* Lowercase: chuyển thành chữ thường
* Bỏ các ký tự đặc biệt như @ ; .... \n

VD: Tôi thích em @@@@ -> tôi thích em

* Gắn nhãn nó

VD:

* D. | N.V.D -> human

D. ăn kem -> human ăn kem

* Dup space
  + a b -> a b
* Dup chữ
  + chủ tịch chủ tịch chủ tịch -> chủ tịch
* word segment
  + học sinh học sinh họp -> học\_sinh học sinh\_học

data\_vi = ["Tôi là Phát", "Tôi là Đại"]

data\_vi = [“Tôi là Phát’’ , “Tôi là Đại’’]

data\_vi = [i.split(" ") for i in data\_vi]

temp = []

for i in data\_vi:

temp.append(i.split(" "))

data\_vi = temp

i = "Tôi là Phát"

i.split(" ") -> ["Tôi", "là", "Phát"]

temp = [["Tôi", "là", "Phát"]]

temp = [["Tôi", "là", "Phát"], ["Tôi", "là", "Đại"]]

Data\_vi = temp