Squad Dataset

노트북:	BIDAF		
만든 날짜:	2019-07-14 오후 5:41	**업데이트:**	2019-07-14 오후 10:08
작성자:	Chungsun Jeong		
URL:	http://localhost:8888/notebooks/data/squad/Untitled.ipynb		

- **요약**
- json 파일(data)
    - keys: version, data
    - data['data']
        - 여러 개의 글들 (lists)
        - data['data'][0]
            - 하나의 글에 대한 하나의 주제와 여러 개의 단락들
            - keys: title, paragraphs
            - data['data'][0]['paragraphs']
                - 여러 개의 단락들 (lists)
                - data['data'][0]['paragraphs'][0]
                    - 하나의 단락에 대한 질문들과 답변 (QAs)
                    - keys: 'context', 'qas'
                    - data['data'][0]['paragraphs'][0]['qas']
                        - 여러 개의 QAs (lists)
                        - data['data'][0]['paragraphs'][0]['qas'][0]
                            - 하나의 QA에 대한 하나의 질문과 여러개의 답변들, 그리고 해당 QA에 대한 identifier
                            - keys: 'question', 'answers', 'id'
                            - data['data'][0]['paragraphs'][0]['qas'][0]['answers']
                                - 여러 개의 답변들 (lists)
                                - data['data'][0]['paragraphs'][0]['qas'][0]['answers'][0]
                                    - 하나의 질문에 대한 가능한 여러 개의 답변들
                                    - train dataset은 1질문 1답변
                                    - test dataset은 1질문 복수 답변
                                    - keys: 'answer_start', 'text'
                                    - 'text': 질문에 대한 답변 내용
                                    - 'answer_start': 해당하는 'context' 내용에서 질문의 답이 시작하는 텍스트의 위치

```
{'title': 'Super_Bowl_50',
 'paragraphs': [{'context': 'Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for t
he 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Caroli
na Panthers 24-10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay A
rea at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initi
atives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have be
en known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.',
   'qas': [{'answers': [{'answer_start': 177, 'text': 'Denver Broncos'},
     {'answer_start': 177, 'text': 'Denver Broncos'},
     {'answer_start': 177, 'text': 'Denver Broncos'}],
    'question': 'Which NFL team represented the AFC at Super Bowl 50?',
    'id': '56be4db0acb8001400a502ec'},
   {'answers': [{'answer_start': 249, 'text': 'Carolina Panthers'},
     {'answer_start': 249, 'text': 'Carolina Panthers'},
     {'answer_start': 249, 'text': 'Carolina Panthers'}],
    'question': 'Which NFL team represented the NFC at Super Bowl 50?',
    'id': '56be4db0acb8001400a502ed'},
   {'answers': [{'answer_start': 403, 'text': 'Santa Clara, California'},
     {'answer_start': 355, 'text': "Levi's Stadium"},
     {'answer_start': 355,
      'text': "Levi's Stadium in the San Francisco Bay Area at Santa Clara, California."}],
    'question': 'Where did Super Bowl 50 take place?',
    'id': '56be4db0acb8001400a502ee'},
```

- **세부**
- json 파일구조 (dev-v1.1.json 기준)
    - dict 구조
    - keys(): dict_keys(['data', 'version'])
    - data['data'] 부분만 보면 된다.
- data['data']
    - list 구조
    - 여러 개의 주제에 해당하는 'paragraph' 들의 dict로 이루어진 list
    - 주제의 총 개수: 48
    - data['data'][0:47] 까지 있음
    - 주제 하나 당 하나의 dict 를 가진다. (data['data'][0])
- data['data'][0]
    - dict 구조
    - keys(): dict_keys(['title', 'paragraphs'])
    - 하나의 주제 ('title') 와 주제에 해당하는 여러개의 'paragraphs' 로 이루어져있다.
    - 'paragraphs'는 글 하나를 여러 개의 단락으로 분리한거라 생각하면 된다.
- data['data'][0]['paragraphs']
    - list 구조
    - 즉, 'paragraphs'의 문장들이 모여서 하나의 주제를 가진 글을 이룬다.
    - 각 주제마다 'paragraphs' 의 수는 다르다.
    - 하나의 글에 대한 하나의 'paragraphs'의 구조를 보면 아래와 같다.
- data['data'][0]['paragraphs'][0]
    - keys(): dict_keys(['context', 'qas'])
    - 'context': paragrapgh 글 내용
    - 'qas': 해당하는 paragraph 에서 답을 얻을 수 있는 질문들
    - 하나의 paragraph 에 여러개의 'qas' 를 가진다.
- data['data'][0]['paragraphs'][0]['qas']
    - list 구조
    - 하나의 paragraph 내용에 대해 여러 개의 질문들을 할 수 있을 것이다.
- data['data'][0]['paragraphs'][0]['qas'][0]
    - 하나의 'qas' 에 대한 내용
    - dict 구조
    - keys(): dict_keys(['answers', 'question', 'id'])
    - 'answers': 주어진 paragraph 내용을 읽고 주어진 질문에 대해 대답 가능한 정답 셋
    - 'question': 주어진 paragraph 내용을 읽고 대답 가능한 질문
    - 'id': identifier
    - 하나의 질문에 여러 개의 정답이 있을 수 있다. (test dataset 한정)
    - train dataset은 1질문 1 답변
    - test dataset은 1질문 복수 답변 (count distrubtion: {1: 3, 2: 136, 3: 8490, 4: 759, 5: 1147, 6: 35})
    - test dataset에서는 똑같은 정답으로 데이터를 구축해놓은 것도 있다. (아래 그림들 참조)
    - 이 'qas'의 answers이 타겟 데이터셋일 것 같다.

- 하나의 paragraph 내용에 대한 하나의 'qas' 은 다음과 같다. (두 가지 예시)
  - ```
    data['data'][2]['paragraphs'][1]['qas'][0]
    ```

    ```
    {'answers': [{'answer_start': 1022, 'text': 'William the Conqueror'},
      {'answer_start': 1022, 'text': 'William the Conqueror'},
      {'answer_start': 1022, 'text': 'William the Conqueror'}],
     'question': 'Who was the duke in the battle of Hastings?',
     'id': '56dddf4066d3e219004dad5f'}
    ```
  - ```
    data['data'][2]['paragraphs'][0]['qas'][0]
    ```

    ```
    {'answers': [{'answer_start': 159, 'text': 'France'},
      {'answer_start': 159, 'text': 'France'},
      {'answer_start': 159, 'text': 'France'},
      {'answer_start': 159, 'text': 'France'}],
     'question': 'In what country is Normandy located?',
     'id': '56ddde6b9a695914005b9628'}
    ```
- data['data'][0]['paragraphs'][0]['qas'][0]['answers']
  - 하나의 질문에 대한 여러 개의 대답 가능한 답변들
  - 각 답변은 'answer_start' 와 'text' 키들을 가진다.
  - 'text': 질문에 대한 답변 내용
  - 'answer_start': 해당하는 'context' 내용에서 질문의 답이 시작하는 텍스트의 위치