

Software Visuali per Analisi Avanzate

Corso sostitutivo di Tirocinio



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

1



Comunicazioni di servizio

- Mancano gli elaborati di alcuni studenti. Pochi giorni ancora per inviare gli elaborati
- Nuovo link slide del corso
 - <https://github.com/valeriovvv/Corso-Software-Visuali-per-Analisi-Avanzate>
- Attestati di partecipazione SaS via mail a breve



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

2

Sommario

Presentazione del corso

Dati Strutturati, Semistutturati, Non Strutturati

Json – Approfondimento

Esercitazione

Installazione ambiente MySql

Obiettivi del Modulo 1



Ponte tra l'Università e le Aziende



Fornire una conoscenza del contesto dell'analitica avanzata



Fornire competenze su tecnologie e prodotti utili alle aziende



Sensibilizzare e stimolare skill utili per lavorare in azienda



Avviare percorsi di certificazioni di prodotto

Obiettivi del Modulo 2



Accedere alle fonti dati disponibili nelle aziende



Approfondire la conoscenza dell'analitica avanzata



Fornire ulteriori competenze su tecnologie e prodotti presenti nel panorama aziendale



Sensibilizzare e stimolare skill utili per lavorare in azienda



Nozioni e tecnologie del contesto Big Data

Skill Modulo 1



Comprendere Requisiti e Necessità del cliente (Interno o Esterno all'azienda)



Risolvere i problemi in modo efficace ed efficiente



Comunicare i risultati ottenuti in modo corretto, comprensibile e interessante



Trovare e comprendere le informazioni rilevanti con cui interagire

Skill Modulo 2



Comprendere Requisiti e Necessità del cliente
(Interno o Esterno all'azienda)



Trovare, comprendere, elaborare, recuperare
le informazioni rilevanti con cui interagire



Risolvere i problemi in modo efficace ed
efficiente



Comunicare i risultati ottenuti in modo
corretto, comprensibile e interessante

Tecnologie



databricks



Calendario del corso

Giorno	Durata	Argomenti
Martedì 5/5 16.00 – 18.00	2 ore	Presentazione del Corso Dati Strutturati, semi-strutturati e non strutturati Software da installare per il corso
Sabato 9/5 9.30 - 13.30	4 ore	Database relazionali MySQL Database, Tabelle, Righe, Colonne Linguaggio SQL: Introduzione, Select, Where, Join
Martedì 12/5 16.00 – 18.00	2 ore	Linguaggio SQL: Creazione, modifica, eliminazione di tabelle
Sabato 16/5 9.30 - 13.30	4 ore	Linguaggio SQL: Query di aggregazione Ottimizzazione delle Query: Gli indici Operazioni sui database Introduzione ai permessi (Grant)
Martedì 19/5	2 ore	Analitica Avanzata in ambienti Big Data Introduzione all'ecosistema Hadoop Apache Spark
Sabato 23/5	4 ore	Databricks e suoi componenti Utilizzo dell'SQL per interrogare Big Data Analitica avanzata con Databricks e R
Martedì 26/5	2 ore	Machine Learning con Databricks e R
Sabato 30/5	4 ore	Introduzione al Deep Learning, frontiera dell'analitica avanzata Conclusione del corso

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020



9

Conseguimento crediti formativi



Rilevamento presenze.

Quando richiesto, inserire in chat:
Nome, Cognome e Matricola



Interazione durante le lezioni;
elaborati da eseguire durante le lezioni



Studio individuale con produzione di elaborati



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

10

Dati
strutturati,
semistutturati,
non strutturati



Capacità di memorizzazione

1956, IBM 350 Disk File (1.000 Kg) 5 MB

2020, Usb DataTraveler (pochi grammi) 2 TB



Il fenomeno della convergenza

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Ecco a voi i **Big Data**



Volume



Varietà



Velocità



Valore

<https://www.linkedin.com/pulse/caratteristiche-dei-big-data-le-6-v-valerio-morfino/>

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Ecco a voi i **Big Data**... continua



Veridicità



Vulnerabilità



Volatility



Visualization

<https://www.linkedin.com/pulse/caratteristiche-dei-big-data-le-6-v-valerio-morfino/>



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

15

Tipologie di Dati per struttura

- Riferendosi alla Varietà dei Big Data, esistono dati di tre tipologie:
 - Dati Strutturati
 - Dati Semi-Strutturati
 - Dati Non-Strutturati
- Tutti possono essere Big Data



16

Tipologie di Dati per struttura

Strutturati

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333

Database, XLS,

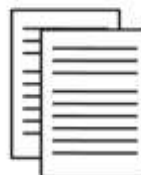
CSV

Semistrutturati



XML, JSON, HTML

Non Strutturati



TESTI

Immagini,
MP3 (audio),
MP4 (video)

Dati Strutturati

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333



Strutturati

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333

Database, XLS,

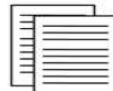
CSV

Semistrutturati



XML, JSON, HTML

Non Strutturati



TESTI

Immagini,
MP3 (audio),
MP4 (video)

Dati Strutturati

- Dati in posizione fissa all'interno di record con un tipo ben definito (es. testo, numero, data, ecc.)
- La struttura dei dati (definizione o metadati) è ben definita e rigida
- Metadati e Dati sono separati
 - Es. Nomi e tipo delle colonne di un file XLS e dati contenuti nelle celle
 - In un Database, tabelle, colonne e tipi di dato e dati contenuti
 - Nomi delle colonne di un CSV e valori delle righe
- Non è possibile avere strutture gerarchiche
 - Una cella di XLS può contenere un testo, un numero, non un elenco
- Alcuni esempi: Database Relazionali, Fogli di calcolo, CSV

Dati Strutturati

- Colonne in posizione fissa
- Netta separazione tra dati e metadati
- Tipi di dato ben definiti
- Non è possibile avere strutture gerarchiche

id-pers	nome	cognome	id-pers	telefono
0000001	Mario	Rossi	0000001	051 1234
0000002	Giorgio	Verdi	0000001	333 3333

	A	B	C	D	E	F
1	Nome	Cognome	Data di Nascita	Media		
2	Mario	Rossi	13/01/1998	28,7		
3	Maria	Verde	24/10/2001	27,3		
4	Luisa	Bianchi	10/09/2000	29		
5	Oreste	Arancio	11/07/1999	24,6		
6						
7						

data;stato;codice_regione;denominazione_regione;lat;long;ricoverati_con_sintomi
 2020-02-24T18:00:00;ITA;17;Basilicata;40.63947052;15.80514834;0
 2020-02-24T18:00:00;ITA;18;Calabria;38.90597598;16.59440194;0
 2020-02-24T18:00:00;ITA;15;Campania;40.83956555;14.25084984;0
 2020-02-24T18:00:00;ITA;08;Emilia-Romagna;44.49436681;11.3417208;10

Dati Non Strutturati



Strutturati

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333

Database, XLS,

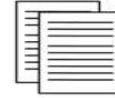
CSV

Semistrutturati



XML, JSON, HTML

Non Strutturati



TESTI

Immagini,
MP3 (audio),
MP4 (video)

Dati Non Strutturati

- Sono dati caratterizzati da assenza di schema (es. Immagini, Video, Audio)
- O da schema molto leggero (es. un testo)
- Se lo schema non è presente, come nel caso di oggetti multimediali e file di solo testo narrativo, le modalità di gestione di questi dati cambiano significativamente rispetto a quelli con uno schema.
- Dati senza schema, o di cui tipicamente non si utilizza lo schema, sono di grandissima importanza: basti pensare a Internet e ai motori di ricerca, che sono per lo più sistemi di Web Information Retrieval.

Estrazione di informazioni dalle immagini



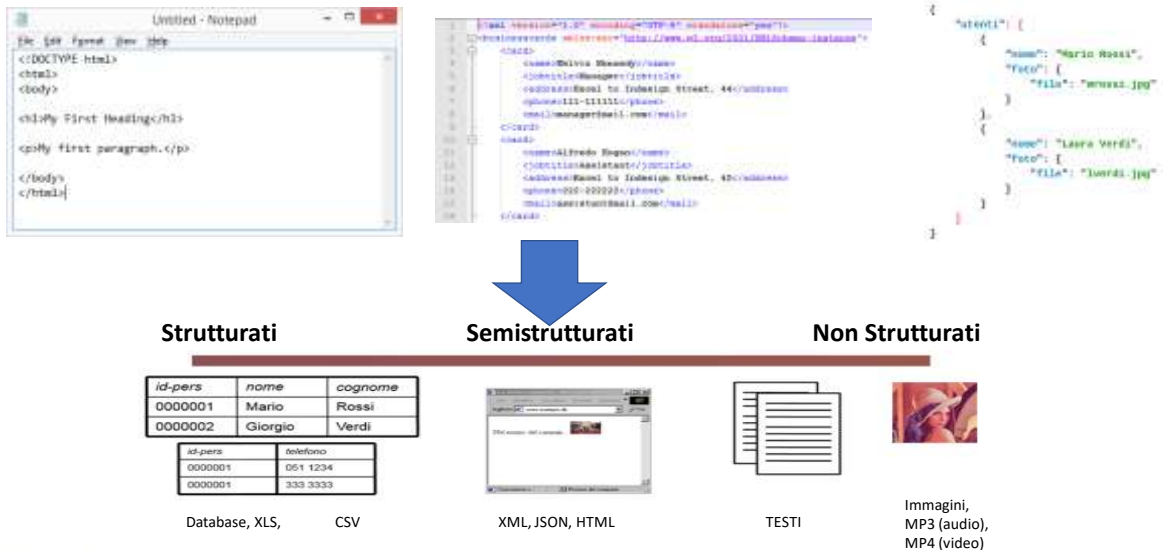
<https://azure.microsoft.com/it-it/services/cognitive-services/face/>

- Grazie ad algoritmi di Machine Learning e Deep Learning è possibile estrarre informazioni strutturate o semistrustrate dalle immagini

Estrazione di informazioni dai testi

- Ricerche sui testi attraverso algoritmi specifici
 - Ranking
 - Lemming
 - Stemming
- Sentimenti Analysis
- Creazione di topic e tag per la classificazione dei testi
- Estrazione automatica di Summary
- Natural Language Processing

Dati Semistruzzurati



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

25

Dati Semistruzzurati

- La struttura dei dati è flessibile
- Metadati e Dati sono tipicamente insieme
 - Ma possono essere presenti degli schemi di validazione
- La struttura è irregolare o parziale
- Sono comuni strutture gerarchiche
- I dati sono separati da TAG, Virgole, caratteri di separazione
- XML e HTML utilizzano una sintassi basata sui TAG
- JSON utilizza una sintassi basata sulla convenzione Javascript
- In genere per XML e JSON si parla di Documenti

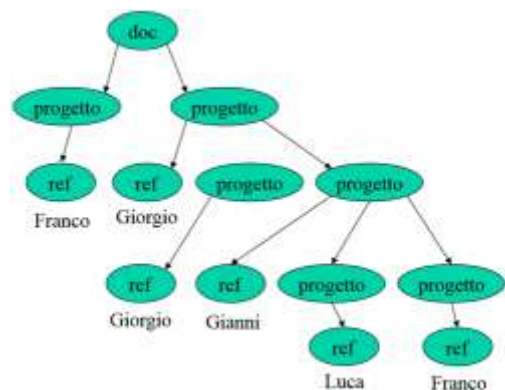
26

XML

- **XML (Extensible Markup Language)**
- Molto utilizzato per rappresentare informazioni in modo portabile (es. tra Windows, Mac, Linux, ecc.)
- Usato nella comunicazione client-server (es. SOAP)
- Ne esistono numerose estensioni
- Si basa sull'utilizzo di tag racchiusi tra un minore e maggiore. Questi ultimi possono essere definiti a piacimento per organizzare le informazioni.

Esempio di struttura gerarchica in XML

```
<doc>
  <progetto><ref>Franco</ref></progetto>
  <progetto>
    <ref>Giorgio</ref>
    <progetto><ref>Giorgio</ref></progetto>
    <progetto><ref>Gianni</ref>
      <progetto><ref>Luca</ref></progetto>
      <progetto><ref>Franco</ref></progetto>
    </progetto>
  </progetto>
</doc>
```



XML

- La struttura di un documento XML può essere più articolata di quella di un documento strutturato
- Può fornire una rappresentazione dell'oggetto modellato più vicina alla realtà:
 - Es. uno studente con tutti gli esami
 - Es. un cliente con tutti i documenti che lo riguardano
 - Es. la cartella clinica di un paziente
- E' possibile navigare un documento XML con il linguaggio Xpath
- E' Molto diffuso anche se oggi si tende a preferire JSON

Fonti e approfondimenti

- http://www.diit.unict.it/users/alongheu/sei2/aa0910/sei2_lezione01_introduzione.pdf
- <http://www.cs.unibo.it/~montesi/CBD/01IntroModelli.pdf>
- <http://www.disit.org/axmedis/ce7/00000-ce7de6e1-9d43-4776-8e5f-38b5de526d2f/3/~saved-on-db-ce7de6e1-9d43-4776-8e5f-38b5de526d2f.pdf>
- <http://docplayer.it/42537526-Argomenti-xml-json-linguaggi-per-la-definizione-e-lo-scambio-di-dati-strutturati-semi-strutturati-non-strutturati-xml-data-model-json.html>
- <https://lorenzogovoni.com/formati-file/>
- <http://reti.di.unimi.it/slide/xmljson.pdf>

JSON

Javascript Object Notation

31

Json

- **Formato JSON (Javascript Object Notation)**
- Basato sulla sintassi Javascript
- Permette di rappresentare gerarchie come XML
- E' più leggero e leggibile rispetto ad XML
- Utilizzato anche nelle comunicazioni client-server
- Usato anche come tipo di documento in database NoSQL, come ad esempio MongoDB

32

Formato di un documento Json

- Per essere validi i documenti devono essere Well Formed
- Ogni documento è racchiuso tra {}
- Ogni elemento del documento è una coppia **chiave : valore**
- Gli elementi sono separati da una **virgola**
- e possono essere in numero qualsiasi
- Tipo di dati ammessi:
 - Stringa
 - Numero
 - Boolean (true o false)
 - Object
 - Array
 - Null

```
{
  "first_name" : "Sammy",
  "last_name" : "Shark",
  "location" : "Ocean",
  "online" : true,
  "followers" : 987
}
```

Tipo di dato Object

- In Json i documenti possono essere annidati: un elemento può avere come valore un altro documento Json
- Questa tecnica è detta embedding

```
{
  "nome": "Maria",
  "cognome": "Rossi",
  "indirizzo": {
    "città": "Benevento",
    "via": "Via Roma, 42",
    "cap": "82100"
  }
}
```

Tipo di dato Array

- In Json i documenti possono contenere Array, ossia liste di elementi
- Gli elementi dell'Array sono racchiusi tra [] e separati da **VIRGOLA**

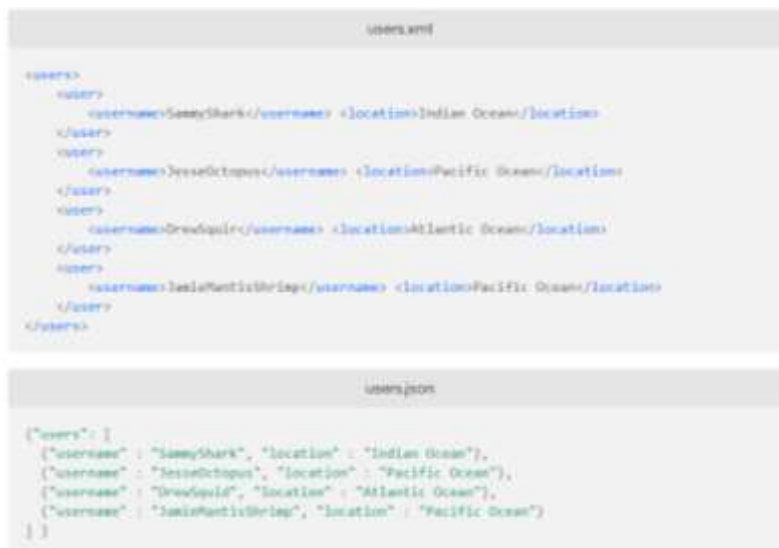
```
{
  "nome": "Maria",
  "cognome": "Rossi",
  "Lingue straniere": [
    "Francese",
    "Inglese"
  ]
}
```

Array ed Oggetti

- Array ed oggetti possono essere usati insieme a formare strutture complesse

```
{
  "nome": "Maria",
  "cognome": "Rossi",
  "Lingue straniere": [
    {
      "Lingua": "Francese",
      "Livello": "B2"
    },
    {
      "Lingua": "Inglese",
      "Livello": "C1"
    }
  ]
}
```

JSON vs XML



The image shows two code snippets side-by-side. The top snippet, titled 'users.xml', is an XML document with a root element 'users' containing five 'user' elements. Each 'user' element has attributes 'username' and 'location'. The bottom snippet, titled 'users.json', is a JSON object with a key 'users' containing an array of five objects. Each object has 'username' and 'location' properties. The data is identical in both formats.

```

users.xml
<?xml version="1.0"?>
<users>
  <user username="SammyShark" location="Indian Ocean"/>
</user>
  <user username="JesseOctopus" location="Pacific Ocean"/>
</user>
  <user username="DrewSquid" location="Atlantic Ocean"/>
</user>
  <user username="JanisPentisWreap" location="Pacific Ocean"/>
</user>
</users>

users.json
{"users": [
  {
    "username": "SammyShark",
    "location": "Indian Ocean"
  },
  {
    "username": "JesseOctopus",
    "location": "Pacific Ocean"
  },
  {
    "username": "DrewSquid",
    "location": "Atlantic Ocean"
  },
  {
    "username": "JanisPentisWreap",
    "location": "Pacific Ocean"
  }
]}

```

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Esercitazione Json

- Json Formatter: <https://jsonformatter.curiousconcept.com/>
- Esercizio 1:
 - Scrivere un file Json Well Formed che rappresenti le seguenti informazioni:
 - Nome, Cognome, Matricola, Esami sostenuti
 - Per ogni esame inserire: Titolo Esame, Data (formato: gg/mm/aaaa), Voto
 - Caricare i dati di uno studente che ha sostenuto 2 esami
- Esercizio 2:
 - Aggiungere al Json precedente un secondo studente che ha sostenuto 1 esame, facendo in modo che il Json rimanga valido
- Esercizio 3:
 - Salvare il file dell'esercizio 2. Aprire Tableau Public e calcolare la media dei voti per anno

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Esercizio 1

```
{
  "Nome":"Antonio",
  "cognome":"Rossi",
  "matricola":"0013424324",
  "esami":[
    {
      "Esame":"Statistica",
      "Data":"10/01/2020",
      "Voto":30
    }
  ]
}
```

Esercizio 2

```
{
  "Studenti":[
    {
      "Nome":"Antonio",
      "cognome":"Rossi",
      "matricola":"0013424324",
      "esami":[
        {
          "Esame":"Statistica",
          "Data":"10/01/2020",
          "Voto":30
        }
      ]
    }
  ],
}
```

CONTINUA



```
{
  "Nome":"Marco",
  "cognome":"Verdi",
  "matricola":"0013425655",
  "esami":[
    {
      "Esame":"Statistica",
      "Data":"10/01/2020",
      "Voto":29
    }, {
      "Esame":"Economia",
      "Data":"11/11/2020",
      "Voto":28
    }
  ]
}
```

Dati semistrutturati e sistemi di analitica

- Tableau, come la maggior parte dei sistemi di Visual Analytics, BI ed Advanced Analytics ha bisogno di dati in formato strutturato
- E' necessario effettuare delle conversioni, ad esempio appiattare il livello gerarchico (Tableau lo fa nativamente)
- Si avrà un numero di righe molto superiore al numero dei documenti Json, perché le testate si ripetono per ogni riga di dettaglio
- Quante righe si sono generate nel nostro JSON degli studenti?

Fonti e approfondimenti

- Json Formatter:
 - <https://jsonformatter.curiousconcept.com/>
- Sample File
 - <https://support.oneskyapp.com/hc/en-us/articles/208047697-JSON-sample-files>
- Tutorial:
 - <https://riptutorial.com/it/json>
 - <https://www.digitalocean.com/community/tutorials/an-introduction-to-json>
- Qualche Dataset:
 - https://catalog.data.gov/dataset?res_format=JSON

Gli strumenti del corso



Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Installazione di MySQL



- Installazione locale sul proprio PC Windows e Mac:
 - OK se il PC è abbastanza performante, si disponi di spazio e dei permessi
 - <https://www.apachefriends.org/it/download.html>
 - Installare seguendo le istruzioni (se viene richiesto, concedere accesso solo su reti private)
 - Avviare Xampp Control Panel
 - Avviare (start) Apache e Mysql
 - Aprire il browser sul link: <http://localhost/phpmyadmin/>
- Utilizzo in Cloud (nessuna installazione sul pc):
 - <https://www.db4free.net/>
 - Creare un account su: <https://www.db4free.net/signup.php>
 - Collegarsi al link: <https://www.db4free.net/phpMyAdmin/>

Dott. Valerio Morfino – Sistemi Visuali per Analisi Avanzate a.a. 2019-2020

Grazie per l'attenzione



<https://it.linkedin.com/in/valerio-morfino>



vmorfino@unisannio.it
