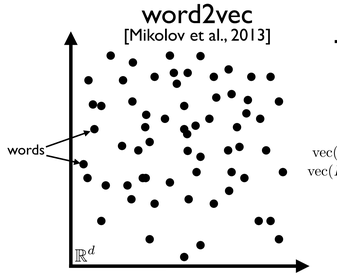
From Word Embeddings To Document Distances

Matt J. Kusner[1], Yu Sun[1,2], Nicholas I. Kolkin[1], Kilian Q. Weinberger[1,2]
[1]Department of Computer Science & Engineering, Washington University in St. Louis, USA
[2]Department of Computer Science, Cornell University, USA

Word Embedding & Document Vectors

word2vec
[Mikolov et al., 2013]



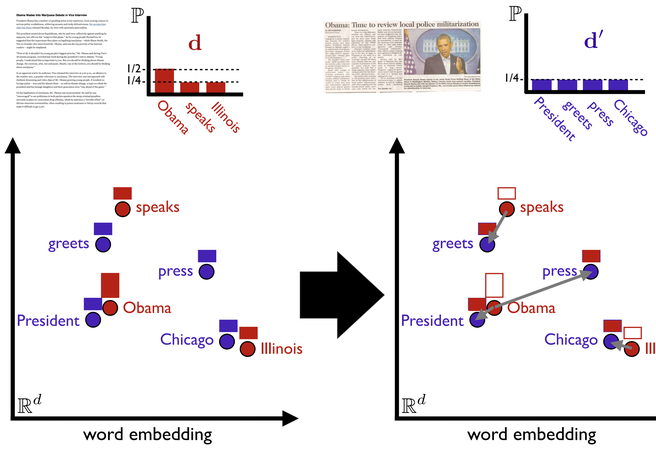
word2vec model
- different from [Collobert & Weston, 2008], [Mnih & Hinton, 2009] word2vec is not deep!
- embeds billions of words / hour with a desktop computer
- learns complex word relationships:

$\text{vec}(Japan)$ - $\text{vec}(sushi)$ + $\text{vec}(Germany) \approx \text{vec}(bratwurst)$
$\text{vec}(Einstein)$ - $\text{vec}(scientist)$ + $\text{vec}(Picasso) \approx \text{vec}(painter)$

Embedded Google News Corpus
- **3 million** different words embedded
- trained on **100 billion** words

How can we leverage this high quality word embedding to compute **document distances?**

The Word Mover's Distance



WMD Optimization

$\text{WMD}(\mathbf{d}, \mathbf{d}') \triangleq$

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$
$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$
$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$

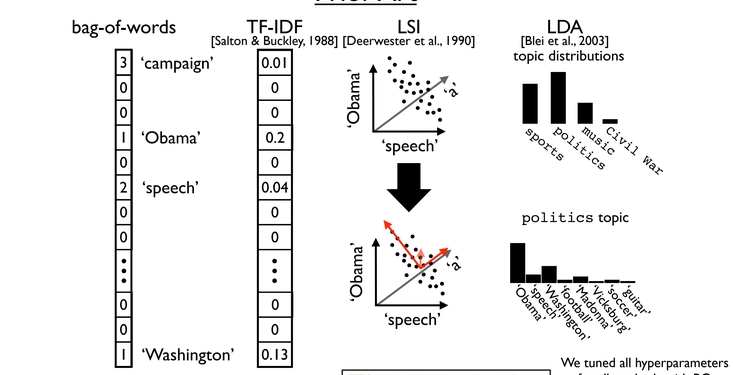
$O(n^3 \log n)$

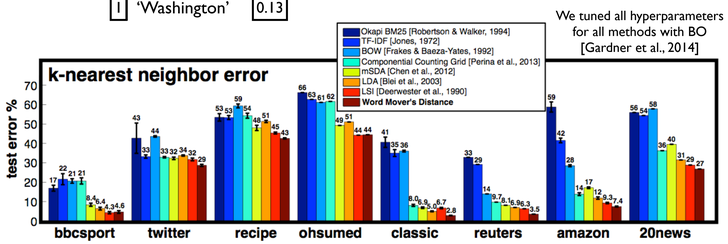
n	number of words in dictionary
d	dimensionality of word embeddings
$\mathbf{d}_i$	normalized BOW (nBOW) weight for word i in first document
$\mathbf{d}'_j$	nBOW weight for word j in second document
$\mathbf{T}_{ij}$	'transport' between word i in **d** and word j in **d'**
$\mathbf{x}_i$	word embedding for word i
$\mathbf{X}$	matrix of all word embeddings (d,n)

in CV this is the Earth Mover's Distance (EMD)
[Rubner et al., 1998]

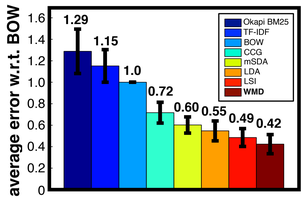
k-Nearest Neighbor Results

Prior Art





average kNN error



different word embeddings

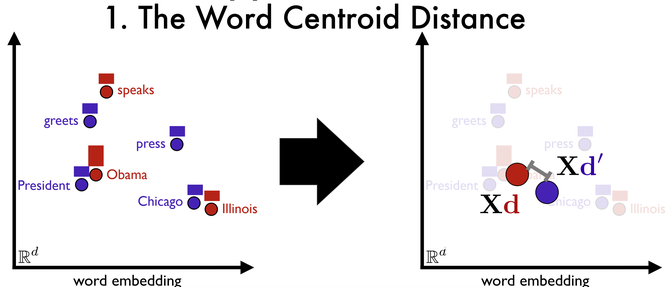
DOCUMENT k-NEAREST NEIGHBOR RESULTS

DATASET	HLBL	CW	NIPS (W2V)	AMZ (W2V)	NEWS (W2V)
BBCSPORT	4.5	8.2	9.5	**4.1**	5.0
TWITTER	33.3	33.7	29.3	**28.1**	28.3
RECIPE	47.0	51.6	52.7	47.4	**45.1**
OHSUMED	52.0	56.2	55.6	50.4	**44.5**
CLASSIC	5.3	5.5	4.0	3.8	**3.0**
REUTERS	4.2	4.6	7.1	9.1	**3.5**
AMAZON	12.3	13.3	13.9	7.8	**7.2**

HLBL: [Mnih & Hinton, 2009]
CW: [Collobert & Weston, 2008]
W2V: [Mikolov et al., 2013]

Approximations

1. The Word Centroid Distance



$$\text{WCD}(\mathbf{d}, \mathbf{d}') \triangleq \|\mathbf{Xd} - \mathbf{Xd}'\|_2$$

$O(nd)$

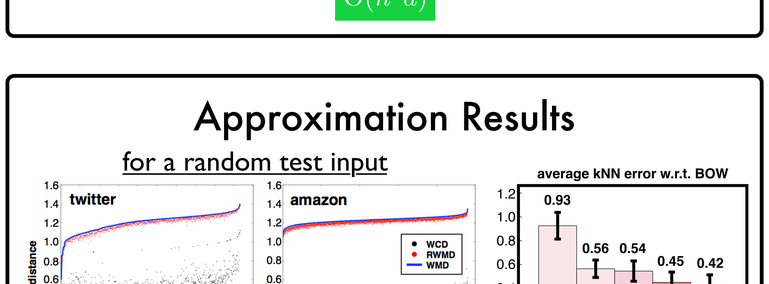
Approximations

2. The Relaxed Word Mover's Distance

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i \quad \sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad s.t. \quad \sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$



$$\text{RWMD}(\mathbf{d}, \mathbf{d}') \triangleq \max \left\{ \sum_{i,j=1}^{n} \mathbf{T}^*_{1,ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2, \sum_{i,j=1}^{n} \mathbf{T}^*_{2,ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \right\}$$

$O(n^2 d)$

Approximation Results

for a random test input

References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In NIPS, pp. 3111– 3119, 2013
2. Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In ICML, pp. 160–167. ACM, 2008.
3. Mnih, A. and Hinton, G. E. A scalable hierarchical dis- tributed language model. In NIPS, pp. 1081–1088, 2009.
4. Rubner, Y., Tomasi, C., and Guibas, L. J. A metric for distributions with applications to image databases. In ICCV, pp. 59–66. IEEE, 1998.
5. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information processing & man- agement, 24(5):513–523, 1988.
6. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.
7. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. Journal of Machine Learning Research, 3: 993–1022, 2003.
8. Gardner, J., Kusner, M. J., Xu, E., Weinberger, K. Q., and Cunningham, J. Bayesian optimization with inequality constraints. In ICML, pp. 937–945, 2014.