**Topic modeling:**

Problem: we have a group of topics (type of stressors/type of emotions) 🡪 We want to measure which topic is more relevant to a document.

**Preparation: Getting Dictionaries:**

Eg. Four stressors

Death of a loved one🡪 bag of keywords

Divorce 🡪bag of keywords

Moving🡪 bag of keywords

Major illness or injury🡪bag of keywords

Job loss🡪 bag of keywords

**Method1: Word2Vec:**

<https://pathmind.com/wiki/word2vec>

🡪get avg distance: words in document 🡪 words in 4 different dictionaries

**Method2: Clustering**

<https://medium.com/@MSalnikov/text-clustering-with-k-means-and-tf-idf-f099bcf95183>

<http://brandonrose.org/clustering>

🡪 clustering all words in the document

🡪 get centers of each clusters

🡪 get centers of words in four dictionaries

🡪 get the nearest distance between document centers to dictionary centers

**Method3: LDA**

<https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>

<https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21>

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>

**Note: we can ask LDA to find num of topics (set the number we wanted) in one document.**

**LDA will return top words for each topic, but will not return a specific topic for these group of words (we need to summarize by ourselves)**

**\*\*We can calculate the distance between top words in each topic and keywords in four stressor dictionaries**