preprocess ข้อมูลดิบ.

ข้อมูลดิบ → อ่านไฟล์

sentiment	review
~~	positive
	negative

เปลี่ยน label positive → 1
negative → 0

sentiment	review
~~	1
	0

↓

lower case ตัวอักษร +
ลบ tag html

↓

ตัดคำด้วยช่องว่าง + ลบ stopword ใน eng

↓

lemmatization. ลดรูปคำเป็น base ของคำๆนั้น.

↓

เอาแต่ละคำมา join กลับเป็นประโยค.

↓

save ลงไฟล์ lemma_result.csv.

cleaned_review	Label
~~	1
	0

สร้าง model.

อ่านไฟล์ที่ผ่านการ preprocess → Tokenizer.

(lemma_result.csv) โดยกำหนดจำนวนคำที่มากสุด = max_features.

↓

fit_on_text.

แปลงข้อมูล "หลายๆประโยค" จากใน column 0 ให้เป็นคำๆ แยกด้วยเว้นวรรคและกำหนด index ให้แต่ละคำโดย "เรียงตามความถี่"

↓

text_to_sequences.

แปลงประโยคให้เป็น Sequence. จากนั้น. หาความยาวของประโยคและหาค่าสูงสุด.(maxlen) หลังจากนั้นเติม padding ให้ทุกประโยคมีความยาวเท่ากัน โดยเติม 0. ด้านหน้าหรือด้านหลังประโยค.

↓

นำข้อมูลมาแยกเป็นฝั่ง train และ test

3 : 1.

กำหนด random_state เป็นค่าคงที่เพื่อให้ random ได้เท่ากันเสมอ.

↓

สร้าง model โดยกำหนดให้

- Embedding
  - มี max-feature เป็น 7000
  - กำหนด dimension = 150
  - กำหนดจำนวน input node = ขนาดความยาวของ sentiment.
- Dropout แรก เป็น 0.2
- LSTM layer โดย.
  - dropout ภายในเป็น 0.2
  - recurrent dropout กรณีนำกลับไปดู output ก่อนหน้า = 0.2
  - hidden node เป็น 200
- Dropout ครั้งที่ 2 เป็น 0.2
- Dense unit เป็น 1 ตาม class output ให้ activation function เป็น 'sigmoid'