

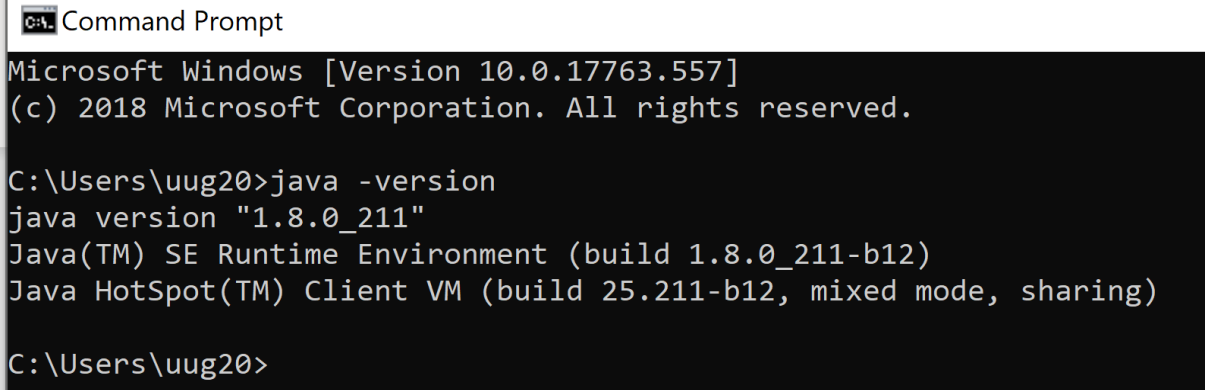
Spark & PyCharm Kurulumu

Windows

1. Adım

PySpark'ı çalıştırabilmemiz için bilgisayarlarımızda Java 7 ve Python 2.6 ya da daha ileri sürümlerinin yüklü olması gerekmektedir. İlk adım olarak Java'nın yüklü olup olmadığı kontrol edilmelidir. Bunun için;

Command Prompt'a girilir ve "java -version" komutu çalıştırılır. Eğer yüklü ise aşağıdaki çıktı görülür.



```
C:\> Command Prompt
Microsoft Windows [Version 10.0.17763.557]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\uug20>java -version
java version "1.8.0_211"
Java(TM) SE Runtime Environment (build 1.8.0_211-b12)
Java HotSpot(TM) Client VM (build 25.211-b12, mixed mode, sharing)

C:\Users\uug20>
```

Eğer "'java' is not recognized as an internal or external command, operable program or batch file." yazıyor ise Java'nın yüklenmesi gerekmektedir. Yüklü değil ise aşağıdaki adımlar izlenir. Yüklü ise 2. Adım'a geçiş yapabilirsiniz.

Java Yükleme

<https://www.java.com/en/download/> sitesinden Java indirilir.

Yükleme gerçekleştirilir.

Yükleme gerçekleştirildikten sonra tekrardan command prompt'a girilerek "java -version" komutu çalıştırılır ve başarılı olarak yüklendiği test edilir.

2. Adım

Bu adımda Python yüklemesini sağlayacağız. İlk olarak Java'nın yüklü olup olmadığını kontrol ettiğimiz gibi Python'a da bakmamız gerekiyor. Tekrardan Command Prompt'a girilerek "python --version" komutu çalıştırılır. Yüklü olması durumunda Python'ın versiyon bilgileri ekrana gelecektir. Eğer "'python' is not recognized as an internal or external command, operable program or batch file." ya da benzeri bir mesaj alınıyor ise Python'ın yüklenmesi gerekmektedir. Yüklü değil ise aşağıdaki adımlar izlenir. Yüklü ise 3. Adım'a geçiş yapabilirsiniz.

Python Yükleme

<https://www.python.org/downloads/windows/> sitesine girilir.

Stable Releases

- [Python 3.9.4 - April 4, 2021](#)

Note that Python 3.9.4 *cannot* be used on Windows 7 or earlier.

→ ▪ Download [Windows embeddable package \(32-bit\)](#)

→ ▪ Download [Windows embeddable package \(64-bit\)](#)

- Download [Windows help file](#)

- Download [Windows installer \(32-bit\)](#)

- Download [Windows installer \(64-bit\)](#)

- [Python 3.9.3 - April 2, 2021](#)

Note that Python 3.9.3 *cannot* be used on Windows 7 or earlier.

3.9.4 versiyonunun ilgili Windows paketi indirilir (Windows 32 bit ya da Windows 64 bit). Yükleme yapılırken dikkat edilmesi gereken bir nokta vardır. Customize Python bölümünde “Add python.exe to Path” seçilmelidir. Aksi takdirde PySpark’ın bazı özellikleri kullanılamayabilir. Yükleme gerçekleştirildikten sonra tekrardan command prompt’a girilerek “python --version” komutu çalıştırılır ve başarılı olarak yüklendiği test edilir.

3. ADIM

Bu adımda Spark’ın Windows makineye yükleme adımları gerçekleştirilir.

<http://spark.apache.org/downloads.html> adresine gidilerek ekran görüntüsündeki versiyonları ile Spark indirilir.

Download Apache Spark™

1. Choose a Spark release:

2. Choose a package type:

→ 3. Download Spark: [spark-3.1.1-bin-hadoop2.7.tgz](#)

4. Verify this release using the 3.1.1 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.

İndirildikten sonra dosya zip’ten çıkartılır. Daha önceki yaptığımız gibi burada bir yükleme söz konusu değildir.

Zip’ten çıkartmış olduğumuz dosyalar “spark-3.1.1-bin-hadoop2.7” adlı bir klasöre çıkartılacaktır. Bu dosyamızı C ya da D sürücüsündeki bir path’e kopyalanır. Örnek olarak “C:\spark\spark-3.1.1-bin-hadoop2.7” diyelim.

Test etmek için Command Prompt’tan ilgili klasöre gidilir. CMD’de bir klasöre gitme komutu cd’dir.

“cd C:\spark\spark-3.1.1-bin-hadoop2.7\bin” dersek ilgili path’e gitmiş oluruz. Path’te olduğumuzdan emin olduktan sonra “pyspark” komutu çalıştırılır. Bu komut PySpark Shell’e giriş yapmamızı sağlayacaktır.

PySpark Shell’de kullanılabilecek komutlar vardır. Örneğin “sc.version” yazıp Enter’a basarsak bize Spark’ın versiyonunu söyleyecektir. Çıkış yapmak için “exit()” yazılır.

Spark’ın kendisi Windows’ta çalışmak üzere dizayn edilmemiştir. Stabil çalışabilmek için winutils.exe adında bir program kullanır. Fakat default’unda winutils.exe’nin path’i atanmaz. Bu yüzden “ERROR Shell: Failed to locate the winutils binary in the hadoop binary path java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.” hatası alınabilir.

Bu hata PySpark Shell’in çalışmasına engel değildir. Fakat bazı fonksiyonları çalışmayacaktır (örneğin spark-submit).

4. ADIM

Winutils Yükleme

Winutils.exe’yi yükleme adımları aşağıdaki gibidir.

“<http://github.com/steveloughran/winutils>” adresinden Hadoop versiyonuna uyumlu olan Winutils indirilir. Biz bu durumda 2.7.1’i indireceğiz. Bu yüzden

“<https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin>” altında bulunan winutils.exe indirilir. İndirilen dosya “C:\spark\spark-3.1.1-bin-hadoop2.7\bin” altına taşınır.

Bu aşamada Hadoop’un kullanabilmesi için bir system environment tanımlamamız gerekmektedir.

Bu pathi user environment kısmında SPARK_HOME ve HADOOP_HOME şeklinde eklememiz gerekiyor.

tugta için kullanıcı değişkenleri

Değişken	Değer
HADOOP_HOME	C:\Users\tugta\spark\spark-3.1.1-bin-hadoop2.7
OneDrive	C:\Users\tugta\OneDrive
OneDriveConsumer	C:\Users\tugta\OneDrive
Path	C:\Users\tugta\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\tugta\AppData\Local\Temp
TMP	C:\Users\tugta\AppData\Local\Temp

Yeni... Düzenle... Sil

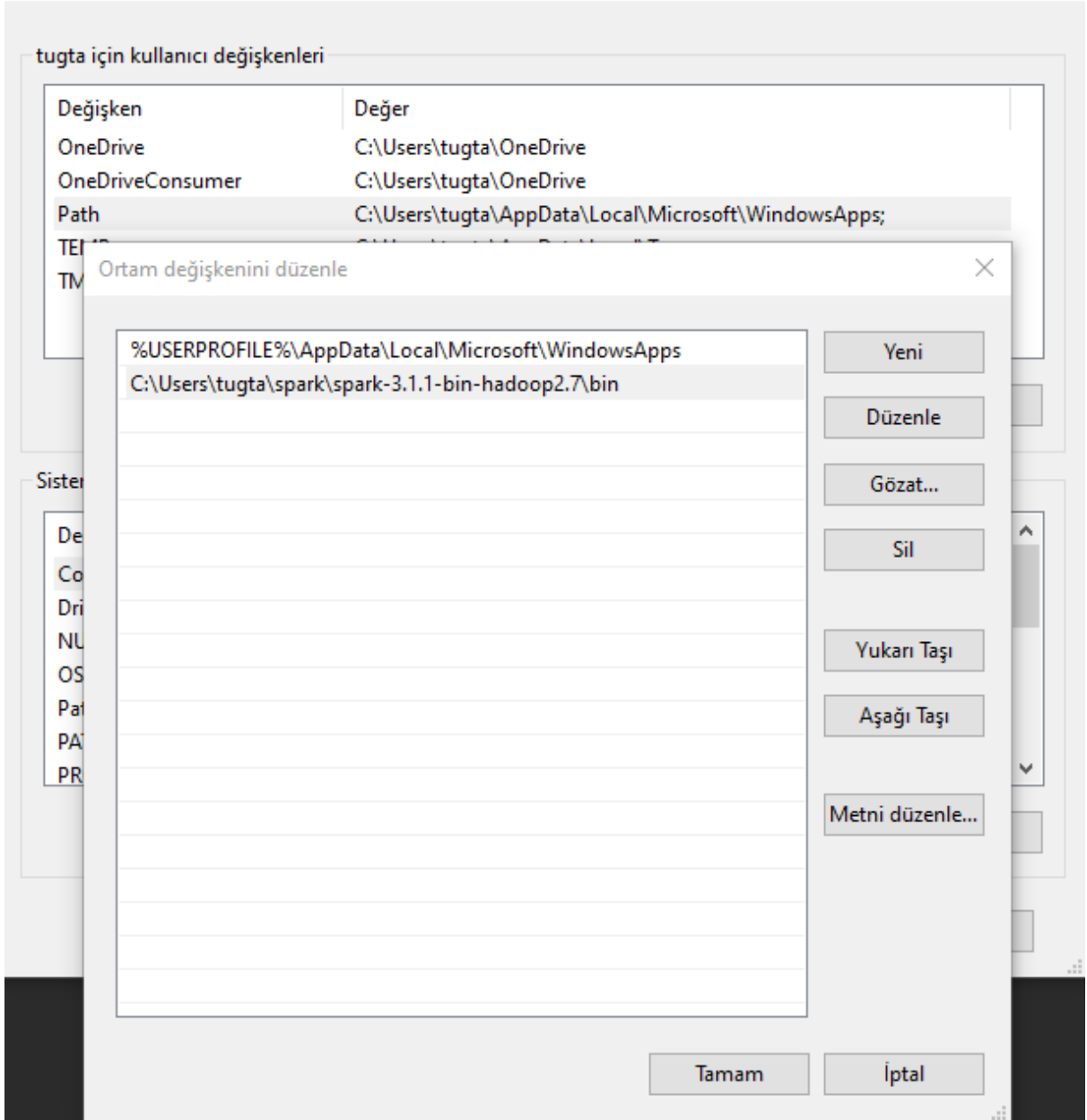
Sistem değişkenleri

Değişken	Değer
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

Yeni... Düzenle... Sil

Tamam İptal

User Enviroment içinde bulunan Path değişkenini de
C:\spark\spark-3.1.1-bin-hadoop2.7\bin pathini ekliyoruz.



Artık “pyspark” komutunu çalıştırdığımız zaman winutils ile alakalı hataların geçmiş olduğunu göreceksiniz Ayrıca pyspark enviroment path’e eklendiği için komut satırına doğrudan pyspark yazarak da spark konsolu başlatabiliriz .

Mesajları Temizleme

PySpark Shell’e girildiği zaman ekrana çokça mesaj basılacaktır. Bu mesajların tipleri INFO, ERROR, WARN’dur. Gösterilen mesajları azaltmak için gerekli işlemler aşağıdaki gibidir.

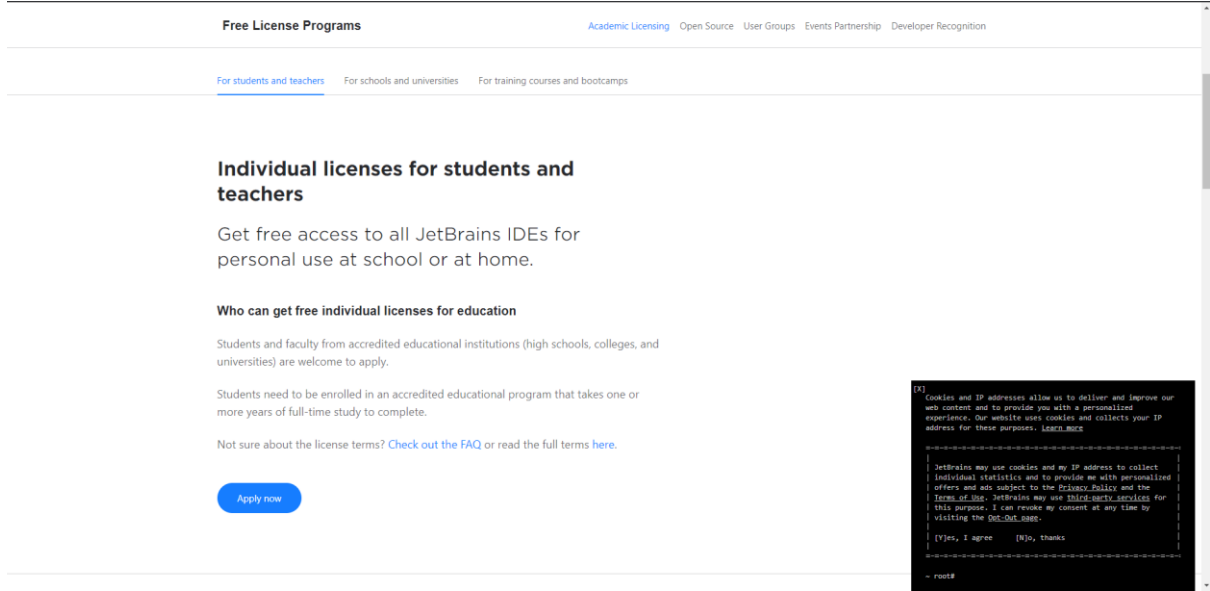
SPARK_HOME\conf altında bulunan “log4j.properties.template” dosyası “log4j.properties” adıyla kopyalanır.

Notepad ile yeni oluşturduğumuz dosya düzenlenir.

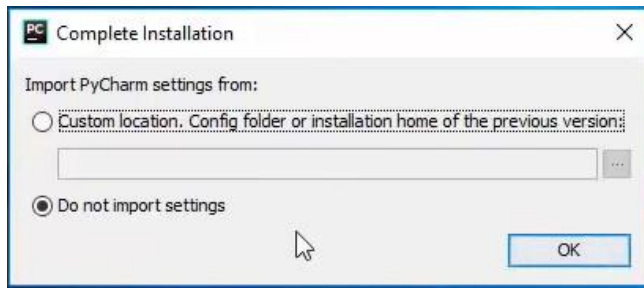
log4j.rootCategory satırındaki değer yerine “WARN” yazılır ve kaydedilir. Bu şekilde shell’i tekrar açtığımız zaman çıkan INFO mesajları temizlenmiş olur.

PyCharm Kurulumu

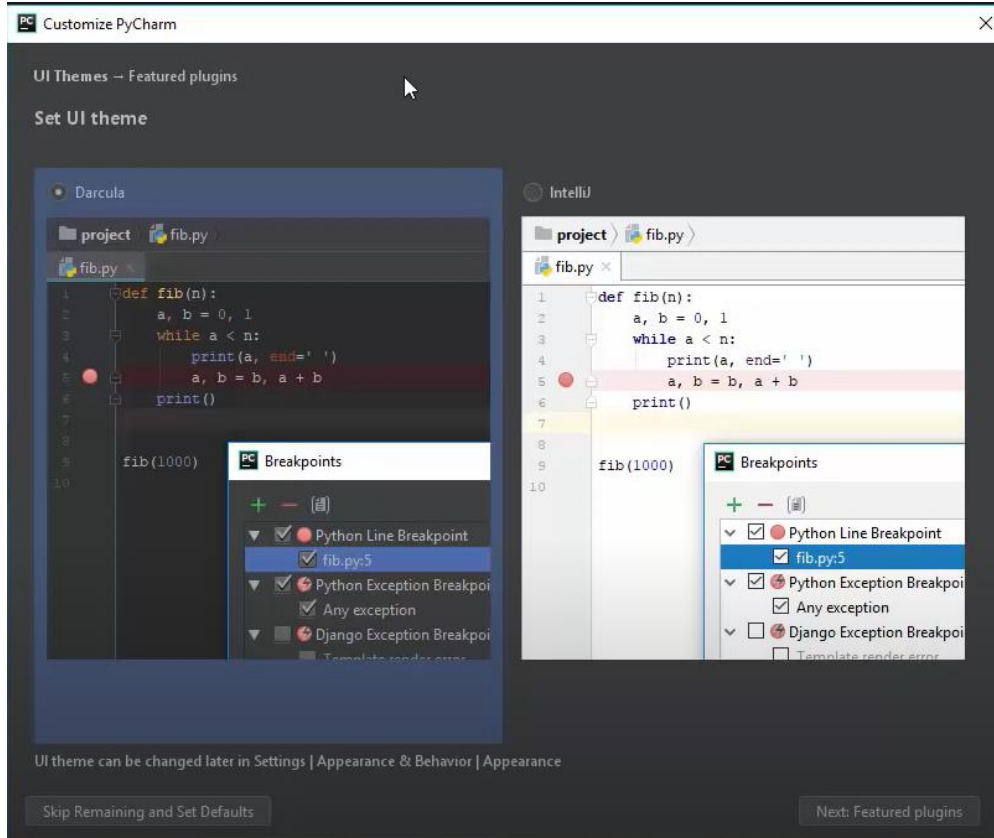
Pycharm üniversite öğrencileri için ücretsiz lisans sağlamaktadır. Aşağıdaki adrese giderek register olmanız yeterli. Sonrasında ihtiyacınız olan pek ürünü ücretsiz kullanma hakkına sahip oluyorsunuz. <https://www.jetbrains.com/community/education/#students> adresine giderek Apply Now diyoruz.



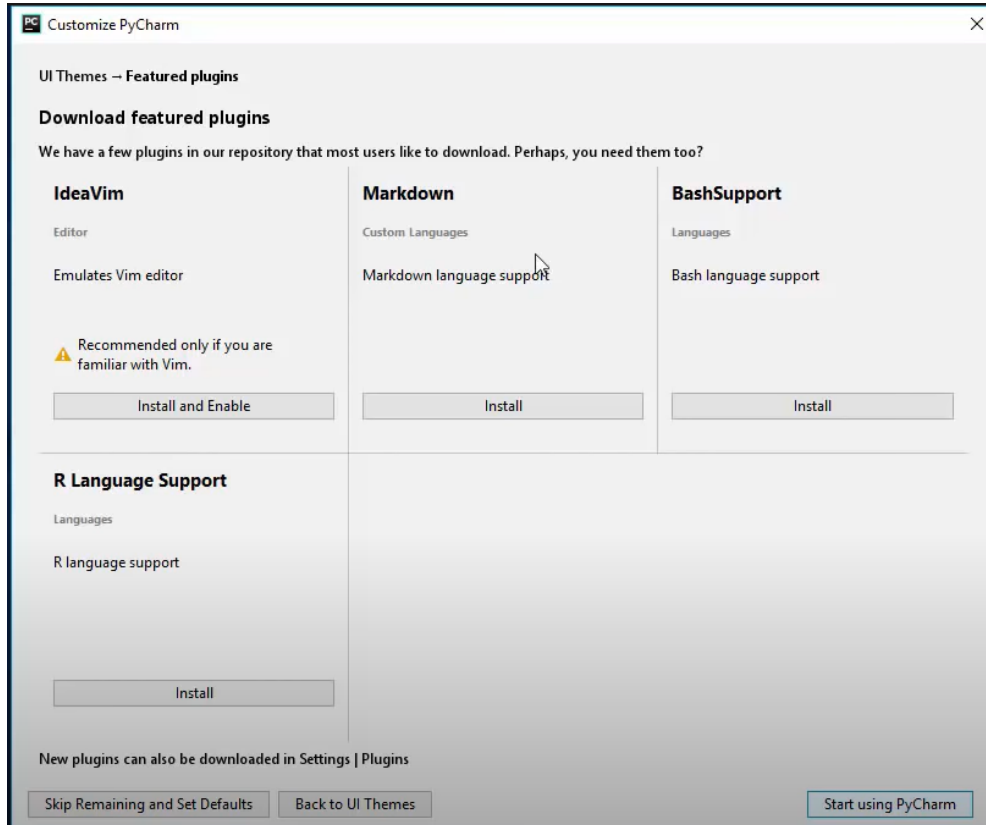
Yeni açılan ekranda istenilen bilgileri girdikten sonra hesabınız oluşturuluyor. Kendi hesabınız üzerinden pycharm indirmeniz gerekmektedir. .exe dosyasını çalıştırdıktan sonra herhangi bir değişiklik yapmadan hep next diyoruz. Yükleme tamamlandıktan sonra windows arama çubuğunda Pycharm yazarak uygulamayı açıyoruz. İlk aşamada sizden bazı ayarları import etmek isteyip istemediğinizi soruyor. Burada "Do not import settings" diyoruz.



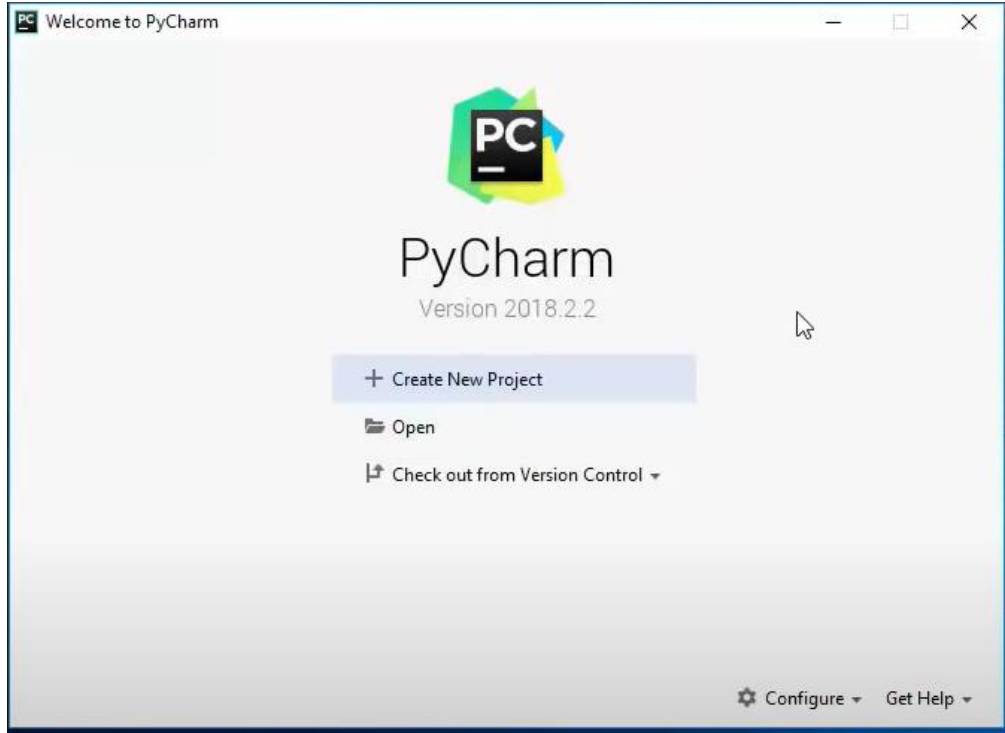
Sonrasında UI Theme seçmenizi istemektedir. Hangisini isterseniz seçebilirsiniz.



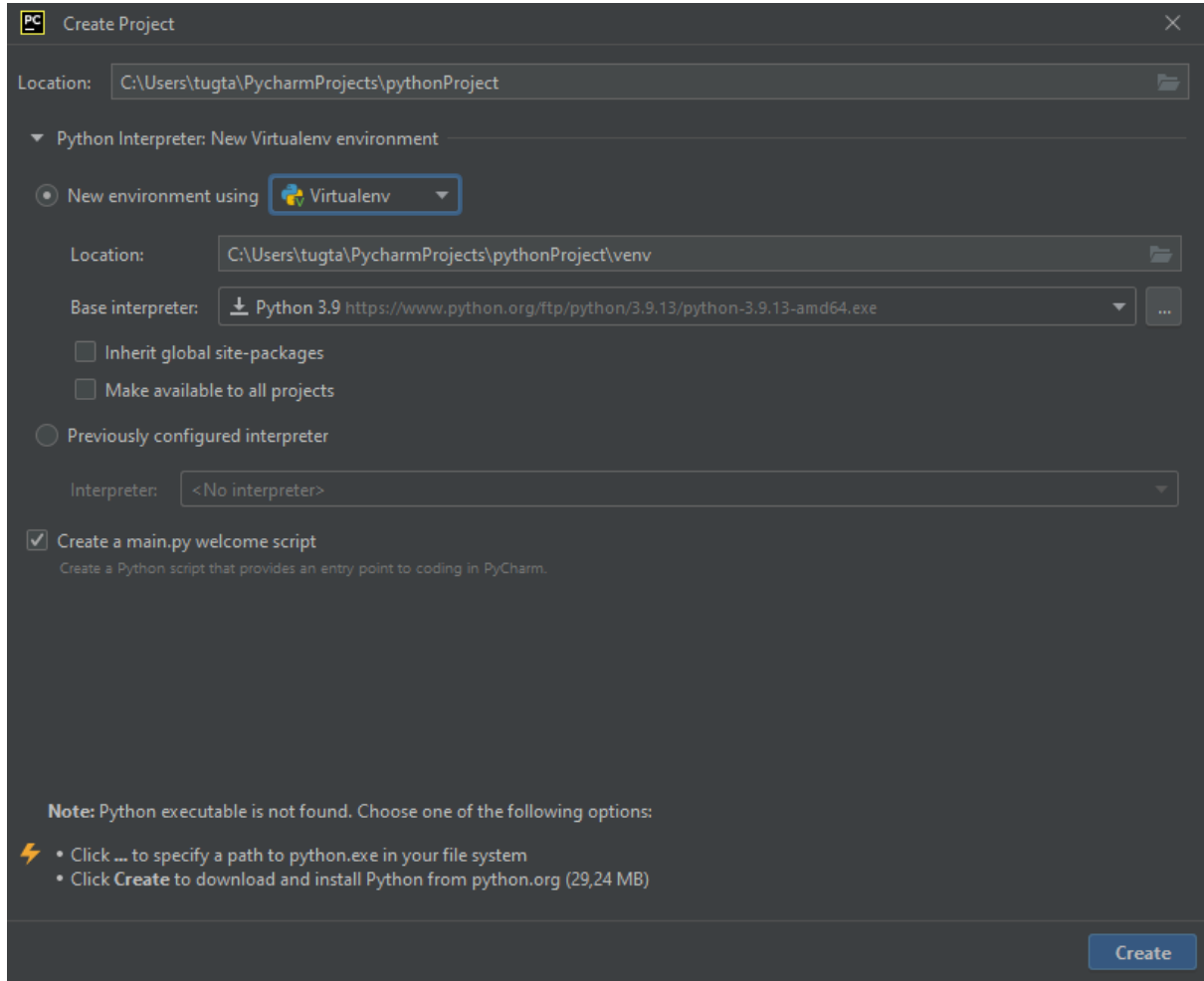
Sonrasında aşağıdaki gibi bir ekran çıkacaktır ve istediğiniz bir paket varsa ekleyebilirsiniz.



Ardından Pycharm kurulumu tamamlanmış oluyor. Aşağıdaki ekranda ilk açılış için proje yaratmanız istenmektedir. “Create New Project” diyoruz.



Aşağıdaki gibi bir ekranda “untitled” yazan yere proje ismini yazıp Environment bilgisi giriyoruz. Burada “VirtualEnv” seçiyoruz.



The image shows the 'Create Project' dialog in PyCharm. The 'Location' field is set to 'C:\Users\tugta\PycharmProjects\pythonProject'. Under 'Python Interpreter: New Virtualenv environment', the 'New environment using' option is selected with 'Virtualenv' chosen from the dropdown. The 'Location' for the virtual environment is 'C:\Users\tugta\PycharmProjects\pythonProject\venv'. The 'Base interpreter' is 'Python 3.9' with the URL 'https://www.python.org/ftp/python/3.9.13/python-3.9.13-amd64.exe'. There are checkboxes for 'Inherit global site-packages' and 'Make available to all projects', both of which are unchecked. The 'Previously configured interpreter' option is also unchecked. The 'Interpreter' field shows '<No interpreter>'. The 'Create a main.py welcome script' checkbox is checked. A note at the bottom states 'Note: Python executable is not found. Choose one of the following options:' followed by two options: 'Click ... to specify a path to python.exe in your file system' and 'Click Create to download and install Python from python.org (29,24 MB)'. A 'Create' button is at the bottom right.

PC Create Project

Location: C:\Users\tugta\PycharmProjects\pythonProject

Python Interpreter: New Virtualenv environment

New environment using Virtualenv

Location: C:\Users\tugta\PycharmProjects\pythonProject\venv

Base interpreter: Python 3.9 <https://www.python.org/ftp/python/3.9.13/python-3.9.13-amd64.exe>

☐ Inherit global site-packages

☐ Make available to all projects

Previously configured interpreter

Interpreter: <No interpreter>

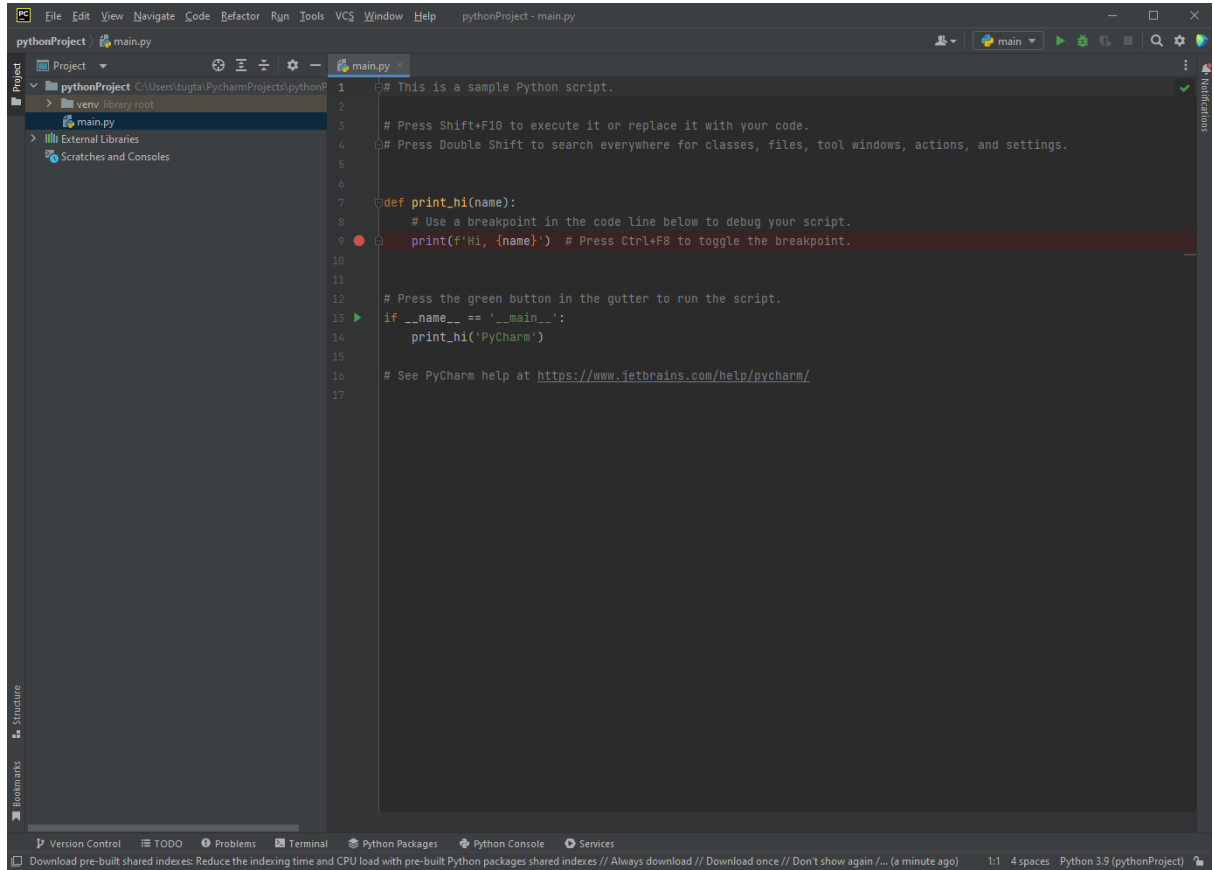
☒ Create a main.py welcome script
Create a Python script that provides an entry point to coding in PyCharm.

Note: Python executable is not found. Choose one of the following options:

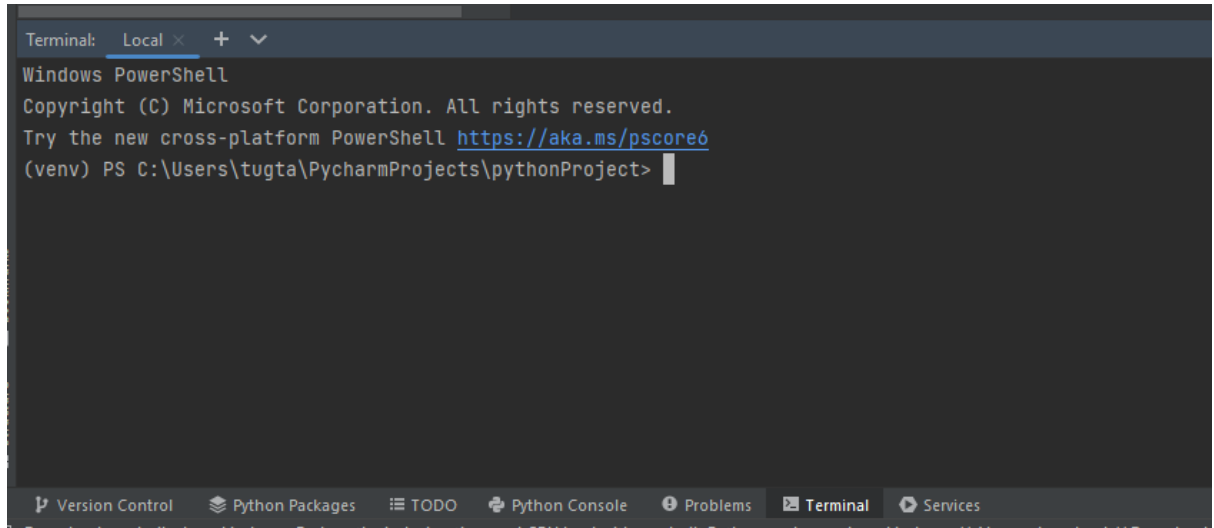
- Click ... to specify a path to python.exe in your file system
- Click Create to download and install Python from python.org (29,24 MB)

Create

Aşağıdaki ekran karşımıza çıktığında pycharm yüklenmiş oluyor.



Bu noktada aşağıda bulunan terminal yazısına basıyoruz ve virtual environment konsolunun açılmasını sağlıyoruz.



Konsolda aşağıdaki komutu yazıp Enter'a basıyoruz.

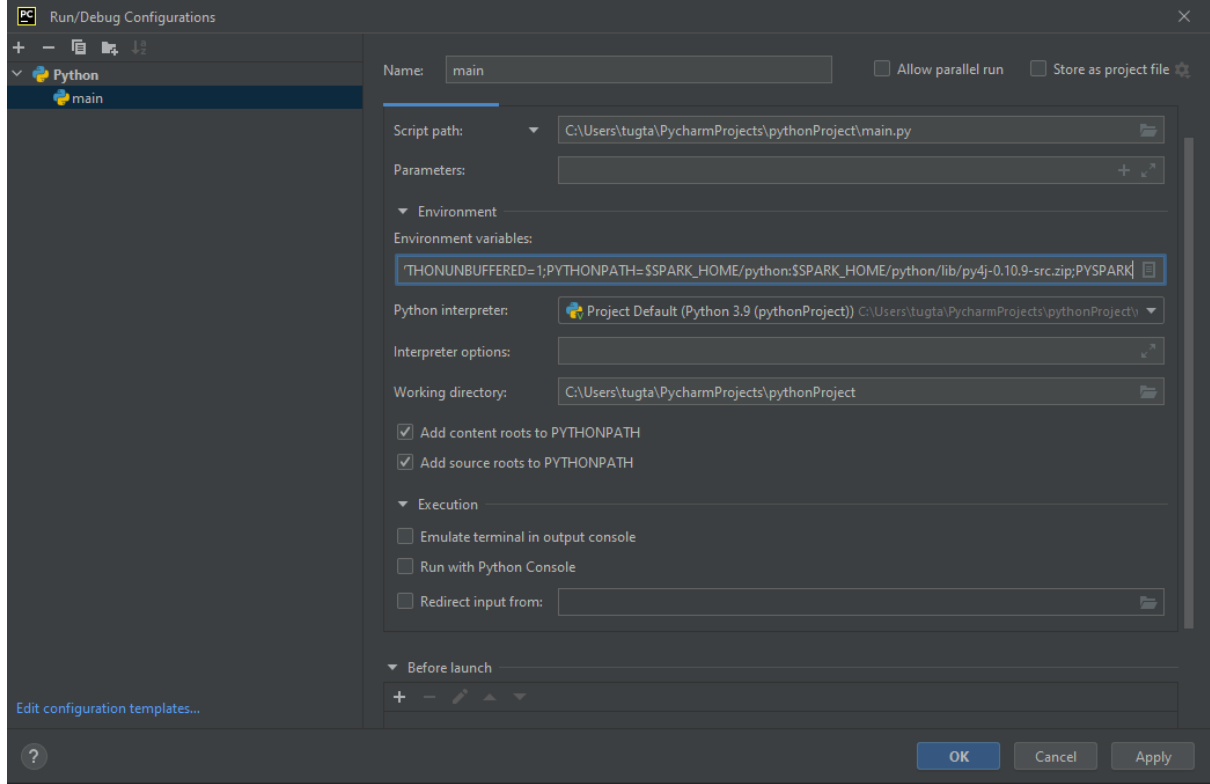
```
pip install pyspark==3.1.1 pyspark[sql]==3.1.1 numpy scipy matplotlib ipython jupyter pandas sympy nose scikit-learn pymongo
```

Proje'nin enviroment variable'ına da aşağıdaki satırları eklemek gerekmektedir.

```
PYTHONUNBUFFERED=1;PYTHONPATH=$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.9-src.zip;PYSPARK_PYTHON=C:\Users\tugta\Python\Development\itu-project\venv\Scripts\python.exe
```

Bunun için sağ üstte bulunan main'e basıp "Edit Configurations" 'a girmemiz lazım.

Aşağıdaki gibi bir ekran açılacak. Burada "Environment Variable" yazan yeri silip yukarıdaki satırı yapıştırıp kaydet deyip ekranı kapatıyoruz.



Test

main.py dosyasının içine aşağıdaki kodu yazıp çalıştığını test edebiliriz.

```
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
```

```
if __name__ == '__main__':
    conf = SparkConf().setAppName("Covid")
    spark = SparkSession.builder.master("local[*]").config(conf=conf).getOrCreate()
    sc = spark.sparkContext
    rdd = sc.parallelize([1, 2, 3, 4, 5, 6, 7, 8])
    print(rdd.collect())
```