IR_LAB.py

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def read_file(file_path):
    """Read the contents of a file."""
    with open(file_path, 'r', encoding='utf-8') as file:
        return file.read()

def compute_cosine_similarity_and_common_words(file1, file2):
    # Read the content from the files
    doc1 = read_file(file1)
    doc2 = read_file(file2)

    # Create the Document list
    documents = [doc1, doc2]

    # Initialize the TF-IDF Vectorizer
    vectorizer = TfidfVectorizer()

    # Transform the documents into TF-IDF vectors
    tfidf_matrix = vectorizer.fit_transform(documents)

    # Compute the cosine similarity between the two documents
    similarity_matrix = cosine_similarity(tfidf_matrix[0:1], tfidf_matrix[1:2])
    cosine_sim = similarity_matrix[0][0]

    # Get feature names (i.e., the words)
    feature_names = vectorizer.get_feature_names_out()

    # Convert the TF-IDF vectors to dense arrays
    tfidf_array1 = tfidf_matrix[0].toarray()[0]
    tfidf_array2 = tfidf_matrix[1].toarray()[0]

    # Identify common words (non-zero in both arrays)
    common_words = []
    for i in range(len(feature_names)):
        if tfidf_array1[i] > 0 and tfidf_array2[i] > 0:
            common_words.append(feature_names[i])

    return cosine_sim, common_words

# Example usage
file1 = 'document1.txt'  # Replace with the path to your first file
file2 = 'document2.txt'  # Replace with the path to your second file

similarity, common_words = compute_cosine_similarity_and_common_words(file1, file2)

print(f"Cosine Similarity: {similarity}\n")
```

```
if common_words:
    print("Common Words:")
    for word in common_words:
        print(f"'{word}'")
else:
    print("No common words found.")
```

OUTPUT:

```
PS C:\Users\saura\Desktop\4th year study material\Lab Program\IR LAB> &
C:/Users/saura/AppData/Local/Programs/Python/Python311/python.exe
"c:/Users/saura/Desktop/4th year study material/Lab Program/IR LAB/Lab_1py"
Cosine Similarity: 0.5479807587583923

Common Words:
'and'
'artificial'
'future'
'intelligence'
'of'
'technology'
'the'
PS C:\Users\saura\Desktop\4th year study material\Lab Program\IR LAB>
```