BÀI TẬP GIỮA KỲ MÔN XỬ LÝ TIẾNG NÓI

1. THÔNG TIN CHUNG

Học viên thực hiện: Nguyễn Thị Ngọc Trâm
Mã số học viên: 21C11036
Khóa: 31 Ngành: Khoa học máy tính

2. NỘI DUNG

Khảo sát từ điển tiếng việt:

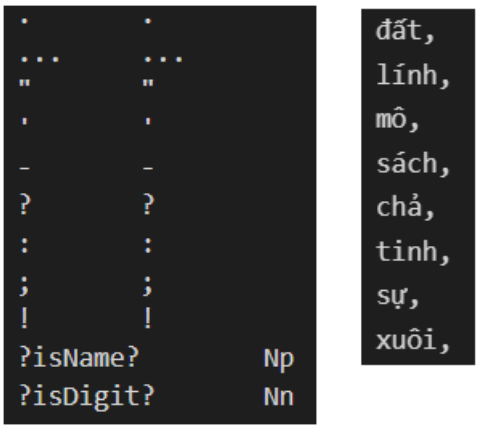
(a) Có bao nhiêu âm tiết tiếng Việt khác nhau?

+ Cắt lấy từng dòng, thay ký tự "-" thành " " rồi split theo khoảng trắng được các từ đơn lẻ.

+ Loại bỏ các ký tự đặc biệt ở đầu câu, các từ trùng do dấu phẩy:

(`bán mặt cho đất bán lưng cho trời`
`bán mặt cho đất, bán lưng cho trời` => tạo ra 2 từ "đất" và "đất," nên phải xóa bớt).

+ Bỏ vào hàm set() để lấy các từ không trùng nhau.



Hình 1: Các trường hợp đặc biệt cần loại bỏ

Đáp án: có 7996 âm tiết trong từ điển

```python
f = open('VDic_uni.txt', encoding="utf8")
wordlist = []
while True:
    line = f.readline()
    if not line:
        break

    tmp = line.split("\t")[0].replace("\n", "")
    tmp = tmp.replace("_", " ")
    w = tmp.split(' ')
    w = [s.lower().replace("\ufeff", "") for s in w if len(s) != 0 and s[0] not in invalid_char and all(char != ',' for char in s)]
    wordlist += w

SyllableLs = set(wordlist)
print("Number of syllable in VDic: ", len(SyllableLs))
```

[5] ✓ 0.2s

```
Number of syllable in VDic:  7996
```

Hình 2: Source code minh họa cho câu (a)

(b) Hãy ước tính số âm tiết có thể có trong tiếng Việt bằng cách tính tổ hợp: [PÂĐ x Đệm x Chính x Cuối x Thanh] có bao nhiêu âm tiết?

```python
phu_am_dau = ['th', 'p', 't', 'tr', 'ch', 'c', 'k', 'q', 'b', 'đ', 'ph', 'x', 's', 'kh', 'h', 'v', 'd', 'gi', 'r', 'g', 'gh', 'm', 'n', 'nh', 'ng', 'ngh', 'l']
print("PAD: ", len(phu_am_dau))
am_dem = ['u', 'o']
print("Am Dem: ", len(am_dem))
am_chinh = ['i', 'y', 'ư', 'u', 'iê', 'ia', 'yê', 'ya', 'ươ', 'ưa', 'uê', 'ua', 'ê', 'ơ', 'â', 'ô', 'e','a', 'ă', 'o' , 'uô', 'oo']
print("Am Chinh: ", len(am_chinh))
am_cuoi = ['m', 'n', 'ng', 'nh', 'p', 't', 'ch', 'c', 'u', 'o', 'i', 'y']
print("Am Cuoi: ", len(am_cuoi))
thanh = 6
numOfSyll = len(phu_am_dau)*len(am_chinh)*len(am_dem)*len(am_cuoi)*thanh
print("So am tiet co the co: ", numOfSyll)
```

```
PAD:  27
Am Dem:  2
Am Chinh:  22
Am Cuoi:  12
So am tiet co the co:  85536
```

Đáp án: Số âm tiết có thể có: 85536

(c) So sánh 2 con số (a) và (b) giải thích lý do tại so có sự chênh lệch này?

- Có những tổ hợp âm không có nghĩa nên không có trong từ điển
- Từ điển chưa đủ phong phú để thể hiện tất cả các từ trong tiếng Việt.

(d) Chỉ rõ các quy luật (có thể có) trong hệ thống âm tiết tiếng Việt về:

- **Bình diện ngữ âm: những âm vị nào thường/phải đii với những âm vị nào?**

  Tạo list phoneme, lập ma trận duyệt hết từ điển, đếm xem âm nào đi với âm nào.

```
1  phonemeLs = phu_am_dau + am_dem + am_chinh + am_cuoi
2  phonemeLs = list(set(phonemeLs))
3  phonemeLs.sort()
4  print(len(phonemeLs))
5  print(phonemeLs)
```

✓ 0.1s Pyth

```
49
['a', 'b', 'c', 'ch', 'd', 'e', 'g', 'gh', 'gi', 'h', 'i', 'ia', 'iê', 'k', 'kh',
'l', 'm', 'n', 'ng', 'ngh', 'nh', 'o', 'oo', 'p', 'ph', 'q', 'r', 's', 't', 'th',
'tr', 'u', 'ua', 'uê', 'uô', 'v', 'x', 'y', 'ya', 'yê', 'â', 'ê', 'ô', 'ă', 'đ',
'ơ', 'ư', 'ưa', 'ươ']
```

```
1  # Create phoneme dataframe
2  lst = [[0]*len(phonemeLs)]*len(phonemeLs)
3  df = pd.DataFrame(lst, columns=phonemeLs, index=phonemeLs)
4  df
```

✓ 0.1s Py

	a	b	c	ch	d	e	g	gh	gi	h	...	yê	â	ê	ô	ă	đ	ơ	ư	ưa	ươ
a	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
ch	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
gh	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
gi	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Viết hàm tách âm vị cho từng từ (chưa hoàn thành)

Kết luận:
Các phụ âm đầu đi với âm đệm hoặc âm chính
Âm chính đi với âm cuối.

- **Bình diện ngữ pháp: những từ có những âm vị nào thường có xu hướng mang từ loại (A, N, V,…) nào?**
  Rút list các từ loại gồm có 49 loại:

```
49
['Aa', 'An', 'Cc', 'Cm', 'E', 'I', 'Ja', 'Jd', 'Ji', 'Jr', 'Jt', 'Jt,Jd', 'Na',
'Nc', 'NcVt', 'Ng', 'Nl', 'Nn', 'Np', 'Nt', 'Nu', 'Pa', 'Pd', 'Pi', 'Pn', 'Pp',
'Va', 'Vb', 'Vc', 'Vim', 'Vit', 'Vitb', 'Vitc', 'Vitim', 'Vits', 'Vla', 'Vo', 'Vs',
'Vt', 'Vta', 'Vtb', 'Vtc', 'Vtim', 'Vto', 'Vts', 'Vtv', 'Vv', 'X', 'đg']
```

Tạo dataframe giống câu bình dị ngữ âm, nhưng do chưa tách được từng âm vị trong từ nên chưa hoàn thiện.

- **Bình diện ngữ nghĩa: những từ có âm vị nào thường có xu hướng mang ý nghĩa nào?**