**BÁO CÁO TIẾN ĐỘ**

**--- Tuần 12-13 ---**

**1. Ý tưởng**

- Trong quá trình khảo sát các tập tin Word đã OCR, nhóm nhận thấy một phần dữ liệu được lưu trữ dưới dạng bảng, là dạng dữ liệu tương đối dễ thao tác và chuyển đổi. Ví dụ:

|  |  |
| --- | --- |
| Bạch - d | *Hơi* |
| Bạch hầu - d | *Pơangeh jĩ ako* |
| Bạch tạng - d | *Mong* |
| Bài - d | *B`ai* |
| Bài giảng - d | *B`ai bơ tho* |
| Bài hát - d | *B`ai hơri* |
| Bài học - d | *Tơdrong pơhrăm; b`ai* |
| Bài sai - d | *B`ai glăi* |
| bài tập - d | *B`ai pơhrăm* |
| Bài thơ - d | *Nỡr pơđỡk* |

- Để tự động hóa việc chuyển đổi các dữ liệu dạng bảng như trên sang tập tin Excel, nhóm quyết định sử dụng ngôn ngữ lập trình **Python** cùng với thư viện **python-docx**, một thư viện cho phép người dùng truy cập, tạo mới hoặc thay đổi một tập tin Word. Hướng dẫn cài đặt và sử dụng thư viện có thể tìm thấy tại đây: [https://python-docx.readthedocs.io/en/latest/](https://python-docx.readthedocs.io/en/latest/%20)

- Bên cạnh việc truy xuất các dữ liệu dạng bảng từ file Word bằng thư viện trên, nhóm cũng cần trích xuất từ loại sang một cột riêng biệt. Để thực hiện, nhóm tiến hành xử lý chuỗi trên cột đầu tiên, nhằm tách riêng phần từ vựng và từ loại thông qua ký tự ngăn cách “-”.

- Quá trình chuyển đổi và xử lý dữ liệu trên file Word diễn ra khá suôn sẻ, nên nhóm đã thu hoạch được một số kết quả khả quan như bên dưới.

**2. Nội dung chi tiết**

**2.1. Source code**

Từ vựng đối chiếu Việt Gia Lai:

# import docx2txt

import docx

import pandas as pd

my\_path = "Word/Tu\_vung\_doi\_chieu\_GiaLai/Viet Ba Na OCR/"

paths = ["tu vung doi chieu p1 done.docx", "tu vung doi chieu p2 done.docx", "tu vung doi chieu p3 done.docx", "tu vung doi chieu p4.docx"]

i = 0

# Vietnamese\_data = []

# Bahnar\_data = []

# word\_type\_data = []

for path in paths:

    file = docx.Document(my\_path + path)

    Vietnamese = []

    Bahnar = []

    word\_type = []

    for table in file.tables:

        for row in table.rows:

            if len(row.cells) != 2:

                continue

            try:

                VN, type = row.cells[0].text.split('-')

                Vietnamese.append(VN.strip())

                word\_type.append(type.strip())

                Bahnar.append(row.cells[1].text.strip())

            except:

                pass

    path = "Excel/df" + str(i) + ".xlsx"

    data = {'language0': Vietnamese,

            'language1': Bahnar,

            'word\_type': word\_type}

    df = pd.DataFrame(data)

    df.to\_excel(path, index = False)

    i += 1

Từ điển Ba Na – Việt – Kon Tum:

import docx

import pandas as pd

import re

my\_path = "E:/Study/CNPM/MR/Word/Tu dien Ba Na - Việt Kon Tum.docx"

file = docx.Document(my\_path)

language0 = []

language1 = []

para = file.paragraphs

for i in range(96, 17126):

    try:

        index = para[i].text.find('x:')

        if index == -1:

            pass

        else:

            para[i].text = para[i].text[:index]

    except IndexError:

        pass

    try:

        x, y = file.paragraphs[i].text.split(':', 1)

        if (x == "" or y == ""):

            continue

        language0.append(x.strip())

        language1.append(y.strip())

    except:

        continue

word\_type = []

for i, word in enumerate(language0):

    try:

        if word[-1] == ')':

            if word[-5] == '(':

                word\_type.append(word[-4:-1])

            else:

                word\_type.append(word[-3:-1])

        else:

            word\_type.append("")

    except:

        word\_type.append("")

    try:

        language0[i] = re.sub("[\(\[].\*?[\)\]]", "", language0[i])

        language0[i] = re.sub('\^', "", language0[i])

        language0[i] = re.sub('®', "", language0[i])

        language0[i] = re.sub('\(', "", language0[i])

        language0[i] = re.sub('\[', "", language0[i])

        language0[i] = re.sub('\)', "", language0[i])

        language0[i] = re.sub('\]', "", language0[i])

        language0[i] = ''.join([x for x in language0[i] if not x.isdigit()])

        language0[i] = language0[i].strip()

    except:

        pass

path = "Excel/df4.xlsx"

print(len(language0))

print(len(language1))

print(len(word\_type))

data = {'language0': language0,

        'language1': language1,

        'word\_type': word\_type}

df = pd.DataFrame(data)

df.to\_excel(path, index = False)

**2.2. Mô tả**

- Trước hết, hai thư viện *docx* và *pandas* được thêm vào chương trình. Thư viện *docx* dùng để tương tác với file Word và *pandas* dùng để chuyển đổi bảng từ vựng sau khi xử lý sang một data frame để truyền vào file Excel.

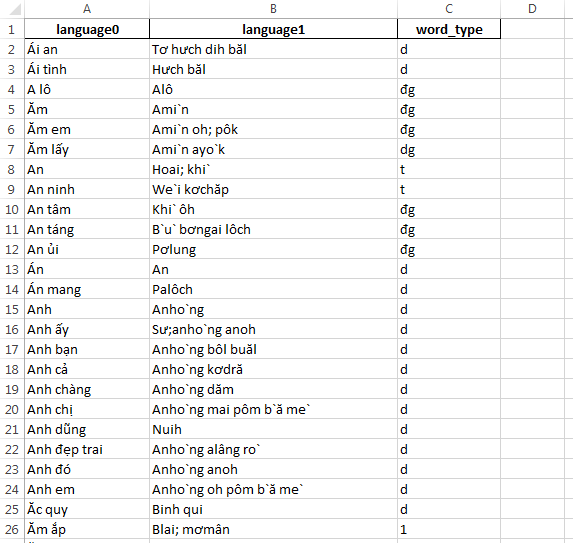
- Các file Word được truyền vào chương trình thông qua đường dẫn đến tập tin.

- Chương trình lần lượt duyệt qua các file. Với mỗi bảng trong file, chương trình sẽ trích xuất dữ liệu trong cột đầu tiên và tách chúng thành hai phần: từ vựng và từ loại.

- Đưa các nội dung đã trích xuất vào các mảng tương ứng và tạo một data frame từ chúng. Truyền data frame đó vào file Excel.

**2.3. Kết quả thu được**

Sau khi hiện thực và chạy chương trình, nhóm thu được các kết quả như sau:



**Nhận xét:**

|  |  |
| --- | --- |
| ***Ưu điểm*** | ***Khuyết điểm*** |
| Sử dụng Python nên có tính đa dụng, dễ hiện thực và mở rộng | Nhóm chưa xử lý được các dữ liệu không phải dạng bảng trong Word |
| Thư viện hỗ trợ tốt, code trong sáng | Quá trình chuyển đổi chưa hoàn toàn chính xác |

**3. Định hướng tiếp theo**

- Nghiên cứu về cách ứng dụng Machine Learning trong việc xử lý dữ liệu không phải dạng bảng từ file Word.

- Tối ưu hóa quá trình chuyển đổi, hạn chế sai sót.