Quantitative Text Analysis. Applications to Social Media Research

Pablo Barberá
London School of Economics
www.pablobarbera.com

Course website:

pablobarbera.com/text-analysis-vienna

Word embeddings

Beyond bag-of-words

Most applications of text analysis rely on a bag-of-words representation of documents

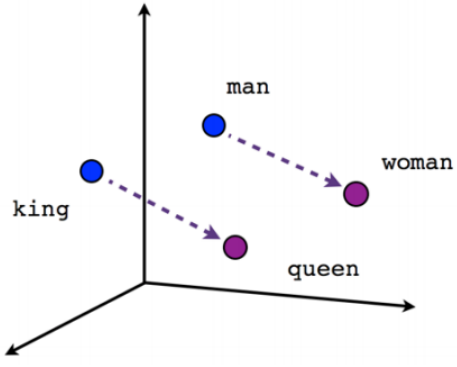
- Only relevant feature: frequency of features
- Ignores context, grammar, word order...
- Wrong but often irrelevant

One alternative: word embeddings

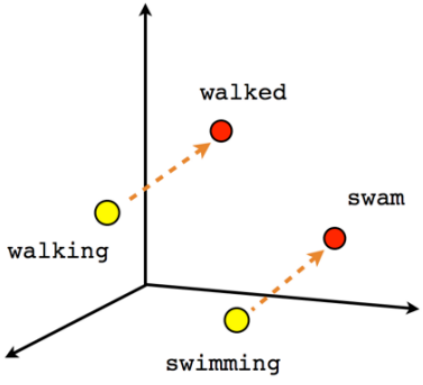
- Represent words as real-valued vector in a multidimensional space (often 100–500 dimensions), common to all words
- Distance in space captures syntactic and semantic regularities, i.e. words that are close in space have similar meaning
  - How? Vectors are learned based on context similarity
  - Distributional hypothesis: words that appear in the same context share semantic meaning
- Operations with vectors are also meaningful

Word embeddings example

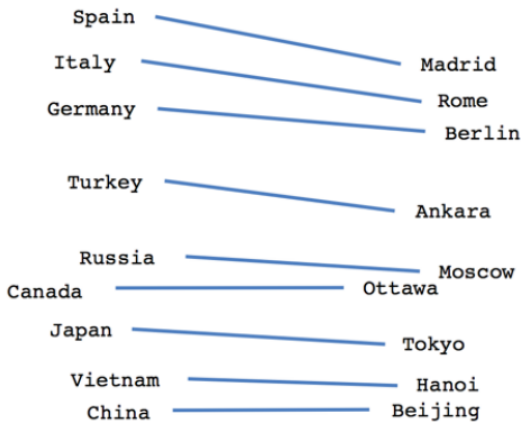
word	D_1	D_2	D_3	. . .	D_N
man	0.46	0.67	0.05	. . .	. . .
woman	0.46	-0.89	-0.08	. . .	. . .
king	0.79	0.96	0.02	. . .	. . .
queen	0.80	-0.58	-0.14	. . .	. . .



Male-Female



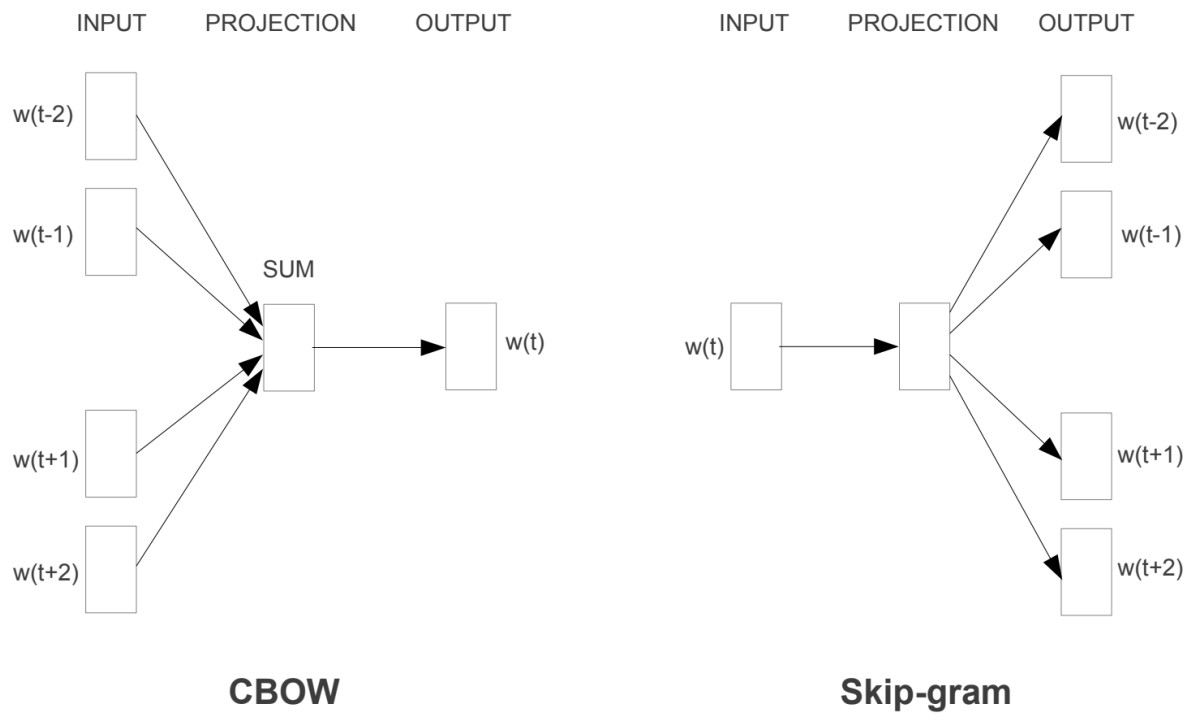
Verb tense



Country-Capital

word2vec (Mikolov 2013)

- Statistical method to efficiently learn word embeddings from a corpus, developed by Google engineer
- Most popular, in part because pre-trained vectors are available
- Two models to learn word embeddings:



Example: Pomeroy et al 2018

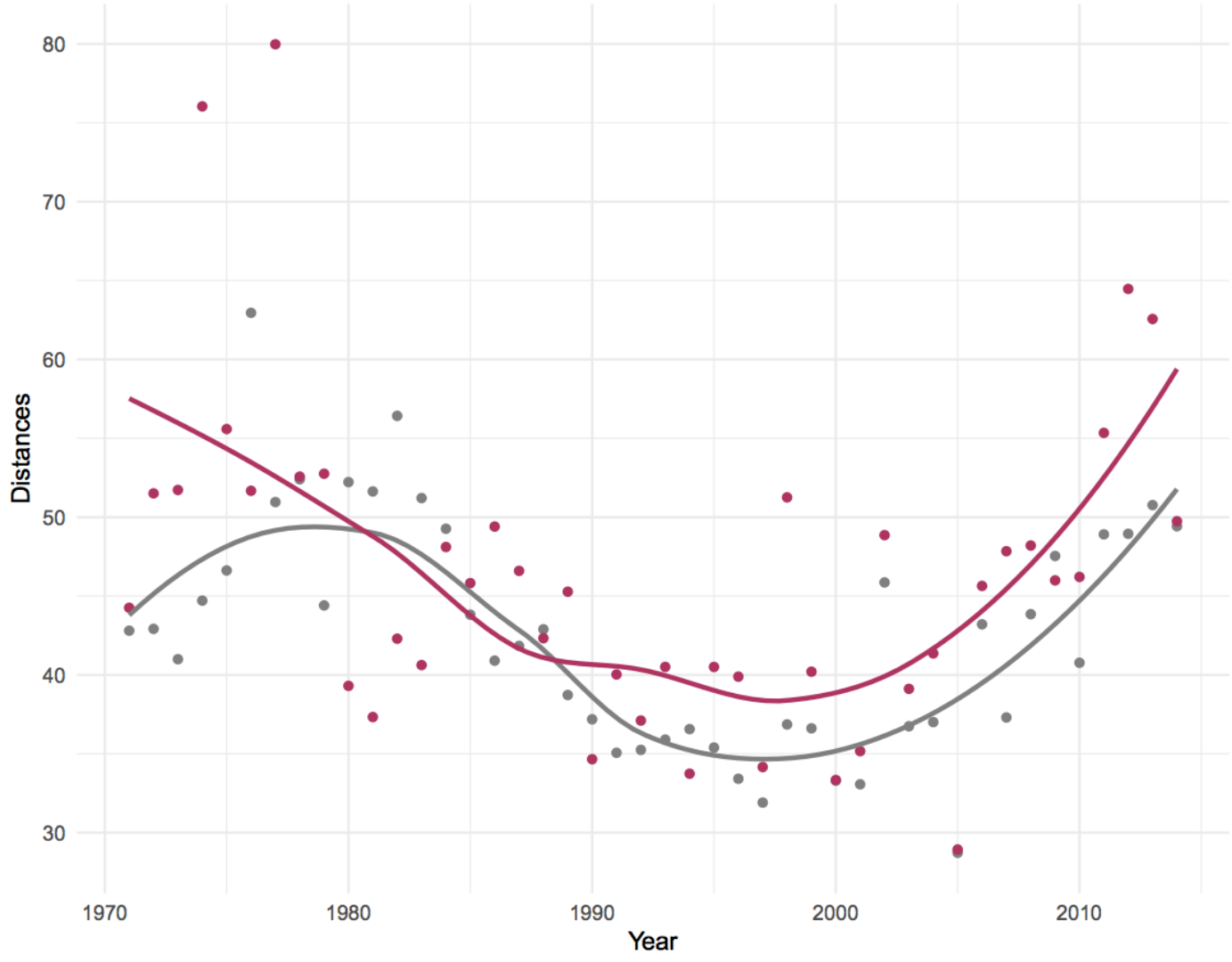


Figure 4: *Distances by core countries*. Plot of Euclidian distances between US and Russia (gray), and US and China (maroon).