정규표현식

표현식 (Expressions)?

- 프로그래밍 언어에서 변수와 상수, 그리고 연산자로 이루어진 식
- 3 + 7, score * 10

정규표현식 (Regular expressions)

`문자열을 취급하는 강력한 도구`

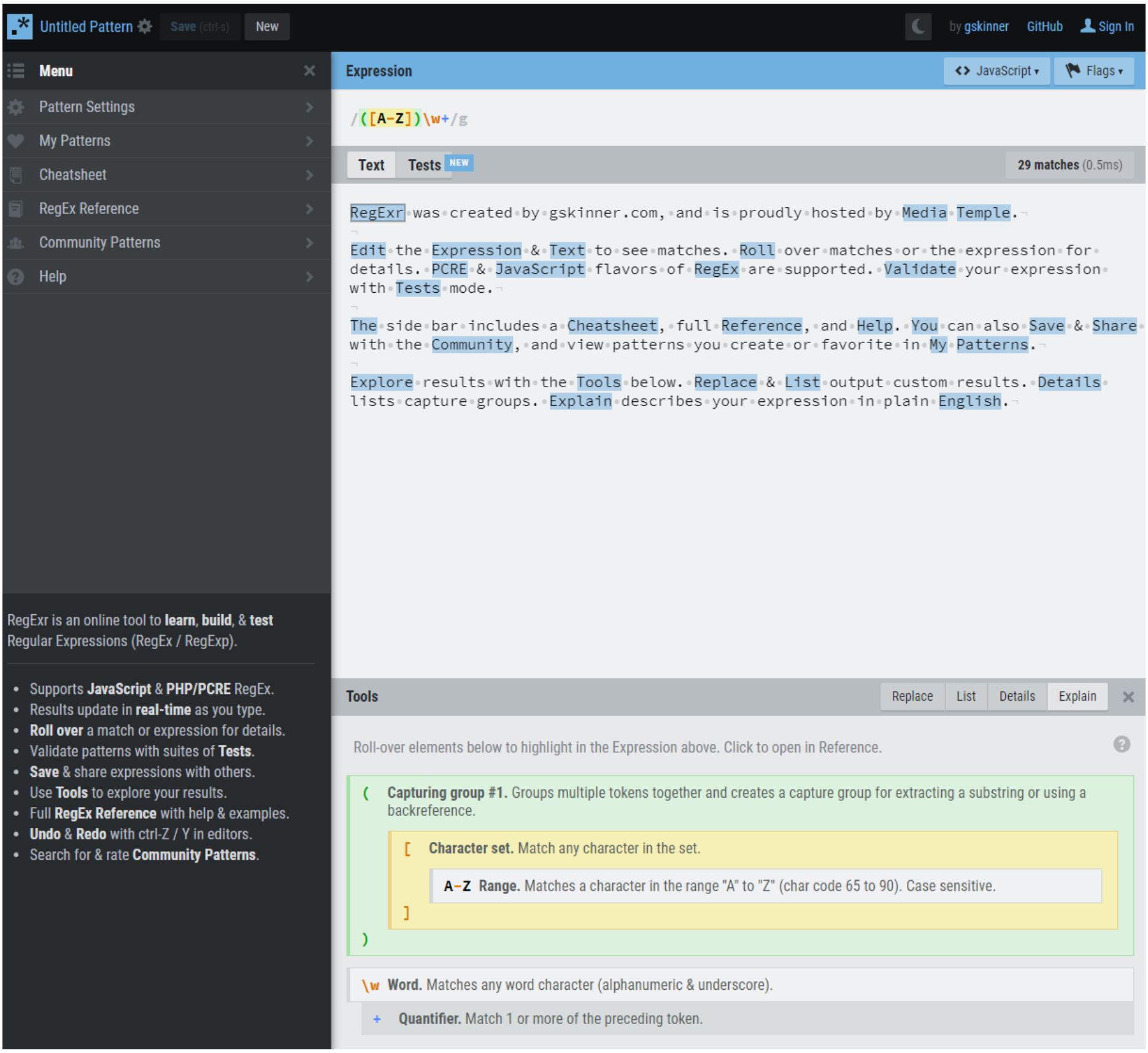
- 완전한 프로그래밍 언어가 아님
- 다른 프로그래밍 언어나 제품에 포함된 `작은 언어` 관점
- 제품마다 상이한 문법
- 텍스트를 찾고 조작하는 데 쓰는 문자열
- 텍스트 검색, 치환에 사용
- 수십 라인의 프로그래밍 없이 정규식 1~2줄로 대부분의 문자열 작업 가능

정규표현식의 사용

- email, 주민번호, 생년월일 등의 형식 검증 (validation)
- 데이터 전처리 작업
- 프로젝트 리팩토링 작업
- Database 검색, 치환 작업
- IDE, editor, 심지어 MS word도 지원

Regex online test

https://regexr.com/



문자 하나 찾기

- 일반적인 문자 그대로 기재
- `.`와 같은 메타문자를 검색하려면 `\`로 이스케이프

문자 집합으로 찾기

- 대괄호를 `[]` 사용하여 문자 집합 표현
- `[]` 집합에 속한 문자 가운데 하나가 일치
- `[]` 내에 `-`은 연속 요소를 표현 `[1-5]` => `[12345]`
- 캐럿 `^` 문자는 집합 안에 있는 문자나 범위를 모두 제외

메타 문자 사용

표 1. 메타 문자

메타 문자	설명	예제
*	0개 이상의 선행하는 atom 인스턴스를 일치시킵니다. 가능한 많은 수의 인스턴스를 일치시킵니다.	goo*는 my godness, my goodness, my goooodness와 일치하지만 my gdness와는 일치하지 않습니다.
+	1개 이상의 선행하는 atom 인스턴스를 일치시킵니다. 가능한 많은 수의 인스턴스를 일치시킵니다.	goo+는 my goodness 및 my goooodness와 일치하지만 my godness와는 일치하지 않습니다.
?	0개 이상의 선행하는 atom 인스턴스를 일치시킵니다.	goo?는 my godness, my goodness, my goooodness와 일치하지만 my gdness와는 일치하지 않습니다. colou?r는 color 및 colour와 일치합니다. end-?user는 enduser 및 end-user와 일치합니다.
$	문자열의 끝을 일치시킵니다.	end$는 the end와 일치하지만 the ending과는 일치하지 않습니다.
^	문자열의 시작을 일치시킵니다. ^ 메타 문자는 대괄호 표현식에도 사용할 수 있습니다.	^severity는 severity level 5와 일치하지만 The severity is 5와는 일치하지 않습니다.
.	단일 문자를 일치시킵니다.	b.at은 baat, bBat, b4at와 일치하지만 bat 또는 bB4at과는 일치하지 않습니다.
()	소괄호 내의 문자를 문자 패턴으로 간주해야 함을 나타냅니다.	A(boo)+Z는 AbooZ, AboobooZ, AbooboobooZ와 일치하지만 AboZ 또는 AboooZ와는 일치하지 않습니다. Jan(uary)?는 Jan 및 January와 일치합니다.
|	파이프 문자의 한 쪽에 있는 atom 중 하나를 일치시킵니다.	A(B|C)D는 ABD 및 ACD와 일치하지만 AD, ABCD, ABBD 또는 ACCD와는 일치하지 않습니다. (AB | CD)는 AB 및 CD와 일치하지만 ABD 및 ACD와는 일치하지 않습니다.
\	뒤에 오는 메타 문자를 일반 문자로 간주해야 함을 나타냅니다. 이 절에 나열된 메타 문자의 특별한 의미를 설정 해제하기 위해 접두어로 백슬래시 문자가 필요합니다. \ 메타 문자는 백슬래시 시퀀스를 구성하는 데에도 사용할 수 있습니다.	*는 * 문자와 일치합니다. \\는 \ 문자와 일치합니다. \. 은 . 문자와 일치합니다.
{m , n}	m개부터 n개까지의 선행하는 atom 인스턴스를 일치시킵니다. 여기서, m은 최소값이고 n은 최대값입니다. 가능한 많은 수의 인스턴스를 일치시킵니다. **참고:** m과 n은 0 - 255 범위의 부호 없는 10진수 정수입니다.	f{1,2}ord는 ford 및 fford와 일치합니다. N/{1,3}A는 N/A, N//A, N///A와 일치하지만 NA 또는 N////A와는 일치하지 않습니다.
{m ,}	m개 이상의 선행하는 atom 인스턴스를 일치시킵니다.	Z{2,}는 2개 이상의 Z 반복을 일치시킵니다.
{m}	정확히 m개의 선행하는 atom 인스턴스를 일치시킵니다.	a{3}은 aaa와 일치합니다. 1{2}는 11과 일치합니다.

반복 찾기

- 파워풀한 정규 표현 패턴의 능력
- `+`: 하나 이상 일치
- : 없거나 하나 이상 일치
- `?`: 없거나 하나 일치
- 중괄호 `{}` 내에 반복 횟수 기재 `{3}`: 3번
- 게으른 수량자로 문자를 최소로 일치

위치 찾기

- 텍스트 영역 내 특정 위치에서 검색 희망

- `\b`: 단어 경계
- `^`: 문자열 경계의 시작
- `$`: 문자열 경계의 끝

하위 표현식 ⭐

- 큰 표현식 안에 속한 일부 표현식을 한 항목으로 다루도록 묶음
- `( )`: 괄호로 묶음 가능
- `$gt;{2,}` vs `(>){2,}`
- 파워풀한 중첩된 하위 표현식

하위 표현식 - 역참조

- 하위 표현식으로 매칭된 타겟을 참조
- 일치한 부분을 반복해 찾거나 치환에 사용
- 텍스트를 검색하고 치환하는데 매우 유용
- `There is a ball on on the table` - 실수로 중복된 전치사?
  - `/(\w+)\s\1/g`

정규 표현식 예제

- 이메일
  - `/^\w+([.-]?\w+)*@\w+([.-]?\w+)*(\.\w{2,3})+$/g`
- 차량번호
  - `/\d{2,3}[가-힣]{1}\d{4}/g`