Automatic Generation of Baseball Articles

김동현, 김용현, 신승우, 이경수

Team Noname

Saturday 15th August, 2015

I. Main Objective

야구 경기 기록에서 야구 기사를 자동으로 작성하는 코드를 짜는 것. 야구 경기 기록은 kbo 홈페이지[1]와 네이버 문자중 계 에서 얻어온다.

II. General Structure

I. Pipeline of Program

1. 데이터 수집

   웹 크롤링을 통해서 기존 야구 경기의 기록과 관련 기사를 수집한다.

2. 학습

   경기의 기록을 분석해서 어떤 부분이 기사에 주로 나오는지를 학습한다.

3. 기사 작성

   기사에 나올 부분을 자연어처리를 통해서 기사로 만들어낸다.

II. Program Structure

II.1 경기 구조

야구 경기는 1구마다, 혹은 교체 등의 예외적인 상황마다 state가 변한다. 이 state가 변하는 과정을 event라고 한다. 따라서 야구 경기는 state들이 변화하는 과정이다. 맨 처음 state는 선발 라인업을 반영해야 하고, 그 이후로는 일어난 event에 따라서 state가 바뀌게 된다. 이 과정을 track함으로써 야구 경기에서 일어난 일을 재현할 수 있다.

State는 다음의 정보를 담고 있다.

- Offense/Defencse Team : 공/수 팀의 포지션별 선수 목록
- ball, out, strike : 각각 볼, 아웃, 스트라이크 카운트
- turn, is_home : 회차와 초/말.

Event의 종류는 아래와 같이 분류된다.

event가 일어날 시에 그 event에 의힌 선수들의 stat 변화량을 계산해서 기록한다. stat들은 크게 수비와 공격 stat으로 나눠서 판단한다. 공격 stat은 wpa를 우선 사용하며, wpa이외의 스탯은 일단 사용하지 않는다. 수비 스탯의 경우는 subgame의 단위로 게임을 나누어서, 각 subgame을 설명하는 지표를 아래와 같이 정한다. [2]

[1] http://www.koreabaseball.com/Schedule/GameList/General.aspx

[2] 스탯 정확히 정해지면 수정 요

Table 1: *Event class*

hitted	fair	hit	single	1루타
			double	2루타
			triple	3루타
			homerun	홈런
			error	실책
		out	singleout	아웃
			doubleplay	병살
			tripleplay	삼중살
			sacrifice fly	희생 플라이
			sacrifice bunt	희생 번트
	foul			파울
Non-hitted	strike	strike		스트라이크
		swing		헛스윙
		strike out		스트라이크 아웃
	ball	ball		볼
		wildpitch		폭투
		hit by pitch		데드볼
		base on balls		사구
steal	safe			스틸 성공
	double steal			더블스틸
	out			도루사
balk				보크
change	PH			대타
	PR			대주자
	Pitcher Change			투수 교체
	Fielder Change			야수 교체
	Penalty			퇴장
Team Change				공수 교대
Bench change				?
Bench Clear				벤치 클리어
hit interference				타격 방해

II.2 야구 기사

야구 기사는 경기에서 일어난 event들 중 특정 event를 기자가 취사선택하여 배열 후 서술한 것이다. 따라서 이 과정을 기계가 대신하게 하는 것으로 기사 작성을 할 수 있다. 이를 위해서는 기존 기사에서 어떤 event를 어떤 순서로 배열하여 작성한지를 알 수 있어야 한다. [3]

III. Details

작업 세부사항.

I. Development Environment

- 언어 : python3, python2
- 필요 모듈 : BeautifulSoup4, KoNLPy(예정), Numpy/Scipy/Matplotlib(예정)
- remote server : 157.7.108.38
- 소스 버젼관리 : 미정
- 소스 공유 툴 : 미정[4]
- 자료 공유 : slack, 카카오톡

II. Build Environment

- 언어 : python2
- 필요 모듈 : BeautifulSoup4, Selenium
- 기타 설치 요 : ChromeDriver

III. Coding Convention

기본적으로 PEP에 나와있는 사항을들 따르되, 아래 사항은 필수적으로 지킬 것.

- Naming
  - 클래스 이름 : 첫 문자 대문자. 이후 단어 단위로 대문자. ex) ExampleClass
  - 함수 이름 : 첫 문자 소문자. 이후 단어 단위로 대문자.[5] ex) exampleFunc
  - 변수 이름 : 모든 문자 소문자. 단어 단위로 _. ex)example_var
  - 전역 변수 : 모든 문자 대문자. 단어 단위로 _. ex)GLOBAL_VAR
- Comment
  모든 주석은 영어로, 아래의 형식으로 달면 됨.
  - 클래스

```
'''
class ExampleClass : class explanation
       att1 : att1 explanation
       att2 : att2 explanation
...

'''
class ExampleClass():
       def __init__(self, att1, att2. ...):
              ....
```

[3] 이 부분에 대한 설계는 아직 정확히 진행하지 않았음. 추가 요망.
[4] 아마 github
[5] Camel Notation

- 함수

```
'''
exampleFunc : type(var1), type(var2), ... -> type(res1), type(res2), ...
        function explanation
        var1 : var1 explanation
        var2 : var2 explanation
        ...
        res1 : res1 explanation
        res2 : res2 explanation
        ...
'''
def exampleFunc(var1, var2, ...):
        ...
```

필요한 경우, 함수 내부에 local variable을 설명하는 주석을 달아주면 좋다.

- 전역변수
전역변수 정의 위에 설명 comment 한 줄.

```
# GLOBAL_VARIABLE : explanation
GLOBAL_VARIABLE = ...
```

IV. Implementation

별지 첨부.

IV. 기타

I. Team Members

모든 사람 명단은 가나다순으로 쓰여졌음.

- 김동현
서버세팅, 코딩 담당.
- 김용현
코딩, 야구 관련 지식 담당.
- 신승우
문서화, 코딩 담당.
- 이경수
문서화, 야구 관련 지식 담당.

II. Project History

2015.07.04 11:00-4:00

- 참가자 : 김동현, 신승우, 이경수
- 한 일 : 전반적인 아키텍쳐 논의. 개발환경 협의. 서버 세팅.
- 합의사항 : 모임장소는 신촌 cafe.blog로 정함. 모임 시간은 매주 토요일 12:00까지 오기로.
- 기타 : 없음.

2015.07.11 11:00-4:00

- 참가자 : 김동현, 신승우, 이경수
- 한 일 : 할 일 분배, pseudo-code implement하기.
- 합의사항 : 스케쥴 정리 - 25일까지. 다음주 미팅 없음.
- 각자 할 일 (due 25일)
  - 김동현 : 크롤러 짜보기
  - 김용현 : 파서 설계해서 pseudocode로 가져올것
  - 신승우 : 파서 설계/evalState 함수 짜오기/eventClass 짜오기
  - 이경수 : 트레이드 정보 포함한 player list 만들기. 각자에게 배분해줄 것.
- 기타 : 없음.

2015.07.26 10:00-6:00

- 참가자 : 김동현, 신승우, 이경수, 김용현
- 한 일 : crawler 작성, 스탯벡터 만드는 전처리
- 합의사항 : Publish 방법 합의 - 후속결정사항 기록 필.
- 각자 할 일 (due 1일)
  - 김동현 : (야구 기사 -> eventClass) Parser 설계 / 크롤러 디버깅
  - 김용현 : (문자중계 -> eventClass) Parser 설계해서 pseudocode로 가져올것 / 수비스탯 합의해서 가져오기
  - 신승우 : (야구 기사 -> eventClass) Parser 설계/event handler 마무리 / 크롤러 디버깅 / 소스트리
  - 이경수 : 수비스탯 합의해서 가져오기
- 기타 : 없음.

2015.07.27 - publish 후속관련사항

- 참가자 : 외부회의
- 합의사항
  - 일정 관련 : 2주 안에 1차적인 학습, 프로토타입 만들어보기. 리그 끝나기 전에 간단한 기사라도 만들어볼 것.
  - 저작권, 수익 : 공동분배. 비율은 차후 합의. 초창기 수익모델은 정해지지 않았음.
  - 기타 : 2주 후 회의.