Phương pháp mã hóa dữ liệu đầu vào

* Tokenizer các từ
  + Tiếng việt underthesea
  + Yếu điểm
    - padding nhiều => giải pháp chia batch
  + Nên remove chấm, chấm phẩy
* Xây dựng embeding
  + Spare vector ( có bao nhiêu từ thì vector sẽ dài bấy nhiêu -> tạo thành vector thưa )
    - Tearm-document
    - Word-vector
    - Word-word
  + Danse vector
    - Embeding tĩnh
      * Word2vec
        + CBOW
        + Skip gram
    - Embeding động
      * RNN
      * HỌ lstm
        + Elmo
      * Họ transformer
        + Bert
        + GPT