Wasserstein distance for document classification

Thierry Klein & Willy Rodríguez
École Nationale de l'Aviation Civile. thierry.klein@math.univ-toulouse.fr; willyrv@gmail.com

Introduction

One of the most common tasks in Natural Language Processing (NLP) is to classify documents according to different criteria. A crucial step in this process is to define a notion of **document similarity** capable of capturing the information we would like take into account for the classification.

Even though it does not give a direct way to compute document similarities, the Latent Dirichlet Allocation (LDA) method [1], is often used to classify texts. In this work we describe a new distance between documents based on LDA, Optimal Transport [4] and Word2Vect [3].

LDA and documents distances

Let β be the topics representation matrix. β_{ij}: probability of the word j appearing in a document of topic i. Shape of β is $K \times V$. K: the number of topics, V: size of the vocabulary. The following distance [2] between topics can be defined:

$$\xi_{ij} = \sum_{v \in V} \mathbb{1}_{\beta_{iv} \neq 0} \mathbb{1}_{\beta_{jv} \neq 0} |\log(\beta_{iv}) - \log(\beta_{jv})|$$

Also, θ: document representation. θ_{ij}: proportion of words from topic j in document i. Distance between documents based on the differences between the proportion of topics they have in common:

$$\sigma_{ij} = \sum_{k \in K} \mathbb{1}_{\theta_{ik} \neq 0} \mathbb{1}_{\theta_{jk} \neq 0} |\log(\theta_{ik}) - \log(\theta_{jk})|$$

Wasserstein distance

Let a and b be two 1-Dimension probability distributions and M be a distance matrix of $n \times n$. The Wasserstein distance between a and b is defined by:

$$\mathcal{W}_{a,b} = \min_{\gamma \in \mathbb{R}_+^{n \times n}} \sum_{i,j} \gamma_{i,j} M_{i,j}$$

$$s.t. \gamma \mathbb{I} = a; \gamma^T \mathbb{I} = b; \gamma \geq 1$$

We compute the distance using [4]

References

[1] Blei, David M and Ng, Andrew Y and Jordan, Michael I, *Latent dirichlet allocation*, Journal of machine Learning research (2003).

[2] Chaney, Allison J.B. and Blei, David M., *Visualizing Topic Models*, Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (2012).

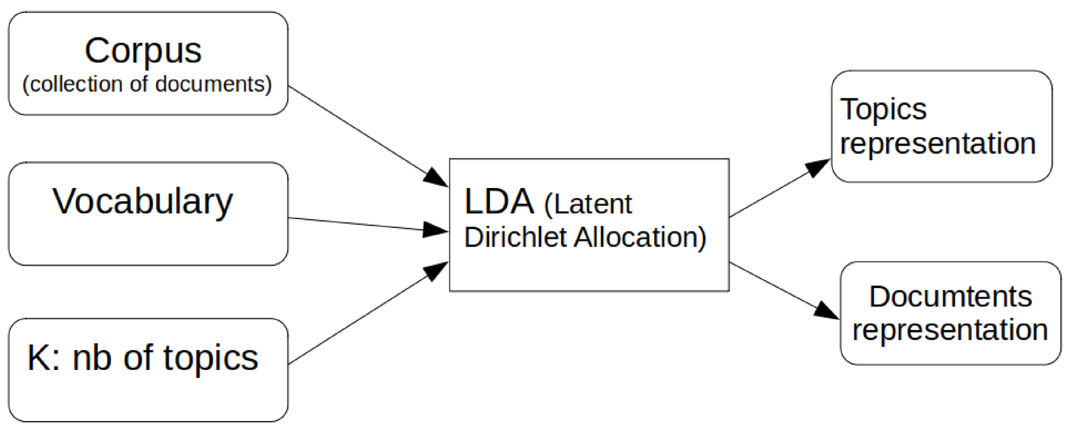
[3] Tomas Mikolov and Kai Chen and Greg Corrado and Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*,preprint arXiv:1301.3781 (2013).

[4] Flamary, Rémi and Courty, Nicolas, *POT Python Optimal Transport library*, https://pythonot.github.io/, (2017).

LDA decomposition applied to the ASRS corpus

The ASRS (Aviation Safety Reporting System) is a reporting system operated by NASA that collects anonymous reports about accidents or incidents in the United States, having a potential impact for aviation safety. The dataset contains more that 300K reports. We applied the LDA algorithm to a subset of this corpus, using 15 for the number of topics.

The LDA outputs



β: Topics representation θ: Documents representation

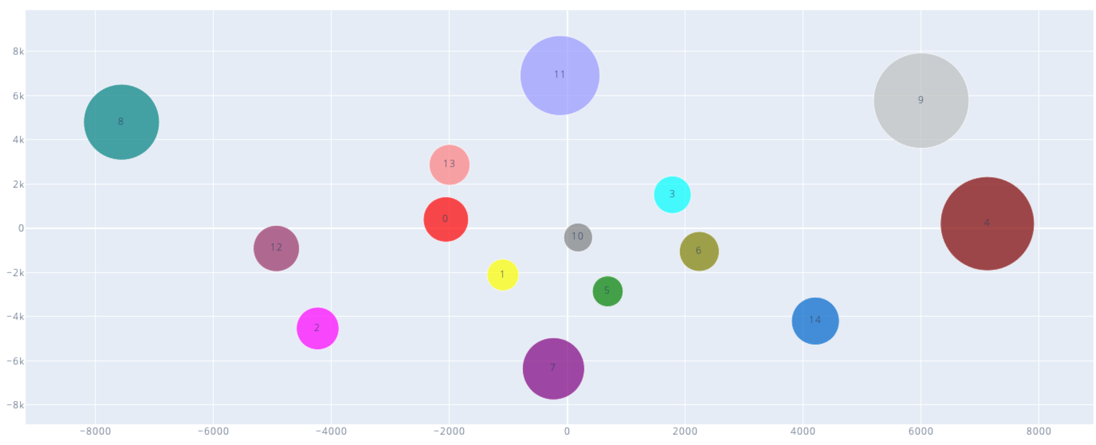
$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1V} \\ \vdots & \ddots & & \\ \beta_{K1} & \beta_{K2} & \dots & \beta_{KV} \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1K} \\ \dots & \ddots & & \dots \\ \theta_{n1} & \theta_{n2} & \dots & \theta_{nK} \end{bmatrix}$$

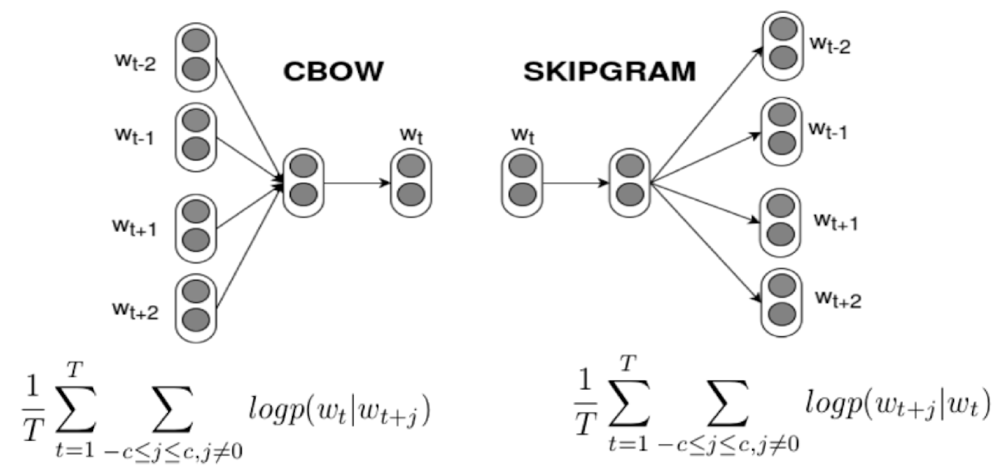
K: Number of topics
V: Number of words
n: Number of documents

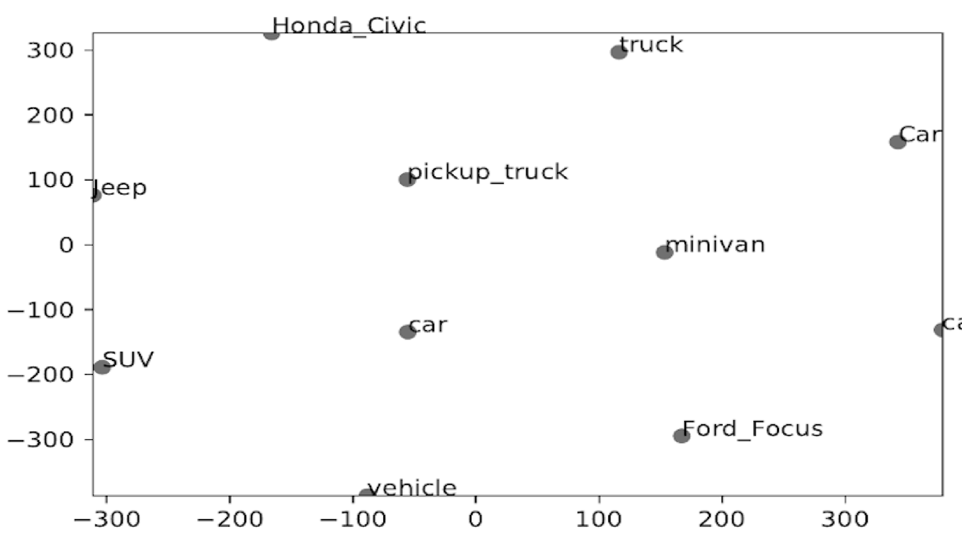
LDA topics decomposition on the ASRS

TOPICS	TOP WORDS	PROPORTION IN THE CORPUS
0	flight, management, officer, arrival, clearance, computer, control, traffic, fix, route	4.12
1	degree, heading, told, officer, turn, zzz, departure, engine, controller, said	2.04
2	damage, fix, aircraft, time, intermediate, pilot, aviation, officer, federal, foreign	3.66
3	aircraft, radar, smt, airport, mile, engine, fix, control, intermediate, falcon	2.85
4	flight, feet, foot, level, traffic, altitude, climb, air, control, rules	17.77
5	aircraft, line, time, zhu, hour, slide, ir, paperwork, flight, sfo	1.92
6	gear, landing, runway, aircraft, nose, approach, did, light, radar, fix	3.2
7	engine, flight, aircraft, landing, crew, captain, information, normal, reference, gate	7.77
8	flight, fuel, aircraft, maintenance, engine, zzz, control, landing, officer, emergency	11.57
9	approach, aircraft, traffic, runway, control, air, feet, turn, tower, visual	18.31
10	speed, airspeed, aircraft, indicated, knots, captain, control, air, emergency, normal	1.69
11	runway, aircraft, tower, ground, taxiway, takeoff, clear, taxi, officer, flight	12.85
12	aircraft, minimum, officer, flight, equipment, list, knot, landing, flying, pilot	4.3
13	rptr, maintenance, flight, aircraft, propeller, minute, information, said, circuit, fix	3.35
14	aircraft, flight, frequency, sector, carrier, air, time, pilot, weather, fix	4.62

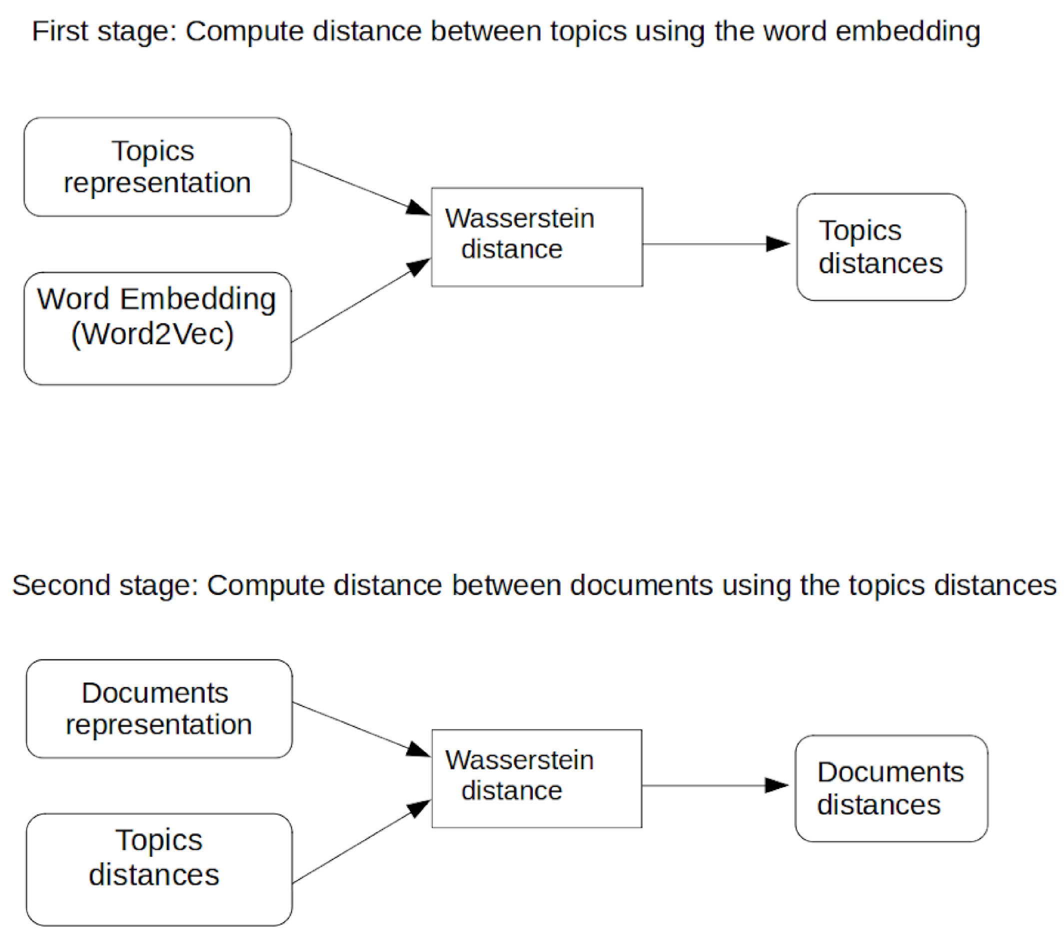


Word Embeddings

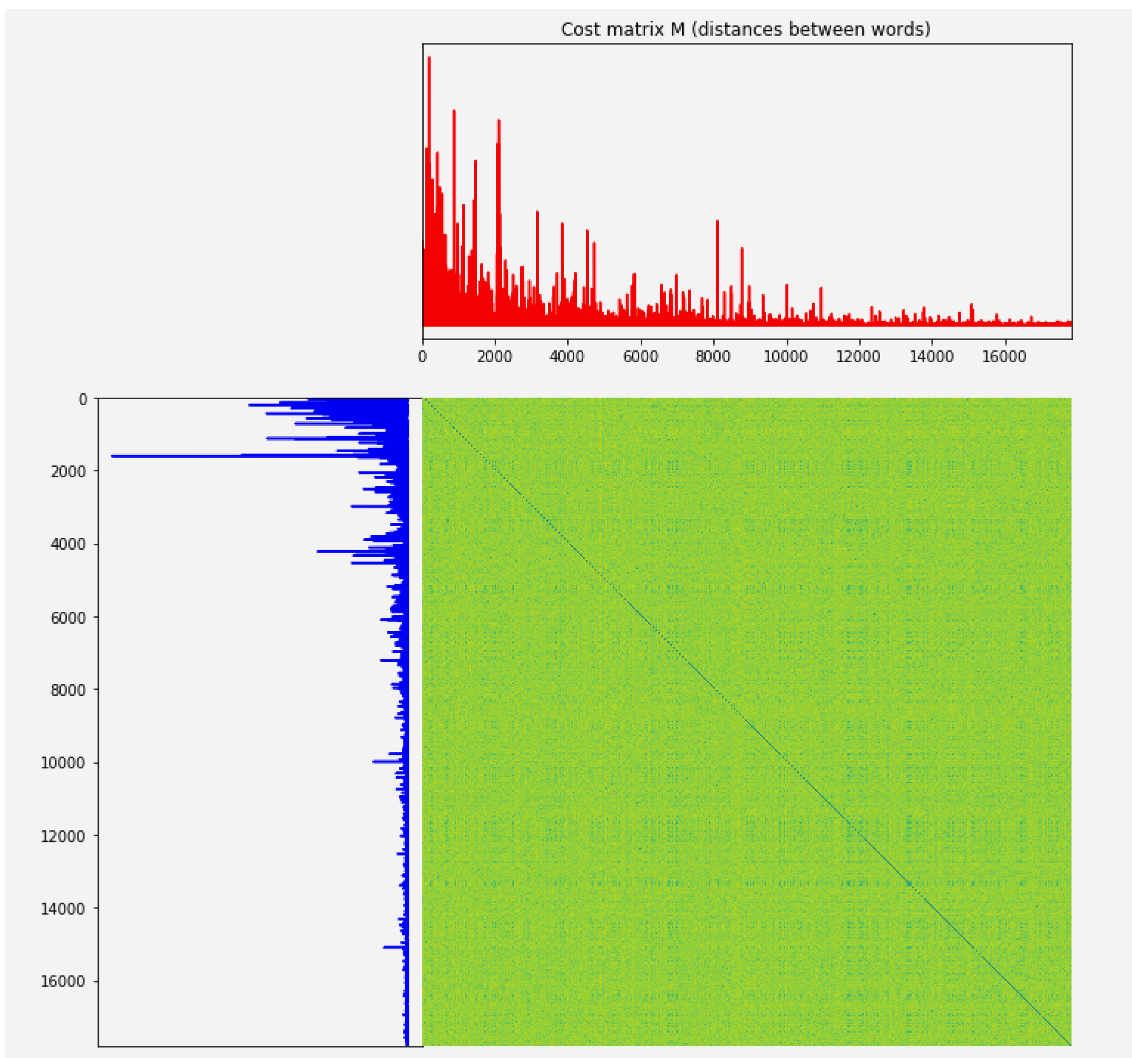


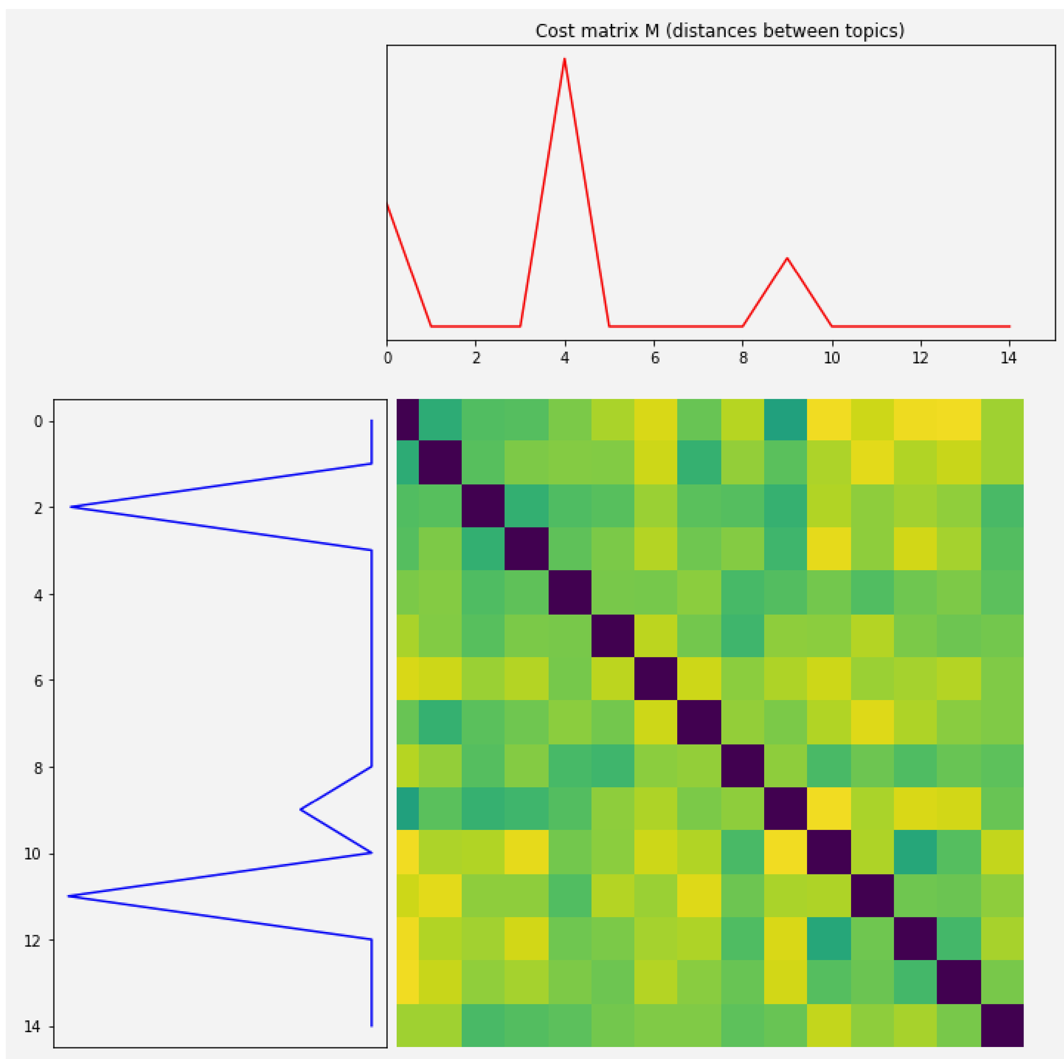


Two stage Wasserstein distance



Applying the new distance to real data





Acknowledgements: This research is supported by the ENAC Safety Management Chair funded by Airbus.