Same content. Different words. Word Mover’s Distance

Rishab Goel

Master’s CS @ IIT Delhi

@RishabGoel

Business Problem

All reviewers are raving about the same thing

“The Sicilian gelato was extremely rich”

“The Italian ice-cream was very velvety”

What about Ambiance, Service and Prices?

Let’s filter “gelato” out and add other aspects!

Credit: Sudeep Das @datamusing applied WMD to restaurant reviews. http://tech.opentable.com/2015/08/11/navigating-themes-in-restaurant-reviews-with-word-movers-distance/

Ways to find similar documents

- Count common words (bag of words, TF-IDF)
    - #Dimensions = #Vocabulary (thousands)

Stuck if no words in common.
“Gelato” != “Ice-cream”

Ways to find similar documents

- Low-dimensional latent features
  - Eigen-values (LSI)
  - Probability (LDA)

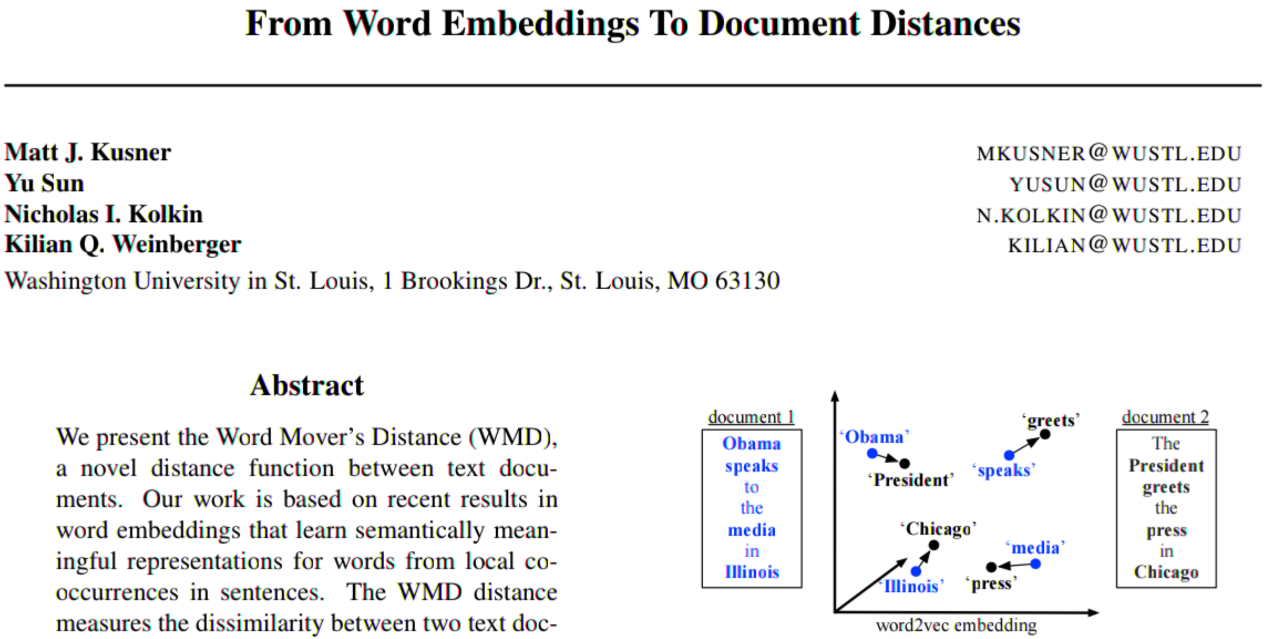
Good representation But …
There is something better now… WMD!

New way to find similar documents

- Word Mover’s Distance
  - Built on top of Google’s word2vec
  - Well-used concept in other fields known as Earth Mover’s Distance

Beats BOW, TF-IDF, LDA, LSI in Nearest Neigbours document classification tasks.

Word Mover’s distance



From Word Embeddings To Document Distances

Matt J. Kusner MKUSNER@WUSTL.EDU
Yu Sun YUSUN@WUSTL.EDU
Nicholas I. Kolkin N.KOLKIN@WUSTL.EDU
Kilian Q. Weinberger KILIAN@WUSTL.EDU
Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130

Abstract

We present the Word Mover’s Distance (WMD), a novel distance function between text documents. Our work is based on recent results in word embeddings that learn semantically meaningful representations for words from local co-occurrences in sentences. The WMD distance measures the dissimilarity between two text doc-

http://jmlr.org/proceedings/papers/v37/kusnerb15.pdf
https://github.com/mkusner/wmd

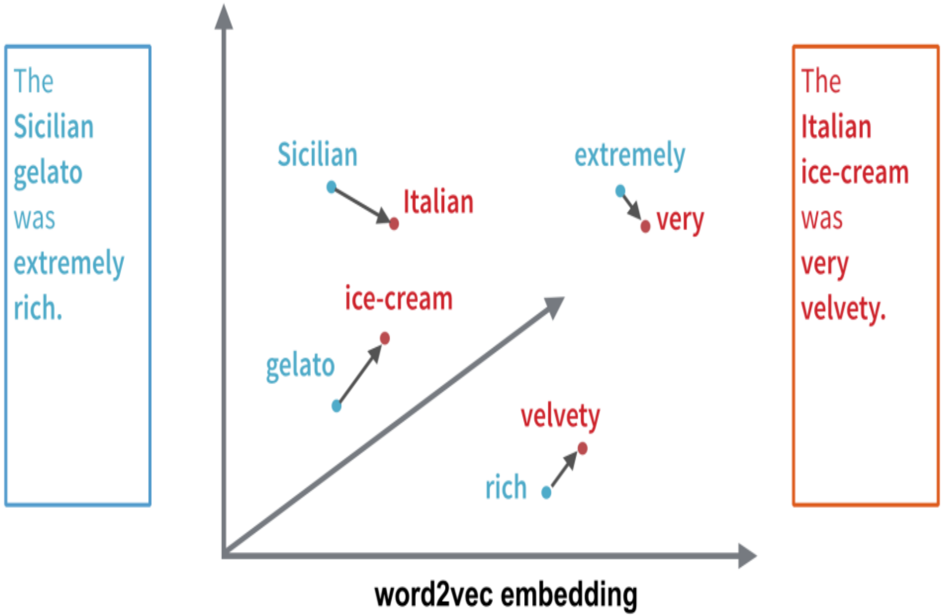
Word Mover’s distance

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to:} \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1,\ldots,n\}$$

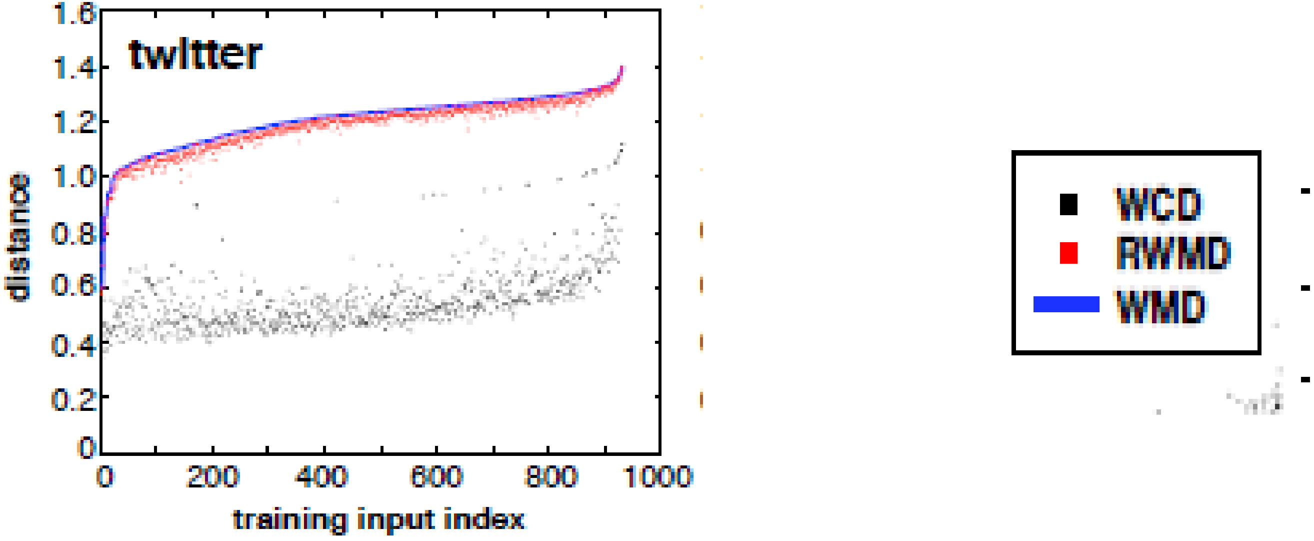
$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1,\ldots,n\}.$$

Optimization Expression


The Sicilian gelato was extremely rich.
Sicilian
Italian
extremely
very
ice-cream
gelato
velvety
rich
word2vec embedding
The Italian ice-cream was very velvety.

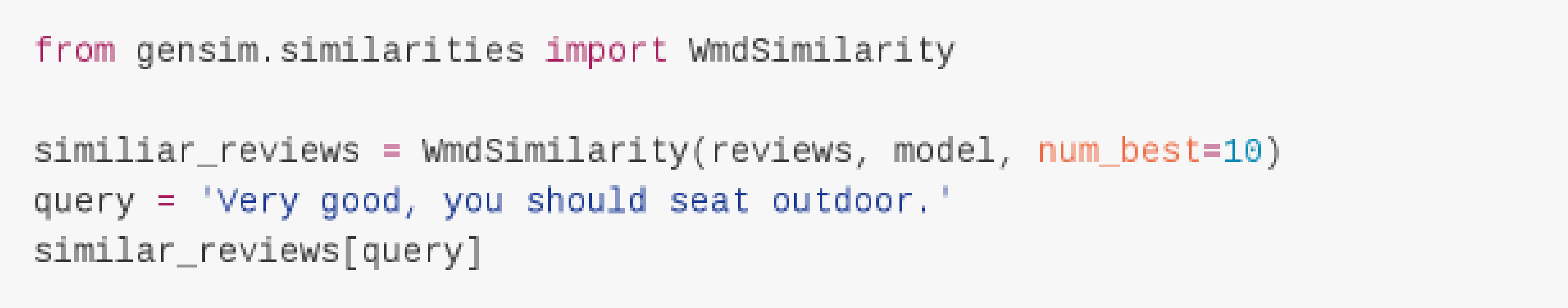
http://tech.opentable.com/2015/08/11/navigating-themes-in-restaurant-reviews-with-word-movers-distance/

Word Centroid Distance is a lower bound
Relaxed Word Mover’s Distance is a tighter bound



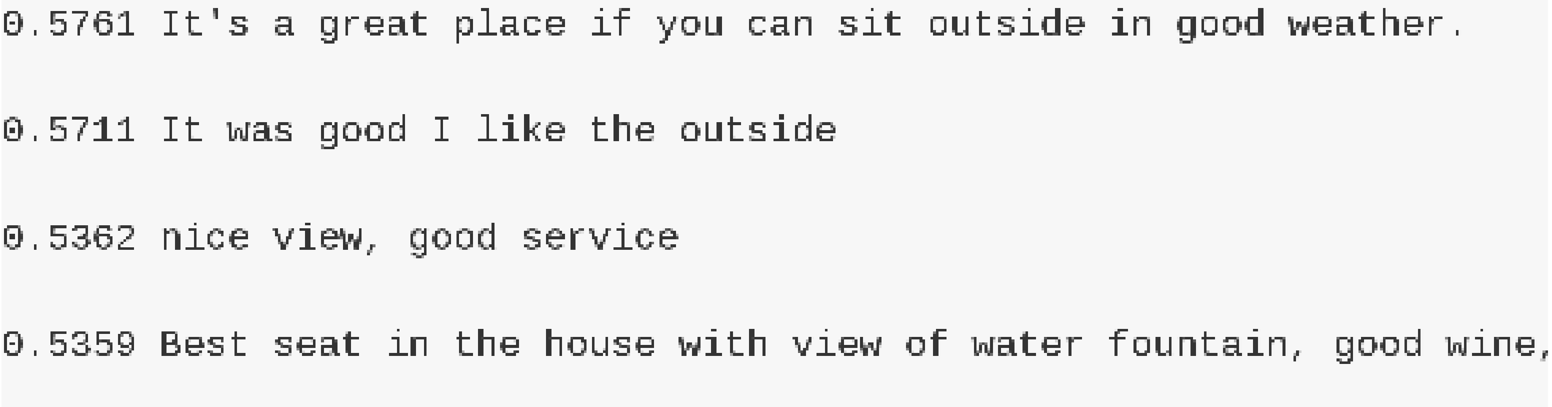
http://jmlr.org/proceedings/papers/v37/kusnerb15.pdf

Finding similar reviews



```
from gensim.similarities import WmdSimilarity

similiar_reviews = WmdSimilarity(reviews, model, num_best=10)
query = 'Very good, you should seat outdoor.'
similar_reviews[query]
```



```
0.5761 It's a great place if you can sit outside in good weather.

0.5711 It was good I like the outside

0.5362 nice view, good service

0.5359 Best seat in the house with view of water fountain, good wine,
```

Thanks!

Link to the Slides

https://github.com/RishabGoel/pycon_india_slides

Extra slides

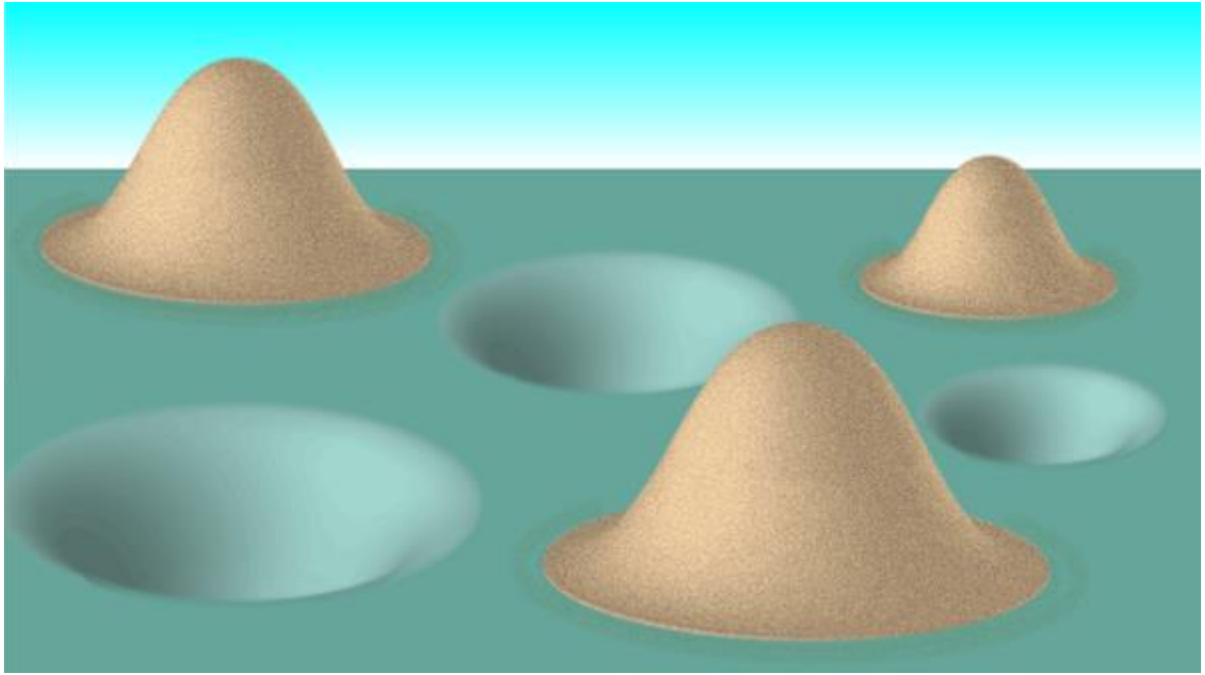
Ways to find similar documents

- Google’s Doc2vec
  - Built on top of word2vec
  - Document tags are just extra words in the document

Hard to tune. Slow inference.

Earth Mover’s Distance

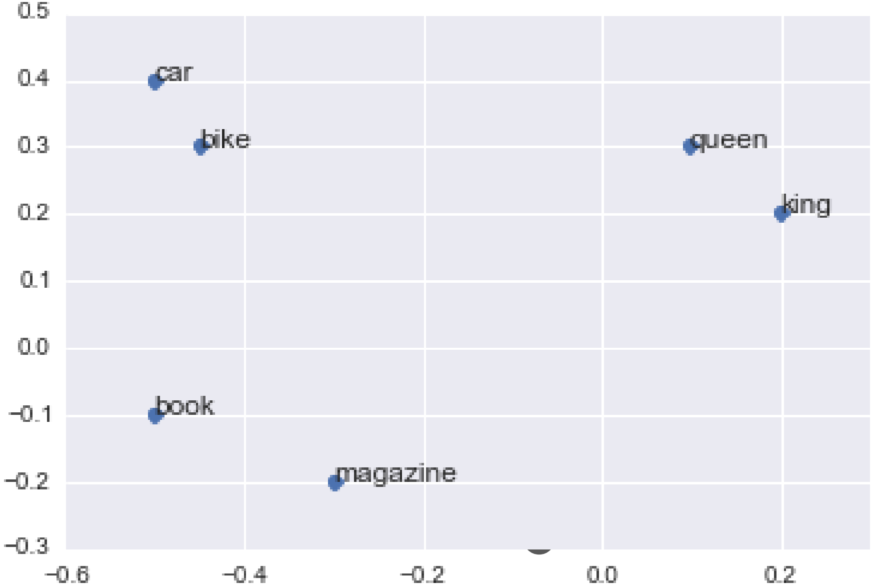
How do you best move piles of sand to fill up holes of the same total volume?



[Image: APS/Alan Stonebraker]

Stated by Monge in 1781. Solved by Kantorovich in 1942

Google’s Word2vec algorithm



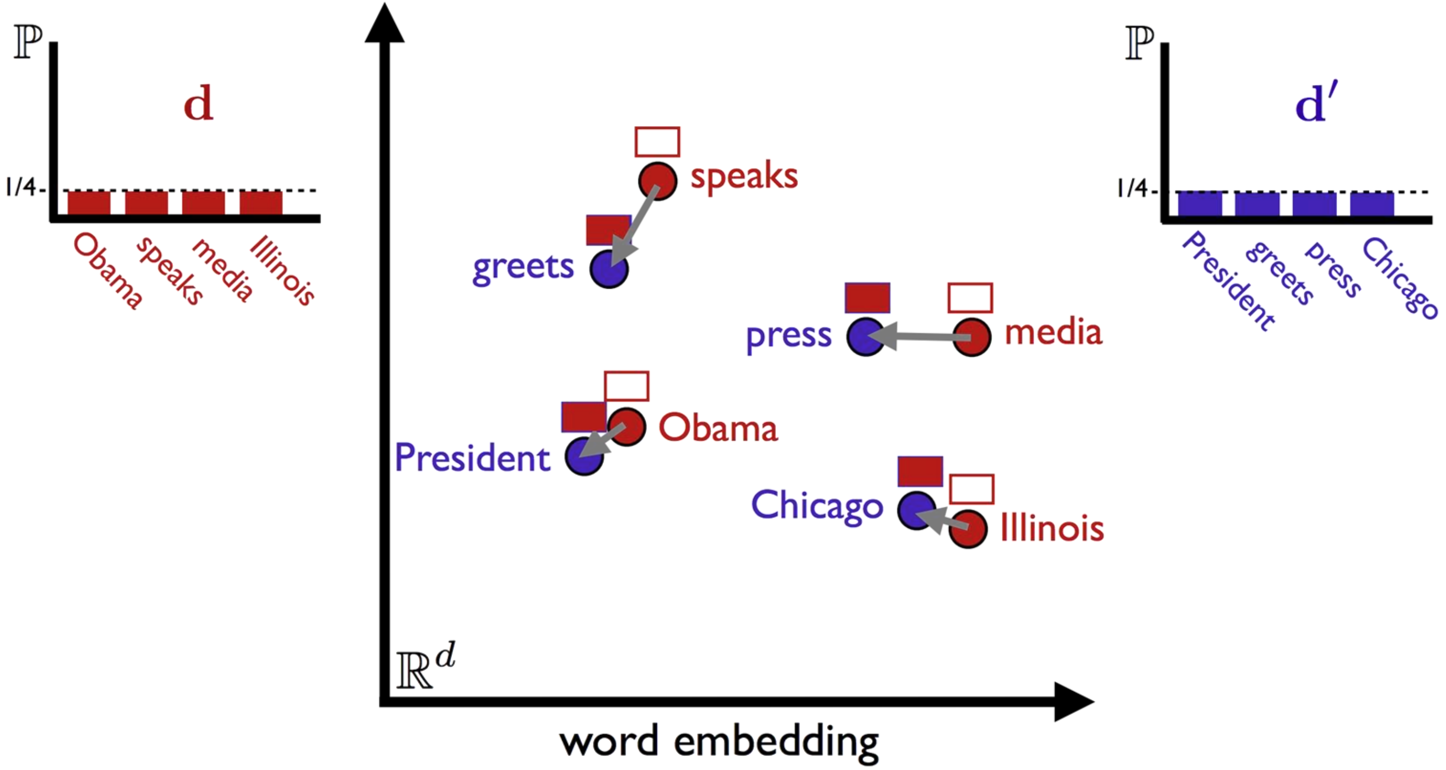
Word becomes a vector in 100-dimensional space.

- king - man + woman = queen

http://nbviewer.jupyter.org/github/fbkarsdorp/doc2vec/blob/master/doc2vec.ipynb

http://radimrehurek.com/2014/02/word2vec-tutorial

Word Mover's distance



http://vene.ro/blog/word-movers-distance-in-python.html