[평생] 멀티미디어 일반형 양식

[원고] 빅데이터 수집 시스템 개발

13회차: 정규식 활용과 문자열

| 내용전문가 | 백현숫 | 교수설계(한기대) | 홍길동 |
|-------|-----|------------|-----|
| 협력업체 | 업체명 | 교수설계(협력업체) | 홍길동 |

| NCS 분류 정보 | 20-01-02-09 | 정보통신 - 정보기술 - 정보기술개발 - 빅데이터플랫폼구축 |
|------------|-------------------|----------------------------------|
| 능력단위 정보 | 03 | 빅데이터 수집시스템 개발 |
| | 2001020903_17v1.1 | 빅데이터 수집시스템 설계하기 |
| 능력단위 요소 정보 | | |
| | | |

| 업무 | 작성자 | 버전 | 작성일 | 특이사항 |
|---------|-----|-------|------------|--------|
| 원고 작성 | 백현숙 | v.1.0 | 2019/09/11 | 2회차 원고 |
| 한기대 피드백 | | v.1.1 | | |
| 원고 보완 | | v.2.0 | | |
| 자문 진행 | | v.2.1 | | |
| 원고 보완 | | | | |



◆ 학습 목차 - 학습 목차를 작성해 주세요. (NCS 비적용 과정일 경우에는 NCS 및 능력단위 정보를 기입하지 않아도 됩니다.)

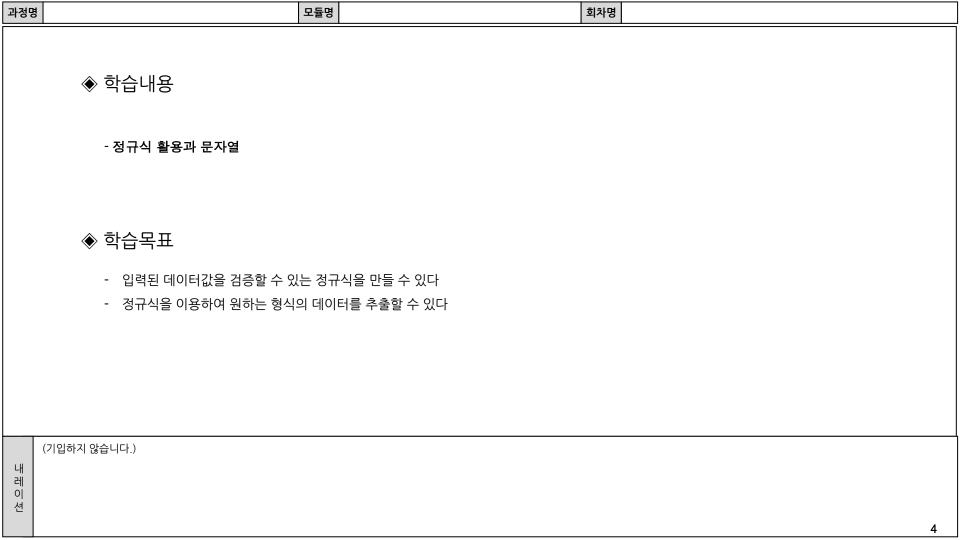
| NCS 분류 정보 | 20-01-02-09 | 정보통신 - 정보기술 - 정보기술개발 - 빅데이터플랫폼구축 |
|-----------|-------------|----------------------------------|
| 능력단위 정보 | 능력단위코드 기입 | 빅데이터 수집시스템 개발 |

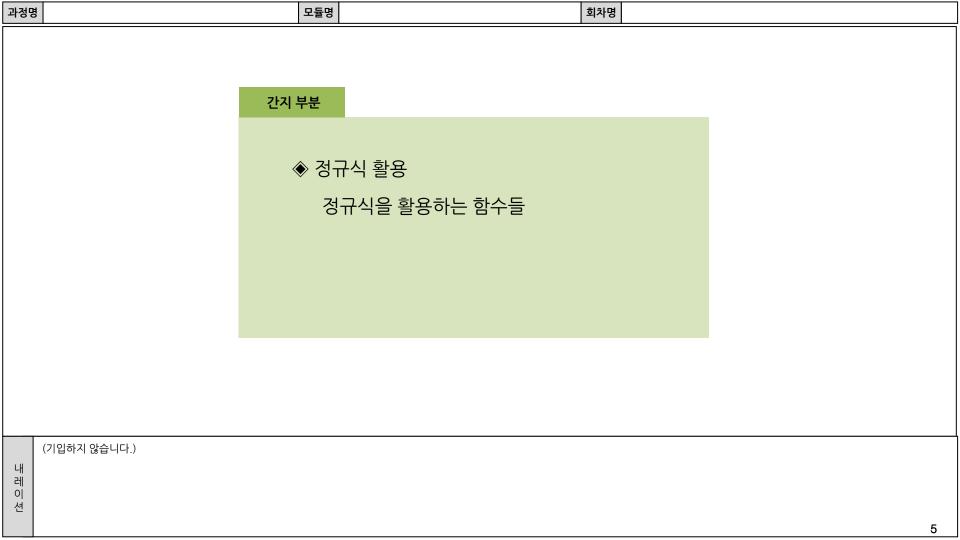
| 모듈명 | 회차명 | | 능력단위 요소 | | 수행준거 |
|-----------|------------|---|------------|------------|--------------------|
| 1. 모듈명 기입 | 01. 회차명 기입 | 1 | 능력단위요소명 기입 | 1.1 | 수행준거 기입 |
| | 02. 회차명 기입 | 1 | 능력단위요소명 기입 | 1.2 | 수행준거 기입 |
| | 03. 회차명 기입 | 1 | 능력단위요소명 기입 | 1.3 1.4 | 수행준거 기입 수행준거 기입 |
| | 04. 회차명 기입 | 1 | 능력단위요소명 기입 | 1.4 | 수행준거 기입 |
| 2. 모듈명 기입 | 05. 회차명 기입 | 2 | 능력단위요소명 기입 | 2.1 2.2 | 수행준거 기입 수행준거 기입 |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | 2 |

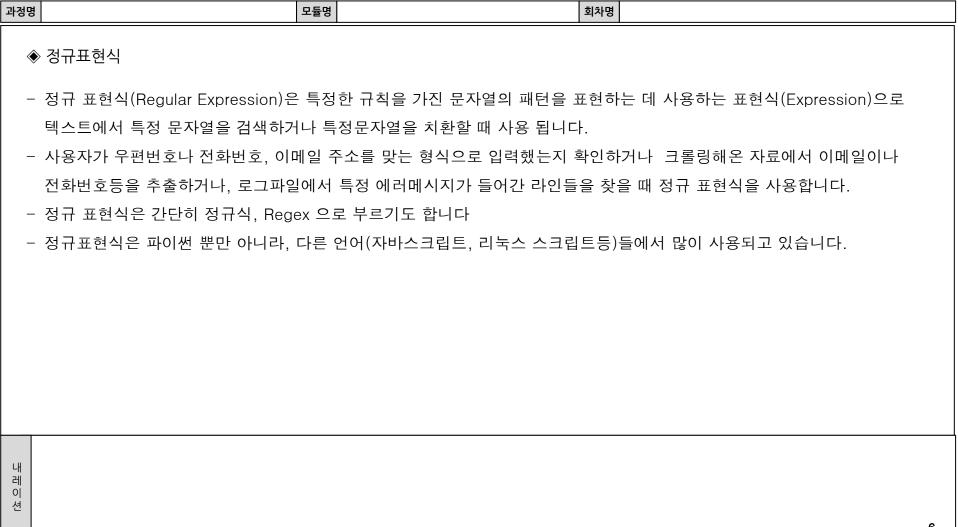
회차명

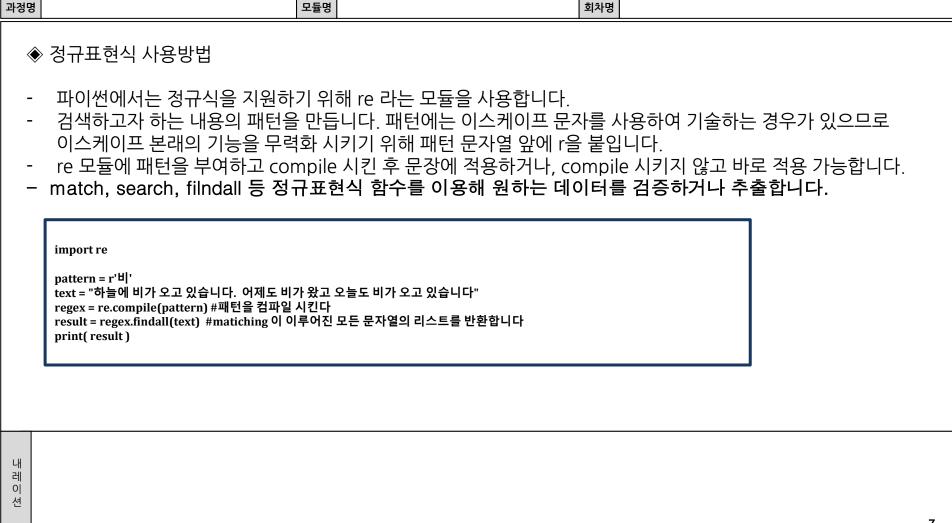
모듈명

과정명





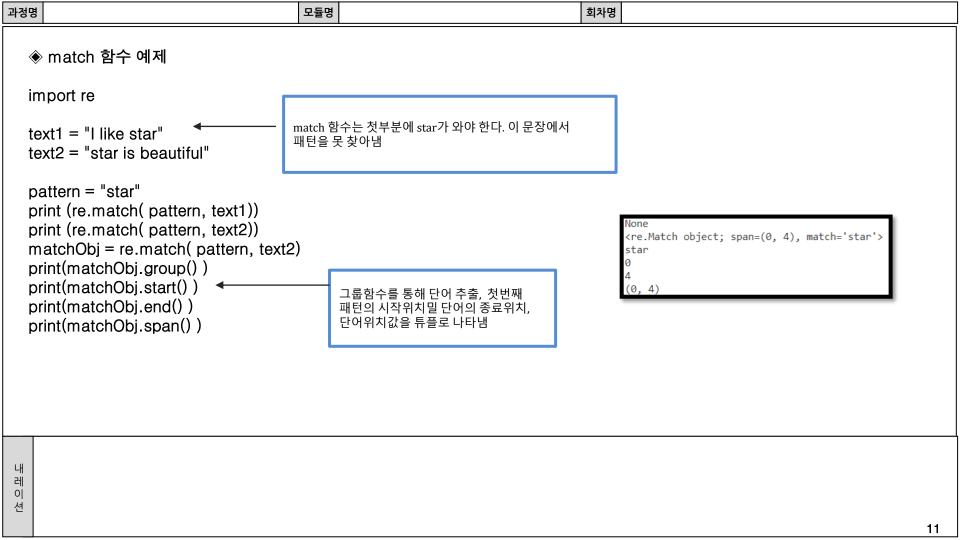




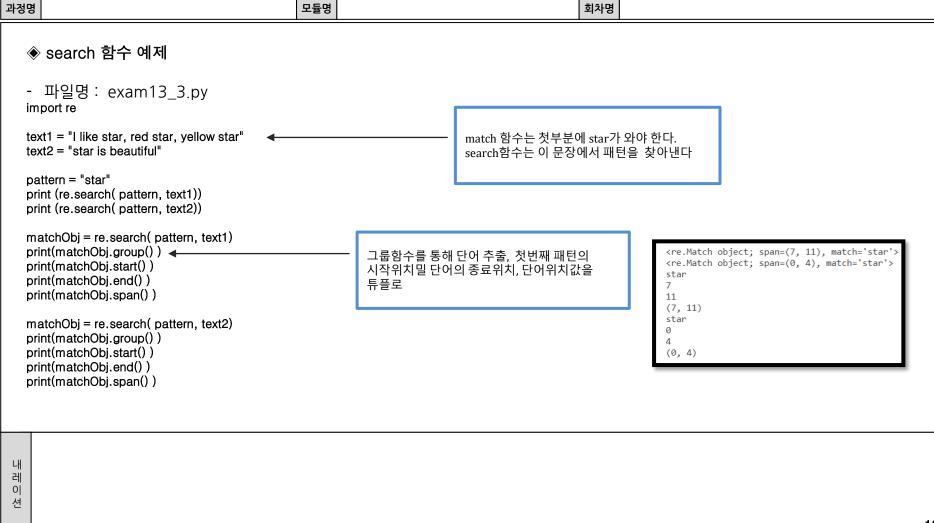
| 과정명 | | 모듈명 | | 회차명 | | | | |
|---------------------------------|---|-----|--|--|--|--|--|--|
| - | ◆ 우편번호 형식 맞추기 정규식 예제 - 우리나라의 우편번호는 6자리에서 5자리로 체계가 바뀌었습니다. 정수 값을 입력을 받아서 이 데이터가 우편번호 형식에 맞는지 확인해보는 예제를 만들어 보겠습니다. - 우편번호 패턴 방식 : \d{5}\$ <- 정수 5개만 가능하다 | | | | | | | |
| ir zi p re re if | 파일명: exam13_1.py nport re ipcode = input("우편번호를 입력하세 attern = r ' ₩d{5}\$' egex = re.compile(pattern) esult = regex.match(zipcode) result != None: print("형식이 일치합니다.") lse: print("잘못된 형식입니다.") | 요") | | 정수 5자리만 입력이 가능합니다. 5자리의 정수를 입력하면 "형식이 일치합니다."로 출력되고 그밖에 문자를 넣거나, 자릿수가 부족하거나 넘치면 "잘못된 형식입니다"라고 출력됩니다 | | | | |
| 내레이션 | | | | 8 | | | | |

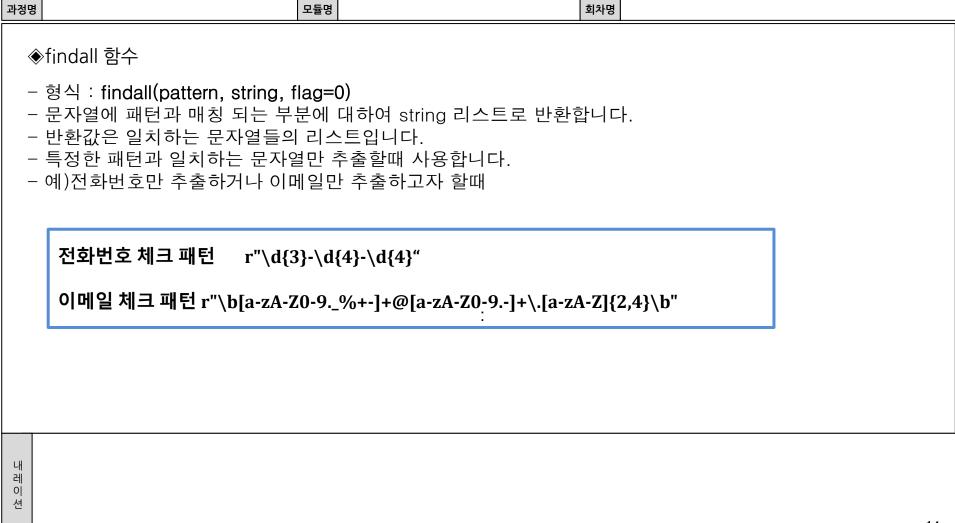
| 과정명 | | 모듈명 | 회차명 |
|------------------|--|--------------------------------------|---------|
| (| 정규식 함수 | | |
| - 5 | natch(pattern, string) 문자열의 시작부분부터 매칭이 되는지 검 매칭후 매칭에 대한 정보를 저장한 객체를 | | |
| | earch(pattern, string) 문자열에 패턴과 매칭되는 곳이 있는지 검 매칭후 매칭에 대한 정보를 저장한 객체를 | | |
| | indall(pattern, string) 정규식과 매치되는 모든 문자열을 리스트 | 로 반환한다 | |
| 1 | inditer(pattern, string) 정규식과 매치되는 모든 문자열을 iterato | or 객체(매칭객체)로 반환한다 | |
| 11 | sub(pattern, replace, string, count=0, 정규식과 매치되는 모든 문자열을 대체 | , flag=0) 문자열로 교체하고, 결과를 str 타입으로 | 로 반환한다. |
| | | | |
| 내 레 이 션 | | | 9 |

| 과정명 | | 모듈명 | 회차명 |
|------------------|--|-----------------------|--------------------|
| • | ≫ match 함수 | | |
| _ | 형식: match(pattern, string, flag | _{j=0}) | |
| _ | 문자열의 시작 부분부터 패턴과 마 | H칭이 되는지 검색하는 함수입니다. | |
| _ | 검색 결과는 matchObject 인스턴 | 트로 반환됩니다. 만일 일치하는 패턴이 | 없을 경우 None을 반환합니다. |
| | - 패턴이 ()등을 이용해 그룹으 - start(): 매칭된 문자열 시즈 - end(): 매칭된 문자열 종료 | | 자열을 반환. |
| 내 레 이 션 | | | 10 |



| 과정명 | 모듈 | 명 회: | 하명 | | |
|------------------|---|------|-----------|--|--|
| - - - - | ◆search 함수 - 형식: search(pattern, string, flag=0) - 문자열에 패턴과 매칭 되는 부분이 있는지 검색하는 함수입니다 match() 함수와의 차이점은 match 함수가 문자열의 시작부분이 일치해야 하는데 반해 search 함수는 문자열의 시작부터가 아니라 중간부터 있더라도 검색을 진행합니다 일치하는 패턴이 여러개 있더라도 맨 처음에 검색된 부분을 반환합니다. | | | | |
| M | MatchObject 주요 멤버함수 - group(): 매칭된 문자열을 반환, 0, 1, 2 파라미터를 주면 해당 문자열을 반환 패턴이 ()등을 이용해 그룹으로 되어 있을 경우 자동으로 분리 - start(): 매칭된 문자열 시작 위치 end(): 매칭된 문자열 종료 위치 span(): 매칭된 문자열 시작과 종료위치를 튜플로 반환. | | | | |
| | | : | | | |
| 내 레 이 션 | | | 12 | | |



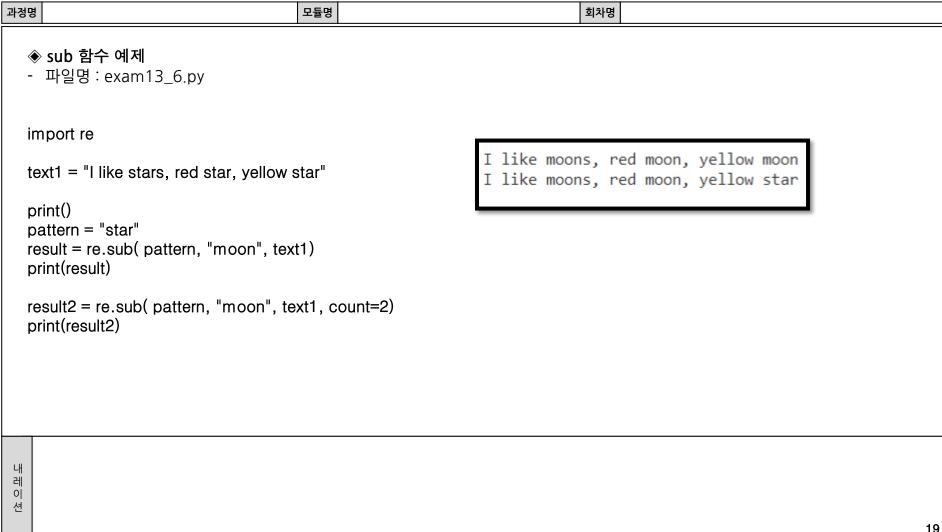


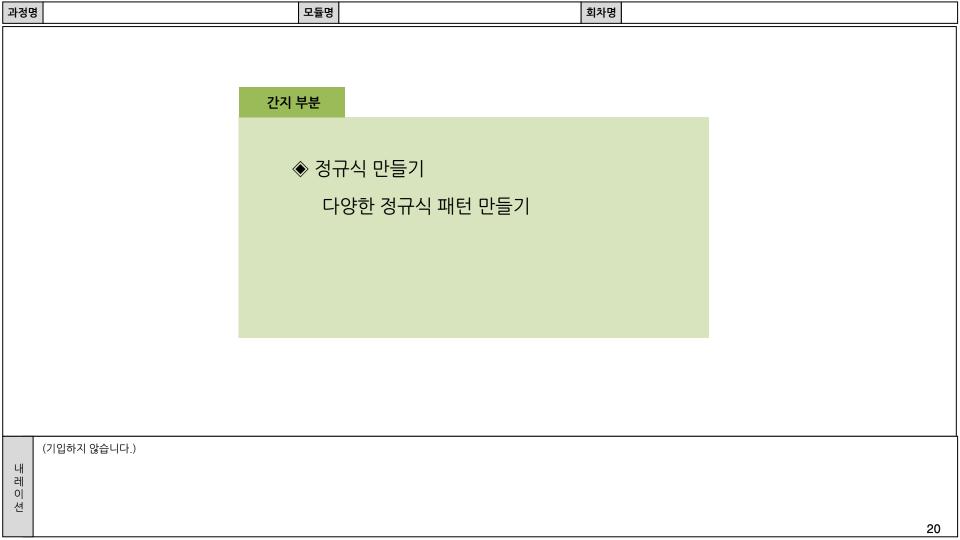
과정명 모듈명 회차명 ◆ findall 함수 예제 파일명:exam13 4.py import re #전화번호만 추출하기 text에 기술된 전화번호와 이메일로 부터 text = """ 원하는 이메일과 전화번호를 추출하는 --- 전화번호 추출하기 --phone: 010-0000-0000 email:test1@nate.com 예제입니다. phone: 010-1111-1111 email:test2@naver.com 010-0000-0000 findall 함수는 원하는 패턴의 데이터를 str의 phone: 010-2222-2222 email:test3@gmail.com list 형태로 반환합니다. 010-1111-1111 010-2222-2222 print() --- 이메일 추출하기 --print("--- 전화번호 추출하기 ---") test1@nate.com phonepattern = r" $\forall d{3}-\forall d{4}-\forall d{4}$ " test2@naver.com matchObj = re.findall(phonepattern, text) test3@gmail.com for item in matchObi: print(item) print("--- 이메일 추출하기 ----") \b가 앞에 있을때는 뒤에 오는 패턴으로 emailpattern = r"₩b[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+₩.[a-zA-Z]{2,4}₩b" ◆ 시작하는이라는 의미이고, \b가 뒤쪽에 있을때는 앞의 패턴으로 끝나라는 의미입니다. matchObj = re.findall(emailpattern, text) 이메일의 경우 test123@hanmail.net 형태인데 for item in matchObi: 앞쪽에는 영문자와 숫자가 하나이상 올 수 있고@ print(item) 와 도메인 그리고 .(도트) 뒤에 영문자가 옵니다. print() 2!4자리정도

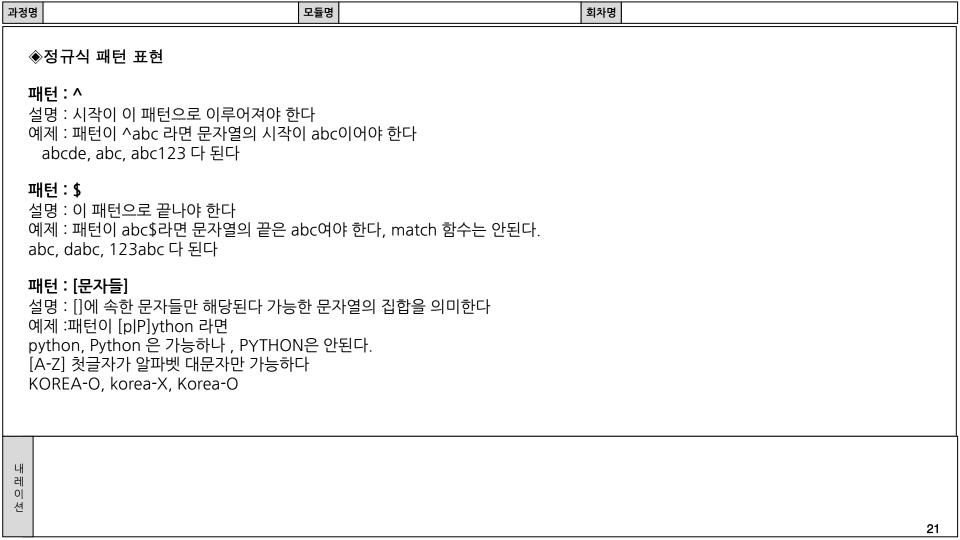
| 과정명 | | 모듈명 | 회차명 |
|------------------|---|--|--|
| - - - - | ●finditer 함수 - 형식: finditer(pattern, string - 문자열에 패턴과 매칭 되는 투 - 반환값은 문자열이 아니라 매 - 특정한 패턴과 일치하는 문자 - 예)전화번호만 추출하거나 이 MatchObject 주요 멤버함수 - group(): 매칭된 문자열 - 패턴이 ()등을 - start(): 매칭된 문자열 - end(): 매칭된 문자열 | g, flag=0) 라분에 대하여 매칭객체를 리스트로 남칭 객체라는 부분에서 find 함수외 남열만 추출할때 사용합니다. 메일만 추출하고자 할때 물을 반환, 0, 1, 2 파라미터를 주면 사장 위치. | 로 반환합니다. 가 차이가 납니다 면 해당 문자열을 반환. |
| | | | |
| 내 레 이 션 | | ÷ | 16 |

| 과정명 | 모듈명 | | 회차명 | |
|--|------------------------------|---------------------------------|-------------------------------|----|
| ● findall 함수 예제 - 파일명: exam13_5.py import re #전화번호만 추출하기 text = """ phone: 010-0000-0000 email:tes phone: 010-1111-1111 email:tes phone: 010-2222-2222 email:tes print() print(" 전화번호 추출하기") phonepattern = r"₩d{3}-₩d{4}-₩d- matchObj = re.finditer(phonepattern, for item in matchObj: print(item.group()) print(item.span()) print() | t2@naver.com t3@gmail.com | 010 (13 010 (60 010 | !환합니다. 이터레이터 통해 일치한 대상에 대한 | |
| 내 레 이 션 | | | | 17 |

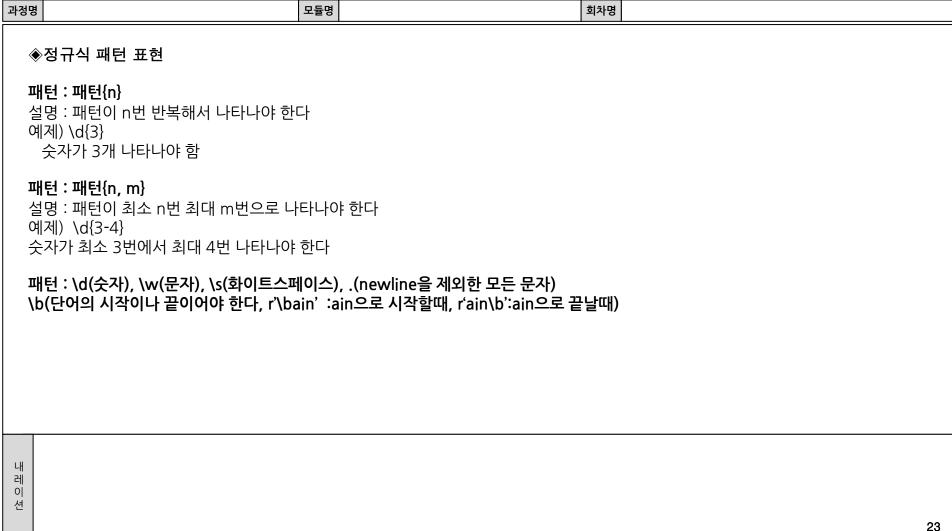
| 과정명 | 모듈명 | 회차명 |
|------------------|--|-------------------------|
| • | sub 함수 | |
| _ | 형식 : sub(pattern, replace, string, count=0, flag=0) - replace는 문자열이 될 수도 있고, 함수가 될 수도 있다 - count는 최대 몇 번까지 교체할 것인가를 설정하는 인자이다. 이 값이 횟수만큼만 교체된다 | 0 이면, 모두 교채, 0보다 크면 지정된 |
| - - - | 문자열에 패턴과 매칭 되는 부분에 대하여 매칭객체를 리스트로 반환합니 반환값은 문자열이 아니라 매칭 객체라는 부분에서 find 함수와 차이가 납 특정한 패턴과 일치하는 문자열만 추출할때 사용합니다. 예)전화번호만 추출하거나 이메일만 추출하고자 할때 | |
| M | atchObject 주요 멤버함수 - group(): 매칭된 문자열을 반환, 0, 1, 2 파라미터를 주면 해당 문경 - 패턴이 ()등을 이용해 그룹으로 되어 있을 경우 자동으로 - start(): 매칭된 문자열 시작 위치. - end(): 매칭된 문자열 종료 위치. - span(): 매칭된 문자열 시작과 종료위치를 튜플로 반환. | |
| 내 레 이 션 | | 18 |

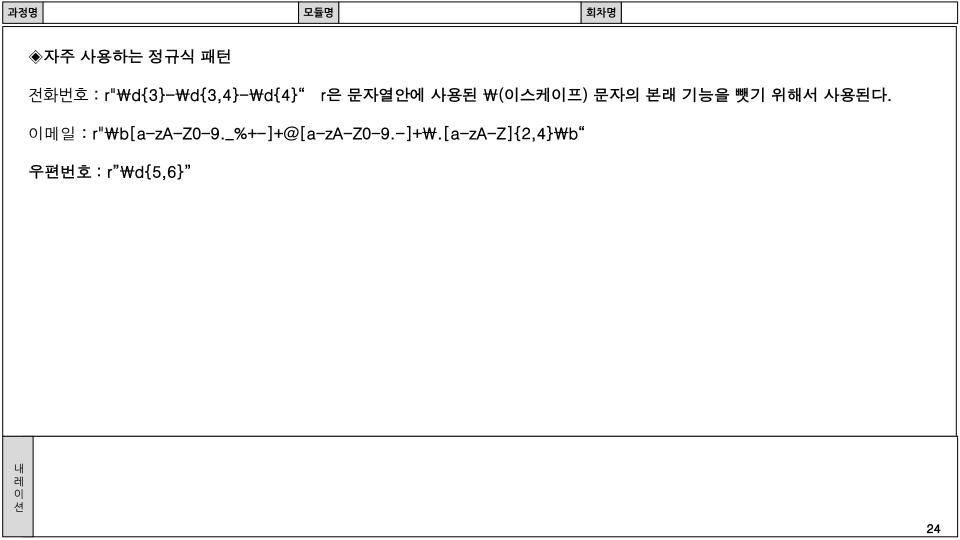


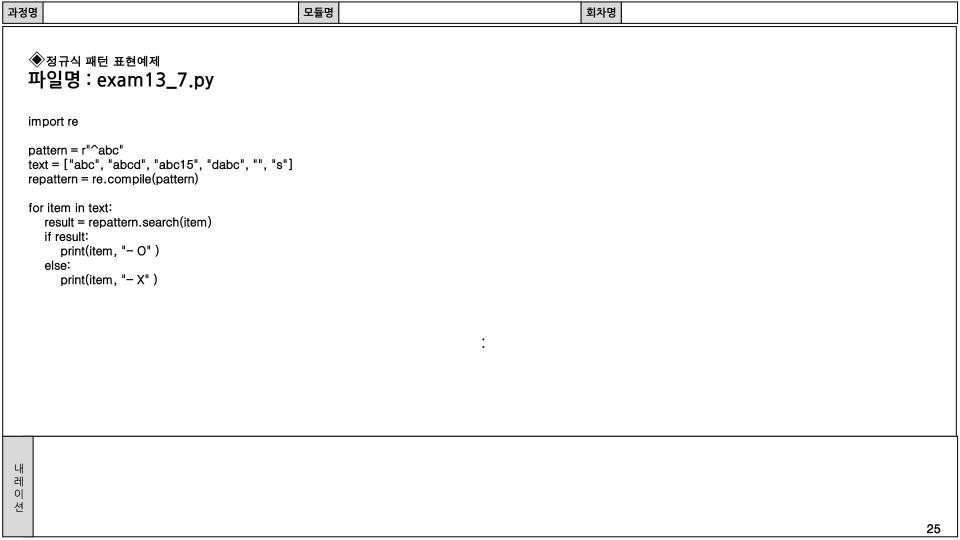


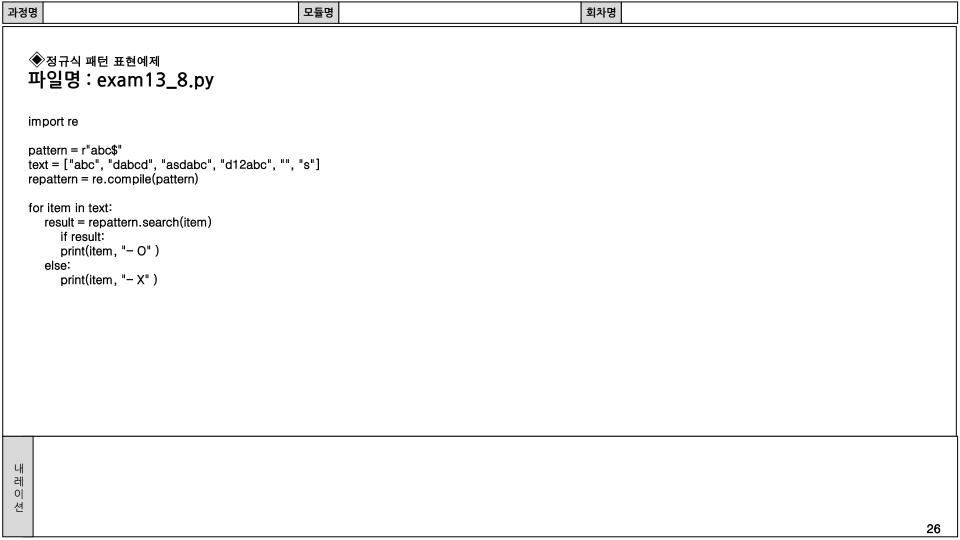


| 과정명 | | 모듈명 | 회차명 |
|------------------|---|------------------------|-----|
| 표 | ◈정규식 패턴 표현 대턴 : [^문자들] 설명 : 피해야할 문자들의 집합이다. 이 문 계제: 패턴이 [^abc] 라면 a, b, c, abc, acb, bca - x, d-o | | |
| 실 0 | 대턴: 설명 : or 연산 둘중 하나만 일치하면 된디 계제) 패턴이 [k K]orea라면 korea, Korea - o, Corea - x | 라 . | |
| 실 | 대턴: ?, +, * 설명: ? 앞의 패턴이 없거나 하나이어야 + 앞의 패턴이 하나 이상 있어야 한 * 앞의 패턴이 0개 이상이어야 함, 계제: \d? - 숫자가 없거나 하나만 있어여 \d+ - 숫자가 하나 이상이어야 함 \d* - 숫자가 없거나 하나 이상이 | 한다 , 반복을 의미 야 한다 | |
| 내 레 이 션 | | | 22 |

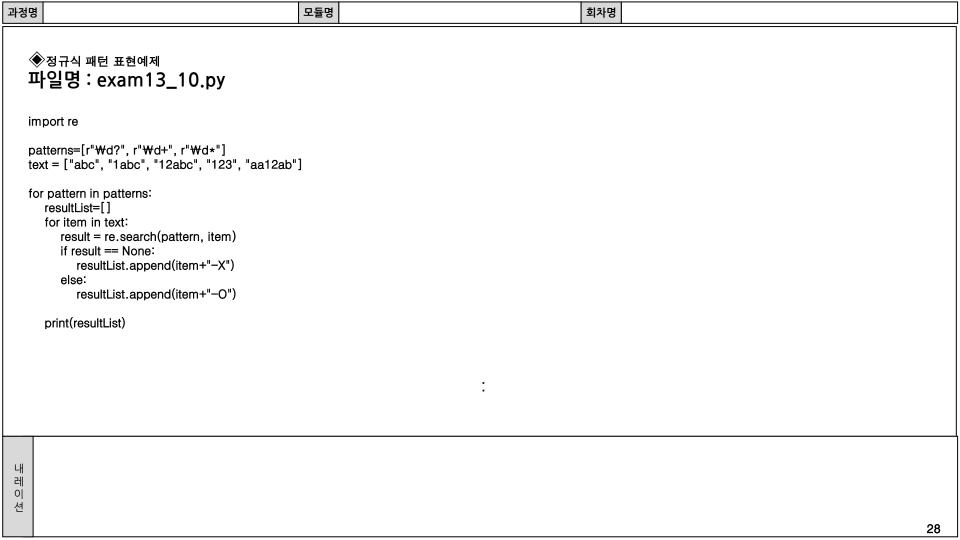




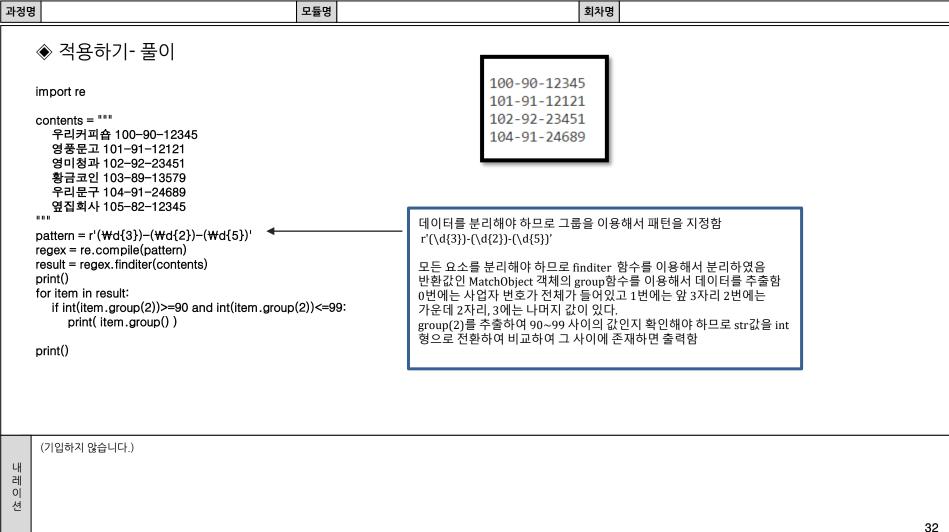


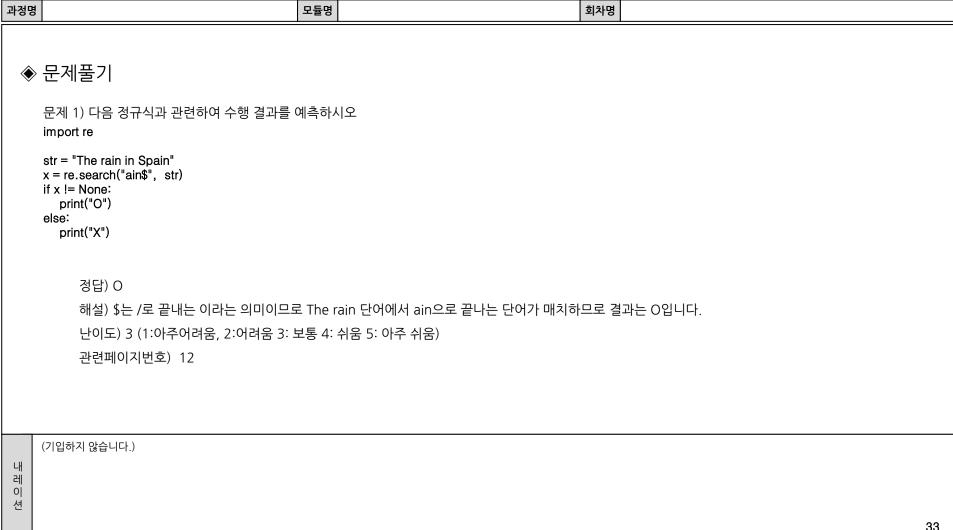


| | 모듈명 | 회차명 | |
|--|--|--|---|
| 다일명:exam13_9.py nport re attern = r"[p P]ython" | | | |
| epattern = re.compile(pattern) or item in text: result = repattern.search(item) if result: print(item, "- O") else: | | | |
| attern = r"[A-Z]" ext = ["python", "Python", "PYTHON", "korea", " epattern = re.compile(pattern) | KOREA", "Korea"] | | |
| | : | | 27 |
| | <pre>if result: print(item, "- O") else: print(item, "- X") attern = r"[A-Z]" ext = ["python", "Python", "PYTHON", "korea", "epattern = re.compile(pattern) or item in text: result = repattern.search(item) if result: print(item, "- O") else:</pre> | 파일명: exam13_9.py Inport re attern = r"[p P]ython" ext = ["python", "Python", "PYTHON"] Expattern = re.compile(pattern) or item in text: result = repattern.search(item) if result: print(item, "- O") else: print(item, "- X") attern = r"[A-Z]" ext = ["python", "Python", "PYTHON", "korea", "KOREA", "Korea"] Expattern = re.compile(pattern) or item in text: result = repattern.search(item) if result: print(item, "- O") else: print(item, "- O") else: print(item, "- X") | 파일명: exam13_9.py nport re attern = r"[p P]ython" ext = ["python", "Python", "PYTHON"] ext = ["python", "Python", "PyTHON"] ext = result = repattern.search(item) if result: print(item, "- O") else: print(item, "- X") attern = r"[A-Z]" ext = ["python", "Python", "PYTHON", "korea", "KOREA", "Korea"] ext = ["python", "Python", "korea", "Korea"] ext = ["python", "python", "pyth |



| 과정명 | ş | 모듈명 | 회차명 | |
|----------------------------------|---|-------|--|---------------------|
| im cc pa re re if | 파일명: exam13_11.py mport re contents = "문의사항이 있으면 010-1234-6789 으 cattern = r'(\d{3})-(\d{4})-(\d{4})' egex = re.compile(pattern) esult = regex.search(contents) result != None: phone1 = result.group(1) phone2 = result.group(2) phone3 = result.group(3) print(phone1) print(phone3) else: print("전화번호가 없습니다.") | 으로 연락 | 주시기 바랍니다." 전화번호 패턴을 괄호를 이용해 그룹화 한다 3개의 구룹으로 나누어졌으므로 group(1), group(2), group(3)로 각각의 번호를 추출한다 | 010 1234 6789 |
| 내 레 이 션 | | | | 30 |





| 번 호 | 문제 | 정답 | 난이도 | 해설 | 관련학습보기 |
|--------|--|----|-----|--|--------|
| 2 | 다음 정규식과 관련하여 수행 결과를 예측하시오 str = "The rain in Spain" x = re.match("ain\$", str) print(x) if x != None: print("O") else: print("X") | X | 2 | match 함수는 단어의 시작부분만 매칭 판단을 하기때문에 이 정규식에 대해서 매칭되는 단어를 검색하지 못합니다 | 14 |
| 3 | str = "The rain in Spain" x1 = re.findall("ain\$", str) x2 = re.findall("ain", str) print(len(x1), len(x2)) 위 코드의 수행 결과는 ?! (1) 1,1 (2)2,2, (3)1,2 (4) 2,1 | 3 | 1 | findall 함수는 패턴과 일치하는 모든 문자셋을 찾아낸다. 이때 패턴에 \$가 있을 경우 문장의 맨끝을 마무리 하는 경우에만 해 당된다. ain 패턴은 두개의 던어 rain과 Spain을 매칭했다고 찾 아내지만 ain\$는 문장 맨 끝의 Spain만 일치한다고 찾아낸다 그래서 1, 2 이다 | 14 |
| 4 | 다음처럼 그룹을 지정했을 경우 패턴이 매칭 되었을 경우 group(0)에 들어간 값은 ? 패턴 (\d{3})-(\d{4})-(\d{4}) 문자열: 내전화 010-1111-2222 ① 010-1111-2222 ② 1111 ④ 2222 | 1 | 4 | 정규화 그룹을 사용할 경우에 그룹의 0번에는 전체 데이터가 1 번부터 그룹내의 내용이 분할하여 들어간다. 그래서 0번에는 전체 번호인 010-1111-2222가 다들어가있다 | 29 |

| 번 호 | 문제 | 정답 | 난이 도 | 해설 | 관련학습 보기 |
|--------|---|----|---------|---|------------|
| 5 | 아래 코드는 모든 공백문자를 9로 바꾸려고 합니다. 적당한 함수는? str = "The rain in Spain" x =re.()("\s", "9", str) print(x) | 4 | 5 | 매칭되어있는 데이터를 찾어서 대체문자열로 교환하는 함수는 sub 함수입니다 | 18 |
| | ① match | | | | |
| | ② search | | | | |
| | ③ findall | | | | |
| | 4 sub | | | | |
| | 패턴이 다음과 같을때 매칭되는 단어가 아닌것은? 패턴 : r"^abc" | 4 | 5 | 단어의 시작이 abc 하고 했으므로 마지막 4번은 d로 시작했음 로 매칭되지 않는다 | 25 |
| | ① abc | | | | |
| 6 | ② abcd | | | | |
| | ③ abc123 | | | | |
| | ④ dabc | | | | |
| | 패턴이 다음과 같을때 매칭되는 단어가 아닌것은? 패턴 : r"abc\$" | 2 | 5 | 단어의 마지막이 abc로 끝나야 한다 | 26 |
| | ① abc | | | | |
| 7 | ② abcd | | | | |
| | ③ 123abc | | | | |
| | ④ dabc | | | | |
| | | • | • | | |

| 번 호 | 문제 | 정답 | 난이 도 | 해설 | 관련학습보기 |
|--------|---|------|---------|---|--------|
| 8 | 패턴 이 다음과 같을 때 올바른 것은 ? 패턴 : r"\d*" ["abc", "1abc", "12abc"] ① "abc"는 매칭되지 않는다 ② "1abc"는 매칭되지 않는다 ③ "12abc"만 매칭된다. | 4 | 1 | * 는 반복되는 패턴이 없어도 된다. 숫자를 요구하지만 *를 추가했으므로 숫자 패턴이 있던 없던 매칭된다. | 28 |
| 9 | 패턴이 다음과 같을 때 매칭이 되는것을 모두 선택한것은? 패턴: r"\d{5,6}" ["12345", "a1234", "123456", "122222222", "1234", "1abc"] ① "12345", "a1234" ② "12345", "123456"' ③ "12345", "123456", "1234" | 2 | 5 | r'\d{5,6} 정수 5자리, 6자리만 해당된다. | 29 |
| 10 | 사원번호가 다음과 같은 형태일때 올바른 패턴 수식을 작성하시오 정규화 그룹을 이용하시오 12-34-999 | 해설참조 | 1 | (\d{2})-(\d{2})-(\d{3}) 각 단위별로 괄호를 해줘야 합니다. | 29 |





- 주요학습내용: 해당 회차 또는 레슨의 주요학습내용을 자세히 기입해 주세요.
- 검색 키워드: 학습자가 검색창에 어떤 검색어를 입력하면 본 회차 또는 본 레슨이 검색될 수 있을지 검색 키워드를 5개 기입해 주세요.

| 제목 | 주요학습내용 | 검색 키워드1 | 검색 키워드2 | 검색 키워드3 | 검색 키워드4 | 검색 키워드5 |
|-----------------|--------|---------|---------|---------|---------|---------|
| 13. 정규식 활용과 문자열 | 정규식 | regex | match | findall | sub | search |