

# CHƯƠNG 6, 7: String and File

## 1. Cấu trúc chuỗi (String Structure)

Chỉ mục (Index): Bắt đầu từ 0. s[0] là ký tự đầu, s[1] là ký tự thứ 2. Độ dài: len(s) trả về số ký tự (số nguyên). Bất biến (Immutable): Bạn không thể thay đổi ký tự trong chuỗi (VD: s[0] = 'J' sẽ báo lỗi). Phải tạo chuỗi mới nếu muốn sửa đổi.

## 2. Cắt chuỗi (Slicing) - Quan trọng

Cú pháp: s[start:end] Quy tắc: Lấy từ start đến sát end (không lấy end). Mặc định: s[:2]: Từ đầu đến index 2 (không bao gồm 2). s[8:]: Từ index 8 đến hết chuỗi. s[:]: Lấy toàn bộ chuỗi.

## 4. Thư viện String (Methods)

.lower() / .upper(): Chuyển thành chữ thường / hoa.  
.strip(): Xóa khoảng trắng (dấu cách, xuống dòng) ở cả 2 đầu.  
(lstrip: bên trái, rstrip: bên phải).  
.find(sub): Tìm vị trí xuất hiện đầu tiên của sub. Trả về -1 nếu không tìm thấy.  
.replace(old, new): Thay thế tất cả old bằng new..startswith(prefix): Kiểm tra chuỗi có bắt đầu bằng prefix không (True/False).

## 1. Mở tệp (File Handle)

```
filename = input('Enter file: ')
try:
    fhand = open(filename)
except:
    print('File cannot be opened:', filename)
    quit()
```

## 3. Cách đọc tệp chuẩn (Pattern)

Cách 1: Đọc từng dòng (Tiết kiệm bộ nhớ) - Khuyên dùng Vòng lặp for sẽ coi file handle là một chuỗi các dòng.  
fhand = open('mbox.txt')  
for line in fhand:  
 print(line)  
Cách 2: Đọc toàn bộ (Read All) Chỉ dùng khi file nhỏ. Đọc cả file vào một chuỗi duy nhất (chứa các ký tự \n bên trong).  
inp = fhand.read()  
print(len(inp))

## 2. Ký tự xuống dòng (Newline \n)

Trong Python, xuống dòng được biểu diễn bằng ký tự \n. Nó được tính là 1 ký tự khi dùng len(). Khi in ra (print), mỗi dòng từ file thường có sẵn \n ở cuối, cộng thêm \n của lệnh print -> Gây ra hiện tượng dòng trắng thừa. -> Giải pháp: Dùng .rstrip().

## 4. Tìm kiếm & Lọc dữ liệu trong tệp

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip() # Xóa ký tự xuống dòng thừa
    if not line.startswith('From:'):
        continue # Bỏ qua dòng không thỏa mãn [cite: 3100]
```

```
# Cách 2: Dùng in (Chứa từ khóa bất kỳ đâu)
if not '@uct.ac.za' in line:
    continue [cite: 3106]
```

```
print(line) # Chỉ in dòng thỏa mãn điều kiện
```