Merkblatt

1 Encoding

Um Textdateien lesen zu können bzw. zu bearbeiten, wird meistens die Zeichenkodierung UTF-8 oder iso_8859_1 benutzt.

UTF-8 kodiert Zeichen mit variabler Byte-Anzahl. Dabei wird ein Unicodezeichen in 1 bis 4 Bytes kodiert. Die Codepoints 0 bis 127, die dem ASCII-Zeichensatz entsprechen, werden in einem Byte kodiert, wobei das höchstwertige Bit stets 0 ist. Mithilfe des achten Bits kann ein längeres Unicode-Zeichen eingeleitet werden, das sich auf 2, 3 oder 4 Byte erstreckt, somit kann UTF-8 so gut wie alle Zeichen, Sprachen und sogar Emojis darstellen.

iso_8859_1 oder auch Latin-1 ist die wohl wichtigste und am häufigsten gebrauchte Kodierung für lateinische Schriften, auch in Teilen Afrikas, in denen nicht die arabische Schrift verwendet wird, ist es weit verbreitet.

2 Zip-Dateien extrahieren

Um Zip-Dateien extrahieren zu können, ist zunächst das Modul zipfile mit dem Befehl `import zipfile` zu importieren. Danach wird der Dateipfad des zu extrahierenden Zipfiles angegeben, das dann in einer Variable (hier z) gespeichert wird. Diese kann mit dem Befehl "extractall" in einen anderen Dateipfad extrahiert werden.

```
import zipfile

with zipfile.ZipFile ("../data/Spam-Emails.zip", "r") as z:
    z.extractall( "../data/Spam-Emails")
```

3 Aus einer E-Mail den Inhalt einlesen

Um aus einer E-Mail den Inhalt einzulesen, müssen zunächst die Kopfzeilen abgetrennt werden. Eine E-Mail besteht in den meisten Fällen aus Kopfzeilen und dem Inhalt. Um nur den Inhalt einzulesen, ist die E-Mail an der Stelle, wo 2 Zeilenumbrüche stehen, mit dem "split" Befehl zu trennen. Dann kann der Inhalt (hier `content`) ausgeben werden.

```
with open (path, mode = "r", encoding = "utf-8") as file:
    email = file.read()
    content = email.split ("\n\n", 1)[1]
    print(content)
```

4 Aus mehrehren E-Mails den Inhalt auslesen

Mit der 1. Funktion geben wird der Inhalt einer E-Mail ausgegeben. (hier `content`).

Mit der 2. Funktion werden zwei Ordner mit E-Mails an die Funktion übergeben.

[ ]:
```
import os

def get_email_content (path):
    with open (path, mode = "r", encoding = "iso-8859-1") as file:
        email = file.read ()
        email_parts = email.split ("\n\n", 1)
        if len (email_parts) != 2:
            return None
        else:
            content = email_parts[1]
            return content


def read_emails (folder):
    contents = []
    for f in os.listdir (folder):
        complete_path = folder + "/" + f
        if os.path.isdir (complete_path):
            continue
        email_content = get_email_content (complete_path)
        if email_content != None:
            contents.append (email_content)

    return contents


spam_contents = read_emails ("../data/Spam-Emails/spam")
ham_contents = read_emails ("../data/Spam-Emails/easy_ham")
```

[ ]:
```

```