1. Word\_Segmentation

Sử dụng để tiền xử lý dữ liệu đầu vào

Các kỷ thuật sử dụng:

* Sử dụng hàm tokenize của thư viện pyvi để tách từ hoặc cụm từ trong câu
* Sau đó dùng một số kỷ thuật khác như:
  + xóa các ký tự đặc biệt
  + xóa các loại text đặc biệt (ngày tháng, email, sđt, số, url)
  + xóa các ký tự cảm xúc
  + xóa tag html
  + chuẩn hóa tiếng việt cho các từ

Input: đường dẫn thư mục chứa tất cả các file txt (input/corpus)

Output: file chưa dữ liệu đã tiền xử lý (output/data\_merge.txt)

1. Sentence2vec

Chức năng chính:

* Mô hình hóa các câu sang vector
* Tạo mô hình dữ liệu huấn luyện
* Chạy thực nghiệm bộ phân lớp

Quy trình thực hiện:

* Tạo mô hình huấn luyện
  + Sử dụng dữ liệu được thu thập từ 14375 văn bản với 27 lĩnh vực khác nhau từ bài báo “Tìm kiếm văn bản tiếng việt theo chủ đề” năm 2007 của nhóm tác giả Nguyễn Đình Thúc và Văn Chí Nam
  + Xử lý dữ liệu (tokenize)
  + Mô hình hóa sang mô hình không gian word2vec (thư viện gensim) với thông số
    - sg=1
    - size=100 (số chiều vector)
    - window=5
    - min\_count=5
    - workers=5
    - iter=10
  + Thu được model train sử dụng ánh xạ dữ liệu thực nghiệm sang sentence2vec
* Chạy thực nghiệm dữ liệu
  + Chọn phương pháp phân lớp (svm, naïve, tree)
  + Sử dụng kỷ thuật shuffle để sốc dữ liệu ngẫu nhiên (mỗi lần chạy)
  + Sử dụng kỷ thuật Kfold chia dữ liệu theo tỷ lệ 9:1, 9 phần dữ liệu huấn luyện và 1 phần dữ liệu test
  + Từ tập dữ liệu huấn luyện đã xác định chạy huấn luyện dữ liệu (gán nhãn)
  + Chạy tập dữ liệu test qua bộ huấn luyện vừa chạy để dự đoán nhãn dữ liệu test
  + Thu được kết quả thực nghiệm

**Kỷ thuật pipeline (không sử dụng model huấn luyện)**

Một *pipeline* trong *sklearn* là một tập các chuỗi thuật toán để trích xuất đặc trưng, tiền xử lý, chuyển hóa và huấn luyện dữ liệu sử dụng các thuật toán học máy cụ thể.

Dưới đây là cách bạn khởi tạo một pipeline trong *sklearn*

pipeline = Pipeline(steps=[

  ('vectorize', CountVectorizer()),

  ('classify', DecisionTreeClassifier())

])

Có nhiều lợi ích khi triển khai Pipeline:

* Tính tiện lợi và tính đóng gói: Chúng tôi gọi fitvà predictchỉ một lần trên dữ liệu để phù hợp với toàn bộ chuỗi các trình ước lượng.
* Lựa chọn tham số chung: Chúng tôi có thể thực hiện tìm kiếm lưới trên các tham số của tất cả các công cụ ước tính trong đường ống.
* Xác thực chéo: Các đường ống giúp tránh rò rỉ dữ liệu từ dữ liệu thử nghiệm vào mô hình được đào tạo trong quá trình xác nhận chéo. Điều này đạt được bằng cách đảm bảo rằng các mẫu giống nhau được sử dụng để huấn luyện máy biến áp và máy dự báo.