**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](#_Toc45989920)

[**2.** **Tạo bộ dữ liệu:** 4](#_Toc45989921)

[**3.** **Huấn luyện mô hình 3-gram với Kenml:** 5](#_Toc45989922)

[**3.1 Tải về và build kenlm:** 5](#_Toc45989923)

[**3.2** **Huấn luyện mô hình:** 6](#_Toc45989924)

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.

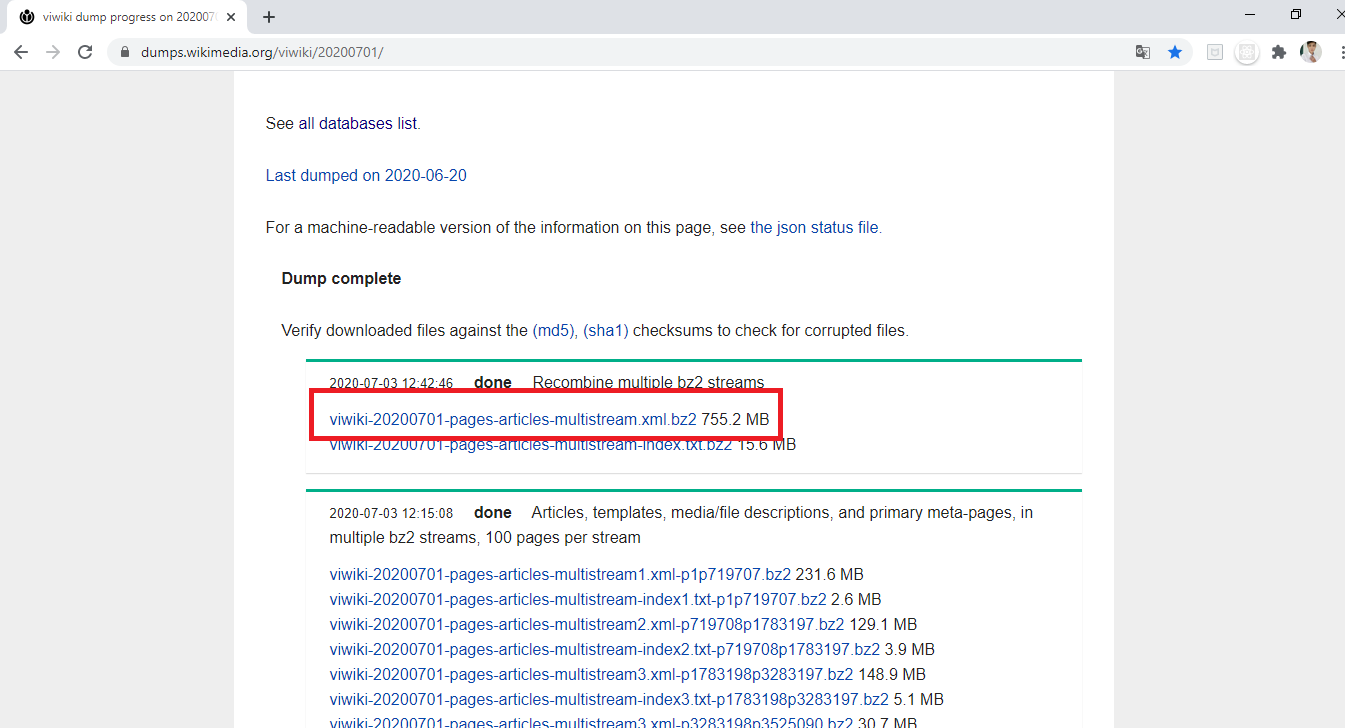
Sau đó huấn luyện Language model 3-gram với bộ dữ liệu này bằng Kenlm trên môi trường google colab.

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

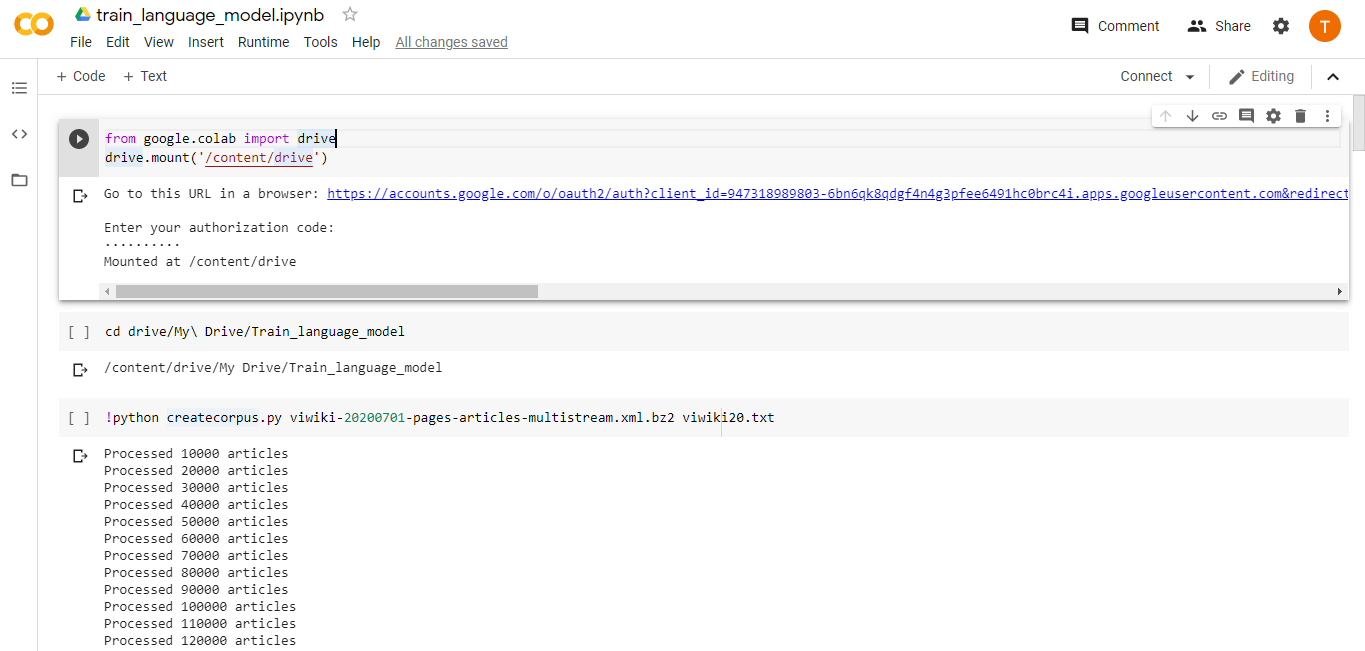
1. **Tạo bộ dữ liệu:**

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](https://drive.google.com/file/d/1oXvbBMZgpkWSlcgxv47CBRynV1A-FaYJ/view?usp=sharing).
* File kiểm tra dữ liệu sau khi được xử lí: [checkwikicorpus.py](https://drive.google.com/file/d/1kpUBbtyXN-qXHZlRNMpamc11Wv46n48p/view?usp=sharing)
* File tiền xử lí dữ liệu khi huấn luyện: [preprocess.py](https://drive.google.com/file/d/1DEBesIAIqsXG3AY-hWKi1LOh5nYhgXaY/view?usp=sharing)



Cd vào thư mục chứa file dữ liệu vừa tải về và chạy file cretecorpus.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)



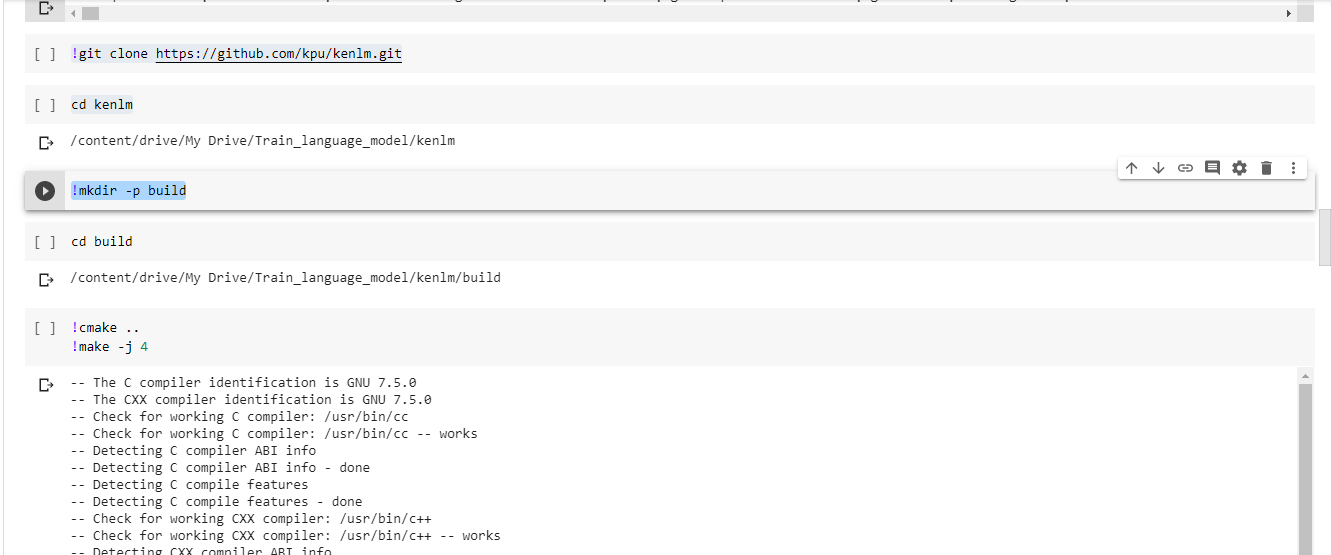
Sau khi ta nhận được file viwiki20.txt ta tiến hành kiểm tra một đoạn dữ liệu bên trong file thử bằng đoạn code sau



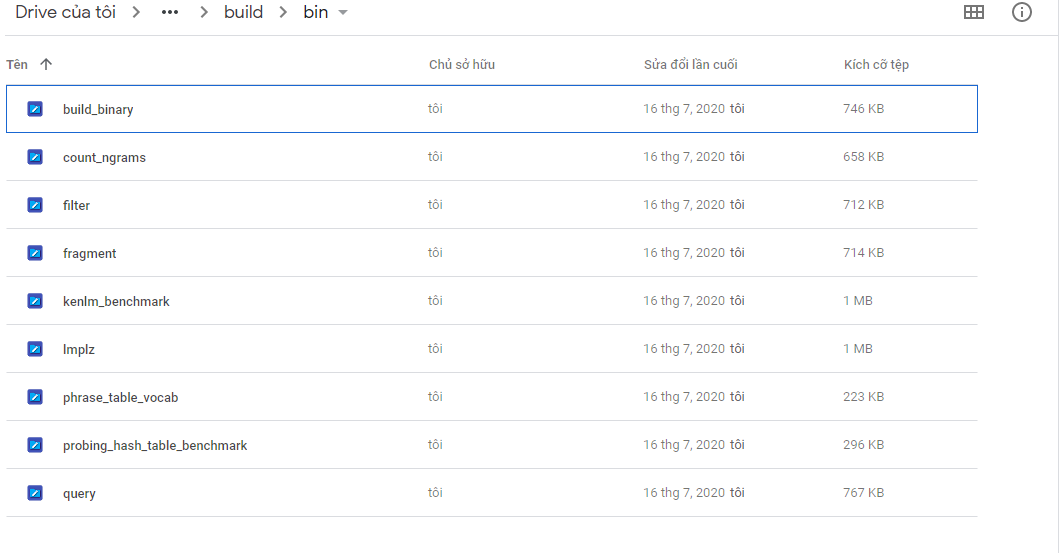
Dữ liệu đúng là các đoạn ngôn ngữ tiếng Việt được viết trên các dòng.

1. **Huấn luyện mô hình 3-gram với Kenml:**
   1. **Tải về và build kenlm:**

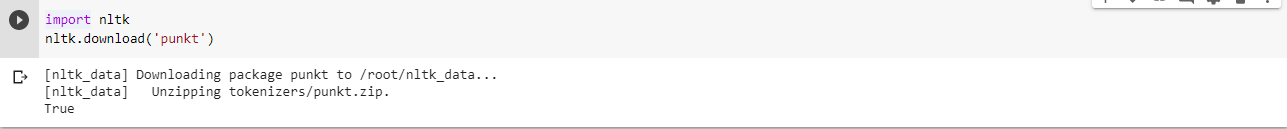
* Tiến hành chạy các đoạn code sau để clone và build kenlm



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



* 1. **Huấn luyện mô hình:**
* Import và tải về thư viện nltk.punkt



* Cấp quyền cho file lmp



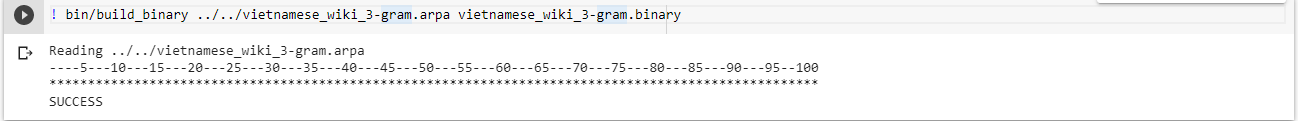
* Nén lại file viwiki20.txt (file dữ liệu đã được trích xuất từ wiki dump) theo định dạng .bz2.



* 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)



* 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.



* Kiểm thử language model

