Hierarchical Optimal Transport for Document Clustering

Pierre Osselin
ENS Paris-Saclay
61 Avenue du Président Wilson, 94235 Cachan
pierre.osselin@gmail.com

Abstract

Measuring distances between documents is a fundamental problem that is crucial for many different tasks. From document retrieval to document classification through document clustering, defining document topologies that make sense is a very difficult exercise that is necessary in order to achieve a computer understanding of human language. While classical distances have been designed through the computation of relevant statistics from corpus such as TF-IDF or Bag of Words, modern approaches learn word embedding in a latent space where euclidean operations matches semantic interpretation. In this paper we refine an approach that leverages the mathematically founded optimal transport methods to conceive meta-distances of document with regards to topics. This describes a hierarchical distance where optimal transport is performed for both the distance between topics and the distance between documents. We use this distance to conceive a new k-means algorithm for clustering where we make use of Wasserstein barycenters and show better performance fur clustering tasks on different data sets.

1 Introduction

Problem. Conceiving meaningful distances between documents is a major challenge for Natural Language Processing (NLP). From Document retrieval [2] to document clustering and news categorization ([11], [6]), through song identification [5], and multilingual document matching [15], translating the semantic understanding of text to mathematical topologies and geometries on documents is the cornerstone of success in building smarter and more efficient document handling systems.

The journey of the development of efficient tools for measuring text distances began with the birth of two (now) classical approaches using Bag Of Words [1] and TF-IDF [9] representation of texts. Though simple and convenient, these representations are very limited, in the sense that they use basic frequency statistics of documents to represent their similarity and thus fails to capture similarities between sentences that express the same meaning with different words or synonyms.

More recently, the statistical machine learning side made huge improvements by introducing statistical generative models that aims at capturing the underlying mechanisms explaining the generation of the data under the form of small latent low-dimensional representations. The Latent Dirichlet Allocation (LDA) [8] is the most famous methods among them. This probabilistic form also allowed for probabilistic modeling refinements such as hierarchical structure with other model such as the Dirichlet Process [16]. This model typically represent a topic as a distribution over words, and documents as distribution over topics.

The emergence of Deep Learning brought major breakthroughs in the problem of word embedding. Word2Vec [19] and GloVe [20] are models that generate word embeddings of unprecedented quality and scales naturally to very large data sets. Furthermore, an impressive result showed by the author is

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

that semantic relationships translates into vector operation. For example, vec(Berlin) - vec(Germany) + vec(France) is close to vec(Paris).

As a result, these recent advances have been the foundation of many techniques for manipulating documents. These representations can be interpreted as distributions of documents over the words embedding, or point cloud of words, and thus can be easily applied to optimal transport techniques. Kusner et al. [22] have very recently leverage the representational power of these embedding to introduce a new distance between documents called *Word Mover's Distance* (WMD). In this setting, the distance between documents is the minimum cumulative distance that words from document A need to travel to match exactly the point cloud of document B. However, this technique still suffers from computational complexity.

Previous works. As previously mentioned, measuring document dissimilarity is capital in many applications and is intrinsically linked to document representation. The classical method to do so relies on Bag-Of-Words where documents are represented as vectors of length $|V|$ where V is the vocabulary, and where the index i represents the number of times the word V_i appears in the document. Some variants include tf-idf where the importance of the word in the whole corpus is taken into account, which enrich the power of the representation.

Similarly, generative models such as LDA model the documents as distributions over topics, and topics as distributions over words. The document is then represented by its amount of involvement in every topics. Depending on which algorithm to use, LDA can be performed using EM-Algorithm [8] or Gibbs Sampling [13]. Equipped with l_2 norm, these representations are an example of dissimilarity score.

More recently, [19] introduced Word2Vec, a procedure trained in an unsupervised setting that learns very rich embeddings of words. This method drastically improved NLP tasks and prompted other variants of Word2Vec such as GloVe [20] to be developped. Based on this approach, [22] introduced Words Mover Distance (WMD) where the authors considered the distance between two documents to be the transport cost from one document, seen as a distribution over words through nBow representation, to the other. In this case, and to leverage the embedding of the words, the cost between two words (points in the cloud representing a document), is the classical l_2 norm between the two embedding vectors. This distance showed better results than the previous methods on many different k-NN tasks. However, this method is limited by its computational cost, as an optimal transport is hard to compute. Other refinements of this distance has been developed, such as [23], [25] or [24] to improve computational complexity but lack topic interpretability.

Finally, [28] introduced a hierarchical optimal distance over documents that makes use of topic modelling coupled with optimal transport to create a metric that uses both embedding, topic representation, and greatly improve computational complexity. In this setting, documents are represented as distributions over topic via the inference of the generative model LDA and topics are represented as distributions over words. Then, distance is computed via optimal transport between distribution of topics where the cost matrix is computed via optimal transport between the topics where the cost matrix is the l_2 norm between vectors embeddings.

The number of topics for this hierarchical distance has to be fixed in advance, or inferred with some methods such as the Dirichlet Process [16], a nonparametric bayesian model. This distance is very interesting because it is highly interpretable, as the distance between two documents can be broken down and explained as the sparse distances between few individual topics, and it naturally incorporates the knowledge encoded in the word2vec space and leads to high retrieval accuracy : it outperforms most of the state-of-the-art alternative document distances in many real world classification task.

Contributions. In this paper we recap the functioning of the HOTT distance on which we will build a new type of k-means clustering using Wasserstein Barycenters. This clustering algorithm combines both the topic modelling and the embedding techniques to achieve better performance than classical clustering techinques for documents.

2 Background

Discrete Optimal Transport Optimal Transport is a very rich mathematical field, see [14] for a complete comprehensive book, that has application in many different domains (see [27] for

applications). In particular, the Word Movers Distance makes use of Discrete Optimal Transport as follows: Given $x = \{x_1, x_2, ...x_n\}$ and $y = \{y_1, y_2, ...y_m\}$ be two sets of points (sites) in a metric space. Let $\Delta^n \subset \mathbb{R}^{n+1}$ denote the probability simplex on n elements, and let $p \in \Delta^n$ and $q \in \Delta^m$ be distributions over x and y. Then, the 1-Wasserstein distance between p and q is :

$$W1(p,q) = \begin{cases} min_{\Gamma \in \mathbb{R}_+^{n \times m}} & \sum_{i,j} C_{i,j} \Gamma_{i,j} \\ s.t & \sum_j \Gamma_{i,j} = p_i \text{ and } \sum_i \Gamma_{i,j} = q_j \end{cases} \tag{1}$$

where the cost matrix C has entries $C(i,j) = d(x_i, y_j)$ where $d(.,.)$ denotes the distance. The constraints allow Γ to be interpreted as a transport plan or matching between p and q. The current state of the art for solving this linear problem exactly is called the Hungarian algorithm [10] with a complexity in $O(l^3 log(l))$ where $l = max(n, m)$.

Entropic regularization Cuturi et al. [17] cope with this complexity problem by introducing entropic regularization that allows for the use of efficient convex optimisation tools. It works by changing the previous objective function by:

$$W_\epsilon(p,q) = \begin{cases} min_{\Gamma \in \mathbb{R}_+^{n \times m}} & \sum_{i,j} C_{i,j} \Gamma_{i,j} - \epsilon E(\Gamma) \\ s.t & \sum_j \Gamma_{i,j} = p_i \text{ and } \sum_i \Gamma_{i,j} = p_j \end{cases} \tag{2}$$

With the entropy defined as:

$$E(\Gamma) = -\sum_{i,j} \Gamma_{i,j} (log(\Gamma_{i,j}) - 1) \tag{3}$$

The regularized transportation problem can be re-written as a projection:

$$W_\epsilon(p,q) = \begin{cases} \epsilon \, min_{\Gamma \in \mathbb{R}_+^{n \times m}} & KL(\Gamma, K) \text{ where } K = e^{-\frac{C}{\epsilon}} \\ s.t & \sum_j \Gamma_{i,j} = p_i \text{ and } \sum_i \Gamma_{i,j} = p_j \end{cases} \tag{4}$$

Where KL is the KullBack-Leibler divergence defined as:

$$KL(\Gamma|K) = \sum_{i,j} \Gamma_{i,j} \left(\log \left(\frac{\Gamma_{i,j}}{K_{i,j}} \right) - 1 \right) \tag{5}$$

This reformulation using entropic regularization can accelerate OT in learning environments. It is most successful when the support of the distributions is large. This algorithm has a computational complexity in $O(\frac{l^2}{\epsilon^2})$. This method will be very useful when we will talk about Wasserstein barycenters in the next section.

Wasserstein Barycenters. In the conception of a k-means algorithm, a fundamental step is to compute the barycenters of the points of the same label. In our case we manipulate documents, and thus this algorothm requires the computation of measure barycenters, called *Wasserstein Barycenters*. This computation is a very complex task, and is studied and solved computationally using Entropic regularization in [21]. This problem formulates as:

$$min_b \sum_k \lambda_k \times W_\epsilon(a_k, b) \tag{6}$$

where $(\lambda_k)_k$ are positive and sum to one, and $(a_k)_k$ is a set of measures from which we want to compute the barycenter b. This problem can be solved fastly using iterative Sinkhorn's Algorithm as explained in [21].

3 Hierarchical Optimal Transport for Clustering

HOTT. The principles of the Hierarchical optimal topic transport is quite simple. Given the output topics given by a topic generative model such as LDA, we can define documents as distributions over topics.

$$\forall i \in \{1, ..., n\}, \; d^i = \sum_{j=1}^{|T|} d^i_j \delta_{t_j} \tag{7}$$

Where $d^i \in \Delta^{|T|}$ and $T = \{t_1, ..., t_{|T|}\}$
Given this representation, we can then define the HOTT as follows:

$$HOTT(d^i, d^j) = W_1\left(\sum_{k=1}^{|T|} d_k^i \delta_{t_k}, \sum_{k=1}^{|T|} d_k^i \delta_{t_k}\right) \tag{8}$$

This representation is justified by the will to exploit the information given by topics on top of words embedding as well as greatly reducing the cost of WMD by computing the optimal transport on a set of $|T|$ points instead of the length of the document. However, in this representation we also need to define the distances between topics to obtain a cost matrix.
The cost matrix of this transport can be pre-computed by scanning the corpus once and is given by:

$$d(t_i, t_j) = WMD(t_i, t_j) \tag{9}$$

Where here topics are distribution over words and can then be treated like documents in the method of WMD conceived in [22]. The term "Hierarchical" resides in these two steps where distances between topics are first computed using optimal transport and then distances between documents is computed using the topics.

The distribution of words in random documents typically follows the Zipf Law as illustrated in [3]. This typically means that a few words account for the entire document. We demonstrated the same phenomena with regards to the considered data set, after stemming and removal of stop words in the next session. This observation suggests that, in order to improve computational efficiency during the evaluation of topic distances, we can truncate the number of words in the computation of the topic distances. Which is justified in [28] paper by performing sensitivity analysis.

Additionally, one can perform the same approximation regarding the number of topics, as shown in the paper [28]. We show our analysis of the statistics of the topics in our data set in the next section. Both previous methods will be implemented and will allow us to improve accuracy.

k-means. K-means algorithm is a very classical non parametric method for clustering points into k clusters. Given a set of points in a metric space, the algorithm alternatively compute the barycenters of the considered clusters and reupdates the entire data set by labeling a point to its closest barycenter. k-means is usually applied in an euclidean space where barycenters are easily computed. In the case of manipulating documents, the problem of constructing a representational space allowing for clustering is very hard. Here we introduce a new clustering algorithm based on the work of [28], where we will consider the HOTT as the distance between documents, and the Wasserstein barycenters of our cluster as the barycenters used in the k-means algorithm.

This algorithm leverage the strength of the accuracy HOTT distance as well as its low computational complexity. Computing the real Wasserstein Barycenters according to the HOTT at every iteration is a very complex task, and we approximate it by using the Wasserstein Barycenter with Entropic Regularization introduces earlier, we will study the impact of this approximation in the next section.

Theory. As shown in [28], we can obviously see that when $|T|$ increases and tends to the size of the vocabulary, we recover the WMD distance. Hence, one can see $|T|$ as the level of granularity for the topics as well as a trade-off between computational complexity and precision, although [28] shows in its sensitivity analysis that the choice of number of topics has not a significant impact on certain data sets. [28] make a further theoretical analysis between WMD and HOTT, more specifically, by the triangle inequality:

$$WMD(d^i, d^j) \leq W_1\left(d^i, \sum_{k=1}^{|T|} d_k^i t_k\right) + W_1\left(\sum_{k=1}^{|T|} d_k^i t_k, \sum_{k=1}^{|T|} d_k^j t_k\right) + W_1\left(\sum_{k=1}^{|T|} d_k^j t_k, d^j\right) \tag{10}$$

The LDA is a generative model whose inference can be made via Gibbs Sampling, the resulting learning aims at minimizing the KL divergence $KL(d^i || \sum_{k=1}^{|T|} d_k^i t_k)$ over the topic proportions d_k^i. Moreover, one can show [4] that:

$$W_1(\mu, \rho) \leq diam(X)\sqrt{\frac{1}{2} KL(\mu||\rho)}$$

Furthermore, the middle term satisfies:

$$W_1\left(\sum_{k=1}^{|T|} d_k^i t_k, \sum_{k=1}^{|T|} d_k^j t_k\right) \leq W_1\left(\sum_{k=1}^{|T|} d_k^i \delta_{t_k}, \sum_{k=1}^{|T|} d_k^j \delta_{t_k}\right)$$

where the right term is $HOTT(d1, d2)$. The optimal topic transport on the right implies an equal-cost transport of the corresponding linear combinations of topic distributions on the left. The inequality follows since W1 gives the optimal transport cost. Combining into a single inequality:

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

Complexity. It has been showed [26] that the computational complexity to compute the Regularized Wasserstein Barycenters from Bregman Iterations is of $\mathcal{O}(\frac{mn^2}{\epsilon^2})$ where m is the number of measures we want to compute the Barycenter from and p is the size of the support of the measures, in our case the number of Topics. Similarly a k-means algorithm alternates between computing Barycenters and reassigning data, which requires pairwise computation between data and the newly formed barycenters. We obtain the final complexity :

$$Complexity = \mathcal{O}(n_{iter} \times \underbrace{|D| \times k \times |T|^3 log(|T|)}_{\text{Pairwise Distances}} \times \quad \underbrace{|D| \times \frac{|T|^2}{\epsilon^2}}_{\text{Computation Barycenters}} \quad) \tag{11}$$

where D is our data set of documents, k the number of clusters ofour algorithm, $|T|$ the number of topics chosen, ϵ the Entropic Regularization for our Barycenter computation and n_{iter} the number of maximum iterations of our k-means algorithm. As we can see, the complexity of our algorithm mainly comes from the number of topics that we choose to take. A quadratic complexity in the data is expected for such k-means algorithm.

3.1 Experiments

Data sets. In this section we test our methods on 6 data sets, the data set bbcsport composed of 737 documents from the BBC Sport website corresponding to sports news articles, and labeled into 5 sport classes. The data set twitter composed of 3108 tweets categorized in 3 sentiments. The data set classic composed of 7093 documents classified into 4 labels. The data set ohsumed composed of 9152 documents categorized in 10 labels. The data set r8 composed of 7674 and 8 labels. Finally, amazon data set is composed of 8000 documents and 4 labels [1].

Topic and Word Truncation. In this subsection we illustrate the behaviour referred in the previous section. We computed the frequency plot of words in our data sets. We obtain the Figures 1 and 2. We can see that the log-log plot of this frequency is sub-linear. In this figure, we can see that very few words accounts for the majority of the documents. Although not sufficient, this indicates a further argument in addition to the sensitivity analysis performed by [28] in the influence of word truncation in HOTT performance. Similarly, we plotted the topic weights in the documents of our data sets and obtained the Figures 3 and 4, which gives insights into our data. We used uncertainty propagation for the computation of the standard deviations.

k-means Clustering To test the performances of our k-means algorithm we perform k-means clustering using different methods. We first use the classical method using tf-idf, where we transform our documents under the form of nBow into Tf-Idf representation and perform k-means clustering with it. The second method used is LDA, where we infer a certain number of topics from the documents using the algorithm used in [28]. We choose a number of topics equals to the number of labels and cluster according the the dominant topic. Finally, we use HOTT and Wasserstein barycenters with regularization with $\epsilon = 0.01$ to leverage our Hierarchical Metric and perform k-means clustering. To evaluate our methods, we compare the clusters inferred with the true labels of the data, using four metrics, The Variation of Information, the Normalized Mutual Information, the Adjusted Mutual Information and the P-value. These metrics are more discussed in appendix. The results of this study are given table 1, 2, 3and 4. To obtain these results, we used a maximum number of iterations in the k-means algorithm equals to 50, and we average out our results on 50 experiments where the k-means algorithm is initialized randomly by selecting k random documents as initial barycenters. The first number comes from empirical convergence of the k-means algorithms on our data sets, the second

[1] The data sets are available at `https://www.dropbox.com/sh/nf532hddgdt68ix/AABGLUiPRyXv6UL2YAcHmAFqa?dl=0` from the paper [22].

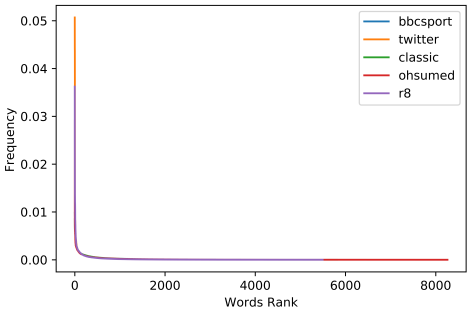

Figure 1: Frequency plot of the data set

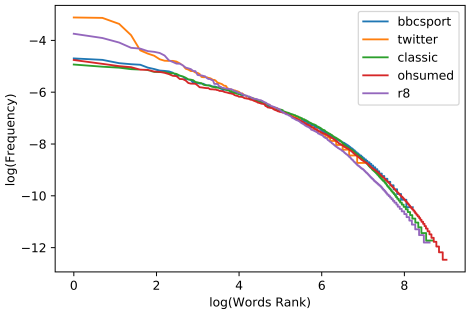

Figure 2: Log-Log Frequency plot of the documents

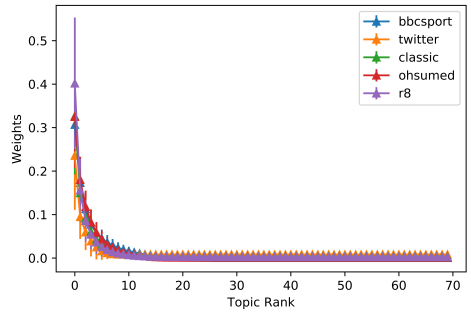

Figure 3: Weight plot of the Topics

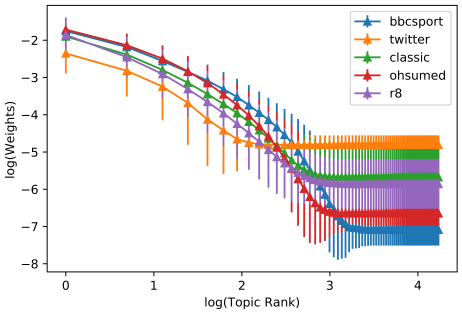

Figure 4: Log-Log Weight plot of the topics

from the constraints arising from our computational power in order to have results in a reasonable amount of time.

From the tables we can see that using HOTT for k-means clustering is almost always equivalent or better than clustering with LDA. This is due to the refinement involved by the use of the Glove embedding to measure distances between topics.

One can notice that when Clustering with Topics gives bad performance, our method gives bad performance as well, as with the case of the twitter data set, which is a very difficult data set for Topic Modelling since it is composed of many documents with very few words and a large vocabulary.

In two of our studied data sets (bbcsport and ohsumed), the improvement from HOTT is significant according to our mutual information measures of performance.

Sensitivity Analysis. We perform sensitivity analysis of our k-means algorithm according to two parameters. The first one is the number of Topics used as intermediary layer of information for our HOTT distance. We plotted in Figures 8, 9 and 10 the influence of the number of topics for our HOTT clustering performances. We took the topics $t \in \{2, 5, 10, 15, 20, 30, 40, 50, 70\}$ and $\epsilon = 0.01$. We can observe that our metrics attain a maximum for few topics (typically 5 to 15 for the data set bbcsport, and that additional topics harm the performances.

The second one is the parameters ϵ that governs our regularization for the computation of our Wasserstein Barycenters. Increasing ϵ increase the computational power but bias the actual barycenter computed. Figures 11, 12 and 13 shows the result for $\epsilon \in \{0.01, 0.05, 0.1, 0.2, 0.35, 0.5, 0.7, 1\}$ averaged over 30 runs. The results confirm our expectations, as ϵ increases, the variation of information increases as well and both Mutual Information scores plummet.

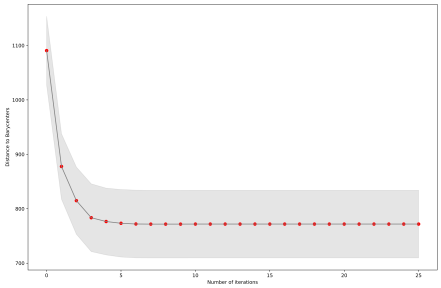

Figure 5: Transport Distance from the Barycenters

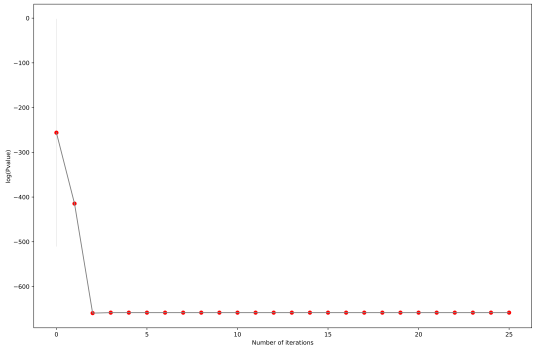

Figure 6: Log P value

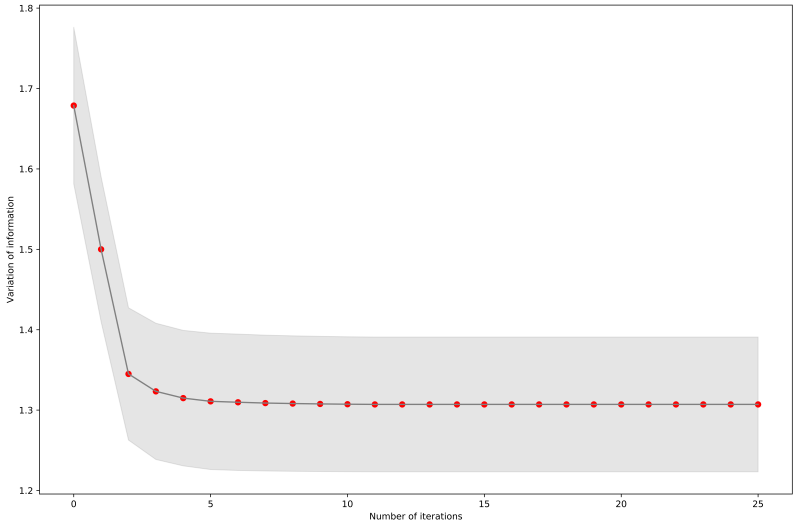

Figure 7: Evolution of the Variation of Information

4 Conclusion

Summary. We built a k-means algorithm on top of the hierarchical optimal distance introduced by [28] that leverages optimal transport, topic modeling, and word embedding. This distance has the advantage of providing global semantic language information, while LDA topic models provide corpus-specific topics and topic distributions. Thus a k-means algorithm based on this distance is expected to give more meaningful clusters. This is shown empirically where our algorithm perform equally or better than the other clustering methods on a wide range of different data sets. HOTT appears to capture differences in the same way a person asked to compare two documents would: by breaking down each document into easy to understand concepts, and then comparing the concepts.

Future Work. There are many avenues for future work. From a theoretical perspective, the results encountered were worse than expected. While HOTT performs better on k-NN tasks compared to LDA, our k-means algorithm only beat the LDA clustering algorithm twice, even on data where the use of word embedding is supposed to bring a significant advantage. One avenue of research would be both 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.

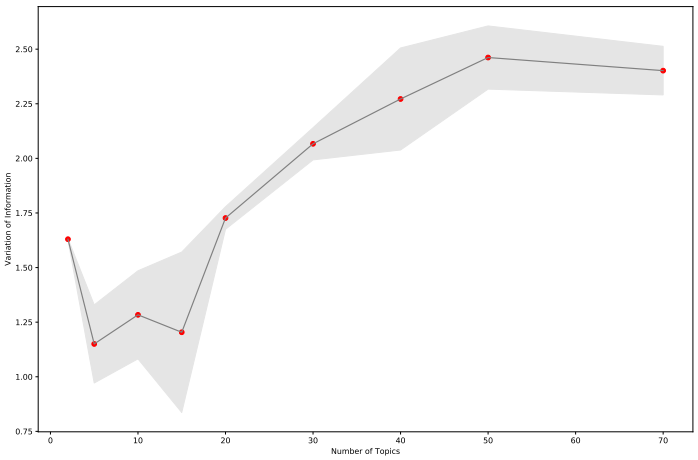

Figure 8: bbcsport VI sensitivity toward topics

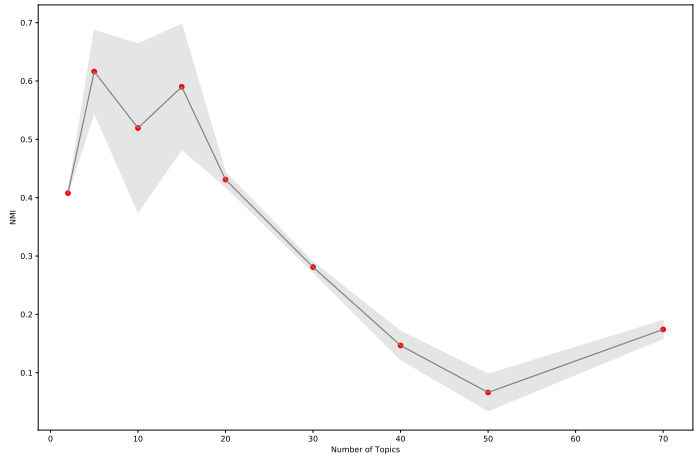

Figure 9: bbcsport NMI sensitivity toward topics

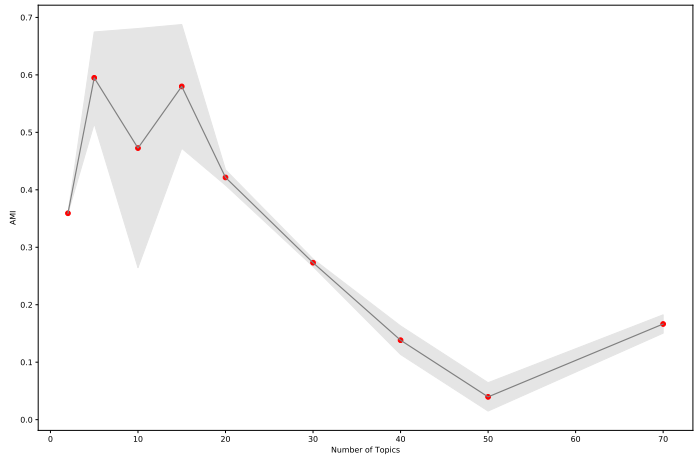

Figure 10: bbcsport AMI sensitivity toward topics

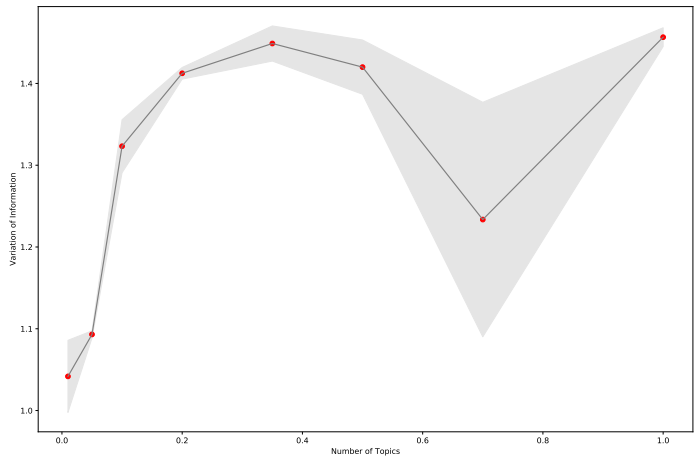

Figure 11: bbcsport VI sensitivity toward ϵ

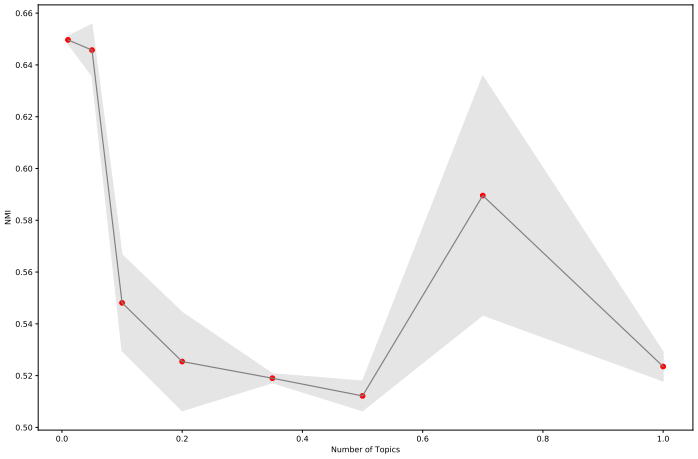

Figure 12: bbcsport NMI sensitivity toward ϵ

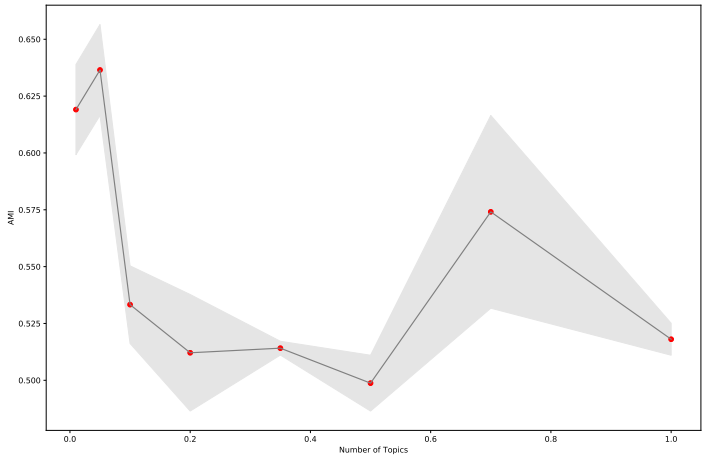

Figure 13: bbcsport AMI sensitivity toward ϵ

Dataset	Tf-Idf	LDA	HOTT
bbcsport	1.41 ± 0.23	1.35	$\mathbf{1.28 \pm 0.20}$
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**	$\mathbf{1.39 \pm 0.12}$
amazon	1.35 ± 0.15	**1.08**	$\mathbf{1.08 \pm 0.02}$

Table 1: Variation of Information

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

Table 2: NMI Score

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 3: AMI Score

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

Table 4: log P Value Score

References

[1] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[2] Gerard Salton et al. *The SMART System—Experiments in automatic document processing*. 1971.

[3] Wentian Li. "Random texts exhibit Zipf's-law-like word frequency distribution". In: *IEEE Transactions on information theory* 38.6 (1992), pp. 1842–1845.

[4] Felix Otto and Cédric Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[5] Eric Brochu and Nando D Freitas. "Name That Song!" In: *Advances in Neural Information Processing Systems*. 2002, pp. 1505–1512.

[6] Jorg Ontrup and Helge Ritter. "Hyperbolic self-organizing maps for semantic navigation". In: *Advances in neural information processing systems*. 2002, pp. 1417–1424.

[7] Alexander Strehl and Joydeep Ghosh. "Cluster ensembles—a knowledge reuse framework for combining multiple partitions". In: *Journal of machine learning research* 3.Dec (2002), pp. 583–617.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[9] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. Piscataway, NJ. 2003, pp. 133–142.

[10] Harold W Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics (NRL)* 52.1 (2005), pp. 7–21.

[11] Derek Greene and Pádraig Cunningham. "Practical solutions to the problem of diagonal dominance in kernel document clustering". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 377–384.

[12] Marina Meilă. "Comparing clusterings—an information based distance". In: *Journal of multivariate analysis* 98.5 (2007), pp. 873–895.

[13] Mark Steyvers and Tom Griffiths. "Probabilistic topic models". In: *Handbook of latent semantic analysis* 427.7 (2007), pp. 424–440.

[14] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.

[15] Novi Quadrianto, Le Song, and Alex J Smola. "Kernelized sorting". In: *Advances in neural information processing systems*. 2009, pp. 1289–1296.

[16] Yee Whye Teh. "Dirichlet process". In: *Encyclopedia of machine learning* (2010), pp. 280–287.

[17] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems*. 2013, pp. 2292–2300.

[18] Mary L McHugh. "The chi-square test of independence". In: *Biochemia medica: Biochemia medica* 23.2 (2013), pp. 143–149.

[19] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[20] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[21] Jean-David Benamou et al. "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[22] Matt Kusner et al. "From word embeddings to document distances". In: *International conference on machine learning*. 2015, pp. 957–966.

[23] Xinhui Wu and Hui Li. "Topic mover's distance based document classification". In: *2017 IEEE 17th International Conference on Communication Technology (ICCT)*. IEEE. 2017, pp. 1998–2002.

[24] Lingfei Wu et al. "Word Mover's Embedding: From Word2Vec to Document Embedding". In: *arXiv preprint arXiv:1811.01713* (2018).

[25] Hongteng Xu et al. "Distilled wasserstein learning for word embedding and topic modeling". In: *Advances in Neural Information Processing Systems*. 2018, pp. 1716–1725.

[26] Alexey Kroshnin et al. “On the complexity of approximating wasserstein barycenter”. In: *arXiv preprint arXiv:1901.08686* (2019).

[27] Gabriel Peyré, Marco Cuturi, et al. “Computational optimal transport”. In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[28] Mikhail Yurochkin et al. “Hierarchical Optimal Transport for Document Representation”. In: *arXiv preprint arXiv:1906.10827* (2019).

A Evaluate Clusters

Variation of Information One metric used to measure the quality of our clusters is a quantity called Normalized Mutual Information [7]. To evaluate the clustering quality we used a distance based on information theory [12]. Given two clusters C and C' we define the variation of Information from changing cluster C into C' by:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \tag{12}$$

Where, with K clusters in C:

$$P(k) = \frac{n_k}{n} \tag{13}$$

$$P(k, k') = \frac{|C_k \cap C'_{k'}|}{n} \tag{14}$$

$$H(C) = -\sum_{i=1}^{K} P(k) log(P(k)) \tag{15}$$

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') log(\frac{P(k, k')}{P(k)P(k')}) \tag{16}$$

One of the great advantage of this method is that it is a true metric over the space of clusterings, which justifies its use and allow for the application of other algorithms.

Normalized Mutual Information Score A second mainstream metric [7] used to compare clusters is the Normalized Mutual Information score (NMI) defined as follow:

$$NMI(C, C') = \frac{2I(C, C')}{H(C) + H(C')} \tag{17}$$

Where $I(C, C')$ and $H(C)$ are defined as in the previous paragraph. This measure gives an indication of how much information is shared between the two clustering. A measure of one gives equal clusters while a measure of zero means that no information is shared.

Adjusted Mutual Information Score The Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. For two clusterings C and C', the AMI is given as:

$$AMI(C, C') = \frac{I(C, C') - \mathbb{E}(I(C, C'))}{(H(C) + H(C'))/2 - \mathbb{E}(I(C, C')} \tag{18}$$

Chi-Square independance test is a statistical test used to assess the degree of independance between two categorical random variables [18]. More specifically, this test considers the two following hypothesis:

$\mathcal{H}_0$: "The two categorical variables X and Y are independent in some population"

$\mathcal{H}_1$: "There is an association between the two categorical random variables X and Y"

The test then compute a statistic, called the *p-value*, that represents the probability of finding our data under the null hypothesis. In our case it means the probability of finding our data if the variables are perfectly independent in the entire population. Usually, the null hypothesis is rejected if $p < 0.05$. The p-value is computed by first building a contingency table between our two categorical variables, then, we compute the chi-square statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Where

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N}$$

And

$$p = \mathbb{P}(X > \chi^2)$$

where X is a chi square distribution with two degrees of freedom.