Assignment

What does tf-idf mean?

Tf-idf stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

How to Compute:

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- **TF:** Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:
  $$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}.$$
- **IDF:** Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:
  $$IDF(t) = \log_e\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}.$$
  for numerical stabiltiy we will be changing this formula little bit
  $$IDF(t) = \log_e\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}+1}.$$

Example

Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e., tf) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12. </p>

Task-1

1. **Build a TFIDF Vectorizer & compare its results with Sklearn:**

- As a part of this task you will be implementing TFIDF vectorizer on a collection of text documents.

- You should compare the results of your own implementation of TFIDF vectorizer with that of sklearns implemenation TFIDF vectorizer.

- Sklearn does few more tweaks in the implementation of its version of TFIDF vectorizer, so to replicate the exact results you would need to add following things to your custom implementation of tfidf vectorizer:
  1. Sklearn has its vocabulary generated from idf sroted in alphabetical order
  2. Sklearn formula of idf is different from the standard textbook formula. Here the constant **"1"** is added to the numerator and denominator of the idf as if an extra document was seen containing every term in the collection exactly once, which prevents zero divisions. $IDF(t) = 1 + \log_e \frac{1 + \text{Total number of documents in collection}}{1 + \text{Number of documents with term t in it}}$.
  3. Sklearn applies L2-normalization on its output matrix.
  4. The final output of sklearn tfidf vectorizer is a sparse matrix.

- Steps to approach this task:
  1. You would have to write both fit and transform methods for your custom implementation of tfidf vectorizer.
  2. Print out the alphabetically sorted voacb after you fit your data and check if its the same as that of the feature names from sklearn tfidf vectorizer.
  3. Print out the idf values from your implementation and check if its the same as that of sklearns tfidf vectorizer idf values.
  4. Once you get your voacb and idf values to be same as that of sklearns implementation of tfidf vectorizer, proceed to the below steps.
  5. Make sure the output of your implementation is a sparse matrix. Before generating the final output, you need to normalize your sparse matrix using L2 normalization. You can refer to this link https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
  6. After completing the above steps, print the output of your custom implementation and compare it with sklearns implementation of tfidf vectorizer.
  7. To check the output of a single document in your collection of documents, you can convert the sparse matrix related only to that document into dense matrix and print it.

Note-1: All the necessary outputs of sklearns tfidf vectorizer have been provided as reference in this notebook, you can compare your outputs as mentioned in the above steps, with these outputs.
Note-2: The output of your custom implementation and that of sklearns implementation would match only with the collection of document strings provided to you as reference in this notebook. It would not match for strings that contain capital letters or punctuations, etc, because sklearn version of tfidf vectorizer deals with such strings in a different way. To know further details about how sklearn tfidf vectorizer works with such string, you can always refer to its official documentation.
Note-3: During this task, it would be helpful for you to debug the code you write with print statements wherever necessary. But when you are finally submitting the assignment, make sure your code is readable and try not to print things which are not part of this task.

Corpus

In [1]:

```
## SkLearn# Collection of string documents

corpus = [
     'this is the first document',
     'this document is the second document',
     'and this is the third one',
     'is this the first document',
]
```

SkLearn Implementation

In [2]:

In [2]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
vectorizer.fit(corpus)
skl_output = vectorizer.transform(corpus)
```

In [3]:

```
# sklearn feature names, they are sorted in alphabetic order by default.

print(vectorizer.get_feature_names())
```

```
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
```

In [4]:

```
# Here we will print the sklearn tfidf vectorizer idf values after applying the fit metho
d
# After using the fit function on the corpus the vocab has 9 words in it, and each has it
s idf value.

print(vectorizer.idf_)
```

```
[1.91629073 1.22314355 1.51082562 1.         1.91629073 1.91629073
 1.         1.91629073 1.        ]
```

In [5]:

```
# shape of sklearn tfidf vectorizer output after applying transform method.

skl_output.shape
```

Out[5]:

```
(4, 9)
```

In [6]:

```
# sklearn tfidf values for first line of the above corpus.
# Here the output is a sparse matrix

print(skl_output[0])
```

```
  (0, 8)	0.38408524091481483
  (0, 6)	0.38408524091481483
  (0, 3)	0.38408524091481483
  (0, 2)	0.5802858236844359
  (0, 1)	0.46979138557992045
```

In [7]:

```
# sklearn tfidf values for first line of the above corpus.
# To understand the output better, here we are converting the sparse output matrix to den
se matrix and printing it.
# Notice that this output is normalized using L2 normalization. sklearn does this by defa
ult.

print(skl_output[0].toarray())
```

```
[[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]]
```

Your custom implementation

In [8]:

```
# Write your code here.
# Make sure its well documented and readble with appropriate comments.
# Compare your results with the above sklearn tfidf vectorizer
# You are not supposed to use any other library apart from the ones given below
```

```python
from collections import Counter
from tqdm import tqdm
from scipy.sparse import csr_matrix
import math
import operator
from sklearn.preprocessing import normalize
import numpy
```

In [9]:

```python
def fit(corpus):
    if(isinstance(corpus,list)): #to check if the parameter passed is of valid datatype
        unique_words=set() #initialise empty set to get unique words only
        for sentence in corpus:
            words=sentence.split()
            for word in words:
                unique_words.add(word)
        vocab=[] #stores all the unique words in alphabetical order
        for word in unique_words:
            vocab.append(word)
        vocab.sort()
        index=[i for i in range(0,len(vocab))]
        vocab_index=dict(zip(vocab,index)) #stores the unique words in sorted order with
corresponding index
        idf_values=[] #contains the idf values of all the words in the vocab list
        for word in vocab:
            count=0 #keeps track of the number of documents in which the word appears
            for sentence in corpus:
                words=sentence.split()
                if(word in words):
                    count+=1
            idf=1+math.log((1+len(corpus))/(1+count))
            idf_values.append(idf)

        return vocab,vocab_index,idf_values

    else:
        print("you need to enter a list of documents as a parameter \n")
```

In [10]:

```python
vocab,vocab_index,idf_values= fit(corpus)
print(vocab) #gives the list of all the unique words in the corpus
print("\n")
print(vocab_index) #gives the list of all the unique words with the corresponding index
print("\n")
print(idf_values) #gives the list of all the idf values of all the unique words
```

```
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']


{'and': 0, 'document': 1, 'first': 2, 'is': 3, 'one': 4, 'second': 5, 'the': 6, 'third':
7, 'this': 8}


[1.916290731874155, 1.2231435513142097, 1.5108256237659907, 1.0, 1.916290731874155, 1.916
290731874155, 1.0, 1.916290731874155, 1.0]
```

In [18]:

```python
def transform(corpus,vocab,idf_values):
    row=[] #stores the row number for sparse matrix representation
    column=[] #stores the columns number for sparse matrix representation
    value=[] #stores the tf-idf values
    if isinstance(corpus,list):
        for sentence in corpus:
            tokens=sentence.split() #list of words in a document
            for word in tokens:
                if word in vocab:
                    index=vocab.index(word)
```

```
                    row.append(corpus.index(sentence))
                    column.append(index)
                    tf_value=tokens.count(word)/len(tokens)
                    tf_idf=tf_value*idf_values[index]
                    value.append(tf_idf)
        csr=csr_matrix((value,(row,column)),shape=(len(corpus),len(vocab)))
        csr=normalize(csr, norm='l2', axis=1, copy=True, return_norm=False) #performing the l2 normalisation on the sparse matrix, the normalise function is from sklearn.preprocessing
        #the l2 normalisation has been performed row wise-for each text document-axis=1
        return csr
    else:
        print("you need to enter a list of documents as a parameter \n")
```

In [19]:

```
feature_matrix=transform(corpus,vocab,idf_values) #the vocab and idf values have already been returned from the fit function
feature_matrix.shape
```

Out[19]:

```
(4, 9)
```

In [20]:

```
print(feature_matrix[0])
```

```
  (0, 1)	0.4697913855799205
  (0, 2)	0.580285823684436
  (0, 3)	0.3840852409148149
  (0, 6)	0.3840852409148149
  (0, 8)	0.3840852409148149
```

In [21]:

```
#printing the first row of feature matrix as a dense matrix
#it is the tf-idf repreenation of the first text document in the corpus
print(feature_matrix[0].toarray())
```

```
[[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]]
```

Task-2

2. Implement max features functionality:

- As a part of this task you have to modify your fit and transform functions so that your vocab will contain only 50 terms with top idf scores.

- This task is similar to your previous task, just that here your vocabulary is limited to only top 50 features names based on their idf values. Basically your output will have exactly 50 columns and the number of rows will depend on the number of documents you have in your corpus.

- Here you will be give a pickle file, with file name **cleaned_strings**. You would have to load the corpus from this file and use it as input to your tfidf vectorizer.

- Steps to approach this task:
    1. You would have to write both fit and transform methods for your custom implementation of tfidf vectorizer, just like in the previous task. Additionally, here you have to limit the number of features generated to 50 as described above.
    2. Now sort your vocab based in descending order of idf values and print out the words in the sorted voacb after you fit your data. Here you should be getting only 50 terms in your vocab. And make sure to print idf values for each term in your vocab.
    3. Make sure the output of your implementation is a sparse matrix. Before generating the final output, you need to

normalize your sparse matrix using L2 normalization. You can refer to this link https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html

4. Now check the output of a single document in your collection of documents, you can convert the sparse matrix related only to that document into dense matrix and print it. And this dense matrix should contain 1 row and 50 columns.

In [22]:

```
# Below is the code to load the cleaned_strings pickle file provided
# Here corpus is of list type

import pickle
with open('cleaned_strings', 'rb') as f:
    corpus = pickle.load(f)

# printing the length of the corpus loaded
print("Number of documents in corpus = ",len(corpus))
```

```
Number of documents in corpus =  746
```

In [56]:

```
# Write your code here.
# Try not to hardcode any values.
# Make sure its well documented and readable with appropriate comments.
def fit_top50(corpus):
    if(isinstance(corpus,list)): #to check if the parameter passed is of valid datatype
        unique_words=set() #initialise empty set to get unique words only
        for sentence in corpus:
            words=sentence.split()
            for word in words:
                unique_words.add(word)
        vocab=[] #stores all the unique words
        for word in unique_words:
            vocab.append(word)
        idf_values=[] #contains the idf values of all the words in the vocab list
        for word in vocab:
            count=0 #keeps track of the number of documents in which the word appears
            for sentence in corpus:
                tokens=sentence.split()
                if(word in tokens):
                    count+=1
            idf=1+math.log((1+len(corpus))/(1+count))
            idf_values.append(idf)
        vocab_idf=dict(zip(vocab,idf_values)) #creates a dict with word as key and idf va
lue as value
        d=Counter(vocab_idf) #counts occurences of each word and sorts in ascending orde
r
        c=d.most_common() #sorts the values in descending order
        vocab=[]
        idf_values=[]
        for i in range(50):
            vocab.append(c[i][0])
            idf_values.append(c[i][1])
        return vocab,idf_values

    else:
        print("you need to enter a list of documents as a parameter \n")
```

In [57]:

```
vocab,idf_values=fit_top50(corpus)
```

In [58]:

```
print(vocab)
print("="*50)
print(len(vocab))
```

```
['major', 'embassy', 'pray', 'exploit', 'boogeyman', 'fundamental', 'hes', 'medical', 'co
mmercial', 'buildings', 'teacher', 'raver', 'juano', 'company', 'road', 'slideshow', 'sca
mp', 'shelves', 'endlessly', 'continue', 'gake', 'joe', 'rejection', 'vitally', 'roeg', '
childrens', 'babysitting', 'warn', 'reasons', 'confuses', 'occurs', 'bitchy', 'wife', 'in
dictment', 'thunderbirds', 'jennifer', 'upper', 'logic', 'surprised', 'buddy', 'range', '
everywhere', 'shenanigans', 'eloquently', 'correct', 'baxendale', 'necklace', 'chills', '
wartime', 'dependant']
==================================================
50
```

In [63]:

```python
print(idf_values)
print("="*50)
print(len(idf_values))
```

```
[6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.9229180045
72872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.9229
18004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872,
6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.92291800457
2872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.92291
8004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6
.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572
872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918
004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.
922918004572872, 6.922918004572872, 6.922918004572872, 6.922918004572872, 6.9229180045728
72, 6.922918004572872, 6.922918004572872, 6.922918004572872]
==================================================
50
```

In [60]:

```python
feature_matrix_top50=transform(corpus,vocab,idf_values)
```

In [72]:

```python
print("documents in corpus: ",len(corpus))
print("shape of feature matrix: ",feature_matrix_top50.shape)
```

```
documents in corpus:  746
shape of feature matrix:  (746, 50)
```

In [73]:

```python
print(feature_matrix_top50)
```

```
  (9, 10) 1.0
  (19, 23) 0.5773502691896258
  (19, 30) 0.5773502691896258
  (19, 47) 0.5773502691896258
  (43, 44) 1.0
  (68, 32) 1.0
  (83, 29) 1.0
  (109, 4) 1.0
  (114, 18) 1.0
  (135, 24) 0.5773502691896258
  (135, 37) 0.5773502691896258
  (135, 43) 0.5773502691896258
  (143, 31) 1.0
  (198, 3) 1.0
  (220, 9) 1.0
  (222, 6) 0.06237828615518054
  (222, 16) 0.9980525784828886
  (228, 34) 1.0
  (247, 13) 1.0
  (295, 19) 1.0
  (324, 14) 1.0
  (333, 36) 1.0
  (338, 0) 1.0
  (347, 8) 1.0
  (360, 46) 1.0
  (368, 15) 1.0
```

```
  (413, 38) 1.0
  (421, 48) 1.0
  (436, 42) 1.0
  (437, 1) 1.0
  (447, 7) 0.7071067811865476
  (447, 49) 0.7071067811865476
  (459, 45) 1.0
  (486, 21) 1.0
  (489, 22) 1.0
  (572, 35) 1.0
  (590, 33) 1.0
  (591, 5) 1.0
  (595, 40) 1.0
  (628, 12) 1.0
  (644, 39) 1.0
  (655, 17) 1.0
  (659, 27) 1.0
  (672, 2) 1.0
  (696, 25) 1.0
  (697, 11) 1.0
  (703, 28) 1.0
  (714, 20) 1.0
  (719, 41) 1.0
  (723, 26) 1.0
```

In [70]:

```
print(feature_matrix_top50[19]) #sparse representation of the 20th document
```

```
  (0, 23) 0.5773502691896258
  (0, 30) 0.5773502691896258
  (0, 47) 0.5773502691896258
```

In [71]:

```
print(feature_matrix_top50[19].toarray()) #dense matrix represenatation of  the 20th docu
ment
```

```
[[0.         0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.57735027
  0.         0.         0.         0.         0.         0.
  0.57735027 0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.
  0.         0.         0.         0.         0.         0.57735027
  0.         0.        ]]
```