```
import gensim
from gensim.models.doc2vec import Doc2Vec, TaggedDocument

import nltk
from nltk.tokenize import word_tokenize, sent_tokenize

import pandas as pd
import re

import time
from tqdm import tqdm
```

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
# Read the CSV file into a DataFrame
df = pd.read_csv('/content/drive/MyDrive/GPT2_summaries.csv')

# Select the first 1000 paragraphs from the 'content' column
selected_paragraphs = df['summary']
len(selected_paragraphs)
```

```
5925
```

```
# Initialize an empty list to store the last sentences
result = []

for paragraph in selected_paragraphs:
    # Tokenize the paragraph into sentences
    if isinstance(paragraph, float):
        continue

    content = paragraph
    result.append(content)

len(result)
```

```
5925
```

```
# Model Training
tagged_data = [TaggedDocument(words=word_tokenize(_d.lower()), tags=[str(i)]) for i, _d in enumerate(result)]
```

```
vector_size = 30
epochs = 80
min_count = 2
```

```
model = gensim.models.doc2vec.Doc2Vec(dm = 0, vector_size=vector_size, min_count=min_count, epochs=epochs)
model.build_vocab(tagged_data)

start_time = time.time()
# model.train(tagged_data, total_examples=model.corpus_count, epochs=epochs)
model.train(tagged_data, total_examples=model.corpus_count, epochs=model.epochs)
end_time = time.time()
model.save(f"d2v.model")
```

Result

```
model = Doc2Vec.load("d2v.model")

similar_doc = model.dv.most_similar('504')
for tag, similarity in similar_doc:
    print(f"Tag: {tag}, Similarity: {similarity}")
```

```
Tag: 1787, Similarity: 0.6370575428009033
Tag: 4474, Similarity: 0.6319007277488708
Tag: 2280, Similarity: 0.628385603427887
Tag: 5422, Similarity: 0.6281779408454895
Tag: 5205, Similarity: 0.6249402761459351
Tag: 3607, Similarity: 0.6230195760726929
Tag: 84, Similarity: 0.6227793097496033
Tag: 3103, Similarity: 0.6199403405189514
Tag: 4310, Similarity: 0.6193053126335144
```

Tag: 3081, Similarity: 0.6181350350379944