In [1]:
```
import torch
import torch.nn.functional as F
import numpy as np
import pysbd
import os
import re
from pathlib import Path

from sentence_transformers import SentenceTransformer
from tqdm.notebook import tqdm
from concurrent.futures import ThreadPoolExecutor, as_completed
```

```
/home/sd205521/anaconda3/envs/rapids-23.12/lib/python3.10/site-packages/sentence_transformers/cross_encoder/CrossEncoder.py:11: TqdmExperimentalWarning: Using `tqdm.autonotebook.tqdm` in notebook mode. Use `tqdm.tqdm` instead to force console mode (e.g. in jupyter console)
  from tqdm.autonotebook import tqdm, trange
```

In [2]:
```
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
```

Functions

In [3]:
```
def process(model_name, input_query, sentences, n=5, device=device):
    # Create a tqdm progress bar with the total number of steps
    progress_bar = tqdm(total=5, desc="Starting")

    # Load the model and set it to the device
    model = SentenceTransformer(model_name).to(device)
    progress_bar.update(1)
    progress_bar.set_description("Model loaded")

    # Encode the input query
    input_embedding = model.encode(input_query, convert_to_tensor=True).to(device)
    progress_bar.update(1)
    progress_bar.set_description("Input query encoded")

    # Filter sentences and encode them
    filtered_sentences = [sentence for sentence in sentences if len(sentence.split()) > 5]
    embeddings = model.encode(filtered_sentences, convert_to_tensor=True).to(device)
    progress_bar.update(1)
    progress_bar.set_description("Sentences encoded")

    # Compute cosine similarity scores
    scores = F.cosine_similarity(embeddings, input_embedding.unsqueeze(0), dim=1)
    progress_bar.update(1)
    progress_bar.set_description("Cosine similarity calculated")

    # Move scores to CPU for further processing
    scores = scores.cpu()

    # Sort and select top n sentences
    top_sentences = sorted(zip(filtered_sentences, scores), key=lambda x: x[1], reverse=True)[:n
    progress_bar.update(1)
    progress_bar.set_description("Sentences sorted and selected")

    # Print top sentences
    print(f"Model name is: {model_name}.\n")
    print(f"Input query is: {input_query}\n")
    for i, (sentence, score) in enumerate(top_sentences):
```

```
        print(f"Ranking: {i+1} | Score: {score:.4f}\nSentence: {sentence}\n")

    progress_bar.close()
```

In [4]:
```
def split_by_character_count(text, chars_per_chunk):
    total_chars = len(text)
    chunks = []
    start = 0

    while start < total_chars:
        # Set the initial end position
        end = min(start + chars_per_chunk, total_chars)

        # Search backwards for a period
        if end < total_chars:
            while end > start and text[end-1] != '.':
                end -= 1

            # If no period is found in the chunk, extend to the next period
            if end == start:
                while end < total_chars and text[end-1] != '.':
                    end += 1

        chunk = text[start:end].strip()
        chunks.append(chunk)
        start = end

    return chunks
```

In [5]:
```
def process_chunk(chunk):
    seg = pysbd.Segmenter(language="en", clean=False)
    return seg.segment(chunk)
```

In [6]:
```
def clean_sentences(sentences):
    cleaned_sentences = []
    modification_count = 0

    for sentence in tqdm(sentences):
        # Remove leading and trailing spaces
        trimmed_sentence = sentence.strip()
        # Replace multiple spaces with a single space
        cleaned_sentence = re.sub(r'\s+', ' ', trimmed_sentence)

        if sentence != cleaned_sentence:
            modification_count += 1

        cleaned_sentences.append(cleaned_sentence)

    print(f"{modification_count} sentences cleaned.")

    return cleaned_sentences
```

Paths

In [7]:
```
base_dir = Path.cwd()
data_dir = base_dir / 'data'

alice_in_wonderland_path = data_dir / 'alice_in_wonderland.txt'
```

Input Data

In [8]:
```
with open(alice_in_wonderland_path, 'r', encoding='utf-8') as file:
    input_text = file.read()
```

In [9]:
```
input_query = "She wonders about things."
```

Processing Data

In [10]:
```
input_text = input_text.replace("\n", " ").replace("-", " ").replace("_", " ")
word_count = len(input_text.split())
```

In [11]:
```
chars_per_chunk = 10000  # Adjust the number of characters per chunk
chunks = split_by_character_count(input_text, chars_per_chunk)
```

In [12]:
```
if len(input_text) > chars_per_chunk * 2:
    num_threads = os.cpu_count()

    with ThreadPoolExecutor(max_workers=num_threads) as executor:
        # Create futures for processing each chunk
        futures = [executor.submit(process_chunk, chunk) for chunk in chunks]

        # Collect all sentences using list comprehension
        sentences = [sentence for future in tqdm(as_completed(futures), total=len(futures), leav

else:
    sentences = process_chunk(input_text)
```

```
 0%|          | 0/17 [00:00<?, ?it/s]
```

In [13]:
```
sentences = clean_sentences(sentences)
```

```
 0%|          | 0/929 [00:00<?, ?it/s]
913 sentences cleaned.
```

Models

all-mpnet-base-v2

In [14]:
```
model_name = 'sentence-transformers/all-mpnet-base-v2'
```

In [15]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
/home/sd205521/anaconda3/envs/rapids-23.12/lib/python3.10/site-packages/huggingface_hub/file_down
load.py:1132: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.
0. Downloads always resume when possible. If you want to force a new download, use `force_downloa
d=True`.
  warnings.warn(
```

```
Model name is: sentence-transformers/all-mpnet-base-v2.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.5630
Sentence: First, she tried to look down and make out what she was coming to, but it was too dark
to see anything; then she looked at the sides of the well, and noticed that they were filled with
cupboards and book shelves; here and there she saw maps and pictures hung upon pegs.

Ranking: 2 | Score: 0.5617
Sentence: Alice asked in a tone of great curiosity.

Ranking: 3 | Score: 0.5466
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 4 | Score: 0.5302
Sentence: How she longed to get out of that dark hall, and wander about among those beds of brigh
t flowers and those cool fountains, but she could not even get her head through the doorway; “and
even if my head would go through,” thought poor Alice, “it would be of very little use without my
shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only knew how t
o begin.”

Ranking: 5 | Score: 0.5151
Sentence: “Thinking again?” the Duchess asked, with another dig of her sharp little chin.
```

bge-large-en-v1.5

```
In [16]: model_name = 'BAAI/bge-large-en-v1.5'
```

```
In [17]: process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
Model name is: BAAI/bge-large-en-v1.5.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.7519
Sentence: “Does the boots and shoes!” she repeated in a wondering tone.

Ranking: 2 | Score: 0.7257
Sentence: Alice asked in a tone of great curiosity.

Ranking: 3 | Score: 0.6796
Sentence: While she was looking at the place where it had been, it suddenly appeared again.

Ranking: 4 | Score: 0.6774
Sentence: “How can I have done that?” she thought.

Ranking: 5 | Score: 0.6752
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.
```

bge-small-en-v1.5

```
In [18]: model_name = 'BAAI/bge-small-en-v1.5'
```

In [19]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
Model name is: BAAI/bge-small-en-v1.5.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.7542
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 2 | Score: 0.7517
Sentence: “Does the boots and shoes!” she repeated in a wondering tone.

Ranking: 3 | Score: 0.7491
Sentence: Alice asked in a tone of great curiosity.

Ranking: 4 | Score: 0.7149
Sentence: “But perhaps he can’t help it,” she said to herself; “his eyes are so very nearly at th
e top of his head. But at any rate he might answer questions.—How am I to get in?” she repeated,
aloud.

Ranking: 5 | Score: 0.7122
Sentence: “Is that the reason so many tea things are put out here?” she asked.
```

paraphrase-multilingual-MiniLM-L12-v2

In [20]:
```
model_name = 'sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2'
```

In [21]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
Model name is: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.5765
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 2 | Score: 0.5761
Sentence: Alice asked in a tone of great curiosity.

Ranking: 3 | Score: 0.5397
Sentence: “I should like to hear her try and repeat something now.

Ranking: 4 | Score: 0.5268
Sentence: “Thinking again?” the Duchess asked, with another dig of her sharp little chin.

Ranking: 5 | Score: 0.5226
Sentence: Lastly, she pictured to herself how this same little sister of hers would, in the after
time, be herself a grown woman; and how she would keep, through all her riper years, the simple a
nd loving heart of her childhood: and how she would gather about her other little children, and m
ake their eyes bright and eager with many a strange tale, perhaps even with the dream of Wonderla
nd of long ago: and how she would feel with all their simple sorrows, and find a pleasure in all
their simple joys, remembering her own child life, and the happy summer days.
```

all-distilroberta-v1

In [22]:
```
model_name = 'sentence-transformers/all-distilroberta-v1'
```

In [23]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
Model name is: sentence-transformers/all-distilroberta-v1.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.5484
Sentence: Alice asked in a tone of great curiosity.

Ranking: 2 | Score: 0.5468
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 3 | Score: 0.4857
Sentence: How she longed to get out of that dark hall, and wander about among those beds of brigh
t flowers and those cool fountains, but she could not even get her head through the doorway; “and
even if my head would go through,” thought poor Alice, “it would be of very little use without my
shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only knew how t
o begin.”

Ranking: 4 | Score: 0.4744
Sentence: “What can all that green stuff be?” said Alice.

Ranking: 5 | Score: 0.4597
Sentence: “How can you learn lessons in here? Why, there’s hardly room for you , and no room at a
ll for any lesson books!” And so she went on, taking first one side and then the other, and makin
g quite a conversation of it altogether; but after a few minutes she heard a voice outside, and s
topped to listen.
```

paraphrase-distilroberta-base-v1

In [24]:
```
model_name = 'sentence-transformers/paraphrase-distilroberta-base-v1'
```

In [25]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
```

```
Model name is: sentence-transformers/paraphrase-distilroberta-base-v1.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.4377
Sentence: “I’m sure I’m not Ada,” she said, “for her hair goes in such long ringlets, and mine do
esn’t go in ringlets at all; and I’m sure I can’t be Mabel, for I know all sorts of things, and s
he, oh! she knows such a very little! Besides, she’s she, and I’m I, and—oh dear, how puzzling it
all is! I’ll try if I know all the things I used to know. Let me see: four times five is twelve,
and four times six is thirteen, and four times seven is—oh dear! I shall never get to twenty at t
hat rate! However, the Multiplication Table doesn’t signify: let’s try Geography. London is the c
apital of Paris, and Paris is the capital of Rome, and Rome—no, that’s all wrong, I’m certain! I
must have been changed for Mabel! I’ll try and say ‘ How doth the little —’” and she crossed her
hands on her lap as if she were saying lessons, and began to repeat it, but her voice sounded hoa
rse and strange, and the words did not come the same as they used to do:— “How doth the little cr
ocodile Improve his shining tail, And pour the waters of the Nile On every golden scale! “How che
erfully he seems to grin, How neatly spread his claws, And welcome little fishes in With gently s
miling jaws!” “I’m sure those are not the right words,” said poor Alice, and her eyes filled with
tears again as she went on, “I must be Mabel after all, and I shall have to go and live in that p
oky little house, and have next to no toys to play with, and oh! ever so many lessons to learn! N
o, I’ve made up my mind about it; if I’m Mabel, I’ll stay down here! It’ll be no use their puttin
g their heads down and saying ‘Come up again, dear!’ I shall only look up and say ‘Who am I then?
Tell me that first, and then, if I like being that person, I’ll come up: if not, I’ll stay down h
ere till I’m somebody else’—but, oh dear!” cried Alice, with a sudden burst of tears, “I do wish
they would put their heads down! I am so very tired of being all alone here!” As she said this sh
e looked down at her hands, and was surprised to see that she had put on one of the Rabbit’s litt
le white kid gloves while she was talking.

Ranking: 2 | Score: 0.4351
Sentence: “Does the boots and shoes!” she repeated in a wondering tone.

Ranking: 3 | Score: 0.3910
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 4 | Score: 0.3788
Sentence: (she couldn’t guess of what sort it was)

Ranking: 5 | Score: 0.3776
Sentence: “I mean, what makes them so shiny?” Alice looked down at them, and considered a little
before she gave her answer.
```

distiluse-base-multilingual-cased-v2

In [26]:
```
model_name = 'sentence-transformers/distiluse-base-multilingual-cased-v2'
```

In [27]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
```

```
Model name is: sentence-transformers/distiluse-base-multilingual-cased-v2.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.5158
Sentence: Alice asked in a tone of great curiosity.

Ranking: 2 | Score: 0.4163
Sentence: She felt very curious to know what it was all about, and crept a little way out of the
wood to listen.

Ranking: 3 | Score: 0.3528
Sentence: “What a funny watch!” she remarked.

Ranking: 4 | Score: 0.3270
Sentence: “Does the boots and shoes!” she repeated in a wondering tone.

Ranking: 5 | Score: 0.3072
Sentence: “Is that the reason so many tea things are put out here?” she asked.
```

msmarco-distilbert-cos-v5

In [28]:
```
model_name = 'sentence-transformers/msmarco-distilbert-cos-v5'
```

In [29]:
```
process(model_name, input_query, sentences)
```

```
Starting:   0%|          | 0/5 [00:00<?, ?it/s]
Model name is: sentence-transformers/msmarco-distilbert-cos-v5.

Input query is: She wonders about things.

Ranking: 1 | Score: 0.4943
Sentence: Alice said; but was dreadfully puzzled by the whole thing, and longed to change the sub
ject.

Ranking: 2 | Score: 0.4883
Sentence: “I can see you’re trying to invent something!” “I—I’m a little girl,” said Alice, rathe
r doubtfully, as she remembered the number of changes she had gone through that day.

Ranking: 3 | Score: 0.4587
Sentence: “Why, she ,” said the Gryphon.

Ranking: 4 | Score: 0.4543
Sentence: (she couldn’t guess of what sort it was)

Ranking: 5 | Score: 0.4460
Sentence: How she longed to get out of that dark hall, and wander about among those beds of brigh
t flowers and those cool fountains, but she could not even get her head through the doorway; “and
even if my head would go through,” thought poor Alice, “it would be of very little use without my
shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only knew how t
o begin.”
```

In [ ]:
```
```