**Document Vectorization with LTUVectorizer and LTUTransformer.**

Vectorization is a key method in Natural Language Processing (NLP) that enables computers to efficiently comprehend and interpret human language. Vectorization makes NLP systems more effective and precise by converting text input into numerical vectors that algorithms can process, understand, and extract valuable language-based information from.

A well-known vectorization formula is TF-IDF vectorization, but this article comes to show another vectorizer called LTU.

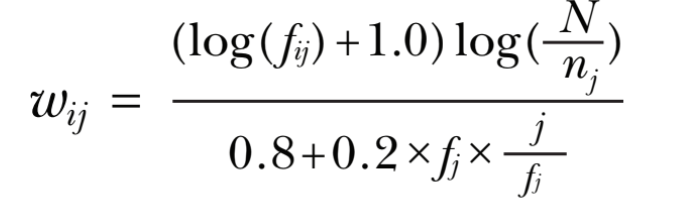
**What is LTUVectorizer and LTUTransformer?**

It is a method for converting a collection of raw documents to a matrix of LTU features.

LTU comes as an extention of BOWs that is compatible with sklearn API. The generated matrix with scores for every word is more informative for the Machine Learning algorithms.

**How is it calculated?**

LTU transformation is calculated using the following formula:



where fij denotes the target term frequency in a particular document, fj the total document frequency, nj the number of documents containing the target term, N the total number of documets.

**How to use the LTUVectorizer?**

First of all, you must install molda if it’s not installed already:

pip install molda

Now you can import the LTUVectorizer or the LTUTransformer:

from molda import LTUVectorizer

from molda import LTUTransformer

The difference between Vectorizer and Transformer is the input. The Transformer uses a count matrix as input while the Vectorizer some text. To vectorize a document, we need that document. So here is a small example we can use:

document = [  
 '''I love my dog  
 and I like pet him.''',  
 'I love my cat!',  
 "I like to pet my cat.",  
 "I enjoy playing with both, cats and dogs."  
]

It’s time to instantiate the vectorizer:

vectorizer = LTUVectorizer()  
sparse\_matrix = vectorizer.fit\_transform(document)

The output is a LTU weighted sparse matrix that can be easily transformed in a simple matrix using numpy library:

print(sparse\_matrix)  
  
(0, 7) 1.4242701385551153  
(0, 11) 1.122910936582831  
(0, 8) 1.122910936582831  
(0, 0) 1.122910936582831  
(0, 4) 1.4242701385551153  
(0, 10) 0.9090931800308315  
(0, 9) 1.122910936582831  
(1, 2) 1.122910936582831  
(1, 10) 0.9090931800308315  
(1, 9) 1.122910936582831  
(2, 13) 1.4242701385551153  
(2, 2) 1.122910936582831  
(2, 11) 1.122910936582831  
(2, 8) 1.122910936582831  
(2, 10) 0.9090931800308315  
(3, 5) 1.4242701385551153  
(3, 3) 1.4242701385551153  
(3, 1) 1.4242701385551153  
(3, 14) 1.4242701385551153  
(3, 12) 1.4242701385551153  
(3, 6) 1.4242701385551153  
(3, 0) 1.122910936582831

LTU is a great alternative to the TF-IDF Vectorizer, the standart aproach from scikit learn library, and it can easily be used with other modules from scikit learn. For more information you can visit the github page: https://github.com/SigmoidAI/molda.

