From Word Embeddings To Document Distances

Matt J. Kusner¹, Yu Sun¹,², Nicholas I. Kolkin¹, Kilian Q. Weinberger¹,²
¹Department of Computer Science & Engineering, Washington University in St. Louis, USA
²Department of Computer Science, Cornell University, USA

Word Embedding & Document Vectors
- Word2vec model
  - different from [Collobert & Weston, 2008]. [Pennington & Morris, 2019] word2vec is not deep!
  - embed billions of words / hour with a desktop computer
- learns complex word relationships:
  - vec(Boston) - vec(x) = vec(Massachusetts) - vec(Boston)
  - vec(Paris) - vec(Parisian) = vec(Prague) - vec(Praha)

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

k-Nearest Neighbor Results

Approximations

2. The Relaxed Word Mover’s Distance

Approximation Results for a random test input

References