TF-IDF — Term Frequency-Inverse Document Frequency

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.datasets import fetch_20newsgroups
from sklearn.metrics.pairwise import cosine_similarity
import math
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
#Dataset
newsgroups = fetch_20newsgroups(subset='all', remove=('headers', 'footers', 'quotes'))
documents = newsgroups.data
```

+ Code + Text

```
# Text cleaning

def clean_text(text):
    cleaned_chars = [char if char.isalnum() or char.isspace() else ' ' for char in text]
    cleaned_text = ''.join(cleaned_chars)
    return cleaned_text

# Tokenization and stop word removal
def tokenize_and_remove_stopwords(text):
    words = nltk.word_tokenize(text)
    stop_words = set(stopwords.words('english'))
    words = [word.lower() for word in words if word.lower() not in stop_words]
    return words
```

```
cleaned_documents = [(tokenize_and_remove_stopwords(clean_text(doc))) for doc in documents]
```

```
documents_str = [' '.join(doc) for doc in cleaned_documents]

vectorizer = TfidfVectorizer()
dtm = vectorizer.fit_transform(documents_str)
```

```
print("Cleaned Documents:")

for i in range(len(cleaned_documents)):
  print("Document", i + 1, ":", cleaned_documents[i])
```

```
Document 9945 : ['following', 'discussions', 'delta', 'clipper', 'program', 'one', 'small', 'question', 'understand', 'de',
Document 9946 : []
Document 9947 : ['vvvvvvvvvvvvvvvvvvvvv', 'vvvvvvvvvv', 'getting', 'close']
Document 9948 : ['may', 'fairly', 'routine', 'request', 'looking', 'fast', 'polygon', 'routine', 'used', '3d', 'game', 'one'
Document 9949 : ['find', 'intriguing', 'remarks', 'could', 'give', 'us', 'bit', 'explanation', 'example', 'religion', 'anti'
Document 9950 : ['sorry', 'card', 'display', '17000k', 'colors', '1700k', 'colors', 'hope', 'one', 'could', 'answer', 'quest
Document 9951 : ['may', 'naive', 'question', 'basis', 'claim', 'cpu', 'get', 'hotter', 'computationally', 'intensive', 'job'
Document 9952 : ['1', 'large', 'padded', 'cordura', 'bag', 'maker', 'unknown', 'nge', 'exterior', 'black', 'straps', 'interi
Document 9953 : ['variance', 'perfect', 'sphericity', 'model', 'earth', 'small', 'enough', 'fit', 'home', 'would', 'probably
Document 9954 : ['wondering', 'mean', 'lectorium', 'rosicrucianum', 'warning', 'point', 'arguing', 'legit', 'golden', 'dawn'
Document 9955 : ['given', 'authority', 'trample', 'snakes', 'scorpions', 'overcome', 'power', 'enemy', 'nothing', 'harm', 'l
Document 9956 : ['know', 'either', 'truth', 'known', 'little', 'known', 'angels', 'even', 'guess', 'really', 'know', 'angels
Document 9957 : ['someone', 'downloaded', 'pctools', 'demo', 'compuserve', 'please', 'upload', 'cica', 'ftp', 'site']
Document 9958 : ['realy', 'like', 'idea', 'would', 'wonderfull', 'see', 'big', 'bright', 'satelite', 'night', 'sky', 'even',
Document 9959 : ['give', 'new', 'viewsonic', '17', 'good', 'look', 'seen', 'side', 'side', 'old', 'viewsonic', '7', 'mag', '
Document 9960 : ['stuff', 'deleted']
Document 9961 : ['hmmm', 'intersting', 'long', 'message', 'twice', 'well', 'care', 'libertarianism', 'philisophical', 'disag
Document 9962 : ['playing', 'centris', '610', '8', '230', 'last', 'couple', 'weeks', 'problem', 'seen', 'couple', 'applicati
Document 9963 : ['good', 'summary', 'posted', 'thanks', 'wanted', 'add', 'another', 'comment', 'remeber', 'reading', 'commer
Document 9964 : ['please', 'excuse', 'redirect', 'already', 'answered', 'small', 'utility', 'switches', 'functionality', 'ca
Document 9965 : ['pardon', 'humble', 'atheist', 'exactly', 'difference', 'holding', 'revealed', 'truth', 'blind', 'faith', '
DocumentIOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

```
print("Vocabulary:")
print()
print(len(vectorizer.get_feature_names_out()))
```

```
Vocabulary:

129906
```

Double-click (or enter) to edit

```
#cosine similarity
def calculate_cosine_similarity(vector1, vector2):
    return cosine_similarity(vector1.reshape(1, -1), vector2.reshape(1, -1))[0][0]

#document similarity
def document_similarity_search(input_document, top_n=5):
    input_vector = vectorizer.transform([input_document])
    similarities = [calculate_cosine_similarity(input_vector, doc_vector) for doc_vector in dtm]
    ranked_indices = sorted(range(len(similarities)), key=similarities.__getitem__, reverse=True)

    # Return top N similar documents
    result = [(documents[i], similarities[i]) for i in ranked_indices[:top_n]]
    return result
```

```
input_doc = "College"
similar_documents = document_similarity_search(input_doc)

for i, (doc, similarity) in enumerate(similar_documents, 1):
    print(i,'Similarity',similarity)
    print(doc)
    print()
```

```
1 Similarity 0.36001385451023843

Ask me whether I'm surprised that you haven't managed to waddle out of
college after all this time.


2 Similarity 0.26326039070130103
Hello,

I am planning on attending Podiatry School next year.

I have narrowed my choices to the Pennsylvania College of Podiatric
Medicine, in Philadelphia, or the California College of Podiatric
Medicine in San Francisco.

If anyone has any information or oppinions about these two schools, please
tell me.  I am having a hard time deciding which one to attend, and must
```

```
make a decision very soon.

thank you, Larry

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Live From New York, It's SATURDAY NIGHT...

3 Similarity 0.2398779889319091
Apparently, the only place to take the MSF course around
here in NC is at a community college.

That woudl preclude some sort of state
subsidation, then, no?


4 Similarity 0.23621977831442106
Could someone please post the rosters for the College Hockey All-Star game East
and West Rosters?  Thanks in advance.


5 Similarity 0.22653384233276685

I would guess that it probably has something to do with the ease of which
ideas and thoughts are communicated on a college campus.  In the real world
(tm) it's easier for theists (well, people in general really) to lock
themselves into a little bubble where they only see and talk to those
people who are of the same opinion as they are.  In college you are
constantly surrounded by and have to interact with people who have
different ideas about life, the universe, and everything.  It is much much
harder to build a bubble around yourself to keep everyone else's ideas from
reaching you.

So, in a world where theists are forced to contend with and listen to
atheists and theists of other religions some are bound to have a change in
their beliefs over four years.  There is nowhere to run.... :-)
```

Start coding or generate with AI.