OTH Regensburg
Fakutät IM
Brijnesh Jain

PGKI
SS 2024
14.05.2024

Assignment 7

Exercise 1: Word Counter

a) Write a program that records the top n most frequently occurring words of a text document. For each of the n words, the program should output the following information: the word, the number of occurrences, the ratio of occurrences to the total number of words, and the length of the word. Test your program on the books `Moby Dick` and `The Crowd`.

Hints:

1. The files of the books are

```
Moby Dick: ./data/books/moby_dick.txt
The Crowd: ./data/books/LeBon.txt
```

2. Use the following code to read the text file line by line:

```
with open('myfile.txt', 'r') as file:
    for line in file:
        print(line)
```

3. Use the following code to strip out all characters from a string `str` that are not alphabetic:

```
import re
str = re.sub(r'[^a-zA-Z]', ' ', str)
```

b) Write a program that takes both sorted lists of exercise 3.a and produces the same output as in 3.a but over all words of both books.

c) Produce the same output as in the previous two exercises, but this time sorted in descending order of word length.

Exercise 1: Dynamic Time Warping Distance

The Dynamic Time Warping (DTW) distance is a time series distance function designed to cope with variations in velocity and phase (Chapter 4.1 of this link provides an introduction to the DTW distance).

The DTW distance $\mathrm{DTW}(\boldsymbol{x}, \boldsymbol{y})$ between time series $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ of length m and n, resp., is defined as follows:

$$\mathrm{dtw}(i,j) = (x_i - y_j)^2 + \min\Big(\mathrm{dtw}(i-1,j),\ \mathrm{dtw}(i,j-1),\ \mathrm{dtw}(i-1,j-1)\Big)$$

where $\mathrm{dtw}(i,j)$ represents the minimum cost of aligning the first i elements of $\boldsymbol{x}$ with the first j elements of $\boldsymbol{y}$. The recursive formula for $\mathrm{dtw}(i,j)$ holds for all $i = 2, \ldots, n$ and $j = 2, \ldots, m$. The initial values are

- $\text{dtw}(1,1) = 0$
- $\text{dtw}(i,1) =$ `float('inf')` for all $i = 2, \ldots, n$
- $\text{dtw}(1,j) =$ `float('inf')` for all $j = 2, \ldots, m$

Then the DTW distance $\text{DTW}(\boldsymbol{x}, \boldsymbol{y})$ between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by $\text{dtw}(m,n)$.

Task: Implement a function that computes the DTW distance between two time series.

Exercise 2: k-NN Algorithm

The `Lightning7` dataset is a collection of time series data from the UCR Time Series Archive. It contains 70 training examples and 73 test examples, each with 319 time points. The data represents recordings of lightning strikes at seven different locations, and the goal is to classify each time series according to its location. The `Lightning7` dataset is provided as two CSV files located under `/data/Lightning7/`:

- `train.tsv` contains the training examples
- `test.tsv` contains the test examples

In the dataset, each row represents a time series along with its associated class label. The first number in each row is the class label, and the remaining values represent the time series data. The values within each row are separated by tabs.

Task: Implement the k-NN algorithm and use it to classify the test examples in the `Lightning7` dataset. Test your k-NN implementation with the Euclidean distance and the dynamic time warping distance (see exercise 1). Compare the results for different values of k.

Hint: You can use `Counter` from the module `collections` to identify the majority class. See notebook `knn.ipynb` for further details.