Tarea 5

NLP

Descripción: para un texto en español, realizar tokenización, lematización y POS tagging con Stanza.

Alumno: Miguel Angel Soto Hernandez

Stanza

```
!pip install stanza
```

```
Requirement already satisfied: stanza in /usr/local/lib/python3.7/dist-packages (1.2)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from stanza) (2.23.0)
Requirement already satisfied: protobuf in /usr/local/lib/python3.7/dist-packages (from stanza) (3.12.4)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from stanza) (1.19.5)
Requirement already satisfied: torch>=1.3.0 in /usr/local/lib/python3.7/dist-packages (from stanza) (1.8.0+cu101)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from stanza) (4.41.1)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->stanza) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->stanza) (20
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->stanza) (3.0
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: six>=1.9 in /usr/local/lib/python3.7/dist-packages (from protobuf->stanza) (1.15.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from protobuf->stanza) (54.1.2)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from torch>=1.3.0->stanza)
```

```
%matplotlib inline
import stanza
import spacy
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# librería para realizar peticiones web
import requests

# librería que nos permite extrar texto de un sitio web y convertirlo a string
from bs4 import BeautifulSoup

# descargando la libreía de stanza en español
stanza.download('es')
```

```
Downloading https://raw.githubusercontent.com/stanfordnlp/stanza-resources/master/resources_1.2.0.json: 128kB [00:00, 2
2021-03-20 00:49:32 INFO: Downloading default packages for language: es (Spanish)...
2021-03-20 00:49:33 INFO: File exists: /root/stanza_resources/es/default.zip.
2021-03-20 00:49:40 INFO: Finished downloading models and saved to /root/stanza_resources.
```

```
def buscarTexto(idioma, tema):
  # enlace para acceder al contenido de la Wikipedia
  response = requests.get(f'http://{idioma}.wikipedia.org/wiki/{tema}')

  # transformar el sitio web a texto plano
  soup = BeautifulSoup(response.text)

  # obtener el texto por párrafos
  parrafos = soup.find_all('p')

  # obtener solo el primer parrafo
  texto = parrafos[0].text
  return texto

# mostrar el texto con el que trabajaremos
texto_esp = buscarTexto('es', 'Google')
print(texto_esp)
```

```
Google, LLC es una compañía principal subsidiaria de la multinacional estadounidense Alphabet Inc., cuya especializació
```

```
# definiendo el idioma y las funciones que se aplicarán
stanza_nlp = stanza.Pipeline('es', processors='tokenize, lemma, pos, mwt, ner')

# convirtiendo el texto a un formato reconocido por stanza
texto_stanza = stanza_nlp(texto_esp)
```

```
2021-03-20 00:49:40 INFO: Loading these models for language: es (Spanish):
=======================
| Processor | Package |
-----------------------
| tokenize  | ancora  |
| mwt       | ancora  |
| pos       | ancora  |
| lemma     | ancora  |
| ner       | conll02 |
=======================

2021-03-20 00:49:40 INFO: Use device: cpu
2021-03-20 00:49:40 INFO: Loading: tokenize
2021-03-20 00:49:40 INFO: Loading: mwt
2021-03-20 00:49:40 INFO: Loading: pos
2021-03-20 00:49:40 INFO: Loading: lemma
2021-03-20 00:49:40 INFO: Loading: ner
2021-03-20 00:49:42 INFO: Done loading processors!
```

```
# tokenización, lematización y POS tagging
print("{:12s}\t{:12s}\t{:6s}".format("Palabra", "Lematizacion",
                                     "POS tagging"))

for i, oracion in enumerate(texto_stanza.sentences):
  print("[Oracion {}]".format(i + 1))
  for palabra in oracion.words:
    print("{:12s}\t{:12s}\t{:6s}".format(palabra.text, palabra.lemma,
                                        palabra.pos))
print("")
```

```
Palabra         Lematizacion    POS tagging
[Oracion 1]
Google          Google          PROPN
,               ,               PUNCT
LLC             LLC             PROPN
es              ser             AUX
una             uno             DET
compañía        compañía        NOUN
principal       principal       ADJ
subsidiaria     subsidiario     ADJ
de              de              ADP
la              el              DET
multinacional   multinacional   NOUN
estadounidense  estadounidense  ADJ
Alphabet        Alphabet        PROPN
Inc.            Inc.            PROPN
,               ,               PUNCT
cuya            cuyo            PRON
especialización especialización NOUN
son             ser             AUX
los             el              DET
productos       producto        NOUN
y               y               CCONJ
servicios       servicio        NOUN
relacionados    relacionado     ADJ
con             con             ADP
Internet        Internet        PROPN
,               ,               PUNCT
software        software        NOUN
,               ,               PUNCT
dispositivos    dispositivo     NOUN
electrónicos    electrónico     ADJ
y               y               CCONJ
otras           otro            DET
tecnologías     tecnología      NOUN
.               .               PUNCT

[Oracion 2]
El              el              DET
principal       principal       ADJ
producto        producto        NOUN
de              de              ADP
Google          Google          PROPN
es              ser             AUX
el              el              DET
motor           motor           NOUN
de              de              ADP
búsqueda        búsqueda        NOUN
de              de              ADP
contenido       contenido       NOUN
en              en              ADP
Internet        Internet        PROPN
,               ,               PUNCT
del             del             ADP
mismo           mismo           DET
nombre          nombre          NOUN
,               ,               PUNCT
aunque          aunque          SCONJ
```

```
      ofrece          ofrecer         VERB
```

```
# named entity recognition
print("{:20s}\t{:12s}".format("Entidad", "Tipo"))
print(*["{:20s}\t{:12s}".format(entidad.text, entidad.type)
        for oracion in texto_stanza.sentences
        for entidad in oracion.ents], sep='\n')
```

```
     Entidad                Tipo
     Google                 ORG
     LLC                    ORG
     Alphabet Inc.          ORG
     Internet               MISC
     Google                 ORG
     Internet               MISC
     Google Drive           MISC
     Gmail                  ORG
     Google Maps            ORG
     Google Street View     ORG
     Google Earth           ORG
     YouTube                MISC
     Google Libros          ORG
     Google Noticias        ORG
     Google Chrome          ORG
     Google+                ORG
     Linux                  MISC
     Android                MISC
     Google Glass           MISC
     «Do the Right Thing»   MISC
     «Haz lo correcto»      MISC
```

Nota: Sólo los idiomas con tokens de varias palabras (MWT), como el alemán o el francés, requieren el MWTProcessor; otros idiomas, como el inglés o el chino, no admiten este procesador en el pipeline.

```
# multi word expression
print("{:20s}\t{:12s}".format("Token", "Palabras"))
for token in texto_stanza.sentences[0].tokens:
  print("{:20s}\t{:12s}".format(token.text,
                        (", ".join([word.text for word in token.words]))))
```

```
     Token                  Palabras
     Google                 Google
     ,                      ,
     LLC                    LLC
     es                     es
     una                    una
     compañía               compañía
     principal              principal
     subsidiaria            subsidiaria
     de                     de
     la                     la
     multinacional          multinacional
     estadounidense         estadounidense
     Alphabet               Alphabet
     Inc.                   Inc.
     ,                      ,
     cuya                   cuya
     especialización        especialización
     son                    son
     los                    los
     productos              productos
     y                      y
     servicios              servicios
     relacionados           relacionados
     con                    con
     Internet               Internet
     ,                      ,
     software               software
     ,                      ,
     dispositivos           dispositivos
     electrónicos           electrónicos
     y                      y
     otras                  otras
     tecnologías            tecnologías
     .                      .
```

FreeLing

FreeLing 4.2 - An Open-Source Suite of Language Analyzers

Enjoy the FreeLing!

Write your sentences

Google, LLC es una compañía principal subsidiaria de la multinacional estadounidense Alphabet Inc., cuya especialización son los productos y servicios relacionados con Internet, software, dispositivos electrónicos y otras tecnologías.

Analysis options

- ☑ Number recognition
- ☑ Date/Time recognition
- ☑ Quantities, ratios, and percentages
- ☑ Named Entity Recognition
- ☑ Multiword detection
- ☐ Phonetic encoding

- ◉ No sense annotation
- ○ WN sense annotation: All senses
- ○ WN sense annotation: UKB disambiguation

Select language: Auto-detect

Select output: PoS Tagging

Submit

Language identification

Identified language is: Spanish (es)

Sentences

Sentence 1

Google	,	LLC	es	una	compañía	principal	subsidiaria	de	la	multinacional	estadounidense	Alphabet_Inc.	,	cuya	especialización	son	los	productos	y	servicios	relacionados	con	Internet	,	software	,	dispositivos	electrónicos	y	otras	tecnologías	.
google	,	llc	ser	uno	compañía	principal	subsidiario	de	el	multinacional	estadounidense	alphabet_inc.	,	cuyo	especialización	ser	el	producto	y	servicio	relacionar	con	internet	,	software	,	dispositivo	electrónico	y	otro	tecnología	.
NP00SP0	Fc	NP00O00	VSIP3S0	DI0FS0	NCFS000	AQ0CS00	AQ0FS00	SP	DA0FS0	NCFS000	AQ0CS00	NP00SP0	Fc	PR0FS00	NCFS000	VSIP3P0	DA0MP0	NCMP000	CC	NCMP000	VMP00PM	SP	NP00V00	Fc	NCMN000	Fc	NCMP000	AQ0MP00	CC	DI0FP0	NCFP000	Fp

▼ CONLL format

1	Google	google	NP00SP0	NP	pos=noun|type=proper|neclass=person B-PER - - - - -
2	,	,	Fc	Fc	pos=punctuation|type=comma - - - - - -
3	LLC	llc	NP00O00	NP	pos=noun|type=proper|neclass=organization B-ORG - - - - -
4	es	ser	VSIP3S0	VSI	pos=verb|type=semiauxiliary|mood=indicative|tense=present|person=3|num=singular - - - - - -
5	una	uno	DI0FS0	DI	pos=determiner|type=indefinite|gen=feminine|num=singular - - - - - -
6	compañía	compañía	NCFS000	NC	pos=noun|type=common|gen=feminine|num=singular - - - - - -
7	principal	principal	AQ0CS00	AQ	pos=adjective|type=qualificative|gen=common|num=singular - - - - - -
8	subsidiaria	subsidiario	AQ0FS00	AQ	pos=adjective|type=qualificative|gen=feminine|num=singular - - - - - -
9	de	de	SP	SP	pos=adposition|type=preposition - - - - - -
10	la	el	DA0FS0	DA	pos=determiner|type=article|gen=feminine|num=singular - - - - - -
11	multinacional	multinacional	NCFS000	NC	pos=noun|type=common|gen=feminine|num=singular - - - - - -
12	estadounidense	estadounidense	AQ0CS00	AQ	pos=adjective|type=qualificative|gen=common|num=singular - - - - - -
13	Alphabet_Inc.	alphabet_inc.	NP00SP0	NP	pos=noun|type=proper|neclass=person B-PER - - - - -
14	,	,	Fc	Fc	pos=punctuation|type=comma - - - - - -
15	cuya	cuyo	PR0FS00	PR	pos=pronoun|type=relative|gen=feminine|num=singular - - - - - -
16	especialización	especialización	NCFS000	NC	pos=noun|type=common|gen=feminine|num=singular - - - - - -
17	son	ser	VSIP3P0	VSI	pos=verb|type=semiauxiliary|mood=indicative|tense=present|person=3|num=plural - - - - - -
18	los	el	DA0MP0	DA	pos=determiner|type=article|gen=masculine|num=plural - - - - - -
19	productos	producto	NCMP000	NC	pos=noun|type=common|gen=masculine|num=plural - - - - - -
20	y	y	CC	CC	pos=conjunction|type=coordinating - - - - - -
21	servicios	servicio	NCMP000	NC	pos=noun|type=common|gen=masculine|num=plural - - - - - -
22	relacionados	relacionar	VMP00PM	VMP	pos=verb|type=main|mood=participle|num=plural|gen=masculine - - - - - -
23	con	con	SP	SP	pos=adposition|type=preposition - - - - - -
24	Internet	internet	NP00V00	NP	pos=noun|type=proper|neclass=other B-MISC - - - - -
25	,	,	Fc	Fc	pos=punctuation|type=comma - - - - - -
26	software	software	NCMN000	NC	pos=noun|type=common|gen=masculine|num=invariable - - - - - -
27	,	,	Fc	Fc	pos=punctuation|type=comma - - - - - -
28	dispositivos	dispositivo	NCMP000	NC	pos=noun|type=common|gen=masculine|num=plural - - - - - -
29	electrónicos	electrónico	AQ0MP00	AQ	pos=adjective|type=qualificative|gen=masculine|num=plural - - - - - -
30	y	y	CC	CC	pos=conjunction|type=coordinating - - - - - -
31	otras	otro	DI0FP0	DI	pos=determiner|type=indefinite|gen=feminine|num=plural - - - - - -
32	tecnologías	tecnología	NCFP000	NC	pos=noun|type=common|gen=feminine|num=plural - - - - - -
33	.	.	Fp	Fp	pos=punctuation|type=period - - - - - -