**COSINE SIMILARITY IN QUESTION ANSWERING**

Cosine similarity is a measure that calculates the cosine angle between two given non-zero vectors. This metric is proven to be very relevant when it comes to comparing words, documents, paragraphs or sentences. Given two documents (represented as a non-zero vector) we can directly calculate the angle between these two vectors which will enable us to determine whether the documents or sentences are similar or not. We can do that using the equation of the dot product which is as follows:

Where **a** is the first vector, **b** is the second vector, |**a**| is the magnitude of vector **a**, |**b**| is the magnitude of vector b and is the angle between vector **a** and vector **b** and ranging from 0 to 180 degrees. If two documents are related, they will have cosine similarity value closer to one; cosine similarity value 1 depicts that the two documents have the same meaning. On the other hand, when the cosine angle is 0, the documents are not related. With that said, we can find the similarity between a user asked question and the questions in our database.

**Normalization of data**

Before converting the sentences or documents to the vectors, we have normalized it by removing byte-like characters, and stop words. Since stop words are commonly used and do not affect the result in anyway, we have removed them.

**Vectorization**

However, we cannot pass the word sentences to the model instead we have to convert them to fixed vectors. We used “tfid” vectorizer to transform all the questions in our training into vectors. We also turned all the test questions into vectors of the same size as the training vectors. We then compute the cosine angle (cosine similarity between) them. The question in our dataset that gives the highest cosine similarity with the test question is probably the question asked. The model then responses to the user with the answer of the most likely similar question.