Keyword Extraction

In this notebook, we will use the Keyword Extraction technique to extract keywords from text. We will use the YAKE! library to extract keywords from text.

YAKE! is a library that can be used to extract keywords from text made by portuguese authors (and a japanese) from Polytechnic Institute of Tomar, University of Beira Interior, University of Porto, INESC TEC (and Kyoto University).

Importing the libraries

Here we import the libraries we will use.

In [ ]:
```
import os
import warnings
import yake
import pandas
```

Functions

Now we define the functions we will use. We will use the following functions:

- **extract_keywords**: to extract keywords from text.
- **extract_keywords_chunks**: to extract keywords from text using chunks.
- **import_csv**: to import a csv file.
- **import_csv_chunks**: to import a csv file using chunks.
- **get_keywords**: to get the keywords from a list of keywords.
- **get_keywords_chunks**: to get the keywords from a list of keywords using chunks.
- **get_keywords_dir**: to get the keywords from a directory of files.
- **get_keywords_dir_chunks**: to get the keywords from a directory of files using chunks.
- **get_keywords_zomato_dir**: to get the keywords from zomato folder.

Function: extract_keywords

This function extracts keywords from text using YAKE! library. It takes as input a text and returns a list of keywords.

In [ ]:
```
def extract_keywords(df):
    keywords = []
    for i in range(0, len(df)):
        review = df["Avaliacoes"][i]
        keywords.append(yake.KeywordExtractor(lan="pt").extract_keywords(review))
    return keywords
```

Function: import_csv

This function imports a csv file. It takes as input a csv file and returns a dataframe.

In [ ]:
```
def import_csv(path):
    df = pandas.read_csv(path, encoding="utf-8")
    return df
```

Function: get_keywords

This function gets all the csv files in a directory and returns a list of keywords from the dataframes.

In [ ]:
```
def get_keywords(path):
    keywords = []
    for file in os.listdir(path):
        if file.endswith(".csv") and not file.startswith("list"):
            df = import_csv(path + "/" + file)
            keywords.append(extract_keywords(df))
    return keywords
```

Function: get_keywords_dir

This function gets all the csv files in a directory and get a list of keywords from the dataframes, then it writes the keywords in a csv file.

In [ ]:
```
def get_keywords_dir(path1, path2, name):
    current_dir = os.getcwd()
    path = current_dir + path1
    keywords = get_keywords(path)
    # print(keywords)
    i = 0
    for keyword in keywords:
        # print(keyword)
        with open(
            current_dir + path2 + name + str(i) + ".csv",
            "w",
        ) as f:
            f.write("Expressao, Frequencia\n")
            for k in keyword:
                for word in k:
                    f.write(
                        str(word)
                        .replace("(", "")
                        .replace(")", "")
                        .replace("\u2010", "-")
                        + "\n"
                    )
        i += 1
```

Function: get_keywords_zomato_dir

This function gets all the csv files from the zomato directory and get a list of keywords from the dataframes, then it writes the keywords in a csv file.

In [ ]:
```
def get_keywords_zomato_dir(path1, path2, name):
    current_dir = os.getcwd()
    path = current_dir + path1
    keywords = get_keywords(path)
    # print(keywords)
    i = 0
    restaurantes = [0, 1, 2, 12, 13, 14, 15]
    for keyword in keywords:
        # print(keyword)
        with open(
            current_dir + path2 + name + str(restaurantes[i]) + ".csv",
            "w",
        ) as f:
            f.write("Expressao, Frequencia\n")
            for k in keyword:
                for word in k:
                    f.write(
                        str(word)
                        .replace("(", "")
                        .replace(")", "")
                        .replace("\u2010", "-")
                        + "\n"
                    )
        i += 1
```

Execution

Now we execute the code. We will use the last set functions to get the keywords from the reviews.

In [ ]:
```
warnings.filterwarnings("ignore")

get_keywords_dir("/../scrapes/booking/hotels", "/booking/hotels/", "hotel")
get_keywords_zomato_dir(
    "/../scrapes/zomato/restaurantes",
    "/zomato/restaurantes/",
    "restaurante",
)
get_keywords_dir(
    "/../scrapes/tripadvisor/restaurants",
    "/tripadvisor/restaurants/",
    "restaurant",
)
get_keywords_dir(
    "/../scrapes/tripadvisor/activities",
    "/tripadvisor/activities/",
    "place",
)
get_keywords_dir("/../scrapes/tripadvisor/hotels", "/tripadvisor/hotels/", "hotel")
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js