TF-IDF_for_Q&A

March 27, 2024

[ ]:
```
'''
Applied TF-IDF and logistic regression to the Q&A json file.
shows super high accuracy but does not really work on real examples.
Could be the problem of the data set or the TF-IDF method or both.

'''

import json
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Load the data
file_path = 'Questions_and_Answers.json'

with open(file_path, 'r') as file:
    data = json.load(file)

# Combine question and answer into a single string, create labels
texts = [entry['question'] + " " + entry['answer'] for entry in data]
labels = [1 if entry['humor'] else 0 for entry in data]

# Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(texts, labels, test_size=0.
 ↪2, random_state=42)

# Apply TF-IDF vectorization
vectorizer = TfidfVectorizer(max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# Train a logistic regression model
model = LogisticRegression(max_iter=1000)
model.fit(X_train_tfidf, y_train)

# Predict and evaluate
```

```
y_pred = model.predict(X_test_tfidf)

accuracy = accuracy_score(y_test, y_pred)

print("Accuracy:", accuracy)

# Function for testing
def predict_humor(question, answer):
    combined_text = question + " " + answer
    transformed_text = vectorizer.transform([combined_text])
    prediction = model.predict(transformed_text)
    return "Humor" if prediction[0] == 1 else "Fact"

# sample question/answer pairs
samples = [
    {"question": "who was the first president of the united states", "answer":␣
 ↪"George Washington."},
    {"question": "What do you call a magic dog?", "answer": "A labracadabrador.
 ↪"}
]

for sample in samples:
    print("Q:", sample["question"])
    print("A:", sample["answer"])
    print("Prediction:", predict_humor(sample["question"], sample["answer"]),␣
 ↪"\n")

# User input test
user_question = input("Enter your question: ")
user_answer = input("Enter the answer: ")
print("The user question was: ", user_question)
print("The user answer was: ", user_answer)

prediction_result = predict_humor(user_question, user_answer)
print(f"The question/answer pair is: {prediction_result}")
```

```
Accuracy: 0.9958624568169037
Q: who was the first president of the united states
A: George Washington.
Prediction: Fact

Q: What do you call a magic dog?
A: A labracadabrador.
Prediction: Fact

The user question was:  How do you find Will Smith in the snow?
The user answer was:  You look for the fresh prints.
```

The question/answer pair is: Humor