Hierarchical Optimal Transport for Document Representation

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin Solomon

Contributions

We introduce *hierarchical* optimal transport to measure dissimilarities between distributions with common structure. Our approach:

- Is **computaionally efficient**;
- Provides **higher level interpretability**;
- Is **practical** for large corpora.

Word Mover's Distance

The 1-Wasserstein distance between p and q is

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

where the cost matrix C has entries $C_{i,j} = d(x_i, y_j)$, where $d(\cdot,\cdot)$ denotes the distance. The constraints allow Γ to be interpreted as a transport plan or matching between p and q.

The Word Mover's Distance (WMD) between documents is then $\mathit{WMD}(d^1,d^2) = W_1(d^1,d^2)$, where d^1 and d^2 are normalized word counts and the ground metric is Euclidean in some embedding space.

Hierarchical Optimal Transport

We define the **hierarchical optimal topic transport distance** (HOTT) between documents d^1 and d^2 as

$$\mathit{HOTT}(d^1,d^2) = W_1\left(\sum_{k=1}^{|T|} \bar{d}^1_k \delta_{t_k}, \sum_{k=1}^{|T|} \bar{d}^2_k \delta_{t_k}\right),$$

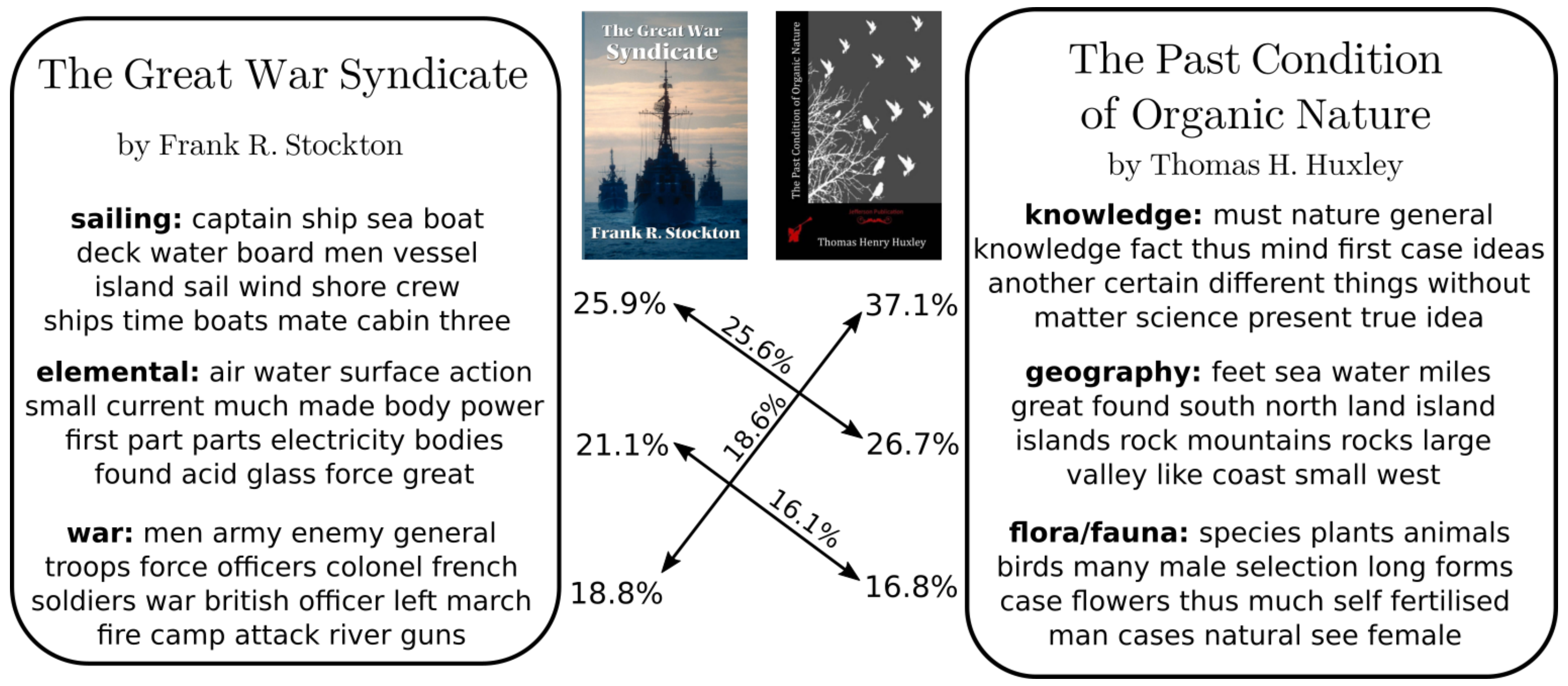
where each Dirac delta δ_{t_k} is a probability distribution only supported on the corresponding topic t_k, yielding the ground metric to be WMD between topics as distributions over words.

Computational Efficiency

	Document pairs per second				
Dataset	RWMD	WMD	WMDT20	HOFTT	HOTT
bbcsport	1494	526	1545	2016	**2548**
twitter	**2664**	2536	2194	1384	1552
ohsumed	454	377	473	829	**908**
classic	816	689	720	980	**1053**
reuters8	834	685	672	918	**989**
amazon	289	259	253	927	**966**
20news	338	260	384	652	**699**
gutenberg	2	0.3	359	1503	**1720**

Interpretability

The additional level of abstraction promotes higher-level interpretability at topic level as opposed to dense word-level correspondences from WMD.



The Great War Syndicate

by Frank R. Stockton

sailing: captain ship sea boat deck water board men vessel island sail wind shore crew ships time boats mate cabin three

elemental: air water surface action small current much made body power first part parts electricity bodies found acid glass force great

war: men army enemy general troops force officers colonel french soldiers war british officer left march fire camp attack river guns

The Past Condition of Organic Nature

by Thomas H. Huxley

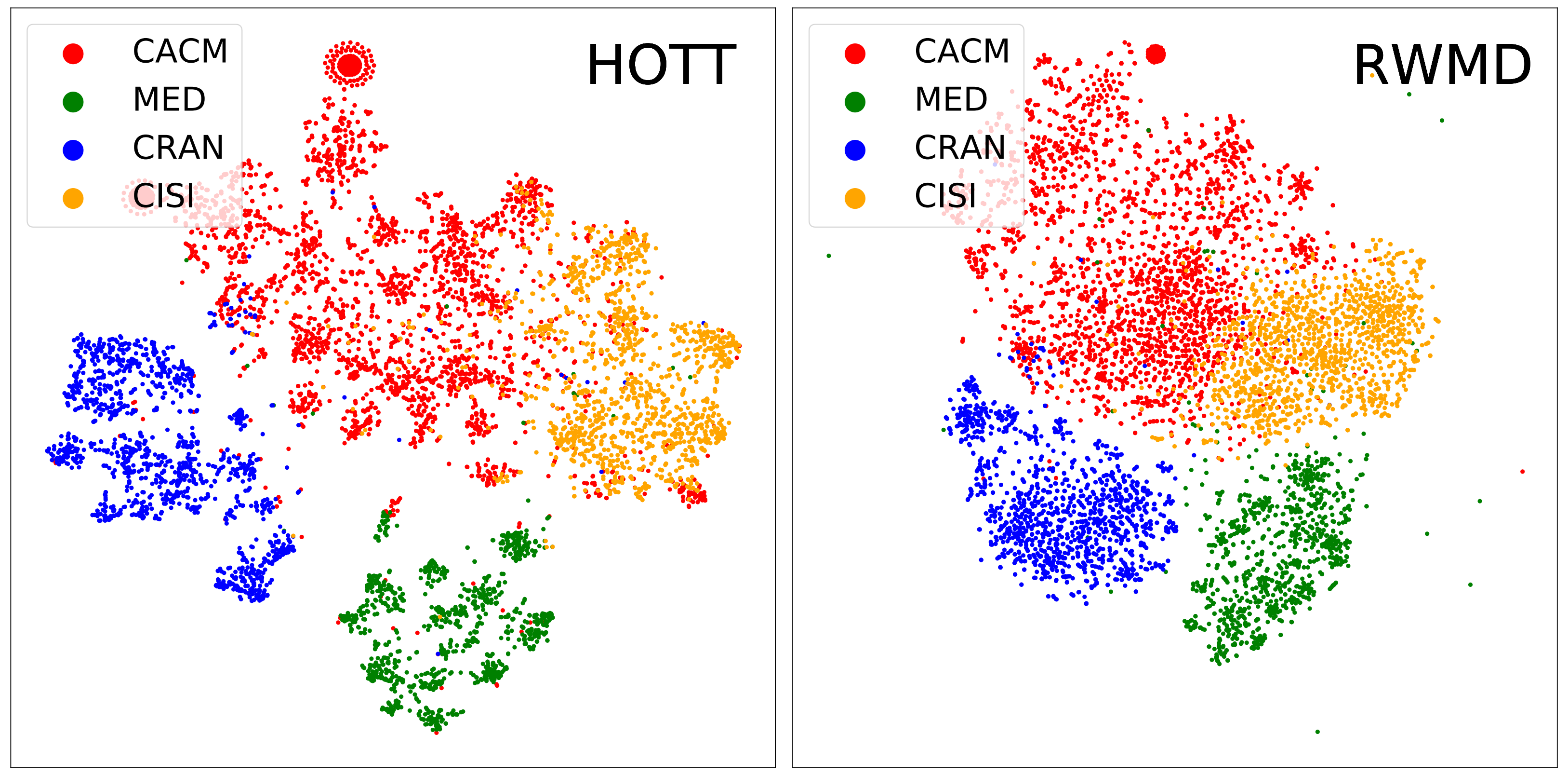
knowledge: must nature general knowledge fact thus mind first case ideas another certain different things without matter science present true idea

geography: feet sea water miles great found south north land island islands rock mountains rocks large valley like coast small west

flora/fauna: species plants animals birds many male selection long forms case flowers thus much self fertilised man cases natural see female

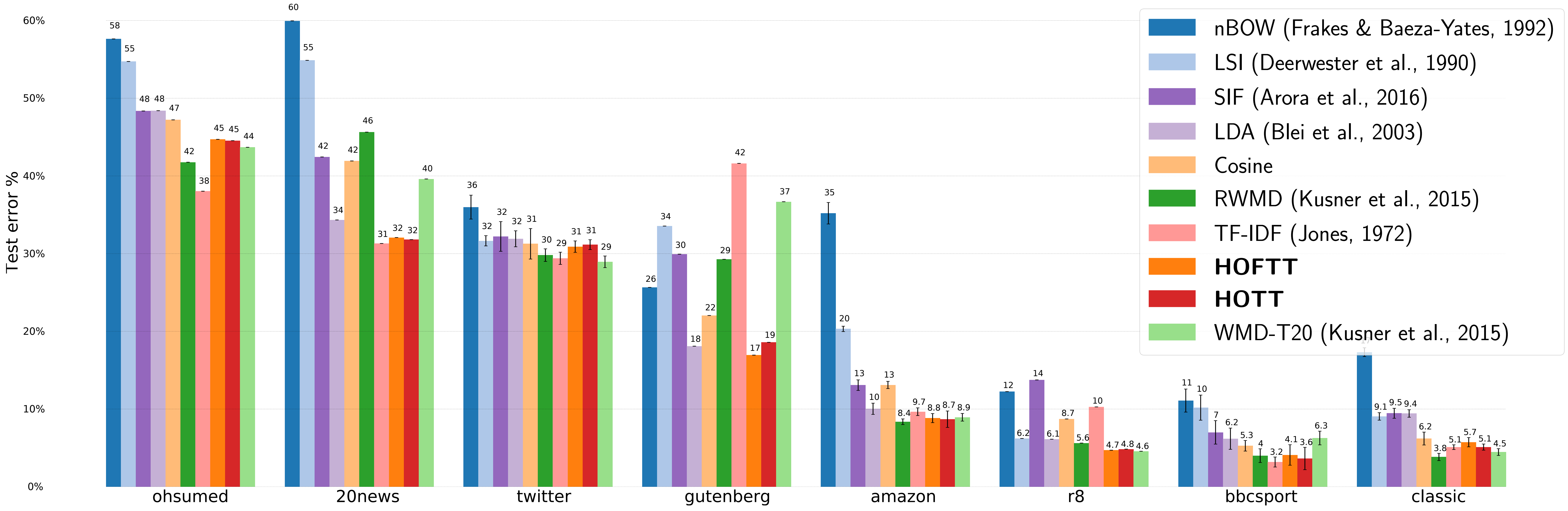
t-SNE Visualization

HOTT is qualitatively better at separating classes under a t-SNE embedding.

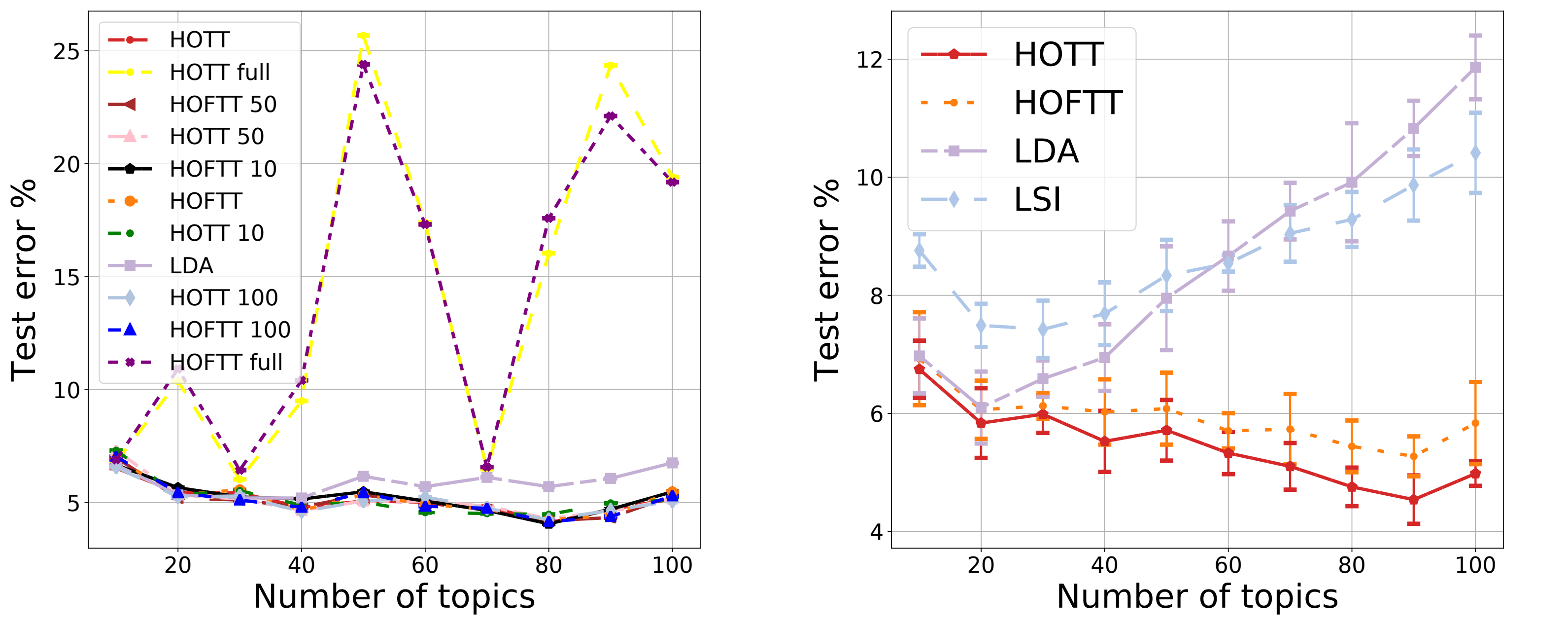


Results

Classification accuracy of a k-NN model on various datasets:



Sensitivity to Parameters



Test error %
Number of topics
HOTT
HOTT full
HOFTT 50
HOTT 50
HOFTT 10
HOFTT
HOTT 10
LDA
HOTT 100
HOFTT 100
HOFTT full
Test error %
Number of topics
HOTT
HOFTT
LDA
LSI