How does a vector have meaning

What words are used a lot?

- Bag of words
- TF-IDF
- Deep Averaging Network

- They don’t consider the order of words.

Bag Of Words

Bag = multiset in mathematics. Allow duplication.

Consider word frequency as a top priority regardless of its order.

Tokenize sentences into words and count word frequency.

(Can be simplified using Boolean)

Still famous in Information Retrieval. Transform user's query into a vector and calculate its cosine similarity over search-counterparts' vectors. Then, display the document which has the most high similarity with the query.

Term Frequency-Inverse Document Frequency

- BOW has huge drawback. (을, 를, 이, 가)
- Cannot estimate the document's subject with these words.

- TF-IDF solve this problem.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- TF denotes the frequency that how many specific word is used in a particular document.
- DF indicates the number of documents containing the word.
- IDF means the total number of docs divided by DF and take a logarithm. → It reduce the weight of words that lack information.

Deep Averaging Network (DAN)

Neural network version of BOW

Sentence's embedding is the mean of words' embedding.

"Man bites dog" vs "Dog bites man" → should be same

Performed poorly on double negation sentence.

https://medium.com/tech-that-works/deep-averaging-network-in-universal-sentence-encoder-465655874a04