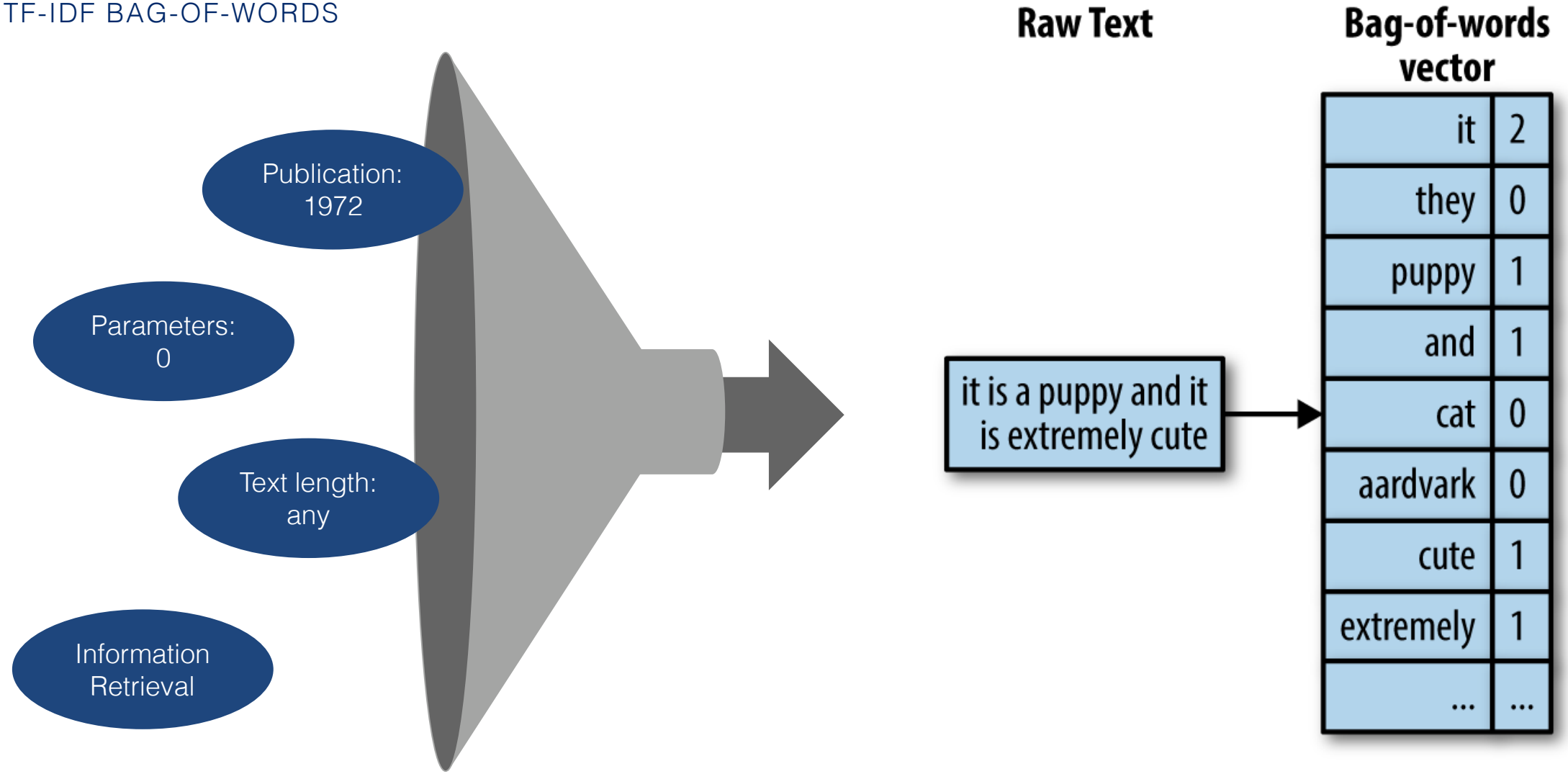
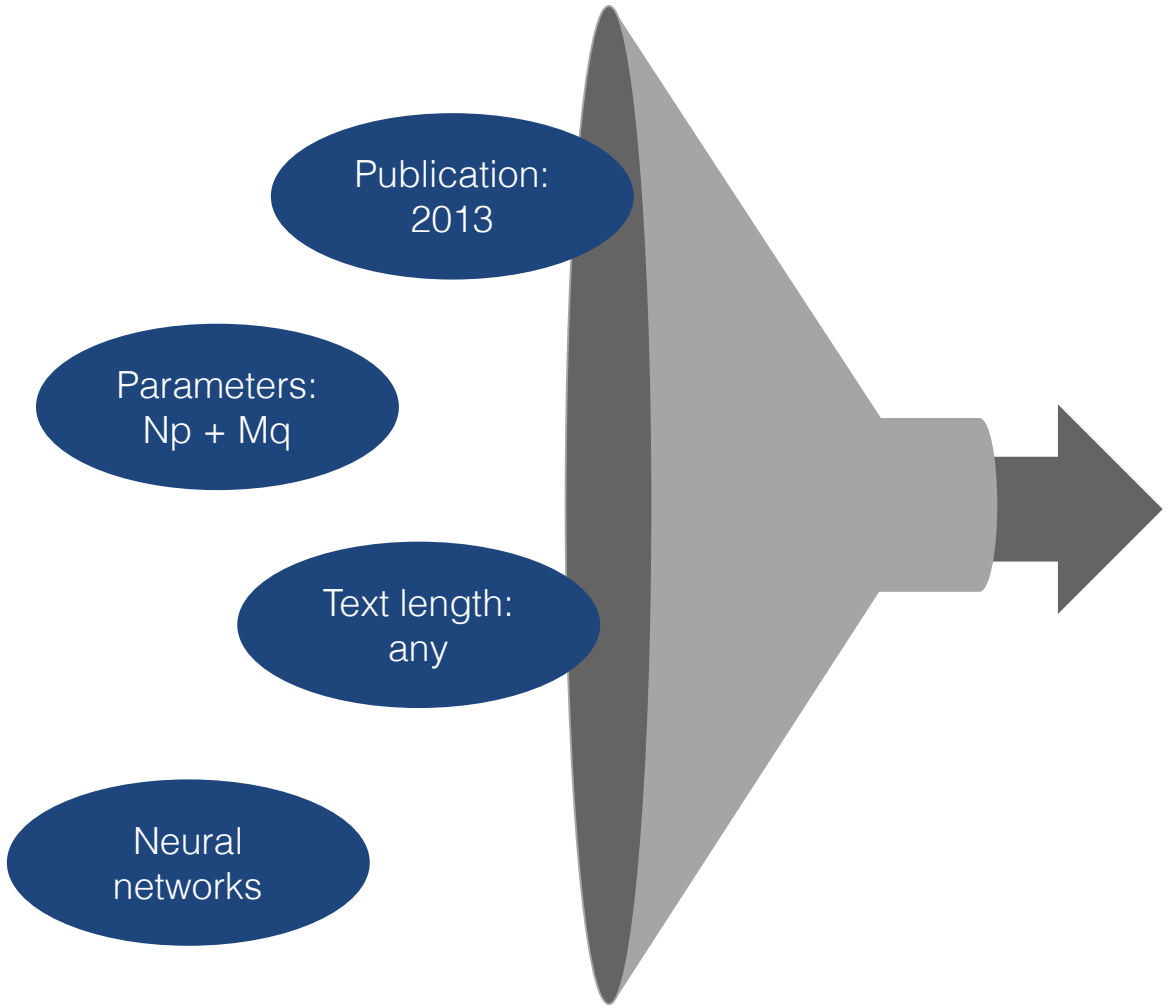
The state of document embedding research

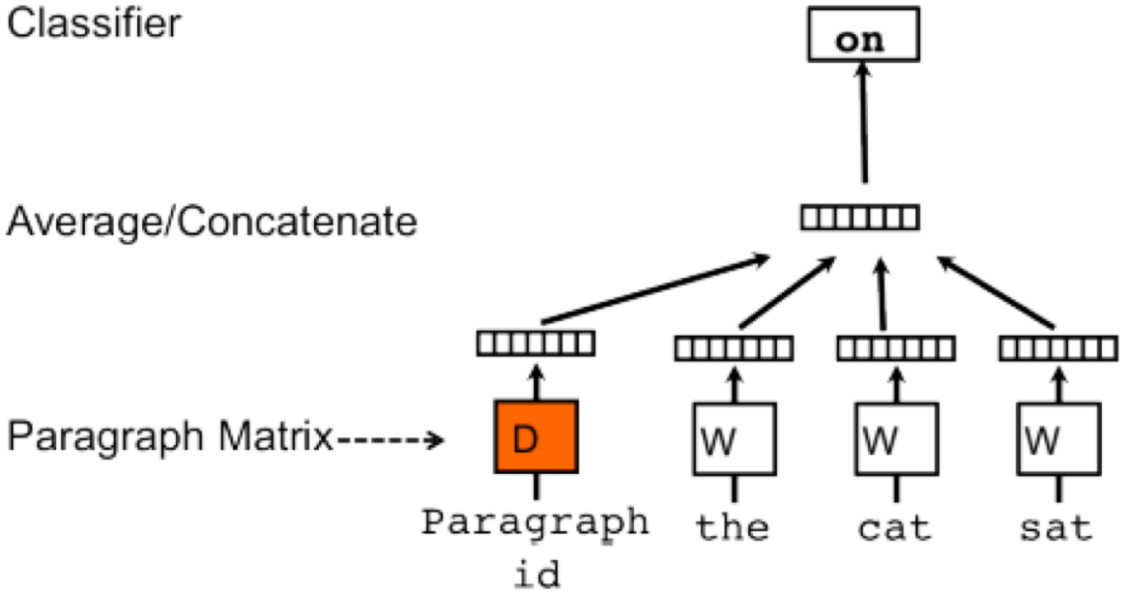
A short story on vectorizing human knowledge

TF-IDF BAG-OF-WORDS

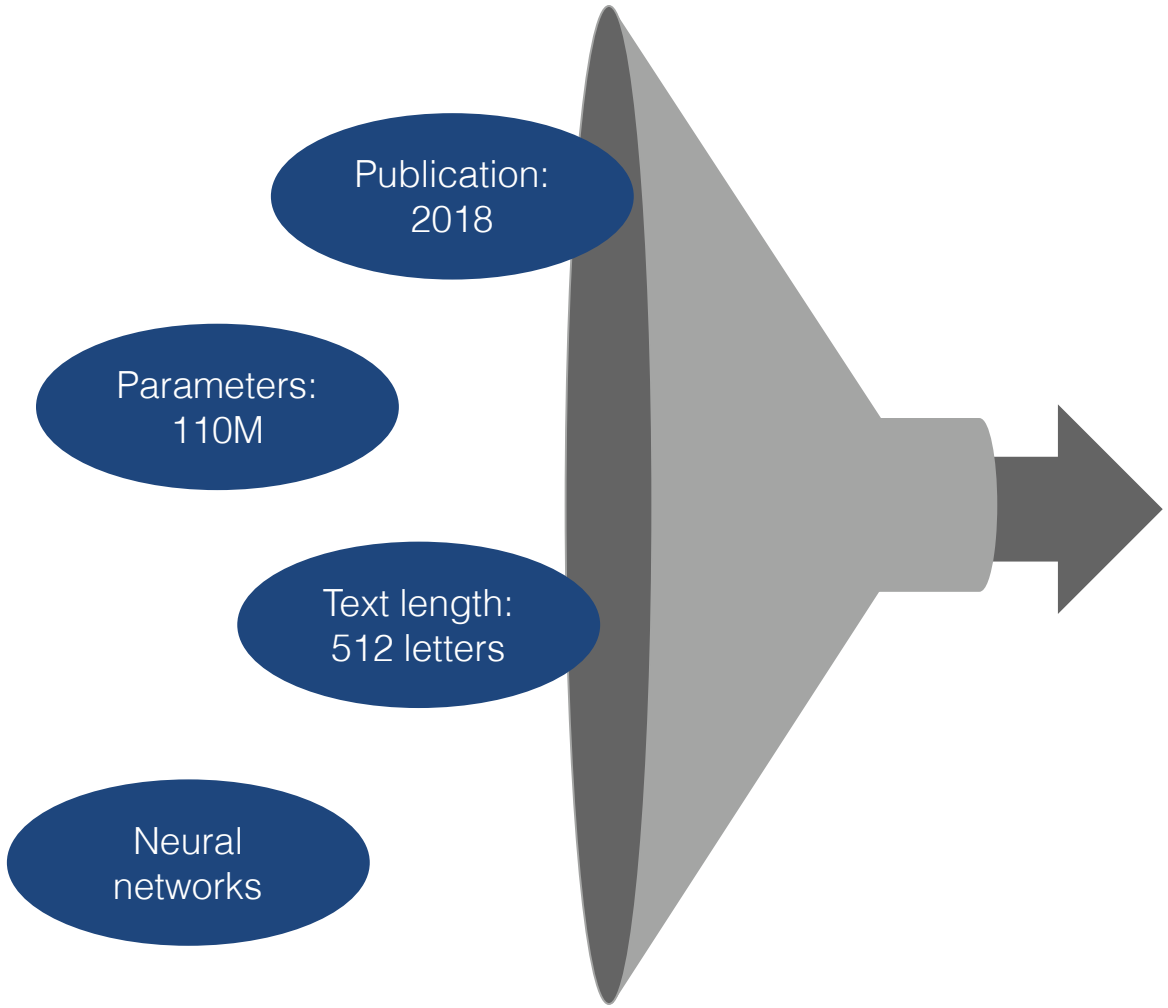


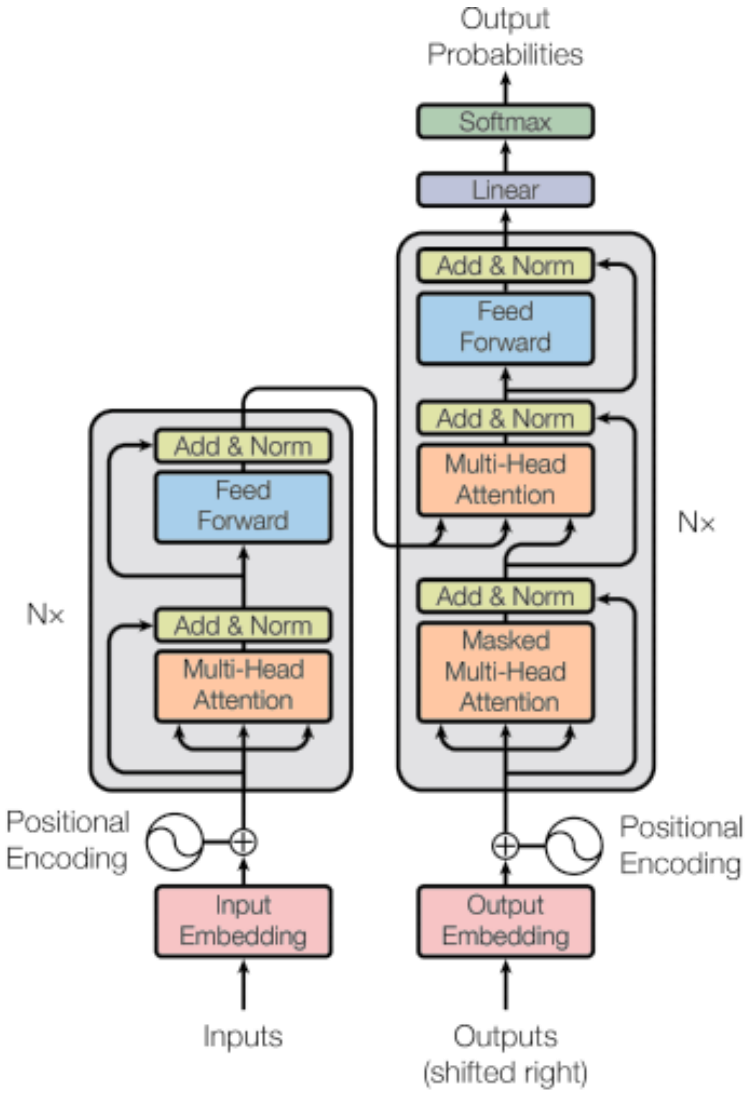
DOC2VEC


Publication: 2013
Parameters: Np + Mq
Text length: any
Neural networks


Classifier
Average/Concatenate
Paragraph Matrix
on
D
W
W
W
Paragraph id
the
cat
sat

TRANSFORMER (BERT)





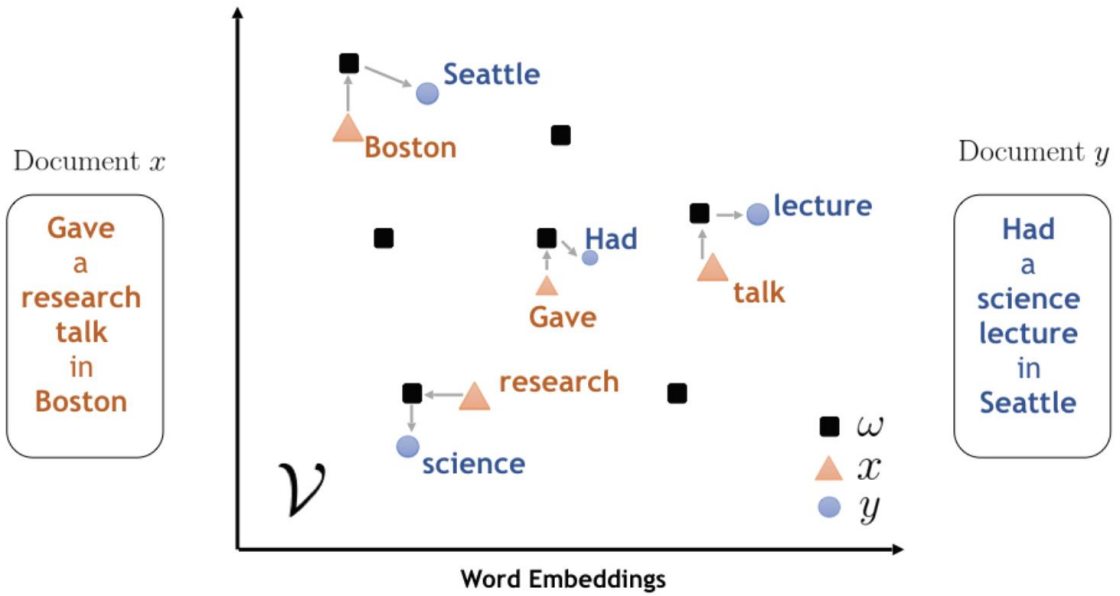
WORD MOVER'S EMBEDDING

Publication: 2018

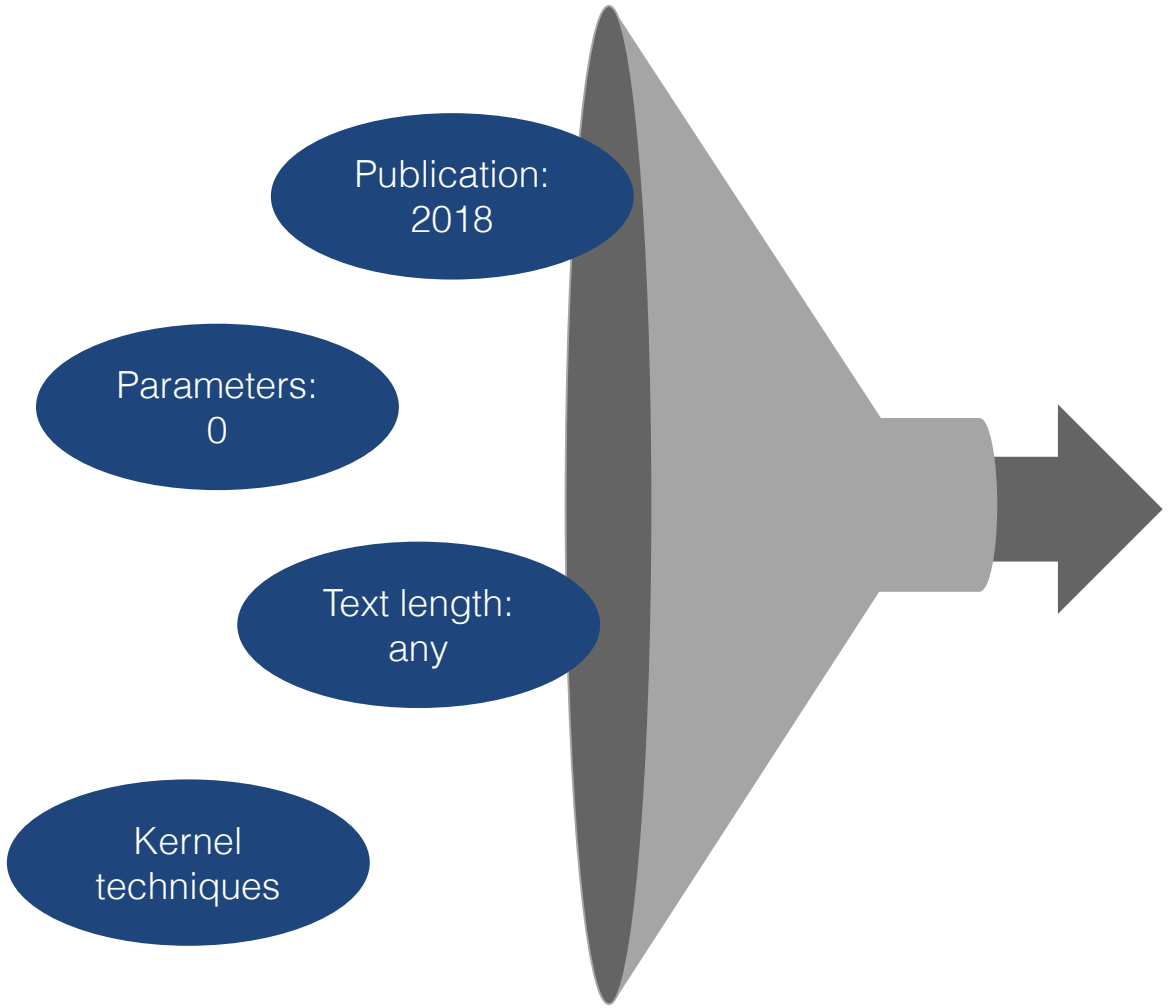
Parameters: 0

Text length: any

Kernel techniques



WORD MOVER'S EMBEDDING



Input: Texts $\{x_i\}_{i=1}^N$, $D_{\max}$, R.
Output: Matrix $Z_{N\times R}$, with rows corresponding to text embeddings.

Compute $v_{\max}$ and $v_{\min}$ as the maximum and minimum values, over all coordinates of the word vectors $\boldsymbol{v}$ of $\{x_i\}_{i=1}^N$, from any pre-trained word embeddings (e.g. Word2Vec, GloVe or PSL999).
for $j = 1, \ldots, R$ **do**
  Draw $D_j \sim \text{Uniform}[1, D_{\max}]$.
  Generate a random document ω_j consisting of D_j number of random words drawn as $\omega_{j\ell} \sim \text{Uniform}[v_{\min}, v_{\max}]^d, \ell = 1, \ldots, D_j$.
  Compute $\boldsymbol{f}_{x_i}$ and $\boldsymbol{f}_{\omega_j}$ using a popular weighting scheme (e.g. NBOW or TF-IDF).
  Compute the WME feature vector $Z_j = \phi_{\omega_j}(\{x_i\}_{i=1}^N)$ using WMD in Equation (2).
end for
Return $Z(\{x_i\}_{i=1}^N) = \frac{1}{\sqrt{R}}[Z_1\ Z_2\ \ldots\ Z_R]$

c/o FU Berlin
Königin-Luise-Straße 24-26
14195 Berlin
www.cct-ev.de

Tim Korjakow