Preprocessing Data

```
import pandas as pd
import re
```

```
df = pd.read_csv('/content/train.csv')
```

```
df.head(5)
```

	id	movie_name	synopsis	genre
0	44978	Super Me	A young scriptwriter starts bringing valuable ...	fantasy
1	50185	Entity Project	A director and her friends renting a haunted h...	horror
2	34131	Behavioral Family Therapy for Serious Psychiat...	This is an educational video for families and ...	family
3	78522	Blood Glacier	Scientists working in the Austrian Alps discov...	scifi
4	2206	Apat na anino	Buy Day - Four Men Widely - Apart in Life - By...	action

Next steps: Generate code with df | View recommended plots | New interactive sheet

Removing inconsistence in synopsis

```
string_to_remove = "... See full synopsis¬†¬M"
df['synopsis'] = df['synopsis'].str.replace(string_to_remove, '').str.strip()
```

Removing noisy data

```
english_alphabet_pattern = re.compile(r'^[a-zA-Z\s]*$')
df['movie_name'] = df['movie_name'].apply( \
    lambda x: x if re.match(english_alphabet_pattern, x) else '')
df = df[df['movie_name'] != '']
```

```
df.head()
```

	id	movie_name	synopsis	genre
0	44978	Super Me	A young scriptwriter starts bringing valuable ...	fantasy
1	50185	Entity Project	A director and her friends renting a haunted h...	horror
2	34131	Behavioral Family Therapy for Serious Psychiat...	This is an educational video for families and ...	family
3	78522	Blood Glacier	Scientists working in the Austrian Alps discov...	scifi
4	2206	Apat na anino	Buy Day - Four Men Widely - Apart in Life - By...	action

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.to_csv('/content/clean_train.csv')
```

Tokenizing

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.preprocessing.text import Tokenizer
```

```
tokenizer = Tokenizer(num_words=10000, oov_token='<OOV>')
```

```
tokenizer.fit_on_texts(df['synopsis'])
```

```
word_index = tokenizer.word_index
```

```
print(word_index)
```

```
{'<OOV>': 1, 'a': 2, 'the': 3, 'to': 4, 'of': 5, 'and': 6, 'in': 7, 'his': 8, 'is': 9, 'an': 10, 'her': 11, 'with': 12, 'on': 13, 't
```

```
sequences = tokenizer.texts_to_sequences(df['synopsis'])
```

```
sequences
```

```
  122,
  292,
  10,
  672,
  12,
  2,
  174,
  1,
  1540,
  8,
  87,
  4,
  728,
  13,
  2,
  5084,
  243,
  3,
  75,
  588,
  231],
 [7,
  10,
  1393,
  1168,
  1,
  662,
  1466,
  541,
  3362,
  4943,
  1,
  332,
  23,
  3,
  1473,
  4329,
  1,
  445,
  5,
  3,
  4476,
  1,
  1549,
  5,
  1,
  1798,
  3,
  1124,
  4,
  2279,
  60,
  5,
  1473,
  3,
  97,
  160],
 ...]
```