Tarea 12 - NLP

April 28, 2021

0.1 # Tarea 12

0.2 ## NLP

Por: Miguel Angel Soto Hernandez
https://pypi.org/project/SNgramExtractor/

[ ]:
```
!pip install SNgramExtractor
```

[ ]:
```
!pip install spacy
```

[28]:
```
import spacy
from SNgramExtractor import SNgramExtractor
import re
from bs4 import BeautifulSoup
import urllib.request as urllib2
import requests
import deplacy
```

[31]:
```
def obtener_texto(tema):
  response = requests.get(f'https://en.wikipedia.org/wiki/{tema}')
  soup = BeautifulSoup(response.text)
  parrafos = soup.find_all('p')
  texto = parrafos[0].text
  return texto


def texto_a_frases(texto):
  documento_stanza = nlp(texto)
  frases = [frase.text for frase in documento_stanza.sentences]
  return frases
```

[32]:
```
texto = obtener_texto('Natural language processing')
texto
```

[32]:
```
'Natural language processing (NLP) is a subfield of linguistics, computer
science, and artificial intelligence concerned with the interactions between
computers and human language, in particular how to program computers to process
and analyze large amounts of natural language data.  The result is a computer
capable of "understanding" the contents of documents, including the contextual
nuances of the language within them. The technology can then accurately extract
```

```
information and insights contained in the documents as well as categorize and
organize the documents themselves. \n'
```

[35]:
```
frases = texto_a_frases(texto)
frases[0]
```

[35]:
```
'Natural language processing (NLP) is a subfield of linguistics, computer
science, and artificial intelligence concerned with the interactions between
computers and human language, in particular how to program computers to process
and analyze large amounts of natural language data.'
```

[44]:
```
for frase in frases:
  text = frase
  SNgram_obj = SNgramExtractor(text, meta_tag='original', trigram_flag='yes',
                               nlp_model=None)
  output = SNgram_obj.get_SNgram()
  print(f'Frase: {text}')
  print('SNGram bigram:', output['SNBigram'])
  print('SNGram trigram:', output['SNTrigram'])
  print('-' * 32)
```

```
Frase: Natural language processing (NLP) is a subfield of linguistics, computer
science, and artificial intelligence concerned with the interactions between
computers and human language, in particular how to program computers to process
and analyze large amounts of natural language data.
SNGram bigram: language_Natural processing_language is_processing processing_(
processing_NLP processing_) subfield_a is_subfield subfield_of of_linguistics
linguistics_, science_computer linguistics_science science_, science_and
intelligence_artificial science_intelligence intelligence_concerned
concerned_with interactions_the with_interactions interactions_between
between_computers computers_and language_human computers_language is_,
program_in in_particular program_how program_to is_program program_computers
process_to program_process process_and process_analyze amounts_large
analyze_amounts amounts_of language_natural data_language of_data is_.
SNGram trigram: subfield_of_linguistics of_linguistics_science
linguistics_science_intelligence science_intelligence_concerned
intelligence_concerned_with concerned_with_interactions
interactions_between_computers with_interactions_between
between_computers_language computers_language_human program_process_analyze
process_analyze_amounts amounts_of_data analyze_amounts_of data_language_natural
of_data_language
--------------------------------
Frase: The result is a computer capable of "understanding" the contents of
documents, including the contextual nuances of the language within them.
SNGram bigram: result_The is_result computer_a is_computer computer_capable
capable_of understanding_" of_understanding understanding_" contents_the
understanding_contents contents_of of_documents contents_, contents_including
nuances_the nuances_contextual including_nuances nuances_of language_the
of_language language_within within_them is_.
```

```
SNGram trigram: computer_capable_of capable_of_understanding
understanding_contents_including of_understanding_contents
contents_including_nuances nuances_of_language including_nuances_of
language_within_them of_language_within
--------------------------------
Frase: The technology can then accurately extract information and insights
contained in the documents as well as categorize and organize the documents
themselves.
SNGram bigram: technology_The extract_technology extract_can extract_then
extract_accurately extract_information information_and information_insights
information_contained contained_in documents_the in_documents as_as as_well
information_as information_categorize extract_and extract_organize documents_the
organize_documents organize_themselves extract_.
SNGram trigram: contained_in_documents in_documents_the
--------------------------------
```

[ ]: