1. **Hoàn thành bộ dữ liệu**

* **Đầy đủ số lượng**
* **Sạch**
* **Đủ số lượng**

1. **Tổng hợp dữ liệu**

* **Gộp lại**
* **Chia dữ liệu thành 10 folds ngẫu nhiên(đảm bảo tỉ lệ)**
* **Tạo thành 10 bộ(train, test) 10-fold cross-validation**

**Xây dựng mô hình phân loại dùng Native Bayes**

1. **Tiền xử lý dữ liệu(vnTagger, Vitk)**

* **Tách câu(mỗi câu trên 1 dòng)**
* **Tách từ, gán nhãn từ loại**
* **Loại bỏ những từ không mang nghĩa(., : “” ! @ .....)**

**VD:**

**Input: Hà Nội là thủ đô của Việt Nam**

**Output: Hà\_Nội/N là/V thủ\_đô của Việt\_Nam**

1. Xây dựng mô hình phân loại

* Xem lại lí thuyết về phương pháp Naïve Bayes(Xem slide bài giảng, Search Google,…)
* Code thuật toán