04 - Text Classification Model Testing

August 3, 2020

1 EA Assignment 04 - Text Classification Model Testing

Authored by: Álvaro Bartolomé del Canto (alvarobartt @ GitHub)

We will start this Jupyter Notebook with a little recap from the previous ones named `02 - Data Preprocessing.ipynb` and `03 - Text Classification Model.ipynb` where we defined the NLP Preprocessing pipeline and the text classification model (which was also trained), respectively.

Reproducibility Warning: this Jupyter Notebook requires some resources that are mandatory in order to use it, this resources may be found inside the `research/resources/` directory but if they are not, you will need to run again the mentioned Jupyter Notebooks so as to automatically generate them.

1.1 Loading Resources

1.1.1 NLP CustomPreProcessor

We will start importing the previously defined `CustomPreProcessor` so as to preprocess the unseen data the same way as the training/validation data has been preprocessed.

If you want more details/insights, please refer to `02 - Data Preprocessing.ipynb`.

[1]:
```
from unidecode import unidecode
```

[2]:
```
import re

URL_PATTERN = re.compile(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:
 ↪%[0-9a-fA-F][0-9a-fA-F]))+')
HTML_PATTERN = re.compile(r'<.*?>|&([a-z0-9]+|#[0-9]{1,6}|#x[0-9a-f]{1,6});')
PUNCTUATION_PATTERN = re.compile(r'[^\w\s]')
NUMBER_PATTERN = re.compile(r'[\d]+')
SPACES_PATTERN = re.compile(r'[ ]{2,}')

BASE_PATTERNS = (
    URL_PATTERN, HTML_PATTERN, PUNCTUATION_PATTERN,
    NUMBER_PATTERN, SPACES_PATTERN
)
```

[3]:
```
from nltk.corpus import stopwords

spanish_stopwords = stopwords.words('spanish')
english_stopwords = stopwords.words('english')
french_stopwords = stopwords.words('french')

STOPWORDS = english_stopwords + spanish_stopwords + french_stopwords

ADDITIONAL_STOPWORDS = [
    'much', 'despues', 'first', 'categoria', 'aqui', 'thumb', 'also', 'tres',␣
 ↪'asi',
    'three', 'one', 'still', 'aquella', 'like', 'aquel', 'mas', 'tal', 'tan',␣
 ↪'hacia',
    'went', 'two', 'new', 'even', 'would', 'tras', 'could', 'pues', 'without',␣
 ↪'category',
    'many', 'twoone', 'tambien', 'well', 'solo', 'dos'
]

STOPWORDS += ADDITIONAL_STOPWORDS
STOPWORDS = set(list(STOPWORDS))
```

[4]:
```
class CustomPreProcessor(object):
    """
    Custom PreProcessor

    Preprocesses the introduced raw text to transform it into clean text. This
    preprocessing pipe is regex based.

        >>> from apinlp.nlp.preprocessing import CustomPreProcessor
        >>> preprocessor = CustomPreProcessor()
        >>> print(preprocessor._preprocess("Visit us at https://www.ea.com/"))
        "visit us"
    """

    def __init__(self, strip_accents=True):
        self.strip_accents = strip_accents

        self.patterns = BASE_PATTERNS
        self.additional_patterns = (SPACES_PATTERN,)

        self.stopwords = STOPWORDS

    def _preprocess(self, text):
        """Cleans and applies a preprocessing layer to raw text"""
        text = text.replace('\t', ' ').replace('\n', ' ')

        if self.strip_accents:
```

```
            text = unidecode(text)

        for pattern in self.patterns:
            text = pattern.sub(' ', text)

        text = text.strip().lower()
        text = text.replace("'", " ")

        text = text.split(' ')

        for word in self.stopwords:
            text = list(filter((word.lower()).__ne__, text))

        text = ' '.join(text)

        for pattern in self.additional_patterns:
            text = pattern.sub(' ', text)

        return text
```

[5]:
```
preprocessor = CustomPreProcessor()
```

1.1.2 ID to Context Dictionary

Since we used a LabelEncoder in order to transform the target variables, which were indeed the document's contexts, we need to retrieve the dictionary which contains the relationship between the assigned/encoded ID and the real value of the context. This is required since the model predict the int value instead of the categorical value (str), so in order to interpret the results we will need to undo/revent the encoding.

If you want more details/insights, please refer to `03 - Text Classification Model.ipynb`.

[6]:
```
import json

with open('resources/id2context.json', 'r') as f:
    ID2CONTEXT = json.load(f)

ID2CONTEXT
```

[6]:
```
{'3': 'wikipedia', '1': 'conference_papers', '0': 'apr', '2': 'pan11'}
```

1.1.3 Trained Text Classification Pipeline

Finally, we will just load the trained pipeline from the .joblib file where it has been dumped previosly. Since the pipeline already includes both the vectorizer and the classifier, there is no need to import any other resource so as to test the trained model.

If you want more details/insights, please refer to `03 - Text Classification Model.ipynb`.

[7]:
```
from joblib import load

pipeline = load('resources/text-classification-pipeline.joblib')
```

1.2 Model Testing

Once we loaded all the required resources, we will just need to retrieve raw data from any of the available context and test both the preprocessing and the text classification pipeline with it.

Note: so as to test it we will be using some pieces of text from Wikipedia written in English, French and Spanish; but you can play around with the text values so as to create your own texts in order to manually evaluate the text classification model.

1.2.1 Spanish Wikipedia

[8]:
```
text = """
Electronic Arts Inc. (EA) es una empresa estadounidense desarrolladora y␣
 ↪distribuidora de videojuegos para ordenador y videoconsolas, fundada por␣
 ↪Trip Hawkins.

Sus oficinas centrales están en Redwood City, California. Tiene estudios en␣
 ↪varias ciudades de Estados Unidos, en Canadá, Suecia, Corea del Sur, China e␣
 ↪Inglaterra. Posee diversas subsidiarias, como EA Sports, encargada de los␣
 ↪simuladores deportivos, EA Games para los demás juegos, y subsidiarias␣
 ↪adquiridas durante el tiempo como Maxis, entre otras. Electronics Arts␣
 ↪también posee la mayor distribución del mundo en este sector, con oficinas␣
 ↪en países como Brasil, Polonia y República Checa.

Actualmente, desarrolla y publica juegos que incluyen los títulos de EA Sports␣
 ↪FIFA, Madden NFL, NHL, NBA Live y UFC. Otras franquicias establecidas por EA␣
 ↪incluyen Battlefield, Need for Speed, Los Sims, Medal of Honor, Command &␣
 ↪Conquer, así como nuevas franquicias como Dead Space, Mass Effect, Dragon␣
 ↪Age, Army of Two, Titanfall y Star Wars: The Old Republic. Sus títulos de␣
 ↪escritorio aparecen en Origin, una plataforma de distribución digital de␣
 ↪juegos en línea para ordenadores.

Actualmente es la segunda third-party más importante de la industria de los␣
 ↪Videojuegos, con un valor de mercado de 33 mil millones de dólares.7
"""
```

[9]:
```
preprocessed_text = preprocessor._preprocess(text=text)
preprocessed_text
```

[9]:
```
'electronic arts inc ea empresa estadounidense desarrolladora distribuidora
videojuegos ordenador videoconsolas fundada trip hawkins oficinas centrales
estan redwood city california estudios varias ciudades unidos canada suecia
```

```
corea china inglaterra posee diversas subsidiarias ea sports encargada
simuladores deportivos ea games demas juegos subsidiarias adquiridas tiempo
maxis electronics arts posee mayor distribucion mundo sector oficinas paises
brasil polonia republica checa actualmente desarrolla publica juegos incluyen
titulos ea sports fifa madden nfl nhl nba live ufc franquicias establecidas ea
incluyen battlefield need speed sims medal honor command conquer nuevas
franquicias dead space mass effect dragon age army titanfall star wars old
republic titulos escritorio aparecen origin plataforma distribucion digital
juegos linea ordenadores actualmente segunda third party importante industria
videojuegos valor mercado mil millones dolares'
```

[10]:
```
ID2CONTEXT[str(pipeline.predict([preprocessed_text])[0])]
```

[10]:
```
'wikipedia'
```

1.2.2 English Wikipedia

[11]:
```
text = """
Electronic Arts Inc. (EA) is an American video game company headquartered in␣
 ↪Redwood City, California. It is the second-largest gaming company in the␣
 ↪Americas and Europe by revenue and market capitalization after Activision␣
 ↪Blizzard and ahead of Take-Two Interactive and Ubisoft as of March 2018.[4]

Founded and incorporated on May 27, 1982, by Apple employee Trip Hawkins, the␣
 ↪company was a pioneer of the early home computer games industry and was␣
 ↪notable for promoting the designers and programmers responsible for its␣
 ↪games. EA published numerous games and productivity software for personal␣
 ↪computers and later experimented on techniques to internally develop games,␣
 ↪leading to the 1987 release of Skate or Die!.

Currently, EA develops and publishes games of established franchises, including␣
 ↪Battlefield, Need for Speed, The Sims, Medal of Honor, Command & Conquer,␣
 ↪Dead Space, Mass Effect, Dragon Age, Army of Two, Titanfall, and Star Wars,␣
 ↪as well as the EA Sports titles FIFA, Madden NFL, NBA Live, NHL, and EA␣
 ↪Sports UFC.[5] Their desktop titles appear on self-developed Origin, an␣
 ↪online gaming digital distribution platform for PCs and a direct competitor␣
 ↪to Valve's Steam and Epic Games' Store. EA also owns and operates major␣
 ↪gaming studios such as EA Tiburon in Orlando, EA Vancouver in Burnaby, DICE␣
 ↪in Sweden and Los Angeles, BioWare in Edmonton and Austin, and Respawn␣
 ↪Entertainment in Los Angeles.[6]
"""
```

[12]:
```
preprocessed_text = preprocessor._preprocess(text=text)
preprocessed_text
```

[12]:
```
'electronic arts inc ea american video game company headquartered redwood city
california second largest gaming company americas europe revenue market
capitalization activision blizzard ahead take interactive ubisoft march founded
incorporated may apple employee trip hawkins company pioneer early home computer
games industry notable promoting designers programmers responsible games ea
published numerous games productivity software personal computers later
experimented techniques internally develop games leading release skate die
currently ea develops publishes games established franchises including
battlefield need speed sims medal honor command conquer dead space mass effect
dragon age army titanfall star wars ea sports titles fifa madden nfl nba live
nhl ea sports ufc desktop titles appear self developed origin online gaming
digital distribution platform pcs direct competitor valve steam epic games store
ea owns operates major gaming studios ea tiburon orlando ea vancouver burnaby
dice sweden angeles bioware edmonton austin respawn entertainment angeles'
```

[13]:
```
ID2CONTEXT[str(pipeline.predict([preprocessed_text])[0])]
```

[13]:
```
'wikipedia'
```

1.2.3 French Wikipedia

[14]:
```
text = """
Electronic Arts ou EA (NASDAQ : EA [archive]) est une société américaine fondée␣
 ↪le 28 mai 1982 et dont le siège se situe à Redwood City en Californie1. EA␣
 ↪est l'un des principaux développeurs et producteurs mondiaux de jeux vidéo.

La société occupe la place de leader sur ce marché jusqu'en 2008, notamment␣
 ↪grâce à des rachats de sociétés et de franchises de jeux, mais aussi en␣
 ↪acquérant les droits de licences sportives, comme celles de la FIFA, la NBA,␣
 ↪la NFL, ou encore celle de la LNH.

Electronic Arts est, en 2013, la troisième plus grande société commercialisant␣
 ↪des jeux vidéo, par chiffre d'affaires, après avoir été la 4e en 2012 et␣
 ↪20113.
"""
```

[15]:
```
preprocessed_text = preprocessor._preprocess(text=text)
preprocessed_text
```

[15]:
```
'electronic arts ea nasdaq ea archive societe americaine fondee mai dont siege
situe redwood city californie ea principaux developpeurs producteurs mondiaux
jeux video societe occupe place leader marche jusqu notamment grace rachats
societes franchises jeux aussi acquerant droits licences sportives comme celles
fifa nba nfl encore celle lnh electronic arts troisieme plus grande societe
commercialisant jeux video chiffre affaires apres avoir ete'
```

[16]:
```
ID2CONTEXT[str(pipeline.predict([preprocessed_text])[0])]
```

[16]: 'wikipedia'