Document text encodings:

**Term frequency (TF) encoding**

tf(w,d)=(number of times word w occurs in d)/(total words in d)

---

**Inverse document frequency (IDF) encoding**

idf(w,D)=log((number of documents in D)/(number of documents in D that contain the word w))

---

**TF.IDF encoding**

tf(w,d,D)=tf(w,d)*idf(w,D)

---

**Context prediction**

Given a set of words in a sentence can we predict the next word? In order to solve this problem we need a model that takes context into consideration. A simple context model would be to predict the fifth word from the first four. For example a model would be given “The cat ate the” and the prediction is “mouse”. To build such a model we need a training dataset (xi,yi) where each xi are four words of a sentence and the yi is the fifth predicted word.

How do we encode the words so we can perform machine learning? Three choices:

1. One hot encoding: high dimensional vectors where each dimension corresponds to a word
2. Label encoding: each word maps to a unique number
3. Word2vec: Use a neural network to the word from a previous one. Use the hidden layer to represent the input word.