Hierarchical Optimal Transport for Document Clustering

Pierre OSSELIN

Computational Optimal Transport

7 Janvier, 2019

Contextualization

NLP Tasks

1. **Natural Language Processing :** Achieve computer understanding of language
2. **Applications for documents :** Includes Document Classification, Document Retrieval, Document Clustering, Sentiment Analysis, Multilingual Document Matching

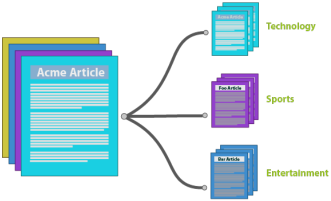


Figure: Document Classification

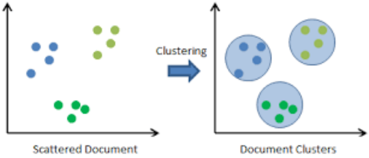
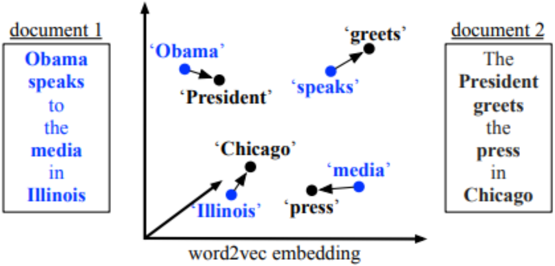


Figure: Document Clustering

Problem

How to Design Meaningful distances between documents ?

1. **Characteristics :** A distance between documents should be meaningful / computationaly viable / Interpretable
2. **Previous Works :** Involves Bag-Of-Words representation, topic Modelling or Word embedding.
3. **Limitations :** Lack of either representational power, computational viability or interpretability



Figure

Proposed method : Hierarchical Optimal Transport

Word Mover's Distance

The 1-Wasserstein distance between p and q is

$$W_1(p,q) = \begin{cases} min_{\Gamma \in \mathbb{R}_+^{n\times m}} & \sum_{i,j} C_{i,j}\Gamma_{i,j} \\ s.t & \sum_j \Gamma_{i,j} = p_i \text{ and } \sum_i \Gamma_{i,j} = p_j \end{cases} \quad (1)$$

1. $C_{i,j} = d(x_i, y_j)$, where $d(.,.)$ denotes the distance
2. Γ can be interpreted as a transport plan
3. The Word Mover Distance (WMD) between documents is then $WMD(d^1, d^2) = W_1(d^1, d^2)$, where d^1 and d^2 are normalized word counts and the ground metric is Euclidean in some embedding space

Proposed method : Hierarchical Optimal Transport

Hierarchical Optimal Transport (HOTT)

Topics t_i are inferred by LDA on the corpus, then the HOTT is defined as:

$$\mathbf{HOTT}(d^i, d^j) = W_1\left(\sum_{k=1}^{|T|} d_k^i \delta_{t_k}, \sum_{k=1}^{|T|} d_k^j \delta_{t_k}\right) \tag{2}$$

The cost matrix is computed one time per corpus with:

$$d(t_i, t_j) = WMD(t_i, t_j) \tag{3}$$

Proposed method : Hierarchical Optimal Transport

Wasserstein Barycenter

Solves, where (a_k) are m measures and $\sum_k \lambda_k = 1$:

$$min_b \sum_k \lambda_k \times W_\epsilon(a_k, b) \tag{4}$$

Where

$$W_\epsilon(p,q) = \begin{cases} min_{\Gamma \in \mathbb{R}_+^{n \times m}} & \sum_{i,j} C_{i,j}\Gamma_{i,j} - \epsilon E(\Gamma) \\ s.t & \sum_j \Gamma_{i,j} = p_i \text{ and } \sum_i \Gamma_{i,j} = p_j \end{cases} \tag{5}$$

And

$$E(\Gamma) = -\sum_{i,j} \Gamma_{i,j}(log(\Gamma_{i,j}) - 1) \tag{6}$$

Problem solved with Bregman Iteration algorithm.

Theoretical analysis

Guarentees

1. **Link between HOTT and WMD :**

$$WMD(d^i, d^j) \leq HOTT(d^i, d^j) + diam(X)\left[\sqrt{\frac{1}{2}KL\left(d^j || \sum_{k=1}^{|T|} d_k^j t_k\right)} + \sqrt{\frac{1}{2}KL\left(d^i || \sum_{k=1}^{|T|} d_k^i t_k\right)}\right] \quad (7)$$

2. **Complexity :** HOTT complexity is $\mathcal{O}(|T|^3 log(|T|))$
K-means clustering has a complexity of

$$\mathcal{O}(n_{iter} \times \underbrace{|D| \times k \times |T|^3 log(|T|)}_{\text{Pairwise Distances}} \times \underbrace{|D| \times \frac{|T|^2}{\epsilon^2}}_{\text{Computation Barycenters}})$$

Numerical findings : K-means Clustering with Wasserstein Barycenters

The metrics are described in annex.

Dataset	Tf-Idf	LDA	HOTT
bbcsport	1.41 ± 0.23	1.35	$\mathbf{1.28} \pm \mathbf{0.20}$
twitter	$\mathbf{1.16} \pm 0.25$	1.80	1.70 ± 0.14
classic	1.40 ± 0.08	**0.62**	0.66 ± 0.15
ohsumed	$\mathbf{2.65} \pm 0.3$	3.76	3.74 ± 0.02
r8	1.69 ± 0.14	**1.38**	$\mathbf{1.39} \pm \mathbf{0.12}$
amazon	1.35 ± 0.15	**1.08**	$\mathbf{1.08} \pm \mathbf{0.02}$

Table: Variation of Information

Dataset	Tf-Idf	LDA	HOTT
bbcsport	0.34 ± 0.18	0.53	$\mathbf{0.648} \pm \mathbf{0.08}$
twitter	0.017 ± 0.015	0.052	$\mathbf{0.053} \pm \mathbf{1e\text{-}17}$
classic	0.11 ± 0.07	0.76	$\mathbf{0.77} \pm \mathbf{0.06}$
ohsumed	0.12 ± 0.03	0.122	$\mathbf{0.131} \pm \mathbf{0.005}$
r8	0.31 ± 0.07	**0.59**	$\mathbf{0.58} \pm \mathbf{0.02}$
amazon	0.17 ± 0.2	**0.58**	$\mathbf{0.57} \pm \mathbf{0.003}$

Table: NMI Score

Numerical findings : K-means Clustering with Wasserstein Barycenters

Dataset	Tf-Idf	LDA	HOTT
bbcsport	0.23 ± 0.17	0.50	$\mathbf{0.639 \pm 0.09}$
twitter	0.011 ± 0.014	0.045	$\mathbf{0.0468 \pm 1e\text{-}20}$
classic	0.06 ± 0.06	**0.75**	0.73 ± 0.07
ohsumed	0.08 ± 0.03	0.113	$\mathbf{0.121 \pm 0.04}$
r8	0.28 ± 0.08	0.49	$\mathbf{0.53 \pm 0.03}$
amazon	0.11 ± 0.17	**0.54**	$\mathbf{0.54 \pm 0.003}$

Table: AMI Score

Dataset	Tf-Idf	LDA	HOTT
bbcsport	-2	$\leq$ -300	$\leq$ -300
twitter	-1	-62	-60
classic	-2	$\leq$ -300	$\leq$ -300
ohsumed	-140	$\leq$ -300	$\leq$ -300
r8	$\leq$ -300	$\leq$ -300	$\leq$ -300
amazon	-1	$\leq$ -300	$\leq$ -300

Table: log P Value Score

Conclusion

1. Leverages optimal transport, topic modeling, and word embedding and provide global semantic language information.
2. The HOTT distance matches our intuition of how humans compare documents : by breaking down each document into easy to understand concepts, and then comparing the concepts
3. Our k-means algorithm performs better or at least equally well on every data set.
4. Necessity to gain insights into the nested metric HOTT to learn its representational capacity, and the nature of the Regularized Wassertstein Barycenters computed with performing k-means. These insights would allow us to design faster and more accurate adaptation of this algorithm.

Annex

We evaluate our clusters with the true labels.

Evaluation of clusters

1. **Variation of Information :** $VI(C, C') = H(C) + H(C') - 2I(C, C')$
2. **Normalized Mutual Information :** $NMI(C, C') = \frac{2I(C,C')}{H(C)+H(C')}$
3. **Adjusted Mutual Information :** $AMI(C, C') = \frac{I(C,C') - \mathbb{E}(I(C,C'))}{(H(C)+H(C'))/2 - \mathbb{E}(I(C,C'))}$
4. **P-value :** P-value Chi-Square independence test.

Where $P(k) = \frac{n_k}{n}$, $P(k, k') = \frac{|C_k \cap C'_{k'}|}{n}$, $H(C) = -\sum_{i=1}^{K} P(k) log(P(k))$ and $I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') log(\frac{P(k,k')}{P(k)P(k')})$