

DSTC2 Vector Table

2015.11.16

실행방법

- 설정사항
 - text_analyzer.py 파일 상단에 있는 경로를 수정하여야 함
 - DATA_FILE : data-train.json 파일의 경로
 - VECTOR_FILE : glove.6B.50d.txt 파일의 경로
- 실행방법
 - python3 text_analyzer.py
 - (python 2 일 경우, python text_analyzer.py)

TextAnalyzer::txt2vectors

- 영어 텍스트 문장을 입력으로 하여 vector table에 있는 값을 받아옴
- 입력
 - `ta = TextAnalyzer(VECTOR_FILE)`
`result = ta.txt2vectors("I'm a sogang university student.")`
- 출력
 - `result => dictionary 타입`
 - key : word
value : numpy array
 - 예 : `[('i', array(1.18910000e-01, 1.52550000e-01, -8.20730000e-02,...)), ('m', array(...)), ('a', array(...)), ('sogang', array(...)), ...]`

GloVe(global vector)

- Stanford에서 제안한 GloVe를 사용하여 텍스트를 훈련할 수 있고 훈련된 결과로 Vector table을 얻음
 - 참고링크 : <http://nlp.stanford.edu/projects/glove/>
- glove.6B.50d.txt
 - Wikipedia 2014 + Gigaword 5로 훈련한 50 dimensions의 vector 테이블(6-Billion tokens으로 훈련)
 - 다운로드 : <http://163.239.169.54/glove.6B.50d.txt.zip>