In [15]:
```
import pandas as pd
```

In [16]:
```
num_samples_en = 84557
lines_en = pd.read_csv('trainen.txt',encoding='utf8', sep='delimiter', names
lines_en = lines_en[0:num_samples_en]
input_texts_en=len(lines_en)
print(input_texts_en)
```

```
C:\Users\sneha\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: ParserWarning: Falling back to t
ngine does not support regex separators (separators > 1 char and different from '\s+' are interprete
ing by specifying engine='python'.
```

```
84557
```

In [17]:
```
lines_en.head()
```

	eng
0	And what is their Sigil?
1	I do not want to die.
2	It's the same country I think.
3	Then they'll be crying like babies.
4	- No, I need power up!

In [18]:
```
lines_en.shape
```

```
(84557, 1)
```

In [19]:
```
num_samples_hi = 84557
lines_hi = pd.read_csv('trainhi.txt',encoding='utf8', sep='delimiter', names
lines_hi = lines_hi[0:num_samples_hi]
input_texts_hi=len(lines_hi)
print(input_texts_hi)
```

```
C:\Users\sneha\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: ParserWarning: Falling back to t
ngine does not support regex separators (separators > 1 char and different from '\s+' are interprete
ing by specifying engine='python'.
```

```
84557
```

In [20]:
```
lines_hi.head()
```

	hin
0	और उनके Sigil क्या है?
1	मैं मरना नहीं चाहता.
2	यह मुझे लगता है कि एक ही देश है.
3	फिर ये नन्हें बच्चों की तरह रोएँगे।
4	नहीं, मुझे पावर की जरुरत है !

In [21]:
```
lines_hi.shape
```

```
(84557, 1)
```

In [22]:
```
data = pd.merge(lines_en, lines_hi, left_index=True, right_index=True)
data.shape
```

```
(84557, 2)
```

In [23]:
```
data.head()
```

	eng	hin
0	And what is their Sigil?	और उनके Sigil क्या है?
1	I do not want to die.	मैं मरना नहीं चाहता.
2	It's the same country I think.	यह मुझे लगता है कि एक ही देश है.
3	Then they'll be crying like babies.	फिर ये नन्हें बच्चों की तरह रोएँगे।
4	- No, I need power up!	नहीं, मुझे पावर की जरुरत है !

In [24]:
```
df = data[~data['hin'].str.contains("[a-zA-Z]").fillna(False)]
df.shape
```

```
(78475, 2)
```

In [25]:
```
df.head()
```

	eng	hin
1	I do not want to die.	मैं मरना नहीं चाहता.
2	It's the same country I think.	यह मुझे लगता है कि एक ही देश है.
3	Then they'll be crying like babies.	फिर ये नन्हें बच्चों की तरह रोएँगे।
4	- No, I need power up!	नहीं, मुझे पावर की जरुरत है !
5	I will not eat him.	मैं उसे नहीं खा जाएगा.

In [26]:
```
df.to_csv('cleaned_data.txt', sep='\t', header=False)
```