

Assignments BLD Papp

Data Engineering

Assignment 1

1.1

Schematisch: Ich arbeite in der Firma an einem Projekt, wo ich Daten von Webservices (SOAP) abrufe. Diese sind in strukturierter Form zu verarbeiten.

Schemalos: Mailverkehr zwischen Kollegen, Kunden und mir. Kommunikationsdaten über Lync.

1.2

Gestreamt: Ich habe mich in der Firma mit Sitecore befasst. Sitecore ist ein CMS, welches auch Kundenaktionen auf der Webseite trackt und in eine NoSQL Datenbank speichert. Diese werden verarbeitet und die Ergebnisse in eine „Reporting“ Datenbank (MSSQL) gespeichert. Ich kann nicht sicher sagen, ob diese Daten tatsächlich gestreamt verarbeitet werden, aber möglich ist es.

Batchverarbeitung: Es werden regelmäßige Backups von Datenbanken gezogen. Dies wird einmal am Tag/Nacht gemacht.

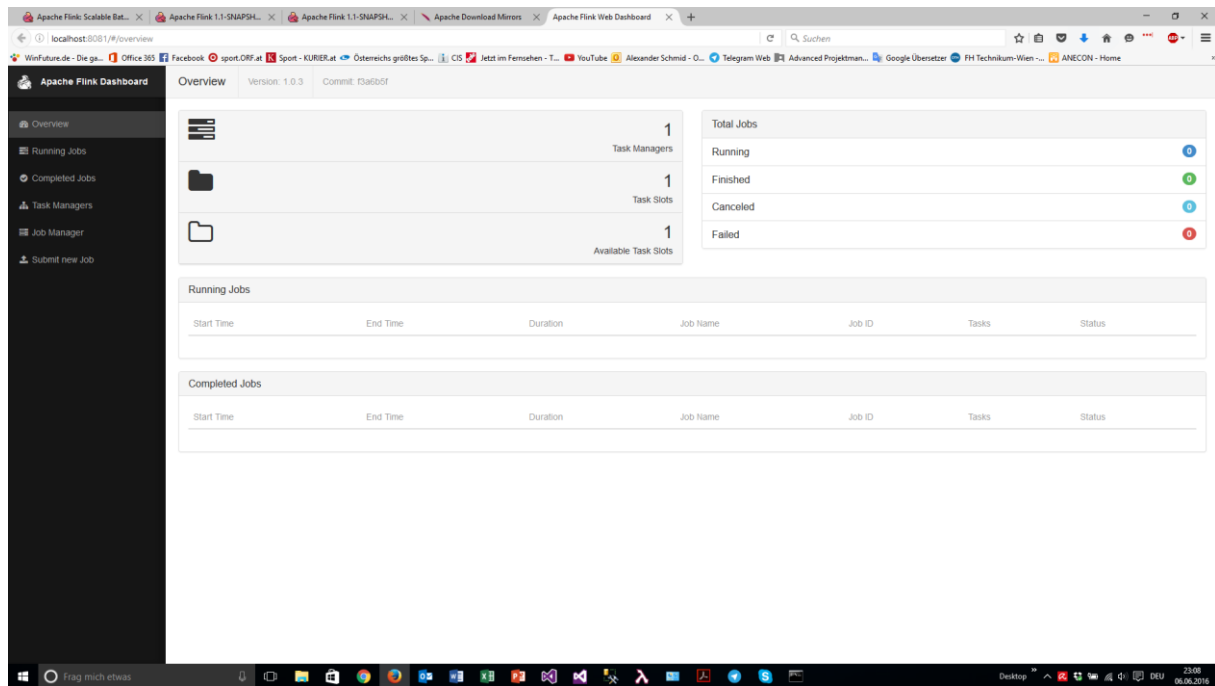
Assignment 2

2.1

Ich habe mich für Apache Flink entschieden. Der Grund dafür ist, dass Apache Flink auch mit Streaming besser umgehen kann (hat Streaming Engine und nicht Batch Engine wie Spark) und somit eine relativ niedrige Latenzzeit erreicht. Das ist ein wesentliches Kriterium und hat mich überzeugt.

2.2

Für die Installation wurde folgende Anleitung verwendet: https://ci.apache.org/projects/flink/flink-docs-master/setup/local_setup.html Es wurde die Source Archiv Datei heruntergeladen, entpackt und über die CMD-Shell die start-local.bat ausgeführt und somit Flink gestartet. Unter folgender Url kann die Startseite nun aufgerufen werden: <http://localhost:8081/>. Der Screenshot der Startseite sieht wie folgt aus:



2.3

Da Flink mit Java verwendet werden kann und nur über Maven Dependencies eingebunden werden muss, würde ich Java verwenden. Meine bevorzugte IDE für Java ist IntelliJ IDEA. Die Abhängigkeiten zu Apache Flink würde ich über Maven einbinden.

Assignment 3

Das Programm liegt im Ordner Flink_Programm. Das Programm kann mit einer IDE (vorzugsweise IntelliJ) geöffnet werden. Der Unittest kann anschließend gestartet werden durch maven test.

Im Unittest wird ein einfacher Wordcount ausgeführt. Die Beispieleingabe besteht aus einem Satz in dem 5 mal das Wort „Alexander“ enthalten ist. Der Unittest überprüft dies. Ebenfalls werden die Ergebnisse ausgegeben.

Data Science

Assignment 1

1.1

Weitere Frameworks um Daten zu analysieren sind Apache Zeppelin, Jupyter, SAS, Julia, SPSS und viele mehr.

1.2

Ich habe bis jetzt noch mit keiner Datasciencetechnologie gearbeitet. Sowohl R als auch Python sagen mir vom Namen her und aufgrund der Vorlesung etwas, aber genauer kann ich die Technologien noch sehr schwer beurteilen. Vom Erlernen der Syntax würde ich mir vermutlich bei Python leichter tun, da diese Technologie im Vergleich zu R mehr Hochsprachen wie C, C# oder Java ähnelt. Im Ernstfall würde ich mich jedoch bei erfahreneren Kollegen besser informieren bzw. diese um Unterstützung bitten. Für dieses Assignment entscheide ich mich für Python.

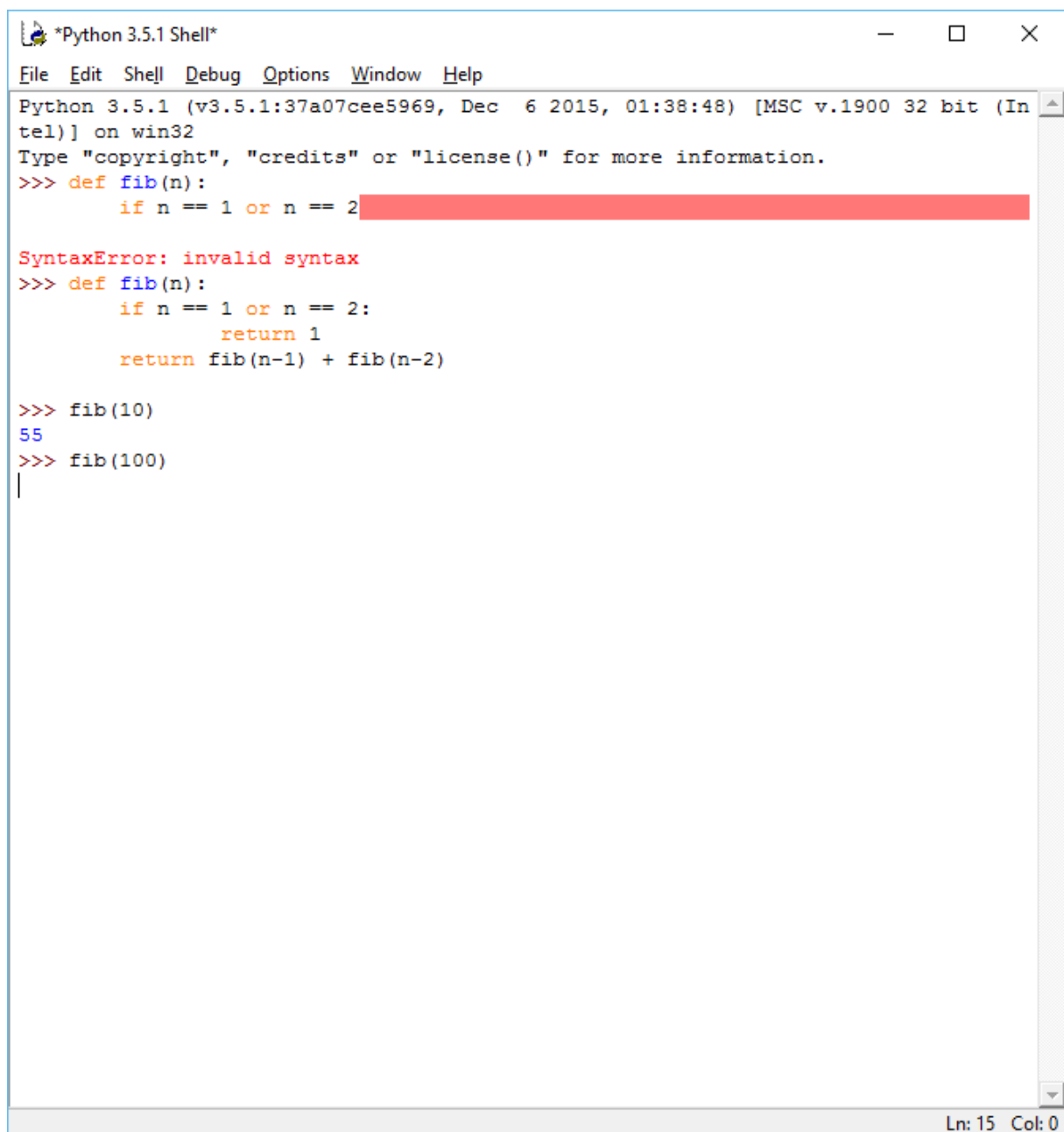
Assignment 2

2.1

Wie bereits unter DataScience 1.2 beschrieben, sind beide Technologien für mich noch ein unbeschriebenes Blatt. Da das Erlernen der Syntax bei Python vermutlich leichter ist, habe ich mich für Python entschieden. Nähere Information siehe DataScience 1.2

2.2

Python kann auf python.org heruntergeladen werden. Nach erfolgter Installation kann mit der Python Shell auf Windows gearbeitet werden. Hier ein Screenshot eines Beispielprogramms (fibonacci – Zahlen) auf meiner Python Shell:



```
Python 3.5.1 Shell
File Edit Shell Debug Options Window Help
Python 3.5.1 (v3.5.1:37a07cee5969, Dec 6 2015, 01:38:48) [MSC v.1900 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> def fib(n):
    if n == 1 or n == 2
SyntaxError: invalid syntax
>>> def fib(n):
    if n == 1 or n == 2:
        return 1
    return fib(n-1) + fib(n-2)
>>> fib(10)
55
>>> fib(100)
|
Ln: 15 Col: 0
```

2.3

Reicht eine Shell aus, würde ich weiterhin diese Shell wie auf dem Screenshot aus Assignment 2.2 verwenden. Habe ich hingegen umfangreichere Vorhaben in Python zu realisieren, würde ich auf PyCharm von JetBrains zurückgreifen, um mit einer bewährten IDE arbeiten zu können.

Assignment 3

Classification: Hier geht es darum, eine Einteilung in vordefinierte Kategorien zu treffen. Ein Beispiel aus meinem beruflichen Umfeld wäre eine Sitecorewebseite mit Tracking der Kundendaten. Bei erneutem Aufruf der Seite wird je nachdem, welche Klicks er auf der Webseite gemacht hat, der Kunde in eine Kategorie eingeteilt und entsprechend angepasste Inhalte gezeigt.

Clustering: Bei Clustering soll ebenfalls gruppiert werden. Es gibt jedoch keine vordefinierten Kategorien, sondern es werden die Daten nach gemeinsamen Merkmalen gruppiert. Ein Anwendungsbeispiel wäre das Clustering von Videos auf Youtube. Je nachdem welches Video ich gerade sehe, werden Videos mit ähnlichen Merkmalen auf der Seite vorgeschlagen.

Regression: Im Gegensatz zu den obigen beiden Algorithmen, soll bei Regression keine Kategorie sondern ein konkreter Wert vorhergesagt werden. Ein Beispiel aus meinem Alltag wäre: Wenn ich bei Amazon bestelle, werden meine Bestellungen für diverse Vorhersagen von Preisen bzw. zukünftige Bestellungen herangezogen.

Dimensional reduction: Dimensional reduction reduziert die Anzahl der Dimensionen um Algorithmen (wie clustering oder classification) einfacher anwenden zu können. Es wird daher vor den eigentlichen Algorithmen verwendet. In manchen Fällen wird der angewandte Algorithmus dadurch auch genauer.