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]:
```
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 method
# After using the fit function on the corpus the vocab has 9 words in it, and each has its 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 dense matrix and printing it.
# Notice that this output is normalized using L2 normalization. sklearn does this by default.

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

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

Custom implementation

In [8]:
```
import warnings
warnings.filterwarnings("ignore")

from collections import Counter
from tqdm import tqdm
from scipy.sparse import csr_matrix
import math
import operator
import numpy
import scipy

import pandas as pd
from sklearn.preprocessing import normalize
```

In [9]:
```
dataset = [
     'this is the first document',
     'this document is the second document',
     'and this is the third one',
     'is this the first document']
```

In [10]:
```
#Tokenization
#Reference: NOTEBOOK Assignment_3_Reference by AAIC

def fit(dataset:'Given Dataset') -> 'It returns Unique Words':
    unique_words = set() # at first we will initialize an empty set
    # check if its list type or not
    if isinstance(dataset, (list,)):
        for row in dataset: # for each review in the dataset
            for word in row.split(" "): # for each word in the review. #split method converts a string into list of words
                if len(word) < 2:
                    continue
                unique_words.add(word)
        unique_words = sorted(list(unique_words))
        vocab = {j:i for i,j in enumerate(unique_words)}

        return vocab
    else:
        print("you need to pass list of sentance")
```

In [11]:
```
vocab = fit(dataset);
print(vocab)
```

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

In [12]:
```
def transform(dataset:'Given Document Corpus',
              vocab:'Unique Words')  -> 'returns tf-idf matrix and idf values':
    rows = []
    columns = []
    values = []
    tf_values = []

    #Reference: NOTEBOOK Assignment_3_Reference by AAIC
    if isinstance(dataset, (list,)):
        for idx, row in enumerate(tqdm(dataset)):
            word_freq = dict(Counter(row.split()))
            N = len(row.split())
            for word, freq in word_freq.items():
                if len(word) < 2:
                    continue
                col_index = vocab.get(word, -1)

        #Computing TF
                if col_index !=-1:
                    rows.append(idx)
                    columns.append(col_index)
                    tf_values.append(freq/N)

        tf_matrix = csr_matrix((tf_values, (rows,columns)), shape=(len(dataset),len(vocab)))

        #Computing IDF
        tf_data = pd.DataFrame(tf_matrix.toarray())
        word_freq_in_corpus = list(tf_data.astype(bool).sum(axis=0))
        idf = list(map(lambda x :(1 + math.log((len(tf_data)+1)/(x+1))), word_freq_in_corpus))
        idfDict = dict(zip(vocab, idf))

        #Computing TF-IDF
        idf_data = pd.Series(list(idfDict.values()), index = tf_data.columns)
        tf_idf_data = tf_data.mul(idf_data)
        tfidf_list = tf_idf_data.values.tolist()


        #Now task is to write sparse metrics
        #Converting dataframe directly to sparse matrics
        output = scipy.sparse.csr_matrix(tf_idf_data.values)

        #Applying L2 normalization
        #Reference: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
        output_normalized = normalize(output, norm='l2', axis=1, copy=True, return_norm=False)
        return idf, output_normalized.toarray()

    else:
        print("you need to pass list of strings")
```

In [13]:
```
idf_values, tfidf_output = transform(dataset,vocab)
print("VOCAB :", vocab)
print("_"*50)
print("IDF VALUES ARE : ", idf_values)
print("_"*50)
print("TFIDF VALUES ARE :",tfidf_output)
```

```
100%|██████████| 4/4 [00:00<00:00, 1328.57it/s]
VOCAB : {'and': 0, 'document': 1, 'first': 2, 'is': 3, 'one': 4, 'second': 5, 'the': 6, 'third': 7, 'this': 8}
__________________________________________________
IDF VALUES ARE :  [1.916290731874155, 1.2231435513142097, 1.5108256237659907, 1.0, 1.916290731874155, 1.916290731874155, 1.0, 1.916290731874155, 1.0]
__________________________________________________
TFIDF VALUES ARE : [[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]
 [0.         0.6876236  0.         0.28108867 0.         0.53864762
  0.28108867 0.         0.28108867]
 [0.51184851 0.         0.         0.26710379 0.51184851 0.
  0.26710379 0.51184851 0.26710379]
 [0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]]
```

Compairing results with sklearn output

In [14]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()

vectorizer.fit(dataset)
sklearn_output = vectorizer.transform(dataset)
```

In [15]:
```
print("SKLEARN TFIDF FEATURES ARE: ", vectorizer.get_feature_names())
print("_"*50)
print("SKLEARN IDF VALUES ARE :",vectorizer.idf_)
print("_"*50)
print("SKLEARN TFIDF VALUES ARE :",sklearn_output.toarray())
```

```
SKLEARN TFIDF FEATURES ARE:  ['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
__________________________________________________
SKLEARN IDF VALUES ARE : [1.91629073 1.22314355 1.51082562 1.         1.91629073 1.91629073
 1.         1.91629073 1.        ]
__________________________________________________
SKLEARN TFIDF VALUES ARE : [[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]
 [0.         0.6876236  0.         0.28108867 0.         0.53864762
  0.28108867 0.         0.28108867]
 [0.51184851 0.         0.         0.26710379 0.51184851 0.
  0.26710379 0.51184851 0.26710379]
 [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 [16]:
```
# 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 [17]:
```
vocab_ = fit(corpus)
```

In [18]:
```
def transform(dataset:'Given Document Corpus',
              vocab:'Unique Words', *,
              max_features:'Top Words' = 50)  -> 'returns tf-idf matrix and top idf values' :

    rows = []; columns = []; values = []; tf_values = []
    rows_new = []; columns_new = []; tf_values_new= []

    #Reference: NOTEBOOK Assignment_3_Reference by AAIC
    if isinstance(dataset, (list,)):
        for idx, row in enumerate(tqdm(dataset)):
            word_freq = dict(Counter(row.split()))
            N = len(row.split())
            for word, freq in word_freq.items():
                if len(word) < 2:
                    continue
                col_index = vocab.get(word, -1)

        #Computing TF
                if col_index !=-1:
                    rows.append(idx)
                    columns.append(col_index)
                    tf_values.append(freq/N)
        tf_matrix = csr_matrix((tf_values, (rows,columns)), shape=(len(dataset),len(vocab)))

        #Computing IDF
        tf_data = pd.DataFrame(tf_matrix.toarray())
        word_freq_in_corpus = list(tf_data.astype(bool).sum(axis=0))
        idf = list(map(lambda x :(1 + math.log((len(tf_data)+1)/(x+1))), word_freq_in_corpus))
        idfDict = dict(zip(vocab, idf))

        #Selecting only top features i.e. max_features based on IDF values
        #Reference : https://stackoverflow.com/questions/16310015/what-does-this-mean-key-lambda-x-x1
        top_idf_dict = dict(sorted(idfDict.items(), key=lambda x : x[1], reverse=True))
        top_idf_dict = dict(Counter(idfDict).most_common(max_features))


        #for word dimension
        vocab_topwords_dimension = dict(zip(top_idf_dict.keys(), list(range(0,max_features))))

        #Computing TF again based on top words
        for idx, row in enumerate(dataset):
            word_freq = dict(Counter(row.split()))
            N = len(row.split())
            for word, freq in word_freq.items():
                if word not in vocab_topwords_dimension.keys():
                    continue
                col_index_idf = vocab_topwords_dimension.get(word, -1)
                if col_index !=-1:
                    rows_new.append(idx)
                    columns_new.append(col_index_idf)
                    tf_values_new.append(freq/N)

        #writing sparse matrics
        tf_matrix_new = csr_matrix((tf_values_new, (rows_new,columns_new)), shape=(len(dataset),len(vocab_topwords_dimension)))
        tf_data_new = pd.DataFrame(tf_matrix_new.toarray())

        #Computing TF-IDF
        idf_data = pd.Series(list(top_idf_dict.values()), index = tf_data_new.columns)
        tf_idf_data = tf_data_new.mul(idf_data)
        tfidf_list = tf_idf_data.values.tolist()


        #Now task is to write sparse metrics
        #Converting dataframe directly to sparse matrics
        output = scipy.sparse.csr_matrix(tf_idf_data.values)

        #Applying L2 normalization
        #Reference: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
        output_normalized = normalize(output, norm='l2', axis=1, copy=True, return_norm=False)

    return output_normalized, top_idf_dict
```

In [19]:
```
max_features = 50
output, top_words = transform(corpus,vocab_, max_features = 50)
```

```
100%|██████████| 746/746 [00:00<00:00, 24880.33it/s]
```

In [20]:
```
output
```

Out[20]:
```
<746x50 sparse matrix of type '<class 'numpy.float64'>'
	with 50 stored elements in Compressed Sparse Row format>
```

In [21]:
```
print(output.shape)
print(output.toarray())
```

```
(746, 50)
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

In [22]:
```
print("Top {} feature:idf values are {} :".format(max_features, top_words))
```

```
Top 50 feature:idf values are {'aailiyah': 6.922918004572872, 'abandoned': 6.922918004572872, 'abroad': 6.922918004572872, 'abstruse': 6.922918004572872, 'aca
demy': 6.922918004572872, 'accents': 6.922918004572872, 'accessible': 6.922918004572872, 'acclaimed': 6.922918004572872, 'accolades': 6.922918004572872, 'accu
rate': 6.922918004572872, 'accurately': 6.922918004572872, 'achille': 6.922918004572872, 'ackerman': 6.922918004572872, 'actions': 6.922918004572872, 'adams':
6.922918004572872, 'add': 6.922918004572872, 'added': 6.922918004572872, 'admins': 6.922918004572872, 'admiration': 6.922918004572872, 'admitted': 6.922918004
572872, 'adrift': 6.922918004572872, 'adventure': 6.922918004572872, 'aesthetically': 6.922918004572872, 'affected': 6.922918004572872, 'affleck': 6.922918004
572872, 'afternoon': 6.922918004572872, 'aged': 6.922918004572872, 'ages': 6.922918004572872, 'agree': 6.922918004572872, 'agreed': 6.922918004572872, 'aimles
s': 6.922918004572872, 'aired': 6.922918004572872, 'akasha': 6.922918004572872, 'akin': 6.922918004572872, 'alert': 6.922918004572872, 'alike': 6.922918004572
872, 'allison': 6.922918004572872, 'allow': 6.922918004572872, 'allowing': 6.922918004572872, 'alongside': 6.922918004572872, 'amateurish': 6.922918004572872,
'amaze': 6.922918004572872, 'amazed': 6.922918004572872, 'amazingly': 6.922918004572872, 'amusing': 6.922918004572872, 'amust': 6.922918004572872, 'anatomis
t': 6.922918004572872, 'angel': 6.922918004572872, 'angela': 6.922918004572872, 'angelina': 6.922918004572872} :
```

In [23]:
```
print("Output of document no 135 is :")
print(output.toarray()[135])
```

```
Output of document no 135 is :
[0.         0.         0.         0.37796447 0.         0.
 0.         0.         0.         0.         0.37796447 0.
 0.         0.         0.         0.         0.         0.
 0.37796447 0.         0.37796447 0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.37796447 0.         0.         0.         0.37796447 0.37796447
 0.         0.         0.         0.         0.         0.
 0.         0.        ]
```
