Big Data and Info Retrieval

Assignment 5

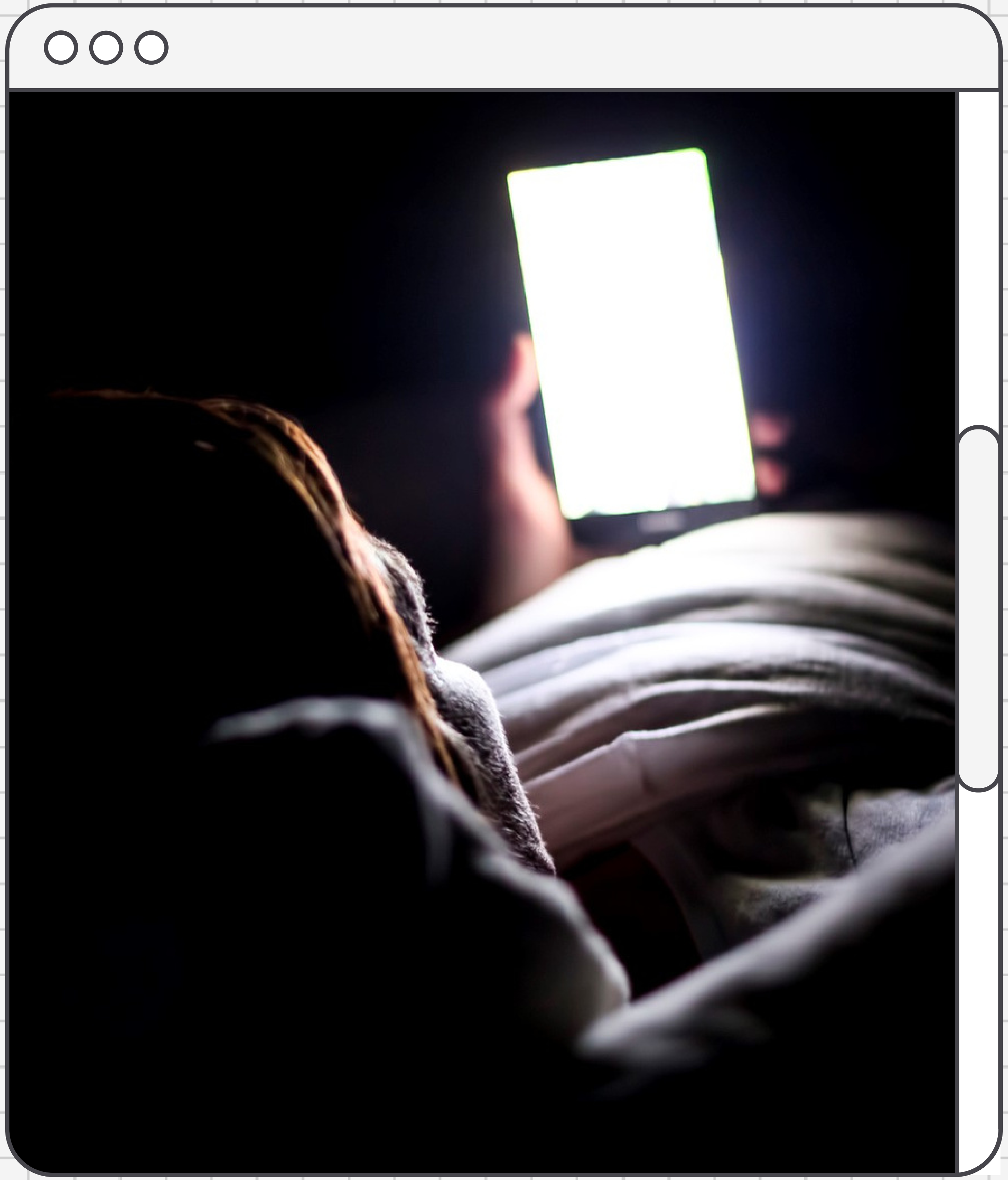
IR from Real Data

Lauren Landa

Code:**Genuine**

The Dataset: Youtube History

- Youtube watch history
  - A good amount of data (2020-Present)
- Hyperfixations
  - Weird patterns
- Known to me
- Easy to understand



Objectives

- Understand watch patterns
- Create predictive model

DA/IR Techniques Used

And Implementation

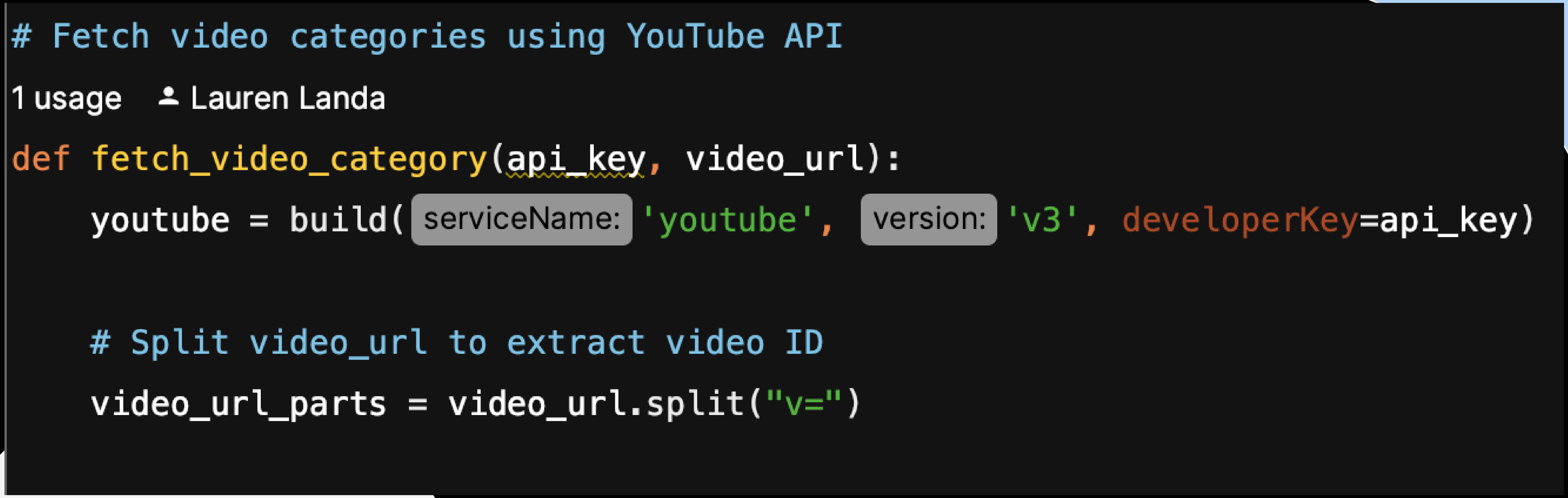
- Youtube API through GCP

- Naive Bayes Classifier

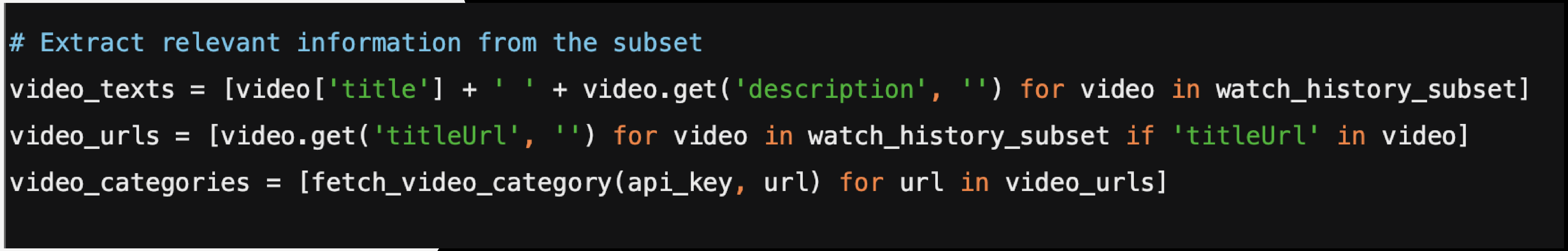
- TF-IDF
    - Applied to video titles and descriptions

Implementation

1. Set up API Key
2. Fetch Youtube categories
3. Load Youtube Watch History
4. Extract Relevant Information
5. Random Sampling
6. Train Classifier
7. Update Category Mapping
8. Predictions
9. Output



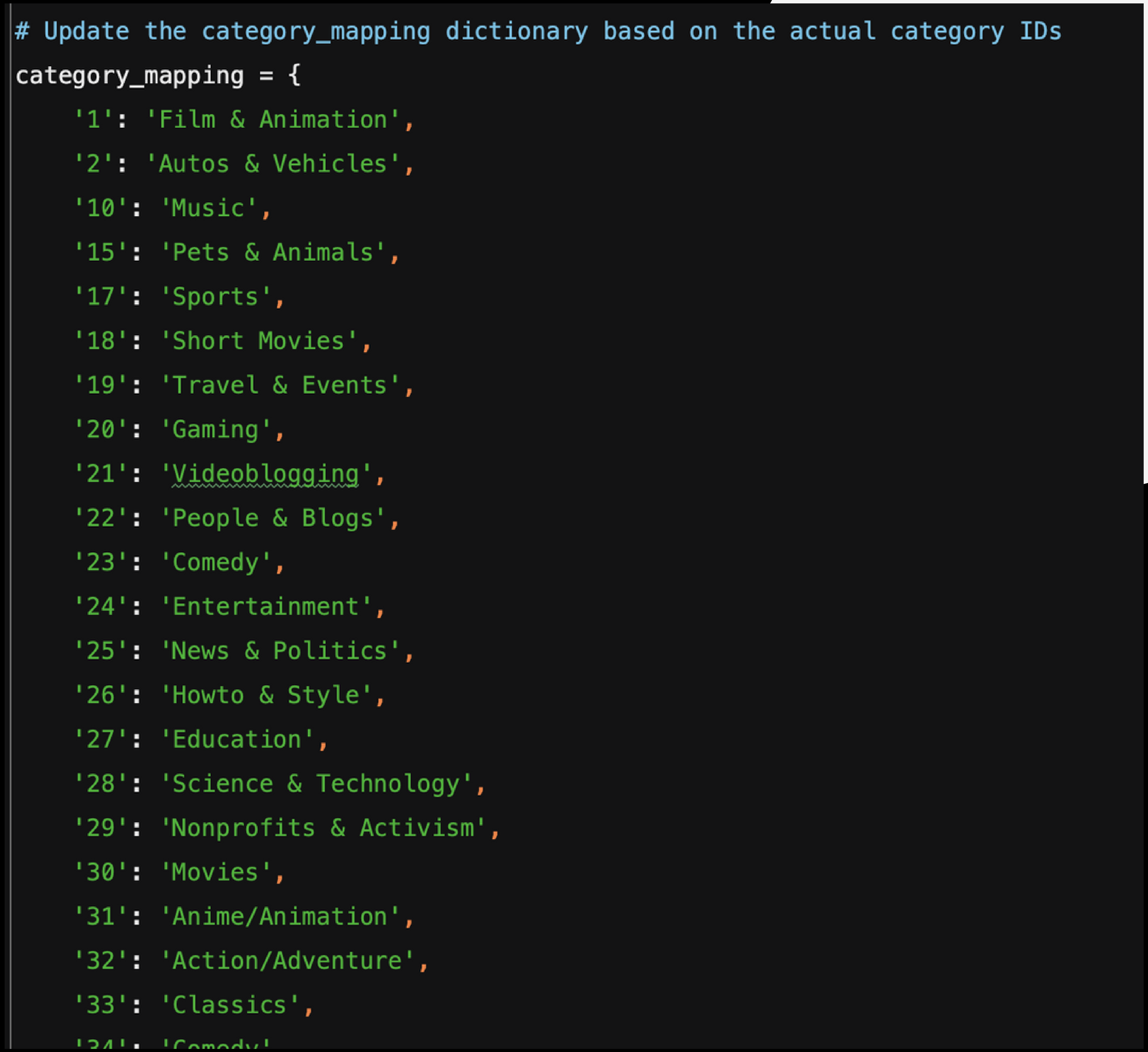
```python
# Fetch video categories using YouTube API
1 usage   Lauren Landa
def fetch_video_category(api_key, video_url):
    youtube = build( serviceName: 'youtube', version: 'v3', developerKey=api_key)

    # Split video_url to extract video ID
    video_url_parts = video_url.split("v=")
```



```python
# Extract relevant information from the subset
video_texts = [video['title'] + ' ' + video.get('description', '') for video in watch_history_subset]
video_urls = [video.get('titleUrl', '') for video in watch_history_subset if 'titleUrl' in video]
video_categories = [fetch_video_category(api_key, url) for url in video_urls]
```

Implementation

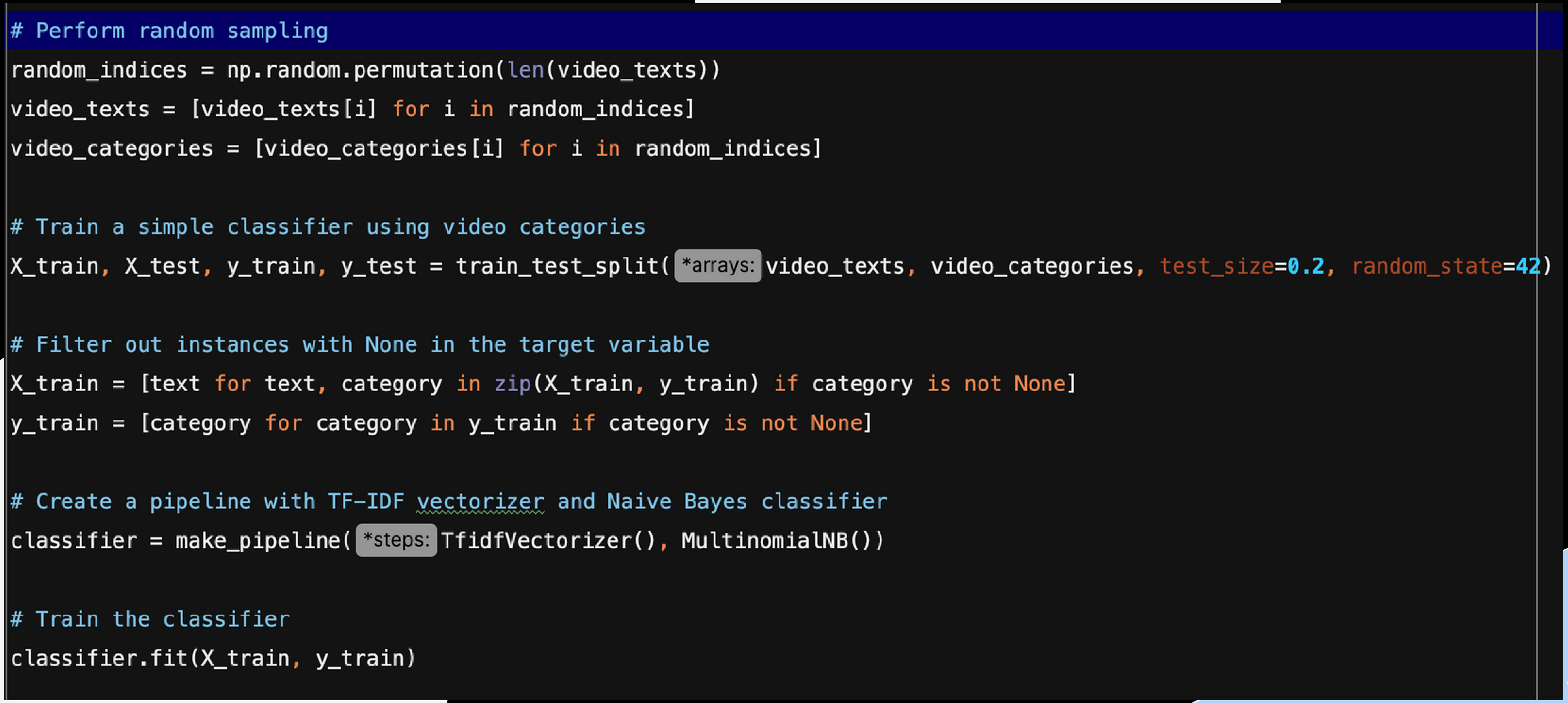
1. Set up API Key
2. Fetch Youtube categories
3. Load Youtube Watch History
4. Extract Relevant Information
5. Random Sampling
6. Train Classifier
7. Update Category Mapping
8. Predictions
9. Output



```python
# Update the category_mapping dictionary based on the actual category IDs
category_mapping = {
    '1': 'Film & Animation',
    '2': 'Autos & Vehicles',
    '10': 'Music',
    '15': 'Pets & Animals',
    '17': 'Sports',
    '18': 'Short Movies',
    '19': 'Travel & Events',
    '20': 'Gaming',
    '21': 'Videoblogging',
    '22': 'People & Blogs',
    '23': 'Comedy',
    '24': 'Entertainment',
    '25': 'News & Politics',
    '26': 'Howto & Style',
    '27': 'Education',
    '28': 'Science & Technology',
    '29': 'Nonprofits & Activism',
    '30': 'Movies',
    '31': 'Anime/Animation',
    '32': 'Action/Adventure',
    '33': 'Classics',
    '34': 'Comedy'
```

Implementation

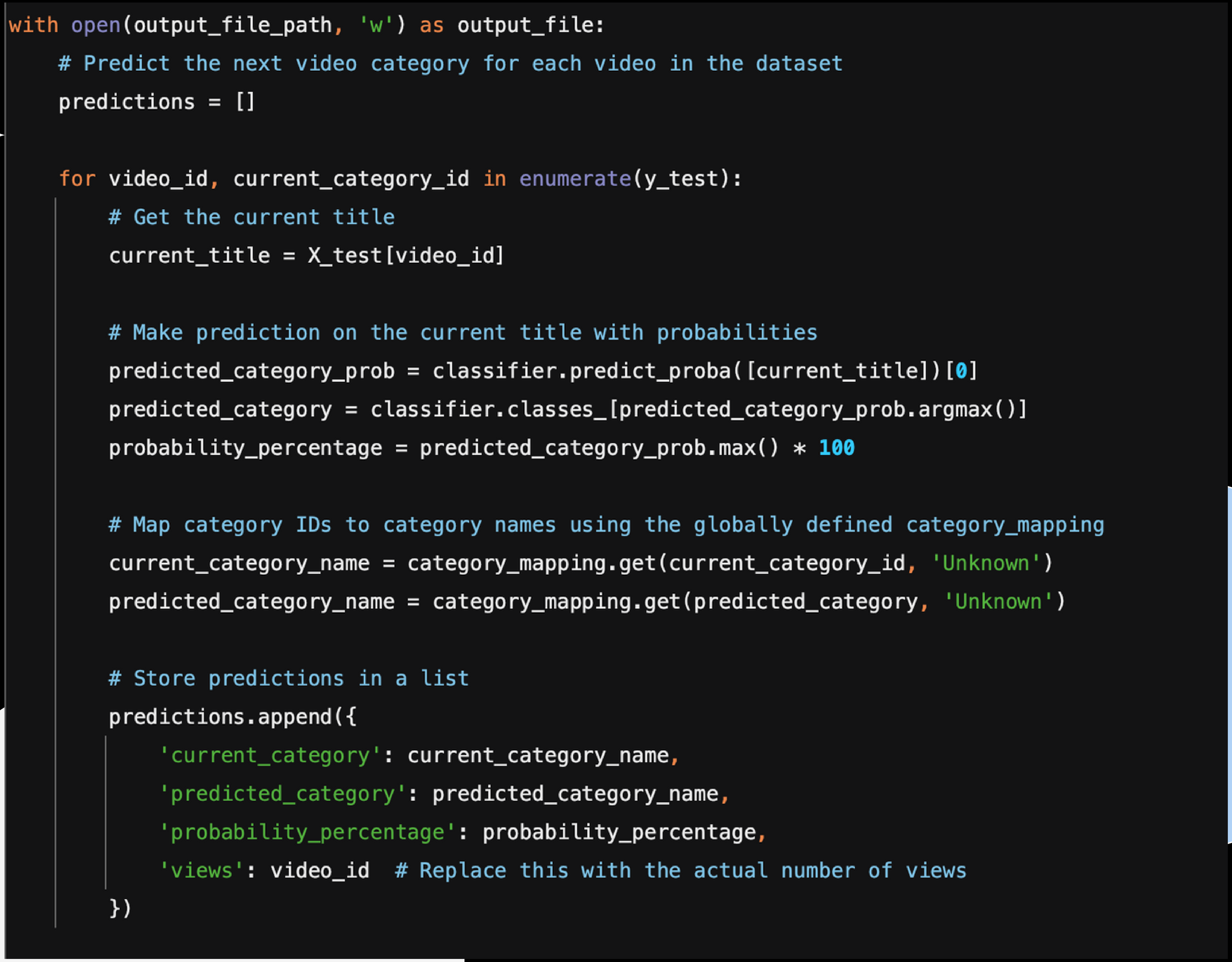
1. Set up API Key
2. Fetch Youtube categories
3. Load Youtube Watch History
4. Extract Relevant Information
5. Random Sampling
6. Train Classifier
7. Update Category Mapping
8. Predictions
9. Output



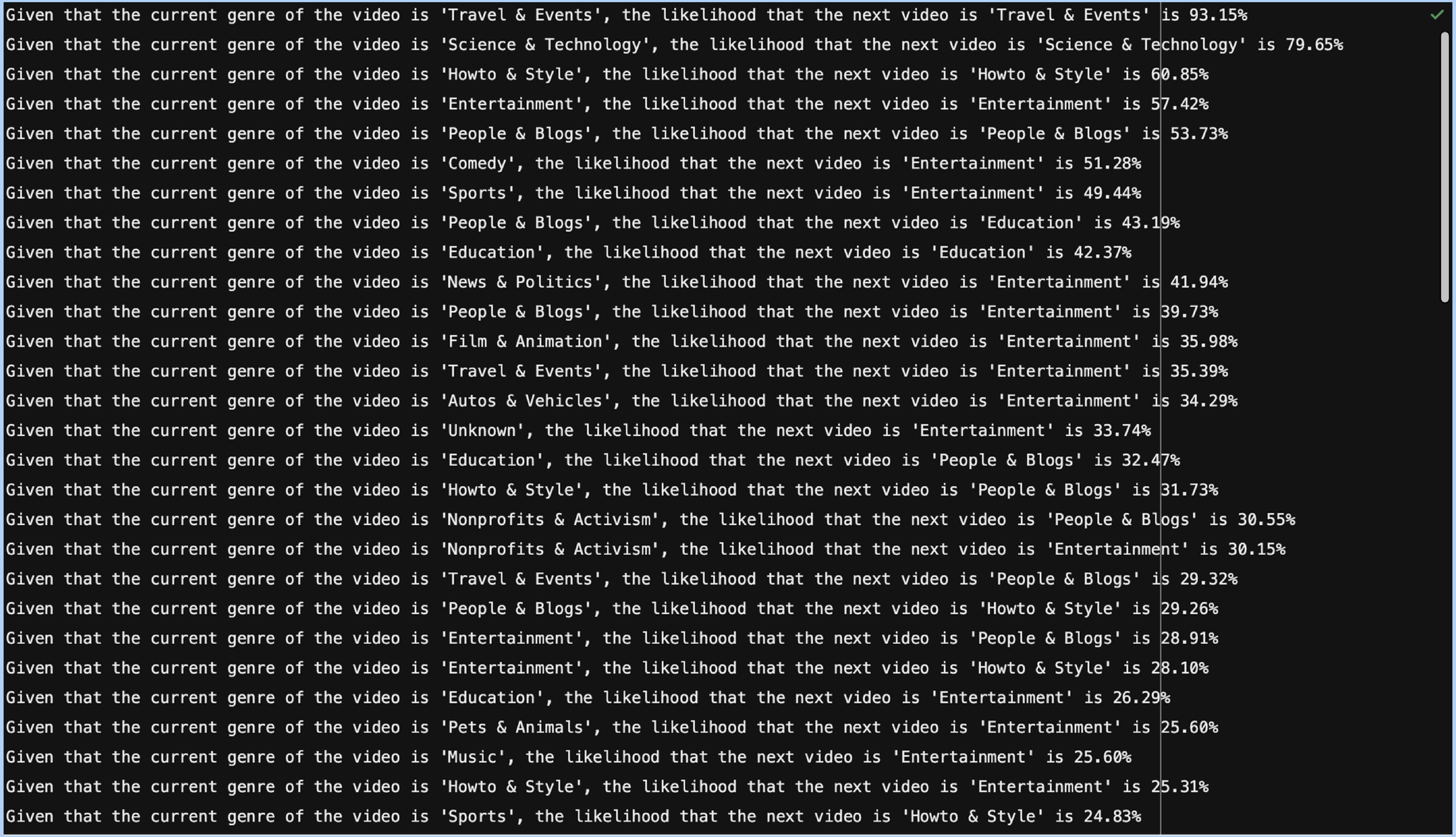
```
# Perform random sampling
random_indices = np.random.permutation(len(video_texts))
video_texts = [video_texts[i] for i in random_indices]
video_categories = [video_categories[i] for i in random_indices]

# Train a simple classifier using video categories
X_train, X_test, y_train, y_test = train_test_split( *arrays: video_texts, video_categories, test_size=0.2, random_state=42)

# Filter out instances with None in the target variable
X_train = [text for text, category in zip(X_train, y_train) if category is not None]
y_train = [category for category in y_train if category is not None]

# Create a pipeline with TF-IDF vectorizer and Naive Bayes classifier
classifier = make_pipeline( *steps: TfidfVectorizer(), MultinomialNB())

# Train the classifier
classifier.fit(X_train, y_train)
```

Implementation

1. Set up API Key
2. Fetch Youtube categories
3. Load Youtube Watch History
4. Extract Relevant Information
5. Random Sampling
6. Train Classifier
7. Update Category Mapping
8. Predictions
9. Output



```
with open(output_file_path, 'w') as output_file:
    # Predict the next video category for each video in the dataset
    predictions = []

    for video_id, current_category_id in enumerate(y_test):
        # Get the current title
        current_title = X_test[video_id]

        # Make prediction on the current title with probabilities
        predicted_category_prob = classifier.predict_proba([current_title])[0]
        predicted_category = classifier.classes_[predicted_category_prob.argmax()]
        probability_percentage = predicted_category_prob.max() * 100

        # Map category IDs to category names using the globally defined category_mapping
        current_category_name = category_mapping.get(current_category_id, 'Unknown')
        predicted_category_name = category_mapping.get(predicted_category, 'Unknown')

        # Store predictions in a list
        predictions.append({
            'current_category': current_category_name,
            'predicted_category': predicted_category_name,
            'probability_percentage': probability_percentage,
            'views': video_id  # Replace this with the actual number of views
        })
```



```
Given that the current genre of the video is 'Travel & Events', the likelihood that the next video is 'Travel & Events' is 93.15%
Given that the current genre of the video is 'Science & Technology', the likelihood that the next video is 'Science & Technology' is 79.65%
Given that the current genre of the video is 'Howto & Style', the likelihood that the next video is 'Howto & Style' is 60.85%
Given that the current genre of the video is 'Entertainment', the likelihood that the next video is 'Entertainment' is 57.42%
Given that the current genre of the video is 'People & Blogs', the likelihood that the next video is 'People & Blogs' is 53.73%
Given that the current genre of the video is 'Comedy', the likelihood that the next video is 'Entertainment' is 51.28%
Given that the current genre of the video is 'Sports', the likelihood that the next video is 'Entertainment' is 49.44%
Given that the current genre of the video is 'People & Blogs', the likelihood that the next video is 'Education' is 43.19%
Given that the current genre of the video is 'Education', the likelihood that the next video is 'Education' is 42.37%
Given that the current genre of the video is 'News & Politics', the likelihood that the next video is 'Entertainment' is 41.94%
Given that the current genre of the video is 'People & Blogs', the likelihood that the next video is 'Entertainment' is 39.73%
Given that the current genre of the video is 'Film & Animation', the likelihood that the next video is 'Entertainment' is 35.98%
Given that the current genre of the video is 'Travel & Events', the likelihood that the next video is 'Entertainment' is 35.39%
Given that the current genre of the video is 'Autos & Vehicles', the likelihood that the next video is 'Entertainment' is 34.29%
Given that the current genre of the video is 'Unknown', the likelihood that the next video is 'Entertainment' is 33.74%
Given that the current genre of the video is 'Education', the likelihood that the next video is 'People & Blogs' is 32.47%
Given that the current genre of the video is 'Howto & Style', the likelihood that the next video is 'People & Blogs' is 31.73%
Given that the current genre of the video is 'Nonprofits & Activism', the likelihood that the next video is 'People & Blogs' is 30.55%
Given that the current genre of the video is 'Nonprofits & Activism', the likelihood that the next video is 'Entertainment' is 30.15%
Given that the current genre of the video is 'Travel & Events', the likelihood that the next video is 'People & Blogs' is 29.32%
Given that the current genre of the video is 'People & Blogs', the likelihood that the next video is 'Howto & Style' is 29.26%
Given that the current genre of the video is 'Entertainment', the likelihood that the next video is 'People & Blogs' is 28.91%
Given that the current genre of the video is 'Entertainment', the likelihood that the next video is 'Howto & Style' is 28.10%
Given that the current genre of the video is 'Education', the likelihood that the next video is 'Entertainment' is 26.29%
Given that the current genre of the video is 'Pets & Animals', the likelihood that the next video is 'Entertainment' is 25.60%
Given that the current genre of the video is 'Music', the likelihood that the next video is 'Entertainment' is 25.60%
Given that the current genre of the video is 'Howto & Style', the likelihood that the next video is 'Entertainment' is 25.31%
Given that the current genre of the video is 'Sports', the likelihood that the next video is 'Howto & Style' is 24.83%
```

What I Found

Search

- Most-Watched Category: Entertainment
- Least-Watched Category: Auto & Vehicles
- Trends
  - Same Category -> Same Category
  - News & Politics -> Education

Save

Cancel

Personal Insights

- Auto & Vehicles?
- Entertainment?
- Same Category is Watched

Areas of Struggle and Improvement

Struggles

- Youtube API
- Size of data
- Category Mapping
- Validating Data

Improvement

- Better prediction
- Refined categories
- Refined features
- Efficiency
- More insights

Thank you!

:)