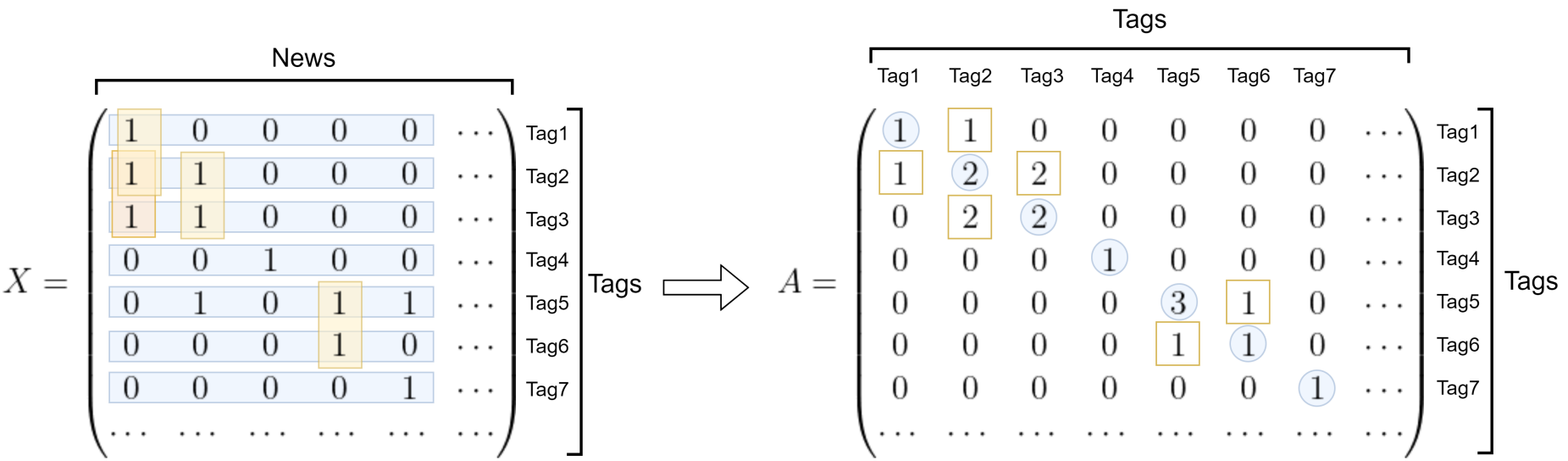
Notation and Problem Definition



- $X = (x_{ij}) \in \mathrm{R}^{n \times m}$: Co-occurrence matrix for hashtags from H.

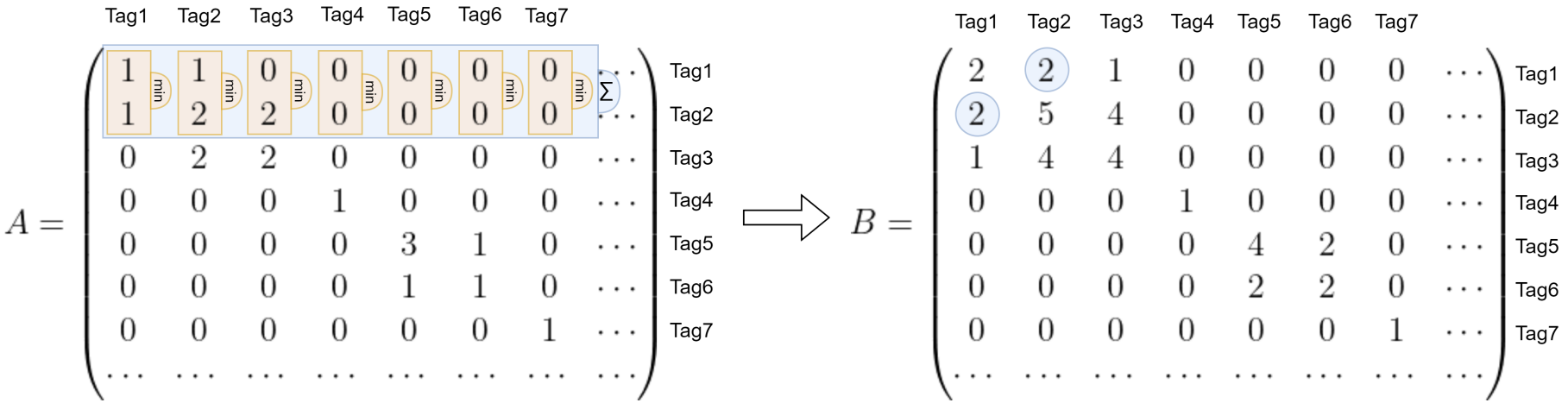
$$x_{ij} = \begin{cases} 1, & \text{if hashtag } i \text{ appear in the news } j \\ 0, & \text{if hashtag } i \text{ don't appear on the news } j. \end{cases}$$

Methodology



- *Approach S.* The approach involves simply counting the number of matching ones for each pair of rows i and j of the matrix X and writing the resulting sum to the corresponding cells $a_{ij} = a_{ji}$ of the matrix A.

Methodology



- *Approach SoS*. The approach uses the adjacency matrix A obtained by the approach S, summing up for each pair of rows i and j the minimum value for each column and writing the resulting sum in the corresponding cell b_{ij} of the matrix B.

Methodology

Co-occurrence Matrix to Adjacency Matrices A and B

- *Approach S.*
- *Approach SoS.*
- The resulting adjacency matrices (tag proximity) were converted into distance matrices $R_1 = (r_{ij}^1)$ and $R_2 = (r_{ij}^2)$ using the transformations $r_{ij}^1 = 1/(1 + a_{ij})$ and $r_{ij}^2 = 1/(1 + b_{ij})$, where $A = (a_{ij})$ and $B = (b_{ij})$.
- The distance matrices $Y = (y_{ij})$ is obtained using Kruskal's Non-metric Multidimensional Scaling, which gives a 2-dimensional configuration.

Methodology

Travel salesman problem

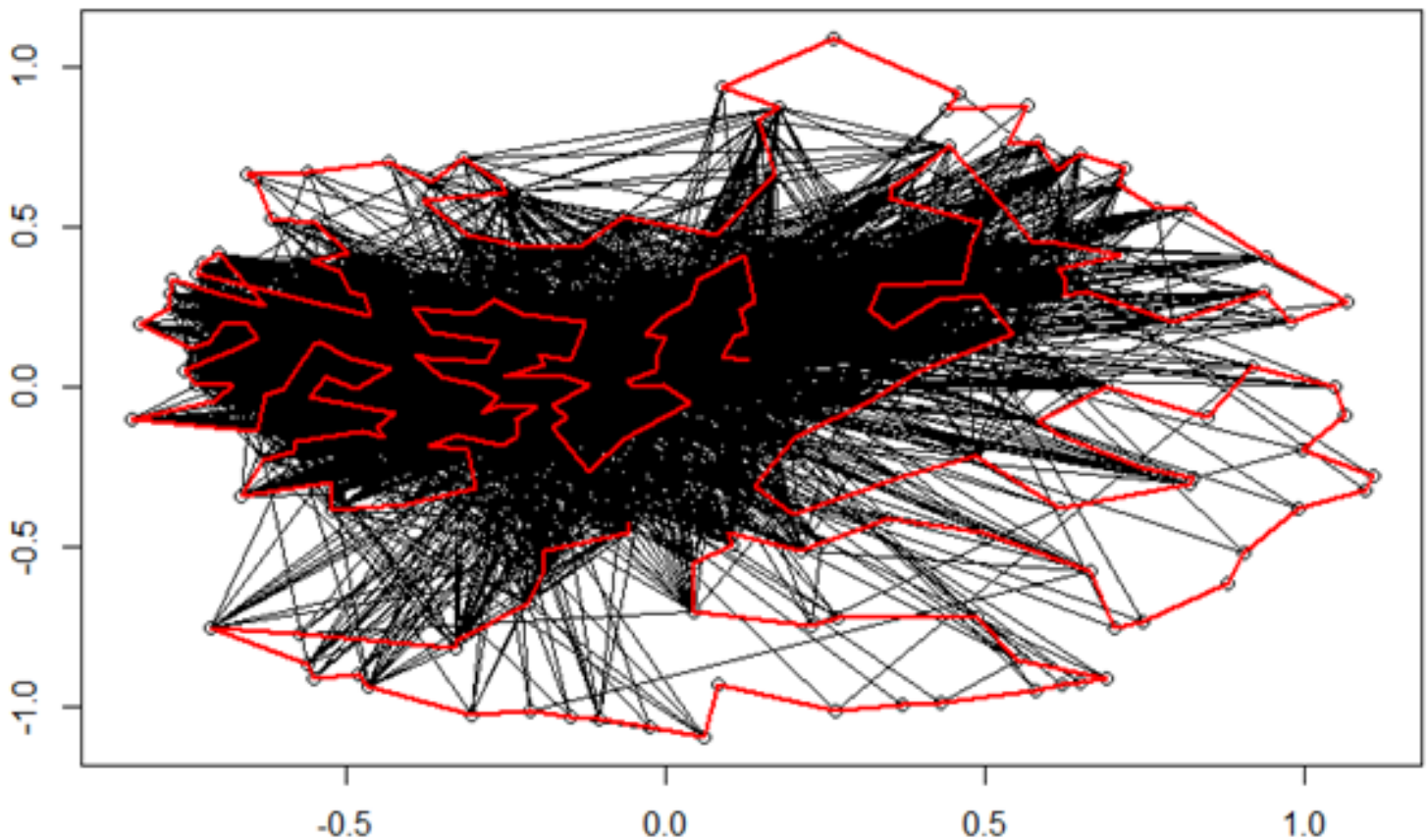
- We formulate word embeddings optimization as a traveling salesman problem and solve it with a high performance solver.
- To find the optimal one-dimensional embedding we solve the following travel salesman problem:

$$\|y_{\sigma_1} - y_{\sigma_n}\| + \sum_{i=1}^{n-1} \|y_{\sigma_i} - y_{\sigma_{i+1}}\| \to \min_{\sigma \in \Phi([n])}, \tag{1}$$

- We use a solver, which implements a procedure to improve the exchange between two edges. This is a tour refinement procedure that systematically swaps two edges in a graph represented by a distance matrix.
- We considered the metrics $\|\cdot\|_1$ and $\|\cdot\|_2$ in the space L_1 and L_2 in problem (1).

Empirical Results

c



Methodology

Criteria for assessing the quality of the received ordered tours.

- **Cor R**. Weighted correlation coefficient as the measure of effectiveness for the ordering.
- **ME**. Measure of effectiveness

$$M(X) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{i,j}(x_{i,j-1} + x_{i,j+1} + x_{i-1,j} + x_{i+1,j})$$

with $x_{0,j} = x_{m+1,j} = x_{i,0} = x_{i,n+1} = 0$. A higher value of ME means a better ordering.

Methodology

Criteria for assessing the quality of the received ordered tours.

- Stress measures the conciseness of the presentation of a matrix and can be seen as a purity function which compares the values in a matrix with its neighbors. The stress measure is computed as the sum of squared distances of each matrix entry from its adjacent entries.

$$L(X) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sigma_{ij},$$

where σ_{ij} may be defined based two types of neighborhoods:

1. Moore comprises the eight adjacent entries.

$$\sigma_{ij} = \sum_{k=\max(1,i-1)}^{\min(n,i+1)} \sum_{l=\max(1,j-1)}^{\min(m,j+1)} (x_{ij} - x_{kl})^2$$

2. Neumann comprises the four adjacent entries.

$$\sigma_{ij} = \sum_{k=\max(1,i-1)}^{\min(n,i+1)} (x_{ij} - x_{kj})^2 + \sum_{l=\max(1,j-1)}^{\min(m,j+1)} (x_{ij} - x_{il})^2$$

Empirical Results

Criterion for a Loss/Merit Function for Data Given a Permutation

Methods	Cor R	ME	Moore stress	Neumann stress
SoS TSP L2	0.03	342	**35534**	17486
S TSP L2	0.07	333	35848	17492
SoS TSP L1	0.02	**373**	35584	**17408**
S TSP L1	**0.10**	353	35704	17412
SoS PCA	0.04	201	36430	17982
S PCA	0.06	211	36754	18090

Examples of segments

Methods	Segments around "Judiciary."	Segments around "NATO Expansion."
SoS TSP L2	Law	Nuclear Safety
	Legal	UK
	Rulings and clarifications of the Supreme Court	Eurasian Economic Union (EAEU)
	Judiciary	NATO Expansion
	Supreme Court	Russia and NATO
	Constitution of Russia	Kazakhstan
	Ministry of Education and Science	"United Russia"

Summary

- The paper proposes a method for dividing a corpus of texts into sets, extracting hashtags or keywords from each set, creating a network of hashtags and publications, and converting this into a matrix of joint mentions.
- The algorithm then ranks the hashtags and represents the space in one dimension. The text also discusses the difficulties in evaluating the methodology and provides examples of the assessed quality of the resulting ordered rings.
- In conclusion, it is challenging to determine the preferable approach, but the SoS TSP L2 approach is considered natural due to semantic similarity between consecutive words.

Future work

- Study the dynamics of media space changes.
- Divide a time interval into several segments in a one-dimensional representation and study how they changed.
- Find stable patterns of tags in the media space.

MANY THANKS!