Lesson 1-7

Files

Tags

Lesson 1. Welcome To Term 2

(내용 없음)

Lesson 2. Intro to Natural Language Processing

structured vs unstructured language

컴퓨터를 전공으로 하는 사람들은 컴파일러 또는 프로그래밍언어 수업에서 이 구조화된 언어 이런 종류의 엄격한 문법을 갖는 언어는

- 명령하는 목적: 프로그래밍언어, SQL
- 데이터를 표현하는 목적: HTML, XML, JSON, YAML 등등

비구조화된 텍스트

- 우리가 텍스트라고 생각하는 텍스트
- context 맥락, structured language 처럼 한번에 하나만 해석되지 않음 (참고로 structured language라고 자유로운건 아님)

NLP pipeline

- text processing: 수집, 정제, 전처리
  - HTML vs plain text
  - 텍스트가 텍스트만 있으면 행복하지...
  - web, 전자문서, 글은 있는데 텍스트는 아닌거, 음성, 테이블같은거
  - 간혹 쓸데없는 단어들

- feature extraction: 벡터로 변환
  - ascii 는 "언어"로서의 수치를 가지고 있지 않음 (rgb vs 유니코드/단어ID)
  - 의미를 가지도록 텍스트를 수치화 시켜져야 하고
  - 이 수치 조차도 목적에 따라서 다르게 사용되어ㅑ ㅇ 함

- modelling: 유용한 결과를 뱉어내는 모델 생성
  - 데이터 넣으면 → 뭔가 멋진 모델이 뭔가 해서 → 유용한 결과 출력
  - 모델이 기능하려면 데이터를 넣어서 fitting 또는 "학습"하는 작업이 필요

Lesson 3. Text processing

- cleanse
  - beautiful soup
  - lxml
  -
  - re module
- normalization
  - upper to lower
- stop words removal
- NER, POS tagging, tokenization, lemmatization
  - NLTK... 여전히 많이 쓰는 도구
  - NER : 의미적 요소(장소 ... etc)
  - POS tagging : 명사, 동사, 대명사...
  - tokenization : "ice americano" or "ice" + "americano"
  - lemmatization : am, are, is... → 동사

용어들에 대한 정리

Lesson 4. Feature Extraction

- Feature extraction
- Bag of words
- One hot encoding
- Word embeddings
- Word2vec
- GloVe: word2vec 알고리즘들이 국지성이 너무 강해서 글로벌 맥락을 도입하려는 시도
- Embeddings for deep learning
  - 번역기든, 감정분석이든, 뉴럴넷에 들어가는 단위는 임베딩들의 연결
- t-SNE
  - 강력한 시각화 알고리즘

Lesson 5. Financial Statements

- 재무재표는... Dart가서 보시고요. 원본 말고 필요한 데이터 가져오는 방법도 있음
- regex 체커
  - 전화번호
  - 이메일
  - url
  - html 태그 제거
  - 보면 알겠지만 regex가 통하는 영역은 대체로 웹문서/전자문서
- html을 쓸 수 있는 상황이면 BS를 쓰는게 더 이로울지도
- requests: 정적페이지 크롤링
- selenium: 동적페이지 크롤링