

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TP.HỒ CHÍ MINH  
KHOA CÔNG NGHỆ THÔNG TIN  
BỘ MÔN CÔNG NGHỆ PHẦN MỀM



# TÀI LIỆU HƯỚNG DẪN THU THẬP DỮ LIỆU VÀ HUẤN LUYỆN LANGUAGE MODEL CHO NGÔN NGỮ TIẾNG VIỆT

**GVHD:**

TS. Ngô Huy Biên

**Sinh viên thực hiện:**

1612689 – Trương Phạm Nhật Tiến

[1612689@student.hcmus.edu.vn](mailto:1612689@student.hcmus.edu.vn)

1612726 – Nguyễn Minh Trí

[1612726@student.hcmus.edu.vn](mailto:1612726@student.hcmus.edu.vn)

**Tp. Hồ Chí Minh, tháng 7 năm 2020**

## Mục lục

1. Giới thiệu .....	3
2. Tạo bộ dữ liệu: .....	4
3. Huấn luyện mô hình 3-gram với Kenml:.....	5
3.1 Tải về và build kenlm:.....	5
3.2 Huấn luyện mô hình: .....	6

## 1. Giới thiệu

Tài liệu hướng dẫn cách tạo bộ dữ liệu gồm các câu tiếng Việt được trích xuất từ viwiki dump progress on 20200701.

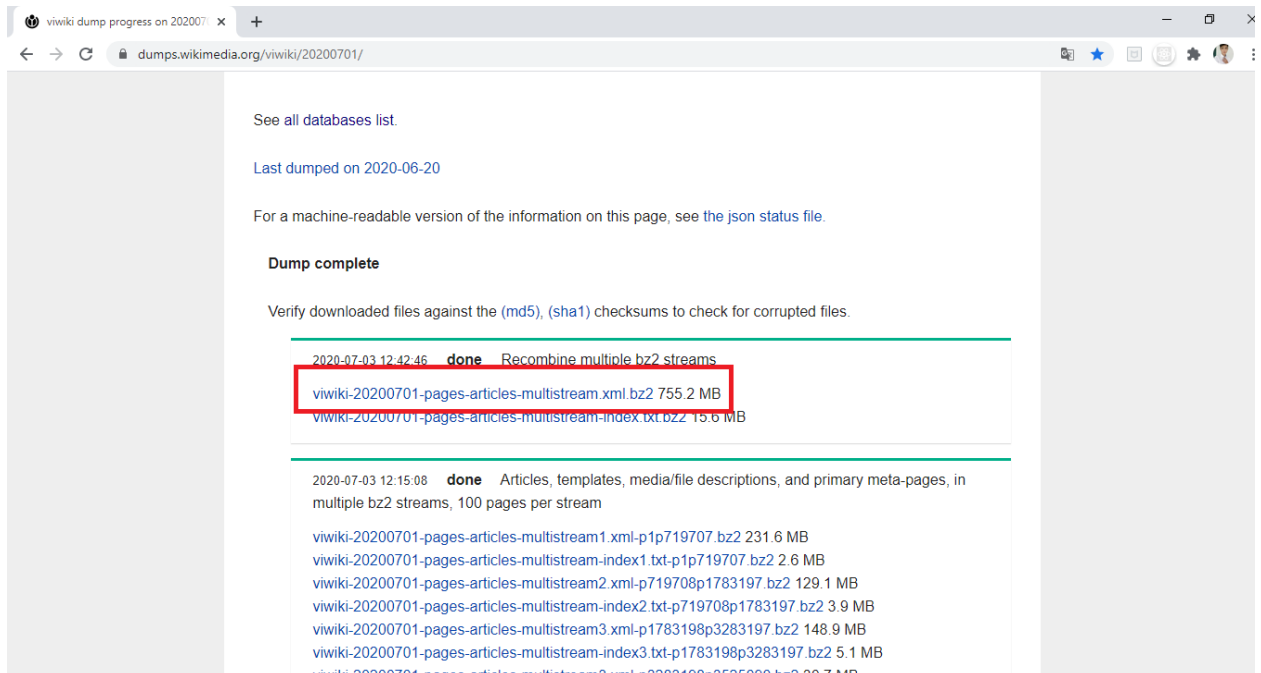
Sau đó huấn luyện Language model 3-gram với bộ dữ liệu này bằng Kenlm trên môi trường google colab.

Kenlm là dụng cụ có thể train mô hình N-gram với N gram tùy chọn.

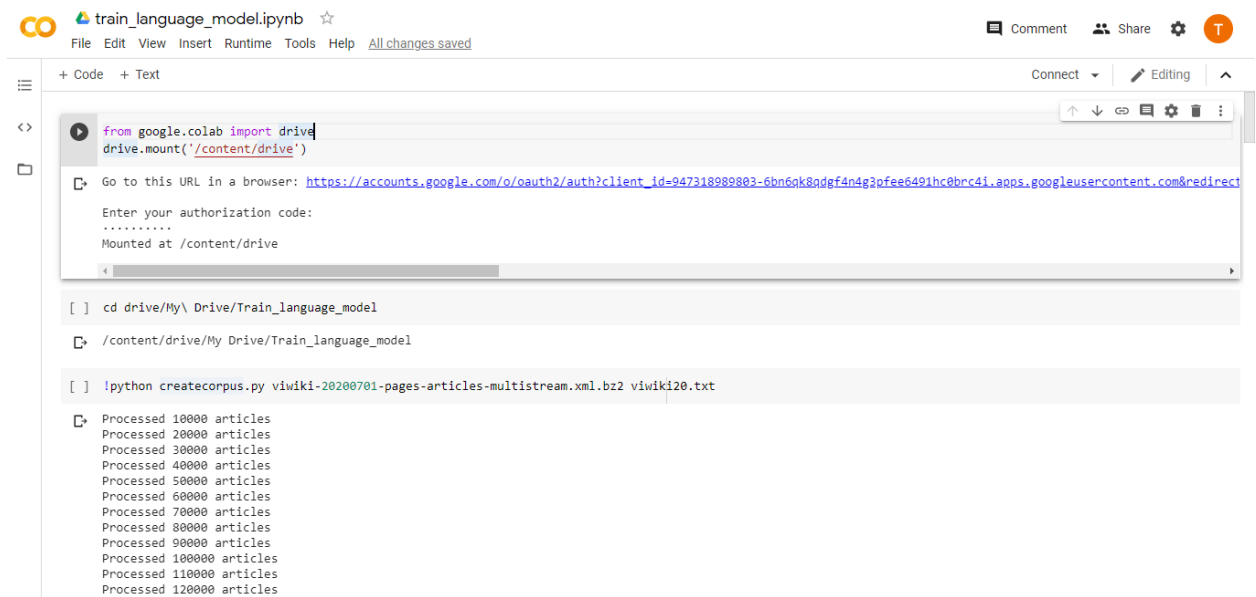
## 2. Tạo bộ dữ liệu:

Đầu tiên tải

- bộ dữ liệu từ link sau : <https://dumps.wikimedia.org/viwiki/20200701/>
- File tạo bộ dữ liệu từ bộ dữ liệu thô ở trên: [createcorpus.py](#).
- File kiểm tra dữ liệu sau khi được xử lí: [checkwikicorpus.py](#)
- File tiền xử lí dữ liệu khi huấn luyện: [preprocess.py](#)



Cd vào thư mục chứa file dữ liệu vừa tải về và chạy file createcorpus.py theo các đoạn code sau trên google colab. (bởi vì file dữ liệu thô hơi lớn, nên thời gian xử lí hơi mất thời gian)



```
from google.colab import drive
drive.mount('/content/drive')

Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989802-6bn6nk8gdf4n4g3ofee6491hc0brc4i.apps.googleusercontent.com&redirect

Enter your authorization code:
.....
Mounted at /content/drive

[ ] cd drive/My Drive/Train_language_model

/content/drive/My Drive/Train_language_model

[ ] !python createcorpus.py viwiki-20200701-pages-articles-multistream.xml.bz2 viwiki20.txt

Processed 10000 articles
Processed 20000 articles
Processed 30000 articles
Processed 40000 articles
Processed 50000 articles
Processed 60000 articles
Processed 70000 articles
Processed 80000 articles
Processed 90000 articles
Processed 100000 articles
Processed 110000 articles
Processed 120000 articles
```

Sau khi ta nhận được file viwiki20.txt ta tiến hành kiểm tra một đoạn dữ liệu bên trong file thử bằng đoạn code sau



```
[ ] !python checkwikicorpus.py viwiki20.txt

Internet society hay isoc là một tổ chức quốc tế hoạt động phi lợi nhuận phi chính phủ và bao gồm các thành viên có trình độ chuyên ngành tổ chức này chú trọng
tiếng việt còn gọi tiếng việt nam tiếng kinh hay việt ngữ là ngôn ngữ của người việt dân tộc kinh và là ngôn ngữ chính thức tại việt nam đây là tiếng mẹ đẻ của
ohio viết tắt là oh viết tắt cũ là là một tiểu bang khu vực trung tây cũ nằm miền đông bắc hoa kỳ tên ohio theo tiếng iroquois có nghĩa là sông đẹp và đó cũng :
california phát âm như ca li pho ní hay ca li phốc ní nếu nhanh ca li phốc nha còn được người việt gọi vắn tắt là ca li là một tiểu bang ven biển phía tây củ
thụy điển tên chính thức là vương quốc thụy điển là một vương quốc bắc âu giáp na uy phía tây và phần lan phía đông bắc nối với đan mạch bằng cầu ðresund phía r
biểu tượng sài gòn thời kỳ liên bang đồng dương thành phố hồ chí minh thường được gọi là sài gòn là một trong hai thành phố lớn nhất việt nam đồng thời cũng là
lào cai là một tỉnh vùng cao biên giới thuộc vùng tây bắc bộ việt nam nằm Lào Cai là đơn vị hành chính việt nam đồng thứ về số dân xếp thứ về tổng sản phẩm trên
world wide web consortium viết tắt là một côngxooxciom lập ra các chuẩn cho internet nhất là cho world wide web chủ tịch của là ngài tim berners lee người sáng
bộ kế hoạch và đầu tư là cơ quan của chính phủ thực hiện chức năng quản lý nhà nước về kế hoạch đầu tư phát triển và thống kê bao gồm tham mưu tổng hợp về chiến
lào lao tên chính thức là cộng hòa dân chủ nhân dân Lào sathalanalat paxathipatai paxaxon lao là quốc gia nội lục tại đông nam và là trung tâm của bán đảo đông
hoa kỳ viết tắt hoặc us hay mỹ america tên đầy đủ là hợp chúng quốc hoa kỳ hoặc hợp chúng quốc mỹ viết tắt usa là một quốc gia cộng hòa lập hiến liên bang gồm 1
hà giang là một tỉnh thuộc vùng đông bắc bộ việt nam nằm Hà Giang là đơn vị hành chính việt nam đồng thứ về số dân xếp thứ về tổng sản phẩm trên địa bàn grdp vì
cao bằng là một tỉnh thuộc vùng đông bắc bộ việt nam nằm Cao Bằng là đơn vị hành chính việt nam đồng thứ về số dân xếp thứ về tổng sản phẩm trên địa bàn grdp xi
iraq tên đầy đủ là cộng hòa iraq phát âm tiếng việt như rắc tiếng rập الجمهورية العراقية al jumhuriyah al iraqiyah tiếng kurd عێراق komara îraqê là một quốc gia khu vực
hà nội là thủ đô của nước cộng hòa xã hội chủ nghĩa việt nam cũng là kinh đô của hầu hết các vương triều phong kiến tại việt nam trước đây do đó lịch sử hà nội
```

Dữ liệu đúng là các đoạn ngôn ngữ tiếng Việt được viết trên các dòng.

### 3. Huấn luyện mô hình 3-gram với Kenml:

#### 3.1 Tải về và build kenml:

- Tiến hành chạy các đoạn code sau để clone và build kenml

```
[ ] git clone https://github.com/kpu/kenlm.git

[ ] cd kenlm

[ ] mkdir -p build

[ ] cd build

[ ] cmake ..
[ ] make -j 4

-- The C compiler identification is GNU 7.5.0
-- The CXX compiler identification is GNU 7.5.0
-- Check for working C compiler: /usr/bin/cc -- works
-- Detecting C compiler ABI info
-- Detecting C compiler ABI info - done
-- Detecting C compile features
-- Detecting C compile features - done
-- Check for working CXX compiler: /usr/bin/c++ -- works
-- Detecting CXX compiler ABI info
```

- Sau khi build thành công ta sẽ có được các file để huấn luyện model như trong thư mục kenlm/build/bin

Drive của tôi > ... > build > bin ▾

Tên ↑	Chủ sở hữu	Sửa đổi lần cuối	Kích cỡ tệp
<input checked="" type="checkbox"/> build_binary	tôi	16 thg 7, 2020 tôi	746 KB
<input checked="" type="checkbox"/> count_ngrams	tôi	16 thg 7, 2020 tôi	658 KB
<input checked="" type="checkbox"/> filter	tôi	16 thg 7, 2020 tôi	712 KB
<input checked="" type="checkbox"/> fragment	tôi	16 thg 7, 2020 tôi	714 KB
<input checked="" type="checkbox"/> kenlm_benchmark	tôi	16 thg 7, 2020 tôi	1 MB
<input checked="" type="checkbox"/> Implz	tôi	16 thg 7, 2020 tôi	1 MB
<input checked="" type="checkbox"/> phrase_table_vocab	tôi	16 thg 7, 2020 tôi	223 KB
<input checked="" type="checkbox"/> probing_hash_table_benchmark	tôi	16 thg 7, 2020 tôi	296 KB
<input checked="" type="checkbox"/> query	tôi	16 thg 7, 2020 tôi	767 KB

## 3.2 Huấn luyện mô hình:

- Import và tải về thư viện nltk.punkt

```
import nltk
nltk.download('punkt')

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
True
```

- Cấp quyền cho file Imp

```
chmod 755 bin/implz
```

- Nén lại file viwiki20.txt (file dữ liệu đã được trích xuất từ wiki dump) theo định dạng .bz2.

```
[ ] !bzip2 ../../viwiki20.txt
```

- Sau đó huấn luyện mô hình (số 3 trong đoạn code tương ứng với 3-gram, ta có thể điều chỉnh tham số này để có thể huấn luyện các mô hình N-gram khác nhau)

```
[ ] !bzipcat ../../viwiki20.txt.bz2 | python ../../preprocess.py | bin/implz -o 3 > ../../vietnamese_wiki_3-gram.arpa
```

- Sau khi có được file Vietnamese\_wiki\_3-gram.arpa (file model), ta sẽ chuyển đổi nó thành định dạng .binary để giảm kích thước file và load nhanh hơn khi sử dụng.

```
! bin/build_binary ../../vietnamese_wiki_3-gram.arpa vietnamese_wiki_3-gram.binary

Reading ../../vietnamese_wiki_3-gram.arpa
-----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
SUCCESS
```

- Kiểm thử language model

```
! pip install kenlm

Collecting kenlm
  Downloading https://files.pythonhosted.org/packages/57/54/0cc492b8d7aceb17a9164c6e6b9c9afc2c73706bb39324e8f6fa02f7134a/kenlm-0.tar.gz (1.4MB)
    1.5MB 7.2MB/s
Building wheels for collected packages: kenlm
  Building wheel for kenlm (setup.py) ... done
  Created wheel for kenlm: filename=kenlm-0.0.0-cp36-cp36m-linux_x86_64.whl size=2274268 sha256=725e43a22f6f5533a16a5d0442455e1c210d844b8a2f58eddcf8757122853b78
  Stored in directory: /root/.cache/pip/wheels/e9/cf/f4/1a1aab56f87f4132667a7a47045a750384f19d646099ab4858
Successfully built kenlm
Installing collected packages: kenlm
Successfully installed kenlm-0.0.0

[ ] import kenlm
    model = kenlm.Model("../../vietnamese_wiki_3-gram.binary")

[ ] sent = "xin chào các bạn"
    senti = "các bạn xin chào"
    print(model.score(sent))
    print(model.score(senti))

-11.640934944152832
-16.874675750732422
```