Computational linguistic : ngôn ngữ học / HCI / CV

NLP : ứng dụng của CL

LLM : mô hình ngôn ngữ lớn

CL -> NLP, info retrieval (trích xuất ttin), DL, neural network,

ứng dụng : phân tích ttin, chatbot, phân tích ngữ nghĩa,

2. word segmentation [pyvi]

- phân tách từ trong câu

- text ambiguity:

3. preprocessing

- tokenization : subword tokenization : tách từng phần : mô hình xác suất - > đoán từ tiếp theo

4. lib

- import nltk

- spaCy

Hugging face tokenize

5. stop word : từ bỏ qua

- tuỳ trg hợp

filtered\_words = [token.text for token in doc if not token.is\_stop]

6. lemmatization : loại bỏ từ theo từ điển

Chuyển phức tạp -> đơn giản

Stemming : cắt bớt để chuyển về từ tương tự

Rule-based : rule có sẵn

Dic-based : tra từ điển

ml-based : từ train data

7. feature extraction

Convert thành các bag of words

Df = N

Idf(t) = len(N)/df(t)

Log2(idf(t))

Tf-idf = tf \* idf

Text -> vector -> khoảng cách -> vector 1-1 -> return text