```
import os
import random
from google.colab import drive
import json
```

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True)
```

```
files = []
for file in os.listdir('/content/drive/MyDrive/hindi_corpus/'):
  if '.dat' in file:
    files.append(file)
```

```
print("Total number of files used =", len(files))
```

```
Total number of files used = 1058
```

```
tagged_Sentences = []
curr = []
for f in files:
  with open('/content/drive/MyDrive/hindi_corpus/' + f, 'r') as file:
    for i in file.readlines():
      i = i.split("\t")
      if len(i) > 5:
        curr.append((i[1], i[4]))
      else:
        tagged_Sentences.append(curr)
        curr = []
print("Total number of sentences =", len(tagged_Sentences))
```

```
Total number of sentences = 18685
```

```
def creating_bigram_sents(current_sents):
  bigram_sents = []
```

```
  bigram_sents = []
  for idx in range(len(current_sents) - 1):
    bigram_sents.append([(current_sents[idx][0], current_sents[idx][1]), (current_sents[idx + 1][0], current_sents[idx + 1][1])])
  return bigram_sents
```

```
def clean_sents(tagged_sentences):
  final_sents = []
  for sents in tagged_sentences:
    new_sents = []
    for word, tag in sents:
      if tag == "SYM" or tag == '' or tag == 'PUNC':
        continue
      new_sents.append((word, tag.lower()))
    final_sents.extend(creating_bigram_sents(new_sents))
  return final_sents
```

```
tagged_sentences = clean_sents(tagged_Sentences)
random.shuffle(tagged_sentences)
print("Total number of sentences after cleaning =", len(tagged_sentences))
```

```
Total number of sentences after cleaning = 331364
```

```
print(tagged_sentences[0])
```

```
[('आग', 'nn'), ('की', 'psp')]
```

```
def removing_tags(tagged_sentences):
  return [word for word, tag in tagged_sentences]
```

```
removed = [removing_tags(i) for i in tagged_sentences]
```

```
tags = []
for i in tagged_sentences:
    tags.extend([i[0][1], i[1][1]])
```

```
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/cleaned_sentences.txt', 'w+') as file:
```

```
    json.dump(tagged_sentences, file)
```

```
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/sentences_without_tags.txt', 'w+') as file:
    json.dump(removed, file)
```

```
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/tags_original.txt', 'w+') as file:
    json.dump(tags, file)
```