

# Data Engineering

## Assignment 1: Big Data in Ihrem Umfeld (4 Punkte)

### 1.1


- Schemalos (unstrukturiert):
  - Wetterdaten, US-Präsidenten Wahl - Prediction (Beispiel von Mario)
- Schematisch (strukturiert):
  - Useraccounts bei Google (Name, email,etc.)

### 1.2

- Stream:
  - Videostreams (Wie orf.at oder twitch.tv)
- Batchverarbeitung:
  - Google Analytics (aus verschiedenen Bereichen in Batch kombinieren)

## Assignment 2: Big Data in Ihrem Umfeld

- Ich habe mich für Apache Spark entschieden, da ich es eher für explorative Daten und Batch processing geeignet ist. Die Test-Daten von Mario würden zum Beispiel eher in diese Kategorien fallen. Aber nach dem wir nur ein Beispiel Programm ausführen sollen, kann man eine wirkliche Entscheidung nicht wirklich treffen.
- Screenshot:

 **Spark Master at spark://192.168.1.12:7077**

URL: spark://192.168.1.12:7077  
REST URL: spark://192.168.1.12:6066 (cluster mode)  
Alive Workers: 1  
Cores in use: 16 Total, 0 Used  
Memory in use: 12.3 GB Total, 0.0 B Used  
Applications: 0 Running, 0 Completed  
Drivers: 0 Running, 0 Completed  
Status: ALIVE

#### Workers

Worker Id	Address	State	Cores	Memory
worker-20160607115918-192.168.1.12-52292	192.168.1.12:52292	ALIVE	16 (0 Used)	12.3 GB (0.0 B Used)

#### Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

#### Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

- IntelliJ IDEA 15.0.2 mit Maven

## Assignment 3: Big Data in Ihrem Umfeld

→ **WordCount\_spark** im GIT

# Data Science

## Assignment 1: Technologien

### 1.1

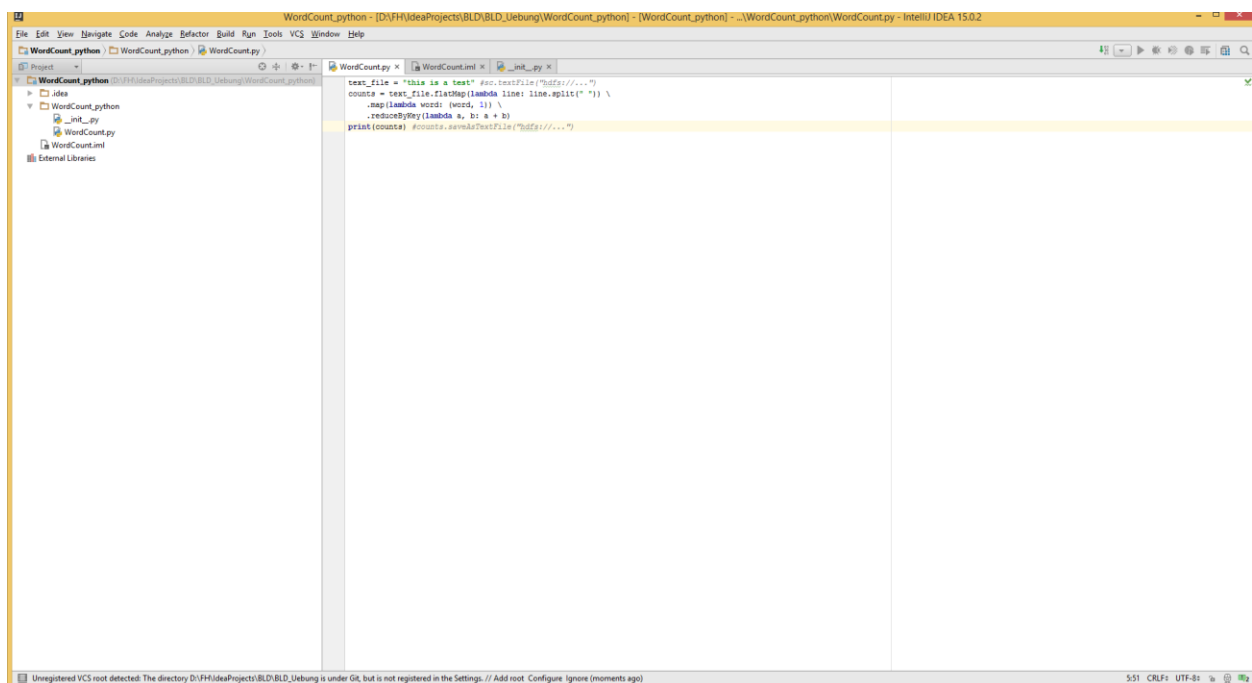
- SCALA
- Matlab
- JULIA

### 1.2

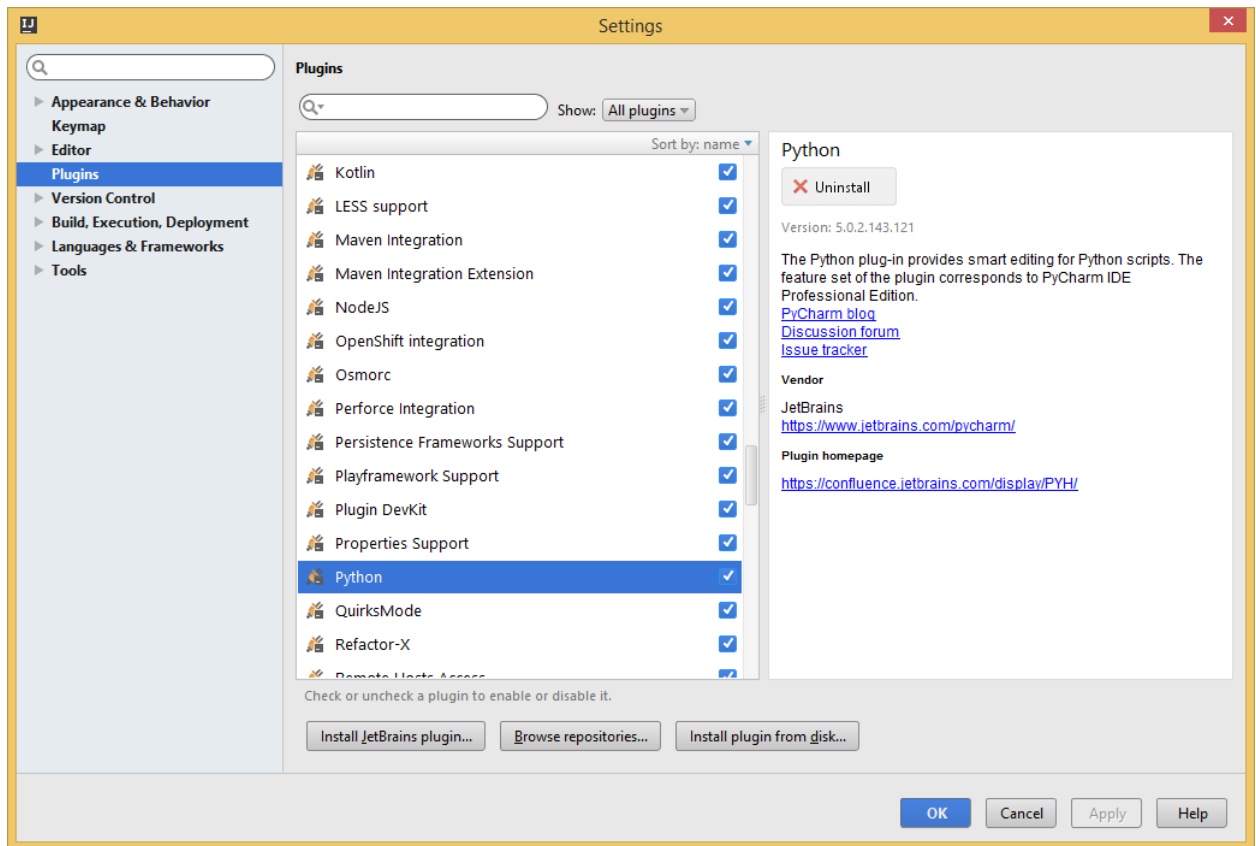
In der engeren Auswahl stehen Matlab (wegen Erfahrung durch FH-LVs) und Python (große Community und sehr Verbreitet). Ich würde mich für Python entscheiden da es viele Informationen, Guides, Foren, etc. gibt und dadurch viel Support.

## Assignment 2: Technologien

- Wie im vorherigen Punkt schon erwähnt, hat Python einen sehr großen Anwendungsbereich, viele Libraries und eine große Community. Deshalb Python.
- Screenshot der IDE:



- IntelliJ 15.0.2 mit Python Plugin (siehe Screenshot)



### Assignment 3: Big Science

- Classification:
  - Daten werden in bekannte Klassen eingeteilt
- Regression:
  - Wie manipuliert eine unabhängige Variabel X eine andere Variable Y
- Clustering:
  - Zusammenfassung von Daten in Gruppen - Cluster. Die Inhalte einer Gruppe sind sich ähnlicher als die Inhalte einer anderen Gruppe. (durch ein oder mehrere Attribute)
- Dimensionality reduction:
  - Verminderung der „Zufall-Variablen“ in den Datensätzen.

Z.B.: Google Ads (Zeigt Produkte die man bei Amazon besucht hat, oder anderen Content aus dem Internet)