Navadeep Munugoti

2403A52015

AIAI 02

Install Required Libraries

```
import nltk
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
```

Import Libraries

```
import pandas as pd
import re
import matplotlib.pyplot as plt

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

from sklearn.feature_extraction.text import TfidfVectorizer
from wordcloud import WordCloud
```

Loading dataset

```
df = pd.read_csv("/content/Tweets.csv")
df
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 14640,\n  \"fields\":
[\n    {\n      \"column\": \"tweet_id\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 779111158481836,\n
\"min\": 567588278875213824,\n        \"max\": 570310600460525568,\n
\"num_unique_values\": 14485,\n        \"samples\": [\n
567917894144770049,\n          567813976492417024,\n
569243676594941953\n        ],\n        \"semantic_type\": \"\",\n
```

```
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"airline_sentiment\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3,\n        \"samples\":
[\n          \"neutral\",\n          \"positive\",\n
\"negative\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"airline_sentiment_confidence\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.1628299590986659,\n
\"min\": 0.335,\n        \"max\": 1.0,\n        \"num_unique_values\":
1023,\n        \"samples\": [\n          0.6723,\n          0.3551,\n
0.6498\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"negativereason\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 10,\n
\"samples\": [\n          \"Damaged Luggage\",\n          \"Can't
Tell\",\n          \"Lost Luggage\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"negativereason_confidence\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.3304397596377413,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 1410,\n        \"samples\": [\n
0.6677,\n          0.6622,\n          0.6905\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
6,\n        \"samples\": [\n          \"Virgin America\",\n
\"United\",\n          \"American\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline_sentiment_gold\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"negative\",\n          \"neutral\",\n          \"positive\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"name\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\":
7701,\n        \"samples\": [\n          \"smckenna719\",\n
\"thisAnneM\",\n          \"jmspool\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"negativereason_gold\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 13,\n        \"samples\": [\n
\"Customer Service Issue\\nLost Luggage\",\n          \"Late Flight\\
nCancelled Flight\",\n          \"Late Flight\\nFlight Attendant
Complaints\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"retweet_count\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0,\n        \"min\": 0,\n
\"max\": 44,\n        \"num_unique_values\": 18,\n        \"samples\":
[\n          0,\n          1,\n          6\n        ],\n
```

```
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 14427,\n
\"samples\": [\n          \"@JetBlue so technically I could drive to
JFK now and put in. Request for tomorrow's flight?\",\n
\"@united why I won't check my carry on. Watched a handler throw this
bag -- miss the conveyer belt -- sat there 10 min
http://t.co/lyoocx5mSH\",\n          \"@SouthwestAir you guys are so
clever \\ud83d\\ude03 http://t.co/qn5odUGFqK\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"tweet_coord\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 832,\n        \"samples\": [\n
\"[40.04915451, -75.10364317]\",\n          \"[32.97609561, -
96.53349238]\",\n          \"[26.37852293, -81.78472152]\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tweet_created\",\n      \"properties\": {\n        \"dtype\":
\"object\",\n        \"num_unique_values\": 14247,\n
\"samples\": [\n          \"2015-02-23 07:40:55 -0800\",\n
\"2015-02-21 16:20:09 -0800\",\n          \"2015-02-21 21:33:21 -
0800\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tweet_location\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3081,\n
\"samples\": [\n          \"Oakland, California\",\n
\"Beverly Hills, CA\",\n          \"Austin, TX/NY, NY\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"user_timezone\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 85,\n        \"samples\": [\n
\"Helsinki\",\n          \"Eastern Time (US & Canada)\",\n
\"America/Detroit\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```

Filter Negative Sentiment Tweets

```python
negative_df = df[df['airline_sentiment'] == 'negative']
negative_df = negative_df[['text']]
negative_df.head()
```

```
{"summary":"{\n  \"name\": \"negative_df\",\n  \"rows\": 9178,\n
\"fields\": [\n    {\n      \"column\": \"text\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 9087,\n        \"samples\": [\n
\"@JetBlue u guys have 2b kidding. No help anywhere. 5 hour delays?
Still no answers. Bad cust service. #idlovetoask
http://t.co/DPX3yoGTEj\",\n          \"@SouthwestAir crazy how every
airline flew out to the northeast tonight except you\",\n
\"@JetBlue what else on this plane is duct-taped?? #ohboy
#shouldigetoutandpush #airplane #flying\\u2026
http://t.co/R9ZsVzuRLw\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"negative_df"}
```

Text Preprocessing

```python
stop_words = set(stopwords.words('english'))
nltk.download('punkt_tab')

def clean_text(text):
    text = re.sub(r"http\S+", "", text)       # remove URLs
    text = re.sub(r"@\w+", "", text)          # remove mentions
    text = re.sub(r"#\w+", "", text)          # remove hashtags
    text = re.sub(r"[^a-zA-Z\s]", "", text)  # remove special chars
    text = text.lower()

    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word not in stop_words]

    return " ".join(tokens)

negative_df['cleaned_text'] = negative_df['text'].apply(clean_text)
negative_df.head()
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
```

```
{"summary":"{\n  \"name\": \"negative_df\",\n  \"rows\": 9178,\n
\"fields\": [\n    {\n      \"column\": \"text\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 9087,\n        \"samples\": [\n
\"@JetBlue u guys have 2b kidding. No help anywhere. 5 hour delays?
Still no answers. Bad cust service. #idlovetoask
http://t.co/DPX3yoGTEj\",\n          \"@SouthwestAir crazy how every
airline flew out to the northeast tonight except you\",\n
\"@JetBlue what else on this plane is duct-taped?? #ohboy
#shouldigetoutandpush #airplane #flying\\u2026
http://t.co/R9ZsVzuRLw\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"cleaned_text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 9051,\n
\"samples\": [\n          \"formally complain customer service handler
misconnected denied boarding amp lost bag help\",\n          \"yep
still waiting bags whats holdup looks like flight got time\",\n
\"hold hours minutes whats going\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"negative_df"}
```

Compute TF-IDF

```
tfidf = TfidfVectorizer(max_features=20)
tfidf_matrix = tfidf.fit_transform(negative_df['cleaned_text'])

tfidf_df = pd.DataFrame(
    tfidf_matrix.toarray(),
    columns=tfidf.get_feature_names_out()
)

tfidf_df.head()
```

```
{"summary":"{\n  \"name\": \"tfidf_df\",\n  \"rows\": 9178,\n
\"fields\": [\n    {\n      \"column\": \"amp\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.16123471561725813,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 192,\n        \"samples\": [\n
0.5650140353603288,\n          0.5056829467393018,\n
0.5726494569529821\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"call\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.15513414206697315,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 183,\n        \"samples\": [\n
0.8013260086824548,\n          0.6079731775194172,\n
0.46471863332624275\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"cancelled\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.17466014371694094,\n        \"min\":
0.0,\n        \"max\": 1.0,\n        \"num_unique_values\": 338,\n
\"samples\": [\n          0.5289445934573577,\n
0.45937952163338047,\n          0.6755211398528438\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"cant\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.16319120399403594,\n
\"min\": 0.0,\n        \"max\": 1.0,\n        \"num_unique_values\":
198,\n        \"samples\": [\n          0.6112290619743166,\n
0.40653597011821896,\n          0.5564624971582296\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"customer\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
0.1679809801780749,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 175,\n        \"samples\": [\n
0.7425107121835887,\n          0.5114682207338577,\n
0.4217348339891991\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"delayed\",\n      \"properties\": {\n        \"dtype\": \"number\",\
n        \"std\": 0.16491639746723846,\n        \"min\": 0.0,\n
\"max\": 1.0,\n        \"num_unique_values\": 170,\n
\"samples\": [\n          0.4090912571588346,\n
0.6533552750471926,\n          0.5095083482085575\n        ],\n
```

```
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"flight\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
0.2999248372438535,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 620,\n        \"samples\": [\n
0.3318650617936042,\n          0.3261570226077109,\n
0.3928358910134839\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"flightled\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.12746964003234199,\n        \"min\":
0.0,\n        \"max\": 0.8680258415275338,\n
\"num_unique_values\": 173,\n        \"samples\": [\n
0.4866107765195581,\n          0.45693121087152255,\n
0.6149107667217942\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"get\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.2150342363288849,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 307,\n        \"samples\": [\n
0.5156079728322264,\n          0.6289478389617591,\n
0.3481119095071017\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"help\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.17633686880687985,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 229,\n        \"samples\": [\n
0.45240585663730726,\n          0.4441979911899029,\n
0.4385214357295621\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"hold\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.16330726518780211,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 246,\n        \"samples\": [\n
0.6857160728663528,\n          0.3986323802546562,\n
0.5616640225621125\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"hour\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.15494312396062354,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 172,\n        \"samples\": [\n
0.41551819809294166,\n          0.5310417113407537,\n
0.5015080574087228\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"hours\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.17249217250184365,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 262,\n        \"samples\": [\n
0.41897695377126676,\n          0.8861288840885093,\n
0.5146767251928148\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"im\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.18059597512587458,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 173,\n        \"samples\": [\n
0.8239477144371351,\n          0.7550173364193317,\n
0.49267100424338145\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"one\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.16407217121846052,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 151,\n        \"samples\": [\n
0.6651622175603299,\n          0.8248063468716108,\n
0.41308502100320266\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"plane\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.17361147477955052,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 157,\n        \"samples\": [\n
0.5746479385074844,\n          0.6699161785564195,\n
0.8140430278937313\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"service\",\n      \"properties\": {\n        \"dtype\": \"number\",\
n        \"std\": 0.187019422240821,\n        \"min\": 0.0,\n
\"max\": 1.0,\n        \"num_unique_values\": 212,\n
\"samples\": [\n          0.5412550923353067,\n
0.5249351548602849,\n          0.4042066968041475\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"still\",\n      \"properties\": {\
n        \"dtype\": \"number\",\n        \"std\":
0.16845686792955383,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 171,\n        \"samples\": [\n
0.46789459533841177,\n          0.4003457800645524,\n
0.5651332587456539\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"time\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.1826274408765689,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 185,\n        \"samples\": [\n
0.6922429480049691,\n          0.632707305973712,\n
0.5421760538621594\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"us\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.17172005108828917,\n        \"min\": 0.0,\n        \"max\":
1.0,\n        \"num_unique_values\": 164,\n        \"samples\": [\n
0.49536291702725754,\n          0.399128096019977,\n
0.5840788024314748\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"tfidf_df"}
```

Identify Top TF-IDF Terms

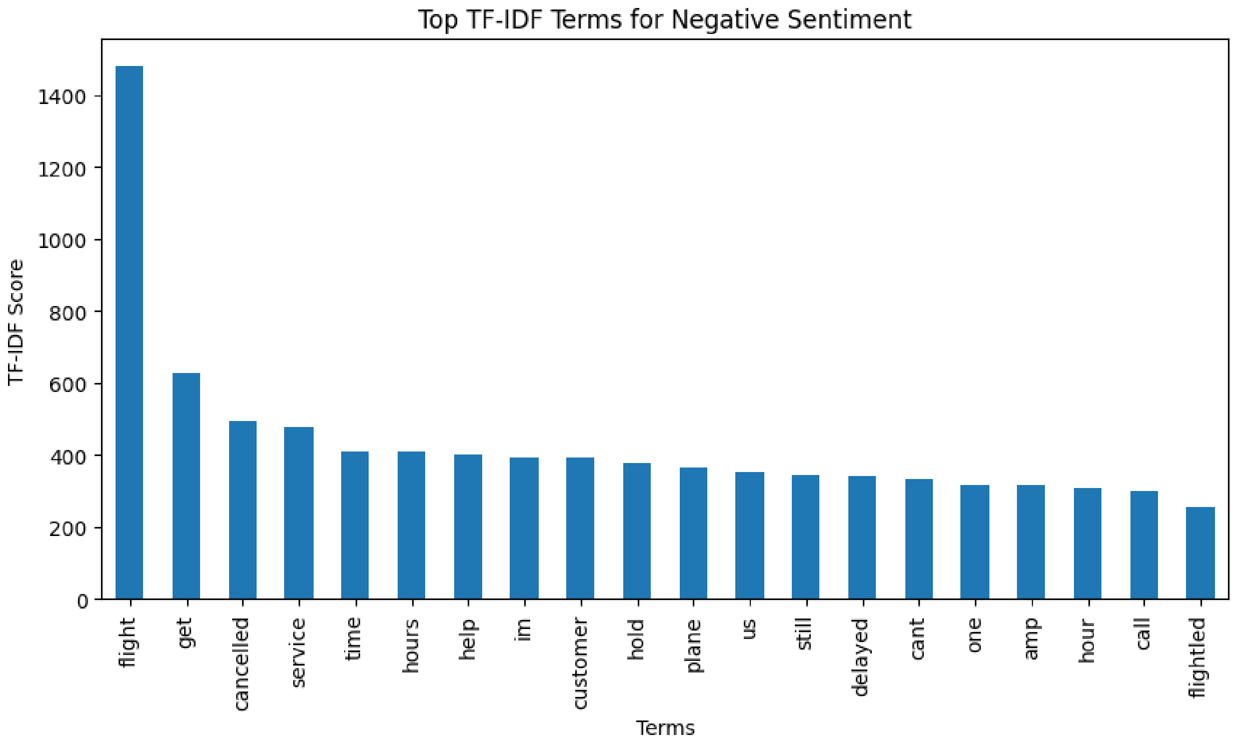
```
tfidf_scores = tfidf_df.sum().sort_values(ascending=False)

tfidf_scores
```

```
flight       1481.000000
get           628.834187
```

```
cancelled     495.030891
service       480.297397
time          410.137949
hours         408.907301
help          400.960459
im            393.561573
customer      391.901305
hold          375.947292
plane         365.325339
us            354.717077
still         346.068281
delayed       340.920914
cant          333.578100
one           317.401553
amp           316.283743
hour          308.308436
call          299.907081
flightled     255.408502
dtype: float64
```

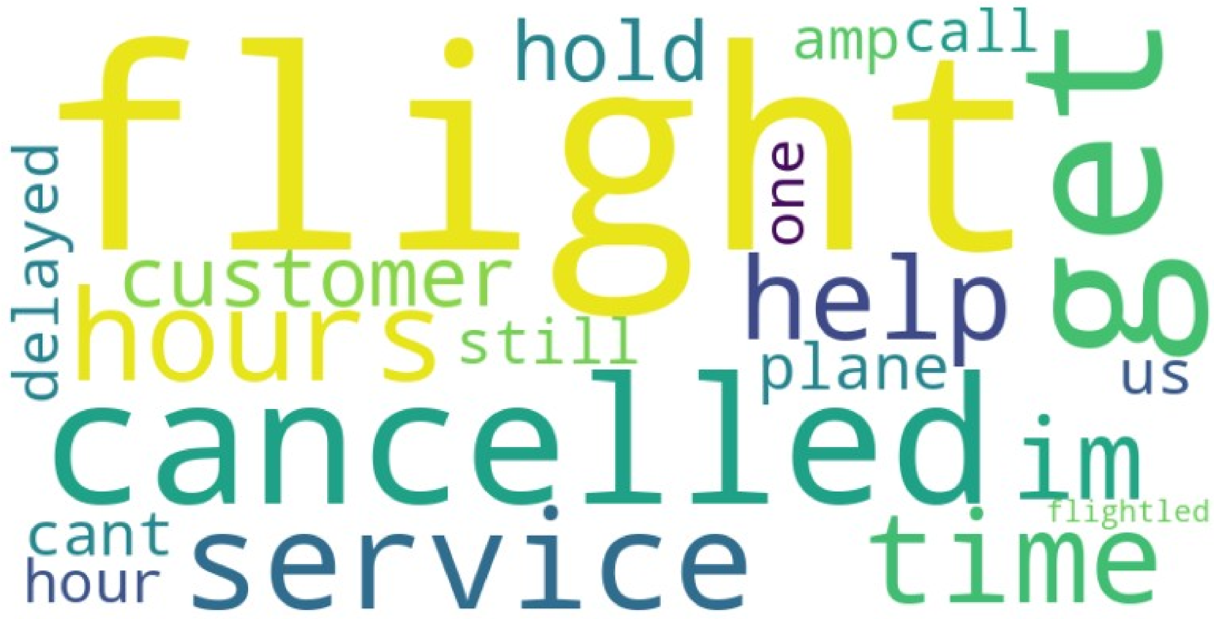
Bar Chart Visualization

```
plt.figure(figsize=(10,5))
tfidf_scores.plot(kind='bar')
plt.title("Top TF-IDF Terms for Negative Sentiment")
plt.xlabel("Terms")
plt.ylabel("TF-IDF Score")
plt.show()
```



Word Cloud Visualization

```
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color='white'
).generate_from_frequencies(tfidf_scores)

plt.figure(figsize=(10,5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Discussion Session

```
## Discussion

# In this experiment, TF-IDF was applied to negative airline tweets to
# identify the most important words contributing to negative
sentiment.
# The analysis highlights frequent complaint-related terms such as
delays,
# service, and cancellations. TF-IDF helps in emphasizing sentiment-
specific
# vocabulary by reducing the impact of commonly occurring words.
```