Sau khi thu thập dữ liệu:

Bước 1: tải mysql

Bước 2: cài đặt các bảng:

Bảng 1: nhóm các kĩ năng lại

|  |
| --- |
| CREATE TABLE skill\_groups (  id INT AUTO\_INCREMENT PRIMARY KEY, -- Mã nhóm kỹ năng (tự tăng)  name VARCHAR(100) NOT NULL -- Tên nhóm (ví dụ: "Git", "Lập trình", "Kỹ năng mềm")  ); |

Bảng 2: Lưu từng kĩ năng cụ thể:

|  |
| --- |
| CREATE TABLE skills (  id INT AUTO\_INCREMENT PRIMARY KEY, -- Mã kỹ năng (tự tăng)  name VARCHAR(100) NOT NULL -- Tên kỹ năng (ví dụ: "Java", "React", "Teamwork")  ); |

Bảng 3: gán kĩ năng vào từng nhóm

|  |
| --- |
| CREATE TABLE skill\_details (  skill\_id INT NOT NULL, -- Mã kỹ năng  group\_id INT NOT NULL, -- Mã nhóm kỹ năng  PRIMARY KEY (skill\_id, group\_id),  FOREIGN KEY (skill\_id) REFERENCES skills(id) ON DELETE CASCADE,  FOREIGN KEY (group\_id) REFERENCES skill\_groups(id) ON DELETE CASCADE  ); |

Bảng 4: bảng công việc chính

|  |
| --- |
| CREATE TABLE jobs (  id INT AUTO\_INCREMENT PRIMARY KEY, -- Mã công việc (tự tăng)  title VARCHAR(255) NOT NULL, -- Tiêu đề công việc  link TEXT, -- Link chi tiết công việc  salary VARCHAR(100), -- Mức lương (ví dụ: "8-18 triệu", "Thỏa thuận")  location VARCHAR(255), -- Tỉnh/thành phố (ví dụ: "Hà Nội")  experience VARCHAR(100), -- Kinh nghiệm yêu cầu (ví dụ: "2 năm")  work\_location\_detail TEXT, -- Địa điểm làm việc cụ thể  working\_time VARCHAR(255), -- Thời gian làm việc (ví dụ: "T2-T6, 8:00-17:00")  deadline DATE -- Hạn nộp hồ sơ  ); |

Bảng 5: Thông tin mô tả chi tiết: mô tả, yêu cầu, phúc lợi.

|  |
| --- |
| CREATE TABLE job\_details (  id INT AUTO\_INCREMENT PRIMARY KEY, -- Mã chi tiết  job\_id INT NOT NULL, -- Mã công việc  description TEXT, -- Mô tả công việc  requirements TEXT, -- Yêu cầu công việc  benefits TEXT, -- Quyền lợi  FOREIGN KEY (job\_id) REFERENCES jobs(id) ON DELETE CASCADE  ); |

Bảng 6: Gán kỹ năng vào công việc (Job Skills)

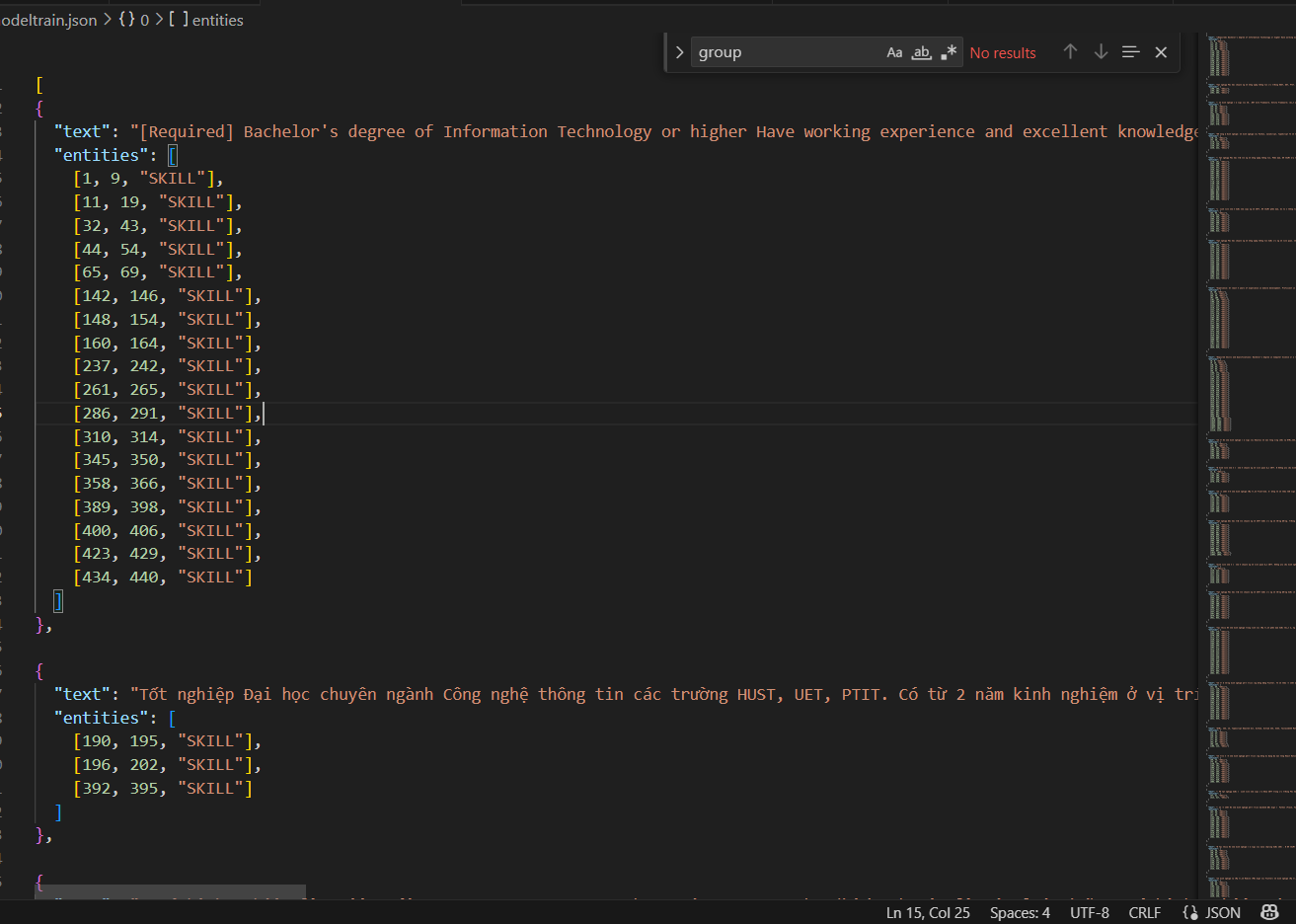
|  |
| --- |
| CREATE TABLE job\_skills (  job\_id INT NOT NULL, -- Mã công việc  skill\_id INT NOT NULL, -- Mã kỹ năng  PRIMARY KEY (job\_id, skill\_id),  FOREIGN KEY (job\_id) REFERENCES jobs(id) ON DELETE CASCADE,  FOREIGN KEY (skill\_id) REFERENCES skills(id) ON DELETE CASCADE  ); |

**Bước 3: làm sạch dữ liệu:**

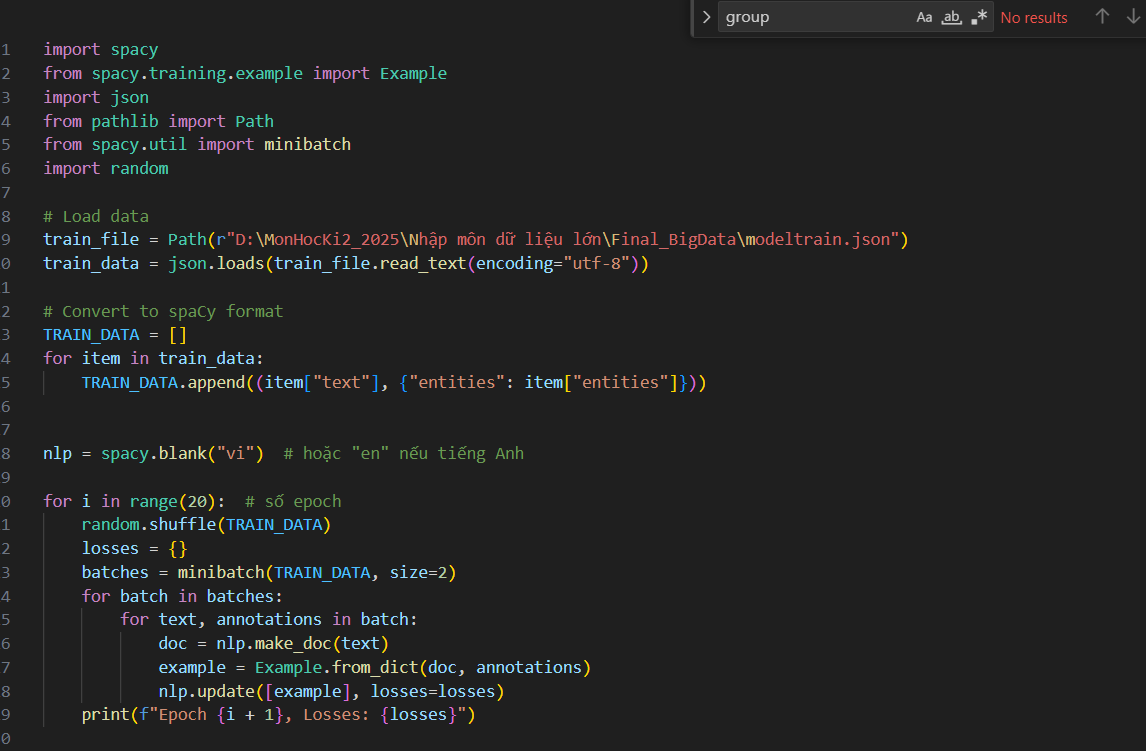
* train dữ liệu để lấy trường dữ liệu name cho bảng Skill. Bằng huấn luyện mô hình nhận diện thực thể (NER) tùy chỉnh với thư viện spaCy. Thì model NER là Nhận diện thực thể có tên là một kỹ thuật trong Xử lý ngôn ngữ tự nhiên (NLP), dùng để tìm và phân loại các cụm từ quan trọng trong văn bản.

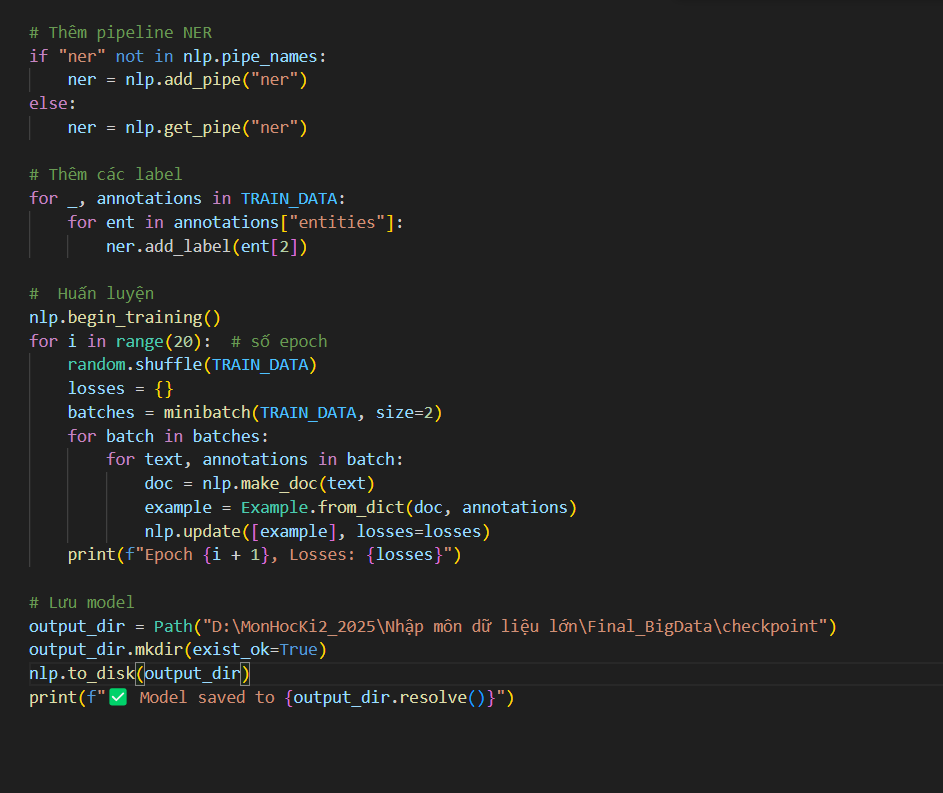
Đầu tiên: tạo file train gồm text, entities.

Có thể áp dụng trang web này để có thể chuyển label: <https://arunmozhi.in/ner-annotator/>



Bước 2: tạo thư mục python train\_model\_ner để có thể lấy dữ liệu cho bảng skill



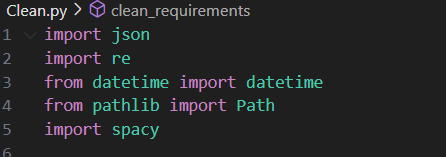


* cài đặt các thư viện cần thiết:

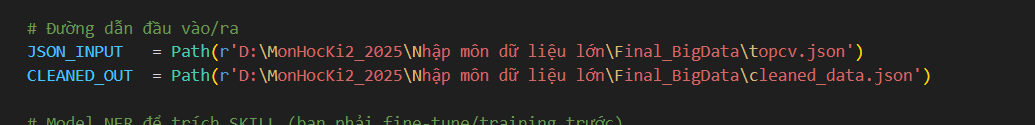
|  |
| --- |
| $ pip install spacy  $ pip install pyvi ==> giúp hỗ trợ tiếng việt |

Bước 3: clean dữ liệu:

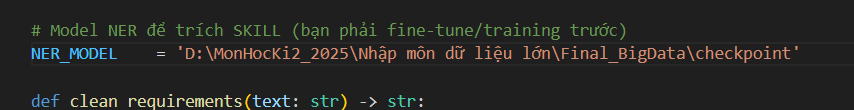
* import thư viện:



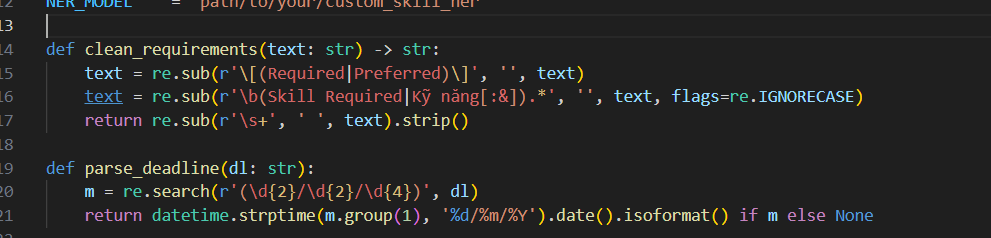
* đưa đường dẫn đầu vào:



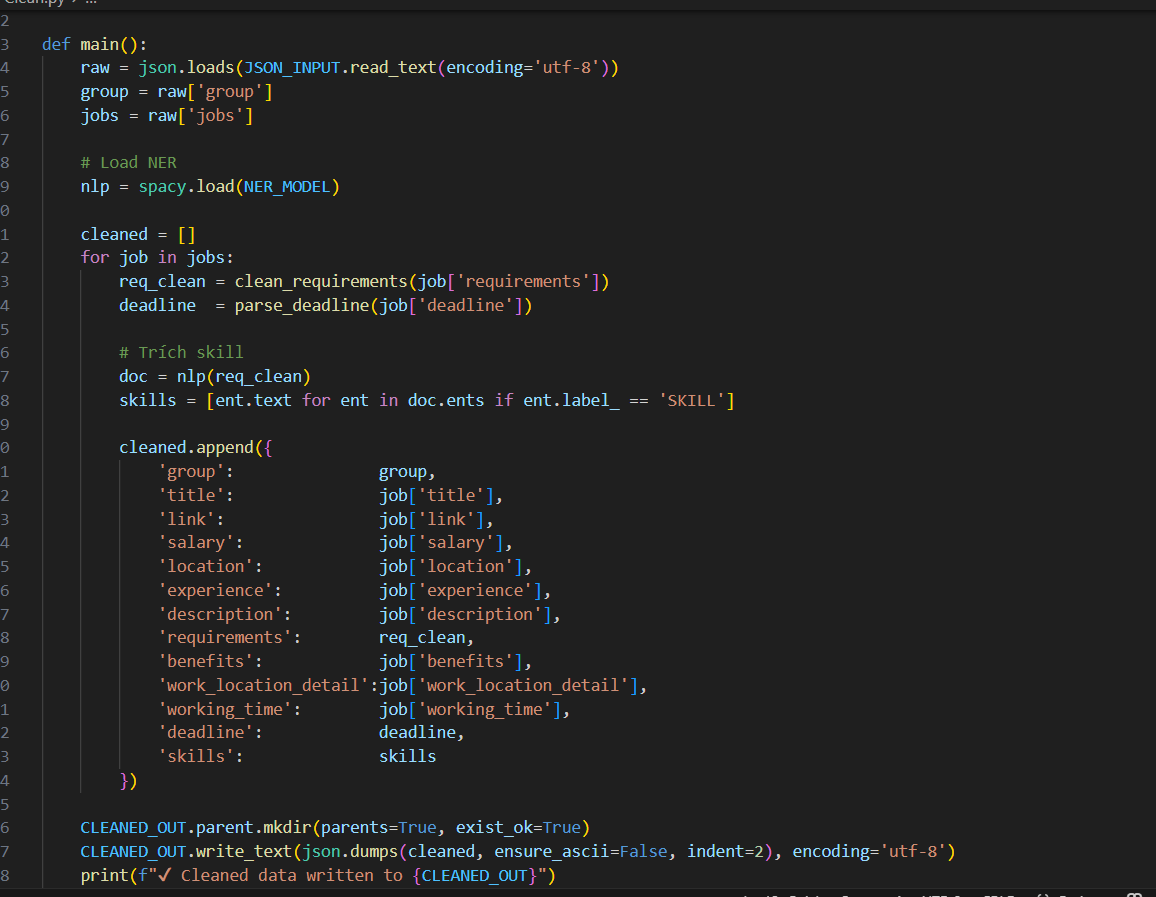
* đưa model đã huấn luyện vào mô hình:

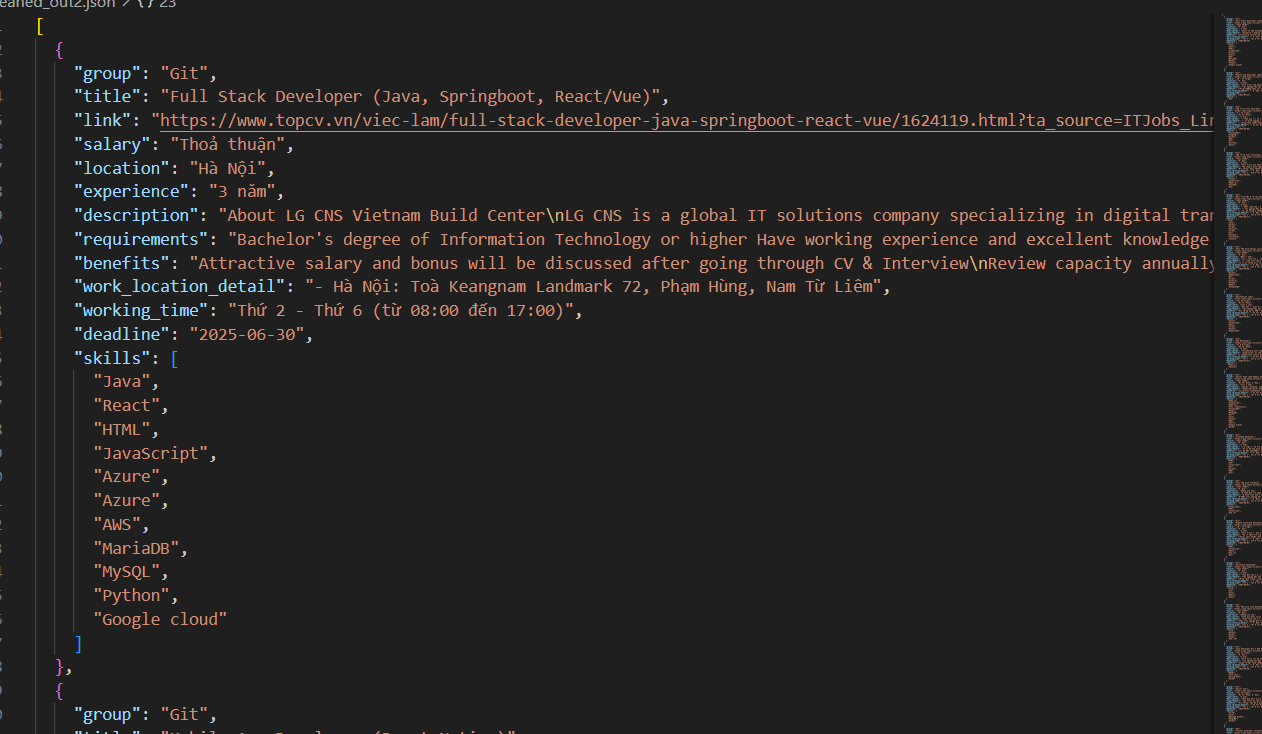


* làm sạch trường dữ liệu requirements, và ngày tháng:



* load model và đưa dữ liệu vào:

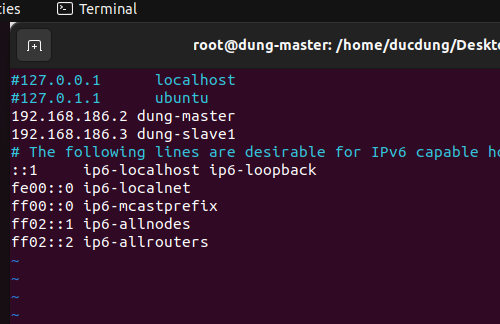


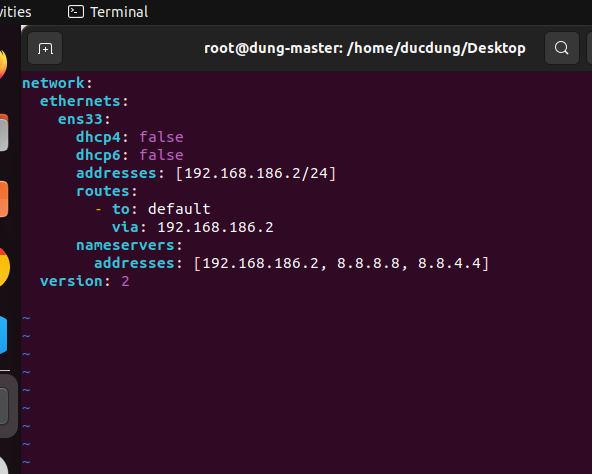
* dữ liệu mẫu làm sạch:  
  

Cấu hình kết nối Hadoop giữa máy Windows và Linux để trao đổi dữ liệu

Mục tiêu: Cần kết nối máy Windows với một cụm Hadoop trên Linux để trao đổi dữ liệu. Máy Windows dùng Java 1.8.0\_441 , máy Linux (Ubuntu) chạy Java 1.8.0\_442 và Hadoop 3.x. Mục tiêu là crawl dữ liệu trên Windows, lưu vào CSDL MySQL hoặc MongoDB, rồi đẩy lên sau đó từ hadoop ở ubuntu cũng có thể lấy dữ liệu từ hadoop của Window. Và ngược lại.

1. Cấu hình lại mạng thành :





1. kết nối mạng từ windows sang linux:

* nếu k ping được mạng!

 Vào **Control Panel** → **Windows Defender Firewall** → **Advanced settings**.

 Chọn **Inbound Rules** → tìm **File and Printer Sharing (Echo Request – ICMPv4-In)**.d

 Nháy đúp, trong tab **Advanced** chọn bật (Enable) cho profile **Private** (và Public nếu cần).

* sau bước đó mà vẫn k được !

tắt tưởng lửa tại win:

netsh advfirewall set allprofiles state off

* nếu ping thành công

b4: thiết lập file core-site.xml trong C:\HADOOP\hadoop-3.3.6\etc\hadoop

|  |
| --- |
| <configuration>    <property>      <name>fs.defaultFS</name>      <value>hdfs://dung-master:9000</value>    </property>  </configuration> |

Bước 5: Nếu muốn ghi lên file hadoop ở windows

1. **Trên Ubuntu**, mở file $HADOOP\_HOME/etc/hadoop/hdfs-site.xml, thêm vào trong <configuration>:

|  |
| --- |
| <property>  <name>dfs.permissions</name>  <value>false</value>  </property> |

1. **Restart HDFS**:

bash

Sao chépChỉnh sửa

$HADOOP\_HOME/sbin/stop-dfs.sh

$HADOOP\_HOME/sbin/start-dfs.sh

1. **Trên Windows**, bạn chỉ có thể trực tiếp đọc:

Nếu muốn ghi phải tạo file ở máy chính là master rồi cấp quyền :

hdfs dfs -chown Admin /win\_only ==> cấp quyền cho admin

hdfs dfs -chmod 700 /win\_only ==> cấp quyền cho user

hdfs dfs -chmod 755 /win\_only ==> cho phép mọi người đọc trên trang UI