28.10 Avg Word2Vec, TF-IDF Weighted Word2Vec

Average Word2Vec

Let us assume we have a document/review 'r_1'. Let its corresponding average vector form be denoted as 'v_1'. Let the total number of words in 'r_1' be 'n_1'.

r_1: w_1 w_2 w_1 w_3 w_4 w_5

So now the average word2vec for 'r_1' is

$$\mathbf{v_1 = (1/n_1)[w2v(w_1) + w2v(w_2) + w2v(w_1) + w2v(w_3) + w2v(w_4) + w2v(w_5)]}$$

$w2v(w_i)$ represents the word vector of the word 'w_i'. This is called the Average Word2Vec representation of 'r1'. Average Word2Vec fairly works well in practice, but is not perfect all the time. It still works well enough.

Average Word2Vec is a simple way to leverage Word2Vector to build sentence vectors.

TF-IDF Weighted Word2Vec

Let us assume we have 7 words (say 'w_1', 'w_2', 'w_3', 'w_4', 'w_5', 'w_6', 'w_7') and the review/document 'r1'.

r_1: w_1 w_2 w_1 w_3 w_4 w_5

The TF-IDF representation of the 'r1' vector is given as below

w1	w2	w3	w4	w5	w6	w7
t1	t2	t3	t4	t5	t6=0	t7=0

Here $t_i \rightarrow$ TF-IDF(w_i,r_i)

Now, let 'v_1' be the TF-IDF Weighted Word2Vec representation of the vector 'r_1', then

$$\mathbf{v_1 = (t_1*w2v(w_1) + t_2*w2v(w_2) + t_3*w2v(w_3) + t_4*w2v(w_4) + t_5*w2v(w_5))/(t_1 + t_2 + t_3 + t_4 + t_5)}$$

It can simply be written as

$$\mathbf{TFIDF\ Weighted\ Word2Vec\ (ri) = \Sigma_{i=1}^{n} (t_i*w2v(w_i))/(\Sigma_{i=1}^{n} t_i)}$$

Note - Special Case:

If t_i=1(ie., $t_1 = t_2 = t_3 = \ldots\ldots = 1$), then the TF-IDF Weighted Word2Vec is the Average Word2Vec.

Average Word2Vec and TF-IDF Weighted Word2Vec are two simple weighting strategies to convert sentences into vectors. They both serve the same purpose. TF-IDF Weighted Word2Vec weights each word differently as compared to Average Word2Vector. In practice, we try both the options and choose the one that performs better at our task.