Doc2Vec

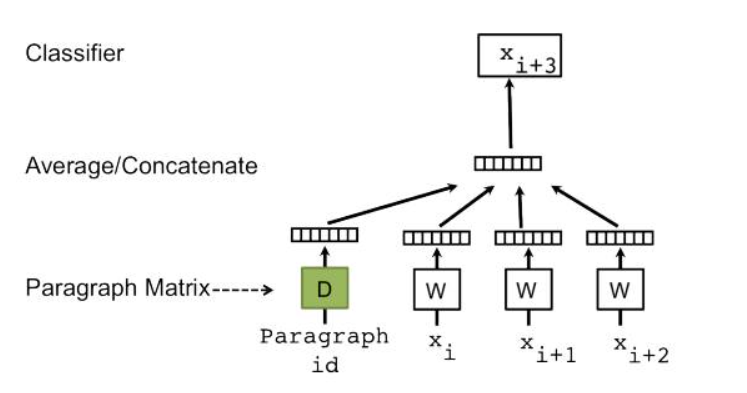
For anyone who knows Word2Vec, Doc2Vec uses similar technique to Word2Vec with some adjustments for text usage.

For now we will explain Doc2Vec model:

Doc2Vec model is model that transform a document into a vector representation. To further understand the model, we need what techniques are used in Doc2Vec:

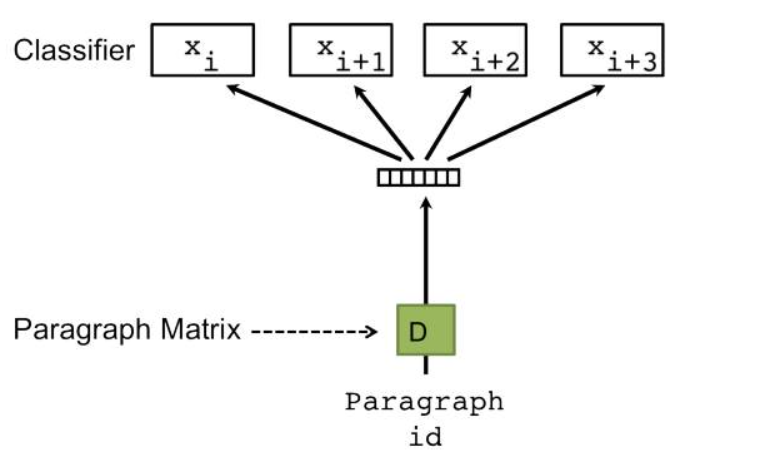
Bài báo Doc2Vec: https://arxiv.org/abs/1405.4053

+ Paragraph Vector - Distributed Memory (PV-DM): Mô hình này đơn giản là không quan tâm thứ tự các từ, training nhanh hơn, không sử dụng local-context/neighboring. Mô hình chèn thêm 1 "word" là ParagraphID, ParagraphID này đại diện cho văn bản được training. Sau khi training xong có thể hiểu các vector ParagraphID này là vector embedded của các văn bản. Hình ảnh được mô tả trong bài báo:



+ Paragraph Vector - Distributed Bag Of Word (PV-BOW):

Mô hình này đơn giản là không quan tâm thứ tự các từ, training nhanh hơn, không sử dụng local-context/neighboring. Mô hình chèn thêm 1 "word" là ParagraphID, ParagraphID này đại diện cho văn bản được training. Sau khi training xong có thể hiểu các vector ParagraphID này là vector embedded của các văn bản. Hình ảnh được mô tả trong bài báo:



**Tóm lại:**ta xem văn bản như là một từ, docID/paragraphID được biểu diễn dạng 1-hot, được embedded vào không gian vector.