AWS EC2 Based ML Lab & DataSet 사용자 가이드

본 교안은 멀티캠퍼스 4차 서비스 산업 데이터를 활용한 머신러닝 분석 과정을 위해 제작되었으며, 본 교육 外 배포/게시/공개를 금합니다.

AWS EC2 Based ML Lab 사용 안내

실습용 서버 사용자 정보

아래 링크로 들어가 안내 드리는 사용자 정보로 접속(비밀번호 변경 권장)

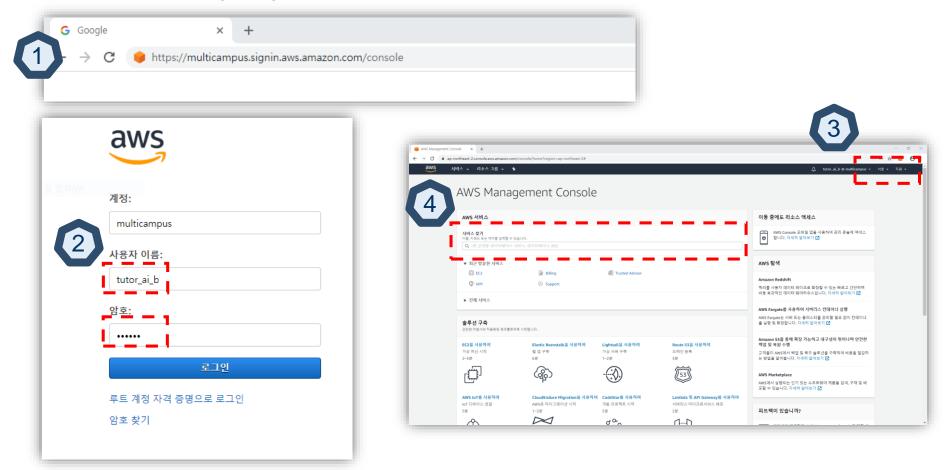
■ https://multicampus-ml.signin.aws.amazon.com/console

NO.	사용자 이름	User name	Password	리전	서버명	사용 login ID
0	(강사)	tutor_ml_b	ml1234	seoul	4th-multi-ml-01~05	tutor
1	고영빈	ybko_ml_b	ml1234	seoul	4th-multi-ml-b-01	lab01
2	구민주	mjkoo_ml_b	ml1234	seoul	4th-multi-ml-b-01	lab02
3	노소영	symo_ml_b	ml1234	seoul	4th-multi-ml-b-01	lab03
4	박소희	shpark_ml_b	ml1234	seoul	4th-multi-ml-b-01	lab04
5	박준원	jwpark_ml_b	ml1234	seoul	4th-multi-ml-b-02	lab05
6	박찬종	cjpark_ml_b	ml1234	seoul	4th-multi-ml-b-02	lab06
7	송이레	yrsong_ml_b	ml1234	seoul	4th-multi-ml-b-02	lab07
8	윤대상	dsyoon_ml_b	ml1234	seoul	4th-multi-ml-b-02	lab08
9	윤석영	syyoon_ml_b	ml1234	seoul	4th-multi-ml-b-03	lab09
10	이다완	dwlee_ml_b	ml1234	seoul	4th-multi-ml-b-03	lab10
11	이리애	layi_ml_b	ml1234	seoul	4th-multi-ml-b-03	lab11
12	이상아	salee_ml_b	ml1234	seoul	4th-multi-ml-b-04	lab12
13	이혁재	hjlee1_ml_b	ml1234	seoul	4th-multi-ml-b-04	lab13
14	이혜준	hjlee2_ml_b	ml1234	seoul	4th-multi-ml-b-04	lab14
15	임다솜	dsim_ml_b	ml1234	seoul	4th-multi-ml-b-05	lab15
16	임지수	jsim_ml_b	ml1234	seoul	4th-multi-ml-b-05	lab16
17	최대건	dgchoi_ml_b	ml1234	seoul	4th-multi-ml-b-05	lab17

- 강사계정은 모든 서버에 접근할 수 있는 권한 부여
 프로젝트 기간에는 팀구성원에 맞게 서버 새로 배치 예정

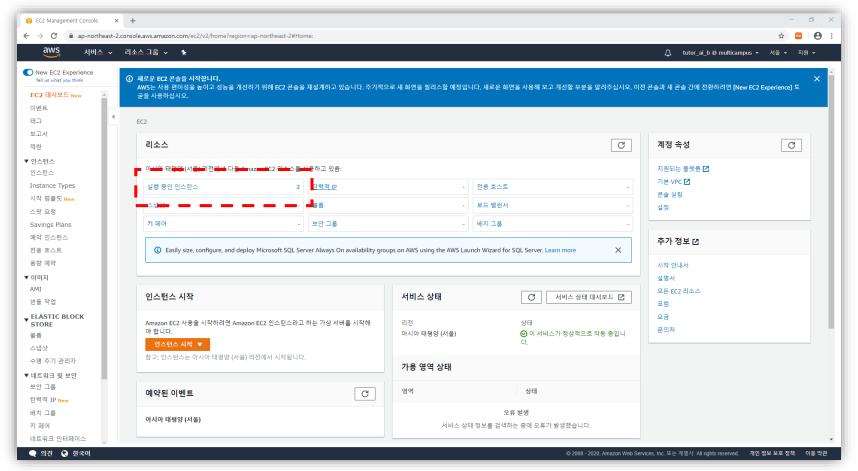
아래 링크로 들어가 안내 드린 사용자 정보로 접속(비밀번호 변경 권장)

■ 우측 상단 리전(서울) 정보 확인 후 서비스 찾기에서 EC2 입력 후 클릭



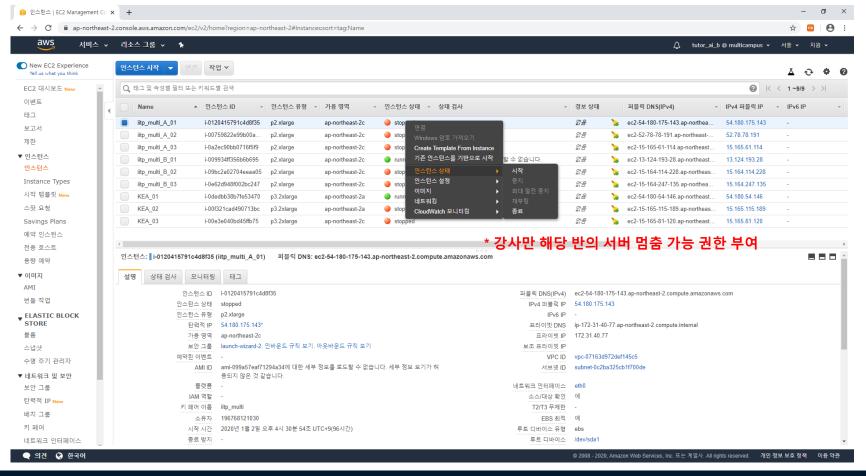
현재 사용 가능한 서버 정보 확인(최소 권한 부여)

■ 실행 중인 인스턴스 클릭



실습용 서버 동작 확인 및 서버 실행(사용자별로 서버 실행권한만 별도부여)

■ 사용할 서버 위에서 마우스 우클릭, 인스턴스 상태에서 시작버튼 클릭



Copyright © 2013 Global Software Campus Co, Ltd. all rights reserved

사용자 로그인 ID 및 정보에 맞는 접속 서버 내용 확인

Putty 접속ID	Putty 접속 Port	RStudio PW	Jupyter notebook 사용Port	서버 접속 Host Name	IP			
tutor	22	multitutor!	8889	모든 서버에 접근 가능				
lab01		multilab01!	8890		15.164.136.150			
lab02		multilab02!	8891	ec2-15-164-136-150.ap-northeast-2.compute.amazonaws.com				
lab03		multilab03!	8892					
lab04		multilab04!	8893					
lab05		multilab05!	8894	ec2-15-164-188-121.ap-northeast-2.compute.amazonaws.com	15.164.188.121			
lab06		multilab06!	8895					
lab07		multilab07!	8896					
lab08		multilab08!	8897					

* Rstudio PW 변경은 서버 접속 후 passwd명령어 입력 후 변경(권장사항)

사용자 로그인 ID 및 정보에 맞는 접속 서버 내용 확인

Putty 접속ID	Putty 접속 Port	RStudio PW	Jupyter notebook 사용Port	서버 접속 Host Name	IP
lab09		multilab09!	8890		15.165.2.211
lab10		multilab10!	8891	ec2-15-165-2-211.ap-northeast-2.compute.amazonaws.com	
lab11		multilab11!	8892		
lab12		multilab12!	8893		15.165.56.198
lab13	22	multilab13!	8894	ec2-15-165-56-198.ap-northeast-2.compute.amazonaws.com	
lab14		multilab14!	8895		
lab15		multilab15!	8896		52.78.167.125
lab16		multilab16!	8897	ec2-52-78-167-125.ap-northeast-2.compute.amazonaws.com	
lab17		multilab17!	8898		

* Rstudio PW 변경은 서버 접속 후 passwd명령어 입력 후 변경(권장사항)

■ 서버 사양:

X HW

- 인텔 제온 E5-2686 v4(Broadwell) 프로세서, DDR4 메모리
- 4vCPU / 30.5GB RAM / 1000GB SSD / Ubuntu 18.04 0S

X SW

- RStudio 1.2.1335 / R 3.6.0 / Anaconda3-5.2.0-Linux-x86 64 (2019-04-04)
- jdk-8u231-linux-x64, spark-3.0.0-preview2-bin-hadoop2.7(단일 머신) 설치
- 기타 : (base) Anaconda 환경에서 TensorFlow, Keras 등 수업에 필요한 머신러닝, 딥러닝 관련 패키지 및 라이브러리 설치 가능

- 교과목 및 프로젝트 기간 서버 운영시간:
 - 평일/주말 : ~18:00**(18시 이후 자동 종료)**
 - 요청사항에 따라 유연하게 확대 운영 가능(평일: ~22:00)
- (요청 시)일정 기간 서버 Full-Day(24시간) 운영
- ※ 확대 운영 요청은 과정 Tutor에게 문의

서버 접속을 위한 사전 준비

원격 서버 접속을 위해 암호화 키를 개인 보관하고 원격 콘솔프로그램인 PuTTY를 다운로드



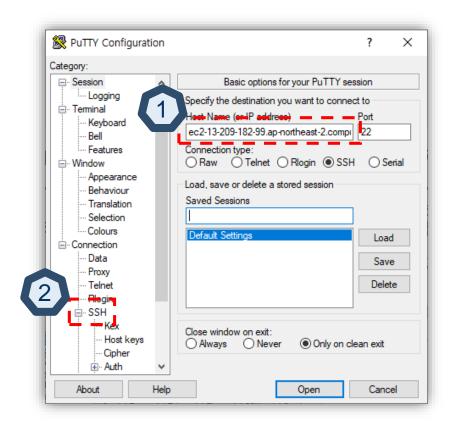
- multi-ml-a.ppk 암호화 키 보관
- www.putty.org 사이트를 통해 PuTTY 64-bit 다운로드

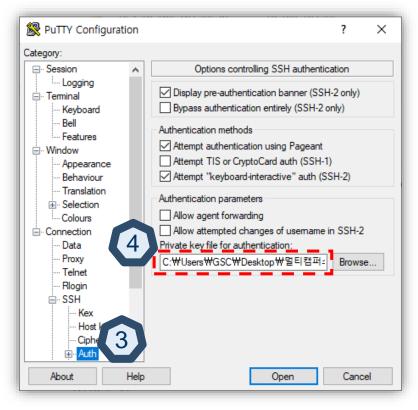


PuTTY 프로그램 실행 및 설정1

사용자별로 부여 받은 Host Name 입력 및 암호화 키 업로드

(암호화 키 : ml-multi-b.ppk)

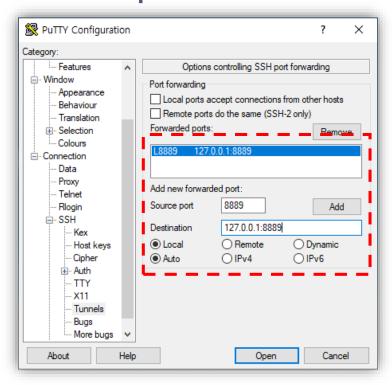




PuTTY 프로그램 실행 및 설정2

원격 서버에서 다른 사용자들과 Jupyter-Notebook을 동시 사용시 접속 충돌 방지를 위해 putty 설정에서 터널링 작업 추가 진행 후 저장

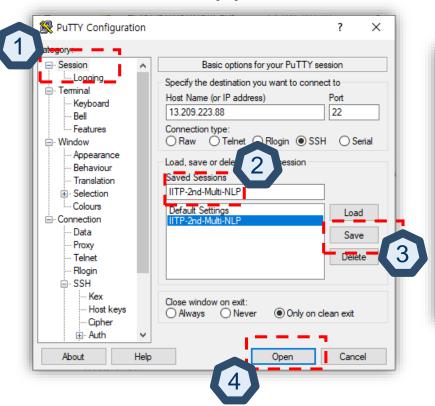
- Source port(8xxx), Destination(127.0.0.1:8xxx) 입력
- 각자 부여된 Jupyter notebook port번호로 터널링 설정(8889 ~ 8901)



PuTTY 설정 내용 저장

반드시 Session으로 돌아와 설정 내용을 저장

- 설정 내용을 저장해 놓으면 프로그램을 열 때마다 재설정할 필요 없음
- 최초 접속 시 예(Y) 클릭





참고 . MAC에서 서버 접속

맥북 사용자가 서버에 접속하는 방법

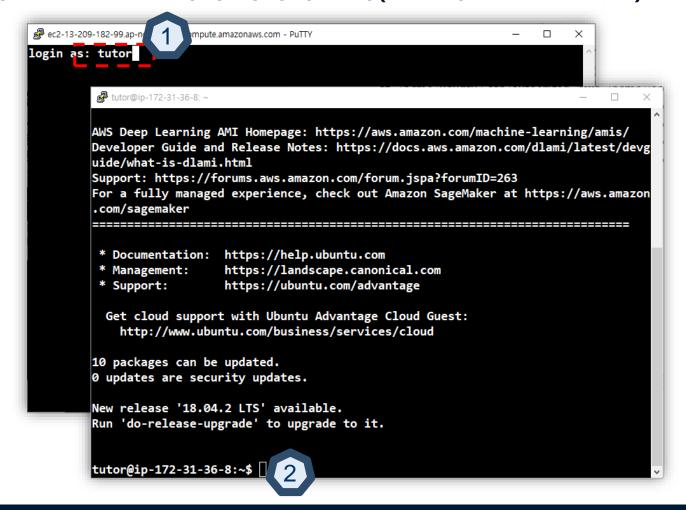
- 기본 터미널 프로그램 실행 후 암호화 키 위치 설정 및 권한 변경 : \$ chmod 400 /암호화 키 저장 위치/ml-multi-b.pem
- 터미널 창에서 실습 서버에 접속

```
$ ssh -i ~/Downloads/ml-multi-b.pem tutor@ec2-13-209-182-99.ap-northeast-2.compute.amazonaws.com -p 22 -L 8xxx:127.0.0.1:8xxx 

맥북 내 암호화 키 저장 경로 부여 받은 ID@서버 접속 Host Name (ssh접속을 위한 port 터널링)
```

서버 접속

부여 받은 ID를 입력해 서버에 접속(별도의 PW필요 없음)



데이터 분석을 위한 SPARK 사용1

SPARK 설치 위치로 이동 후, Spark shell 사용을 위한 명령어 입력

- SPARK 설치 위치 이동을 위한 명령어 : cd /opt/spark/
- 원격 서버환경에서 Spark-shell 실행 명령어 : ./bin/spark-shell

```
ubuntu@ip-172-31-42-212: /opt/spark
 * Documentation: https://help.ubuntu.com
                   https://landscape.canonical.com
 * Management:
                   https://ubuntu.com/advantage
   Support:
  System information as of Tue Jul 16 11:53:54 KST 2019
                                                        114
  System load: 0.0
                                   Processes:
  Usage of /: 5.0% of 387.70GB Users logged in:
                                   IP address for ens5: 172.31.42.212
  Memory usage: 0%
  Swap usage: 0%
 * MicroK8s 1.15 is out! Thanks to all 40 contributors, you get the latest
   greatest upstream Kubernetes in a single package.
     https://github.com/ubuntu/microk8s
8 packages can be updated.
0 updates are security updates.
Last login: Tue Jul 16 09:59:34 2010 from 103.90:180:109
ubuntu@ip-172-31-42-212: $ cd /opt/spark/
ubuntu@ip-172-31-42-212:/opt/spark$ ./bin/spark-shell
```

데이터 분석을 위한 SPARK 사용1

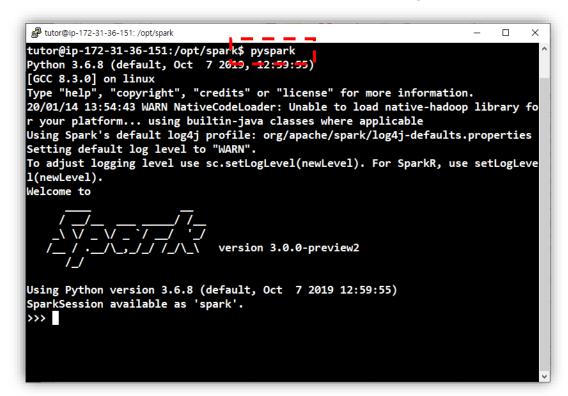
JAVA, Spark 설치 위치 및 환경변수 정보

- JAVA_HOME=/usr/local/java/jdk1.8.0_231
 JRE_HOME=\$JAVA_HOME/jre
 PATH=\$PATH:\$JAVA_HOME/bin:\$JRE_HOME/bin
- SPARK_HOME=/opt/spark
- PYSPARK_PYTHON=/usr/bin/python3.6

데이터 분석을 위한 SPARK 사용2

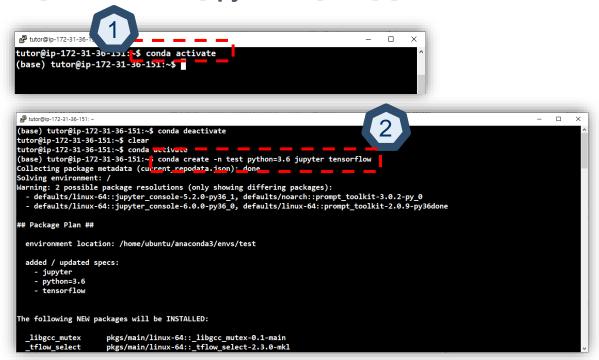
Python3.6 버전의 pySpark 사용

■ pyspark 명령어 입력 시 python3.6버전의 pySpark 바로 사용 가능 (환경변수 및 alias 등록 등은 사전에 설정해 놓음)



- 기 설치된 (base) 아나콘다 가상 환경 실행 및 개인적으로 생성
- conda activate : (base) 아나콘다 가상 환경 실행
- 원하는 패키지 활용 개인생성 예시

conda create -n [가상환경 이름] python=[버전] [설치하고자 하는 패키지]



주피터 노트북 사용을 위한 명령어 입력

■ Jupyter-notebook --ip=0.0.0.0 --no-browser --port=8xxx

각자 부여 받은 port번호 사용

```
(tensorflow_p36) tutor@ip-172-31-1-173:~$ jupyter
jupyter
                          jupyter-labextension
                                                    jupyter-qtconsole
jupyter-bundlerextension jupyter-labhub
                                                    jupyter-run
jupyter-console
                          jupyter-migrate
                                                    jupyter-serverextension
jupyter-kernel
                          jupyter-nbconvert
                                                    jupyter-troubleshoot
jupyter-kernelspec
                          jupyter-nbextension
                                                    jupyter-trust
jupyter-lab
                          jupyter-notebook
(tensorflow_p36) tutor@ip-172-31-1-173:~$ jupyter-notebook --ip=0.0.0.0 --no-browser --port=8889
```

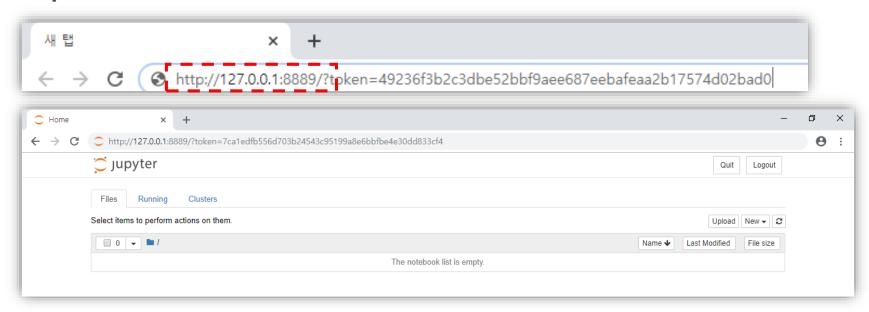
주피터 노트북 접속 URL 복사

■ http://127.0.0.1로 시작하는 URL주소 드래그 & 마우스 우클릭(복사기능)

```
tutor@ip-172-31-36-151: ~
CondaSystemExit: Exiting.
(base) tutor@ip-172-31-36-151:~$ jupyter-notebook --ip=0.0.0.0 --no-browser --port=8889
/home/ubuntu/anaconda3/lib/python3.7/site-packages/notebook/services/kernels/kernelmanager.py:19: VisibleDeprecationWarning: zmq.eve
ntloop.minitornado is deprecated in pyzmq 14.0 and will be removed.
    Install tornado itself to use zmg with the tornado IOLoop.
  from jupyter_client.session import Session
 I 17:15:33.894 NotebookApp]
                             JupyterLab extension loaded from /home/ubuntu/anaconda3/lib/python3.7/site-packages/jupyterlab
 I 17:15:33.894 NotebookApp]
                             JupyterLab application directory is /home/ubuntu/anaconda3/share/jupyter/lab
                             Serving notebooks from local directory: /home/tutor
 I 17:15:34.715 NotebookApp]
[I 17:15:34.715 NotebookApp] The Jupyter Notebook is running at:
                             http://ip-172-31-36-151:8889/?token=49236f3b2c3dbe52bbf9aee687e<u>ebafeaa2b17574d02bad0</u>
[I 17:15:34.715 NotebookApp]
I 17:15:34.715 NotebookApp]
                             or http://127.0.0.1:8889/?token=49236f3b2c3dbe52bbf9aee687eebafeaa2b17574d02bad0
[I 17:15:34.715 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 17:15:34.720 NotebookApp]
   To access the notebook, open this file in a browser:
        file:///home/tutor/.local/share/jupyter/runtime/nbserver-5921-open.html
   Or copy and paste one of these URLs:
        http://ip-172-31-36-151:8889/?token=49236f3b2c3dbe52bbf9aee687eebafeaa2b17574d02bad0
     or http://127.0.0.1:8889/?token=49236f3b2c3dbe52bbf9aee687eebafeaa2b17574d02bad0
 http://127.0.0.1:8889/?token=49236f3b2c3dbe52bbf9aee687eebafeaa2b17574d02bad0
```

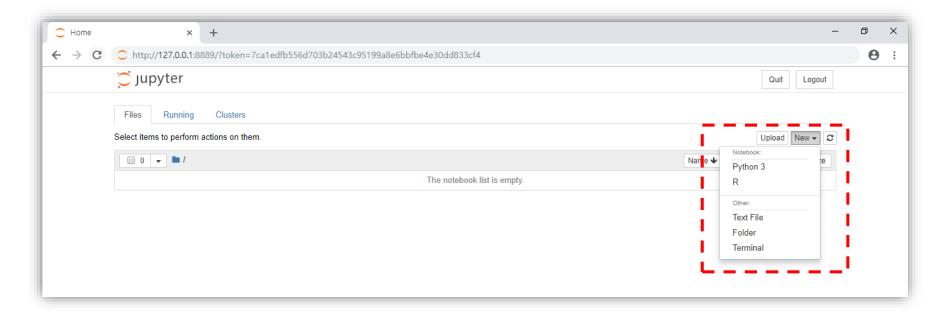
개인 로컬 PC 웹브라우저에서 원격 서버에 있는 주피터 노트북 실행

- (주의) URL 주소는 달라지는 port번호에 유의(8889~8901)
- port번호 초과 시 PuTTY설정에서 터널링 작업 추가 진행(권장x)



개인 로컬 PC와 동일한 방식으로 주피터노트북 사용 가능

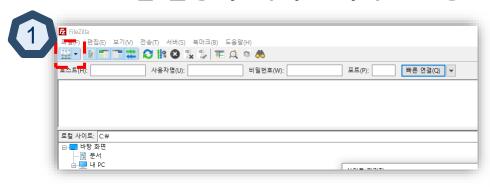
- Upload : 예제코드 및 데이터 파일 일괄 업로드 가능(폴더는 따로 생성)
- New : 주피터 노트북 파일(.ipynb) 및 폴더 생성
- 업로드한 파일의 저장 위치는 서버에 위치한 각 개인 홈 폴더(/home/사용자명)

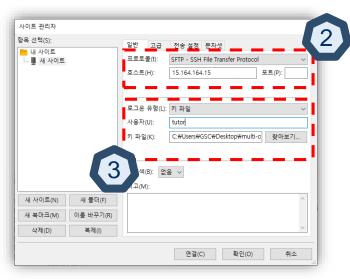


참고. 원격 서버로 파일 전송

로컬 PC에서 원격 서버로 파일을 전송하는 방법(window 기준)

- FileZilla Client 프로그램 다운로드
 - : https://filezilla-project.org/download.php?platform=win64
- 프로그램 실행 후 아래 순서대로 진행

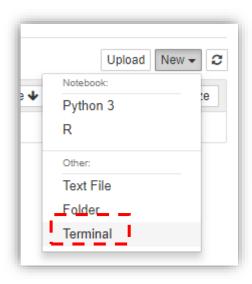


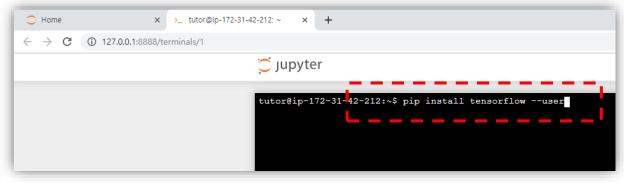


- 위 버튼 클릭 후 사이트 관리자 추가
- 프로토콜 선택 후 putty설정과 유사하게 호스트, 로그온 유형 진행
- 사용자에 맞게 사용자명은 변경해서 접속 후 파일 전송

(base) 환경에서 원하는 패키지는 터미널 창을 통해 각자 설치

- New -> Terminal 을 통해 기본 라이브러리 이외의 패키지 설치
- Terminal에서 pip install tensorflow --user 등의 명령어를 통해 설치 (주의) 반드시 --user 옵션을 붙여 줘야함

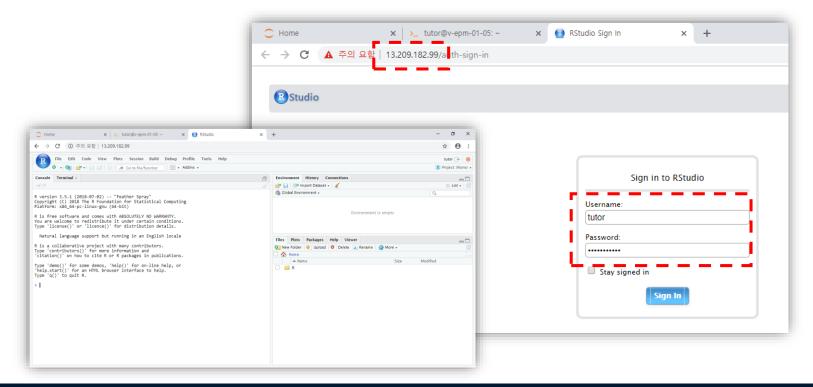




데이터 분석을 위한 R Studio 사용

원격 서버에서 R Studio 사용가능

- 앞서 부여 받은 IP주소를 웹브라우저 창에 바로 입력할 경우 R Studio 서버 버전 바로 사용 가능
- 로그인 ID와 PW는 부여 받은 그대로 입력



DataSet 사용 안내

DataSet 정보

키워드 기반 SNS 수집 데이터

- 트위터(키워드 5개)
- 인스타그램(키워드 5개)
- 블로그(키워드 5개)
- 뉴스(카테고리 10개)

<mark>구분</mark>	데이터명	데이터 구분	키워드 및 카테고리	데이터 량(건)	수집기간	속성	
	트위터	텍스트 비정형 데이터 ·	맛집	469,000	20190728~20190730		
			은행	167,000	20190720~20190828		
1			편의점	363,000	20190307~20190323	트윗 날짜, 답글 수, 트윗 내용 길이, html, 트위터 닉네임, 트윗내용, 리트윗 수, 키워드, aid, 유저명, 좋아요 수	
			호텔	344,000	20190318~20190515	夾게ㅇ, 니ㅡ 夾 ㅣ, 시꿔ㅡ, alu, ㅠ싱, ᇂ어쇼 ㅜ _	
			화장품	94,000	20190614~20190802		
	인스타그램	텍스트 비정형 데이터	맛집	3,100,000	20190617~20190823		
2			은행	14,000	20190614~20190822		
			편의점	75,000	20190810~20190823	인스타 날짜, 비디오 콘텐츠 여부, 답글 수, 인스타 내용 길이, l 인스타 내용, 키워드, aid, 유저명(고유식별번호), 좋아요 수	
			호텔	149,000	20190421~20190822		
			화장품	312,000	20190821~20190823		
		블로그 텍스트 비정형 데이터	맛집	118,419	20190103~20190821		
			은행	118,326	20190103~20190821	740 530 477 48	
3	블로그		편의점	118,359	20190101~20190821	글쓴이, 등록일, 블로그 본문, url, 사진, 본문 제목, 미디어, 카테고리 분류	
			호텔	118,327	20190103~20190820		
			화장품	118,346	20190101~20190821		
4	뉴스	텍스트 비정형 데이터	연예, 사회, 정치, IT, 기타, 경제, 오피니언, 스포츠, 생활, 세계	3,588,000	20190621~20190806	뉴스 제목, 본문, 뉴스 url, 언론사, 기사 게재일, 기사 종류, 기사 id	

DataSet 从용

원격 서버에서 Jupyter Notebook활용 pySpark로 DataSet 불러오기(예)

