Text Similarity in Vector Space Models: A Comparative Study

Authors : Omid Shahmirzadi, Adam Lugowski, Kenneth Younge

Abstract : Automatic measurement of semantic text similarity is an important task in natural language processing. In this paper, we evaluate the performance of different vector space models to perform this task. We address the real-world problem of modeling patent-to-patent similarity and compare TFIDF (and related extensions), topic models (e.g., latent semantic indexing), and neural models (e.g., paragraph vectors). Contrary to expectations, the added computational cost of text embedding methods is justified only when: 1) the target text is condensed; and 2) the similarity comparison is trivial. Otherwise, TFIDF performs surprisingly well in other cases: in particular for longer and more technical texts or for making finer-grained distinctions between nearest neighbors. Unexpectedly, extensions to the TFIDF method, such as adding noun phrases or calculating term weights incrementally, were not helpful in our context.

Keywords : big data, patent, text embedding, text similarity, vector space model
Conference Title : ICMLA 2019 : International Conference on Machine Learning and Applications
Conference Location : Copenhagen, Denmark
Conference Dates : June 11-12, 2019

Open Science Index, Computer and Information Engineering Vol:13, No:06, 2019 publications.waset.org/abstracts/102930.pdf