summary generator spacy nlp

May 11, 2023

[3]:
```
import spacy
from spacy.lang.en import English
from sklearn.feature_extraction.text import TfidfVectorizer
import numpy as np
```

[4]:
```
nlp = spacy.load("en_core_web_sm")
```

[5]:
```
nlp = English()
```

[6]:
```
nlp.add_pipe('sentencizer')
```

[6]:
```
<spacy.pipeline.sentencizer.Sentencizer at 0x261ecee2ac0>
```

[7]:
```
text_corpus = """
Google celebrated British illustrator and artist Sir John Tenniel's
200th birth anniversary with a doodle on February 28. An acclaimed
Victorian painter, Tenniel is celebrated for his illustrations for
Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-Glass.
Tenniel was born in Bayswater, West London in 1820. At the age of 20, Tenniel
received a major eye injury and eventually, lost sight in his right eye.
From a very early age, Tenniel was appreciated as a humorist and soon after,
also cultured his talent for scholarly caricature.
His first illustration was for Samuel Carter Hall's The Book of British
Ballads in 1842. Eight years later, he joined the historic weekly magazine
Punch as a political cartoonist. Lewis Carroll noticed Tenniel's distinct style
of work and in 1864, approached the artist to illustrate his book, Alice's
Adventures in Wonderland. This association marked Carroll and Tenniel's␣
 ↪creative
partnership and continued with Through the Looking Glass in 1872. "The result:
a series of classic characters, such as Alice and the Cheshire Cat, as depicted
in the Doodle artwork's rendition of their iconic meeting-characters who, along
with many others, remain beloved by readers of all ages to this day," the␣
 ↪Google
Doodle page says. After working with Lewis Carroll, Tenniel resumed his work␣
 ↪with
Punch. For his work, Tenniel also received a knighthood in 1893.
Sir John Tenniel died on February 25, 1914. He was 93.
"""
```

[8]:
```
doc = nlp(text_corpus.replace("\n", ""))
```

[9]:
```
sentences = [sent.text.strip() for sent in doc.sents]
```

[10]:
```
doc.sents
```

[10]:
```
<generator at 0x261ea892900>
```

[11]:
```
sentences
```

[11]:
```
["Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.",
 "An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
 'Tenniel was born in Bayswater, West London in 1820.',
 'At the age of 20, Tenniel received a major eye injury and eventually, lost
sight in his right eye.',
 'From a very early age, Tenniel was appreciated as a humorist and soon after,
also cultured his talent for scholarly caricature.',
 "His first illustration was for Samuel Carter Hall's The Book of British
Ballads in 1842.",
 'Eight years later, he joined the historic weekly magazine Punch as a political
cartoonist.',
 "Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
the artist to illustrate his book, Alice's Adventures in Wonderland.",
 'This association marked Carroll and Tenniel\'s creative partnership and
continued with Through the Looking Glass in 1872. "',
 'The result: a series of classic characters, such as Alice and the Cheshire
Cat, as depicted in the Doodle artwork\'s rendition of their iconic meeting-
characters who, along with many others, remain beloved by readers of all ages to
this day," the Google Doodle page says.',
 'After working with Lewis Carroll, Tenniel resumed his work with Punch.',
 'For his work, Tenniel also received a knighthood in 1893.Sir John Tenniel died
on February 25, 1914.',
 'He was 93.']
```

[13]:
```
sesis = [ ses.text.strip() for ses in doc.sents]
```

[14]:
```
sesis
```

[14]:
```
["Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.",
 "An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
 'Tenniel was born in Bayswater, West London in 1820.',
 'At the age of 20, Tenniel received a major eye injury and eventually, lost
```

```
sight in his right eye.',
 'From a very early age, Tenniel was appreciated as a humorist and soon after,
also cultured his talent for scholarly caricature.',
 "His first illustration was for Samuel Carter Hall's The Book of British
Ballads in 1842.",
 'Eight years later, he joined the historic weekly magazine Punch as a political
cartoonist.',
 "Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
the artist to illustrate his book, Alice's Adventures in Wonderland.",
 'This association marked Carroll and Tenniel\'s creative partnership and
continued with Through the Looking Glass in 1872. "',
 'The result: a series of classic characters, such as Alice and the Cheshire
Cat, as depicted in the Doodle artwork\'s rendition of their iconic meeting-
characters who, along with many others, remain beloved by readers of all ages to
this day," the Google Doodle page says.',
 'After working with Lewis Carroll, Tenniel resumed his work with Punch.',
 'For his work, Tenniel also received a knighthood in 1893.Sir John Tenniel died
on February 25, 1914.',
 'He was 93.']
```

1 Creating sentence organizer

[15]:
```
# Let's create an organizer which will store the sentence ordering to later␣
↪reorganize the
# scored sentences in their correct order
sentence_organizer = {k:v for v,k in enumerate(sentences)}
```

[16]:
```
sentence_organizer
```

[16]:
```
{"Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.": 0,
 "An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.": 1,
 'Tenniel was born in Bayswater, West London in 1820.': 2,
 'At the age of 20, Tenniel received a major eye injury and eventually, lost
sight in his right eye.': 3,
 'From a very early age, Tenniel was appreciated as a humorist and soon after,
also cultured his talent for scholarly caricature.': 4,
 "His first illustration was for Samuel Carter Hall's The Book of British
Ballads in 1842.": 5,
 'Eight years later, he joined the historic weekly magazine Punch as a political
cartoonist.': 6,
 "Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
the artist to illustrate his book, Alice's Adventures in Wonderland.": 7,
 'This association marked Carroll and Tenniel\'s creative partnership and
```

```
continued with Through the Looking Glass in 1872. "': 8,
 'The result: a series of classic characters, such as Alice and the Cheshire
Cat, as depicted in the Doodle artwork\'s rendition of their iconic meeting-
characters who, along with many others, remain beloved by readers of all ages to
this day," the Google Doodle page says.': 9,
 'After working with Lewis Carroll, Tenniel resumed his work with Punch.': 10,
 'For his work, Tenniel also received a knighthood in 1893.Sir John Tenniel died
on February 25, 1914.': 11,
 'He was 93.': 12}
```

2 Creating TF-IDF model

[17]:
```
# Let's now create a tf-idf (Term frequnecy Inverse Document Frequency) model
tf_idf_vectorizer = TfidfVectorizer(min_df=2,  max_features=None,
                                    strip_accents='unicode',
                                    analyzer='word',
                                    token_pattern=r'\w{1,}',
                                    ngram_range=(1, 3),
                                    use_idf=1,smooth_idf=1,
                                    sublinear_tf=1,
                                    stop_words = 'english')
```

[18]:
```
# Passing our sentences treating each as one document to TF-IDF vectorizer
tf_idf_vectorizer.fit(sentences)
```

[18]:
```
TfidfVectorizer(min_df=2, ngram_range=(1, 3), smooth_idf=1,
                stop_words='english', strip_accents='unicode', sublinear_tf=1,
                token_pattern='\\w{1,}', use_idf=1)
```

[19]:
```
# Transforming our sentences to TF-IDF vectors
sentence_vectors = tf_idf_vectorizer.transform(sentences)
```

[23]:
```
sentence_vectors
```

[23]:
```
<13x35 sparse matrix of type '<class 'numpy.float64'>'
        with 88 stored elements in Compressed Sparse Row format>
```

[21]:
```
# Getting sentence scores for each sentences
sentence_scores = np.array(sentence_vectors.sum(axis=1)).ravel()
```

[22]:
```
sentence_scores
```

[22]:
```
array([3.69831017, 4.08663958, 1.        , 1.94791376, 1.35058657,
       1.70557256, 1.        , 4.08573981, 2.77457248, 1.88083155,
       2.60622417, 3.15900587, 0.        ])
```

[24]:
```
# Sanity checkup
print(len(sentences) == len(sentence_scores))
```

```
True
```

[25]:
```
# Getting top-n sentences
N=3
top_sentences = [sentences[i] for i in np.argsort(sentence_scores, axis=0)][::
 ↪-1][:3]
```

[26]:
```
top_sentences
```

[26]:
```
["An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
 "Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
the artist to illustrate his book, Alice's Adventures in Wonderland.",
 "Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28."]
```

3 Performing final summarization

[27]:
```
# Let's now do the sentence ordering using our prebaked sentence_organizer
# Let's map the scored sentences with their indexes
mapped_top_sentences =[(sentence,sentence_organizer[sentence]) for sentence in␣
 ↪top_sentences]
```

[28]:
```
mapped_top_sentences
```

[28]:
```
[("An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
  1),
 ("Lewis Carroll noticed Tenniel's distinct style of work and in 1864,
approached the artist to illustrate his book, Alice's Adventures in
Wonderland.",
  7),
 ("Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.",
  0)]
```

[29]:
```
for element in mapped_top_sentences:
    print(element)
```

```
("An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.", 1)
("Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
```

```
the artist to illustrate his book, Alice's Adventures in Wonderland.", 7)
("Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.", 0)
```

[30]:
```
# Ordering our top-n sentences in their original ordering
mapped_top_sentences = sorted (mapped_top_sentences, key = lambda x: x[1])
```

[31]:
```
mapped_top_sentences
```

[31]:
```
[("Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.",
  0),
 ("An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
  1),
 ("Lewis Carroll noticed Tenniel's distinct style of work and in 1864,
approached the artist to illustrate his book, Alice's Adventures in
Wonderland.",
  7)]
```

[33]:
```
ordered_scored_sentences = [element[0] for element in mapped_top_sentences]
```

[34]:
```
ordered_scored_sentences
```

[34]:
```
["Google celebrated British illustrator and artist Sir John Tenniel's 200th
birth anniversary with a doodle on February 28.",
 "An acclaimed Victorian painter, Tenniel is celebrated for his illustrations
for Lewis Carroll's Alice's Adventures in Wonderland and Through the Looking-
Glass.",
 "Lewis Carroll noticed Tenniel's distinct style of work and in 1864, approached
the artist to illustrate his book, Alice's Adventures in Wonderland."]
```

[35]:
```
# Our final summary
summary = " ".join(ordered_scored_sentences)
```

[36]:
```
summary
```

[36]:
```
"Google celebrated British illustrator and artist Sir John Tenniel's 200th birth
anniversary with a doodle on February 28. An acclaimed Victorian painter,
Tenniel is celebrated for his illustrations for Lewis Carroll's Alice's
Adventures in Wonderland and Through the Looking-Glass. Lewis Carroll noticed
Tenniel's distinct style of work and in 1864, approached the artist to
illustrate his book, Alice's Adventures in Wonderland."
```

[37]:
```
print("Summary: \n", summary)
```

```
Summary:
 Google celebrated British illustrator and artist Sir John Tenniel's 200th birth
```

```
anniversary with a doodle on February 28. An acclaimed Victorian painter,
Tenniel is celebrated for his illustrations for Lewis Carroll's Alice's
Adventures in Wonderland and Through the Looking-Glass. Lewis Carroll noticed
Tenniel's distinct style of work and in 1864, approached the artist to
illustrate his book, Alice's Adventures in Wonderland.
```

4 Creating a function template using above steps:

[38]:
```
def summarizer(text, tokenizer, max_sent_in_summary=3):
    # Create spacy document for further sentence level tokenization
    doc = nlp(text_corpus.replace("\n", ""))
    sentences = [sent.string.strip() for sent in doc.sents]
    # Let's create an organizer which will store the sentence ordering to later␣
 ↪reorganize the
    # scored sentences in their correct order
    sentence_organizer = {k:v for v,k in enumerate(sentences)}
    # Let's now create a tf-idf (Term frequnecy Inverse Document Frequency)␣
 ↪model
    tf_idf_vectorizer = TfidfVectorizer(min_df=2,  max_features=None,
                                        strip_accents='unicode',
                                        analyzer='word',
                                        token_pattern=r'\w{1,}',
                                        ngram_range=(1, 3),
                                        use_idf=1,smooth_idf=1,
                                        sublinear_tf=1,
                                        stop_words = 'english')
    # Passing our sentences treating each as one document to TF-IDF vectorizer
    tf_idf_vectorizer.fit(sentences)
    # Transforming our sentences to TF-IDF vectors
    sentence_vectors = tf_idf_vectorizer.transform(sentences)
    # Getting sentence scores for each sentences
    sentence_scores = np.array(sentence_vectors.sum(axis=1)).ravel()
    # Getting top-n sentences
    N = max_sent_in_summary
    top_n_sentences = [sentences[ind] for ind in np.argsort(sentence_scores,␣
 ↪axis=0)[::-1][:N]]
    # Let's now do the sentence ordering using our prebaked sentence_organizer
    # Let's map the scored sentences with their indexes
    mapped_top_n_sentences = [(sentence,sentence_organizer[sentence]) for␣
 ↪sentence in top_n_sentences]
    # Ordering our top-n sentences in their original ordering
    mapped_top_n_sentences = sorted(mapped_top_n_sentences, key = lambda x:␣
 ↪x[1])
    ordered_scored_sentences = [element[0] for element in␣
 ↪mapped_top_n_sentences]
    # Our final summary
```

```
    summary = " ".join(ordered_scored_sentences)
    return summary
```

5

[ ]: