Facultatea de Electronică,
Telecomunicații și
Tehnologia Informației

SISTEME INTELIGENTE DE SUPORT DECIZIONAL

Ș.l.dr.ing. Laura-Nicoleta IVANCIU

Seminar 3 – Prelucrarea datelor. Big Data. Generarea limbajului natural.

Cuprins

- Prelucrarea datelor - regresie
- Big Data
- Generarea limbajului natural

Regresie – tehnică de Data Mining de tip predictiv (P)

Regresie liniară simplă

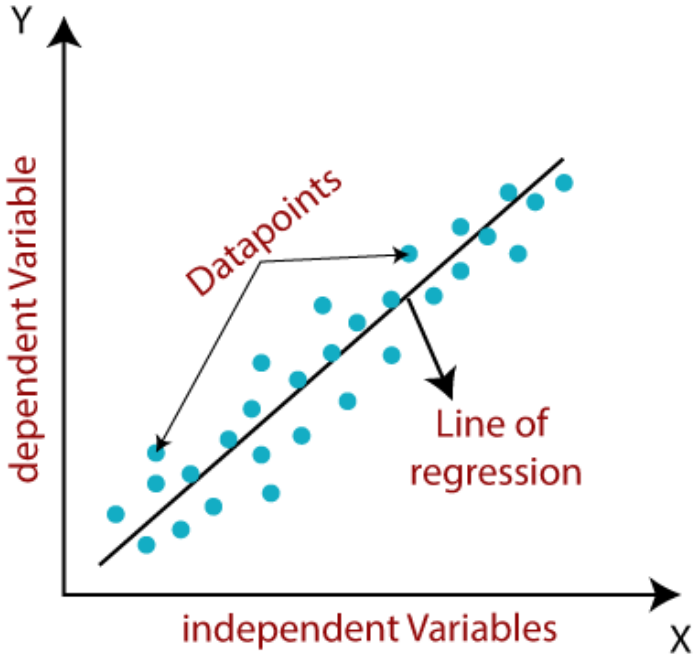
$$\widehat{y} = ax + b$$

X – variabilă independentă (predictor)

Y- variabilă dependentă (regresor)

a – panta

b – termen liber



Sursă

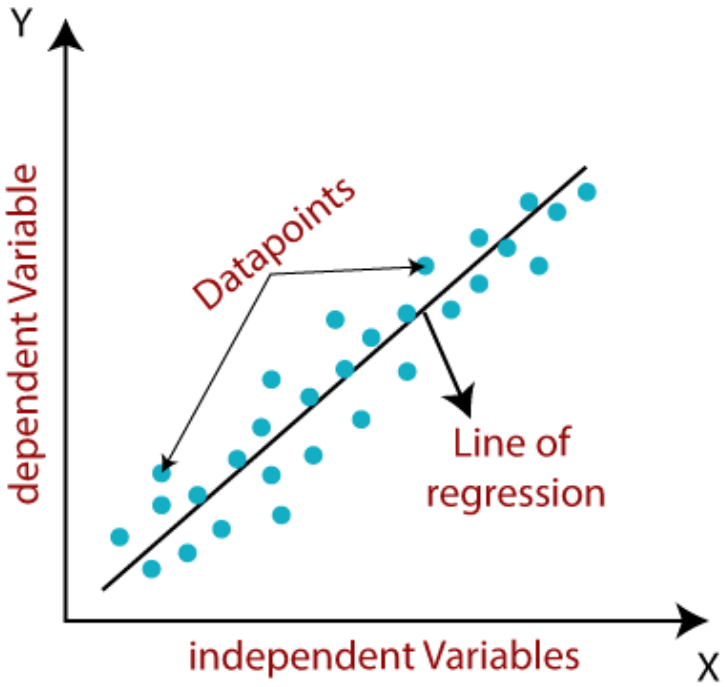
Eroare = valoare estimată (prezisă) – valoare cunoscută

Regresie

Regresie liniară multiplă

$$\hat{y} = a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n + b$$

X_i – variabile independente (predictori)
Y- variabilă dependentă (regresor)
a_i – coeficienți
b – termen liber
i = 1, 2, ..., n



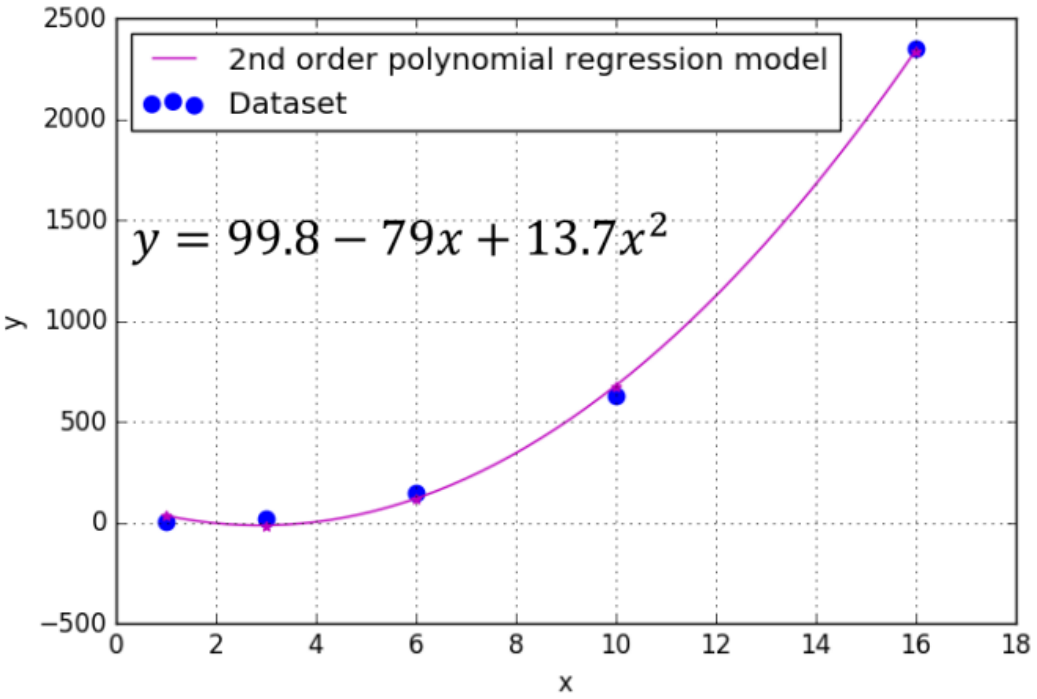
Sursă

Regresie

Regresie polinomială

$$\hat{y} = a_1 x + a_2 \cdot x^2 + a_3 \cdot x^3 + \ldots + a_n \cdot x^n + b$$

X – variabilă independentă
Y- variabilă dependentă
a_i – coeficienți
b – termen liber
i = 1, 2, ..., n



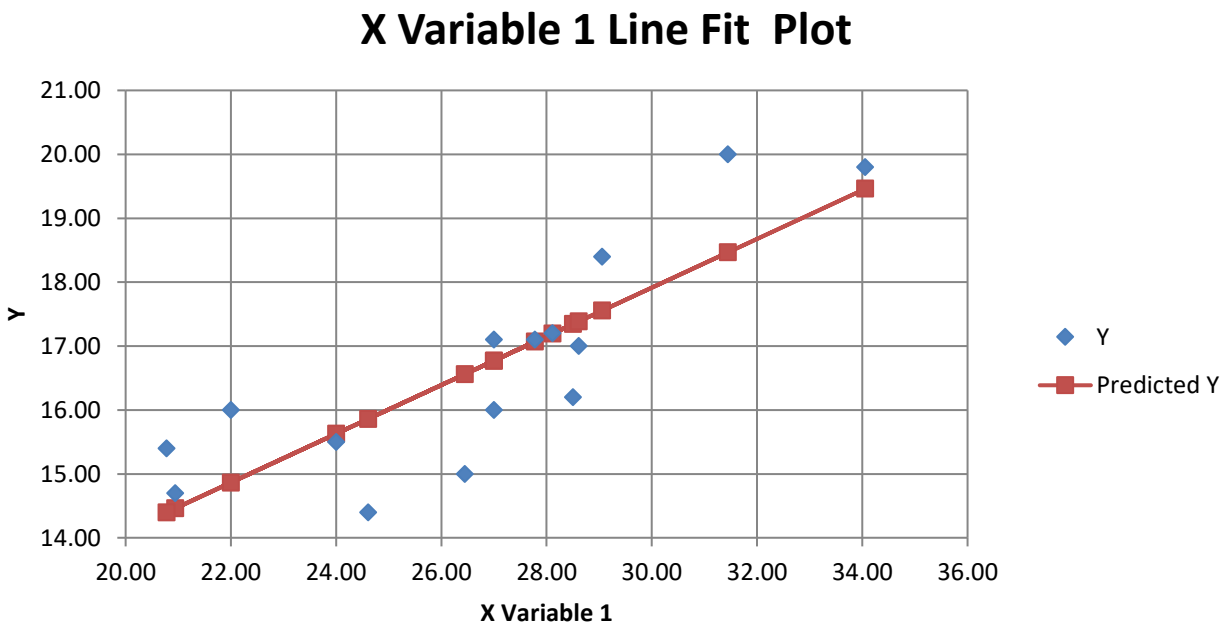
Sursă

Evaluarea calității regresiei

- Eroare $\hat{y}^{(i)} - y^{(i)}$
- Eroare pătratică $\left(\hat{y}^{(i)} - y^{(i)}\right)^2$
- Eroare medie pătratică (MSE) $\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$
- Root mean squared error (RMSE) $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2}$

Exemplu

Temp [Celsius]	Cricket chirps/s	Predicted
31.44	20.00	**18.47**
22.00	16.00	**14.86**
34.06	19.80	**19.46**
29.06	18.40	**17.56**
27.00	17.10	**16.77**
24.00	15.50	**15.63**
20.94	14.70	**14.46**
27.78	17.10	**17.07**
20.78	15.40	**14.40**
28.50	16.20	**17.34**
26.44	15.00	**16.56**
28.11	17.20	**17.20**
27.00	16.00	**16.77**
28.61	17.00	**17.39**
24.61	14.40	**15.86**

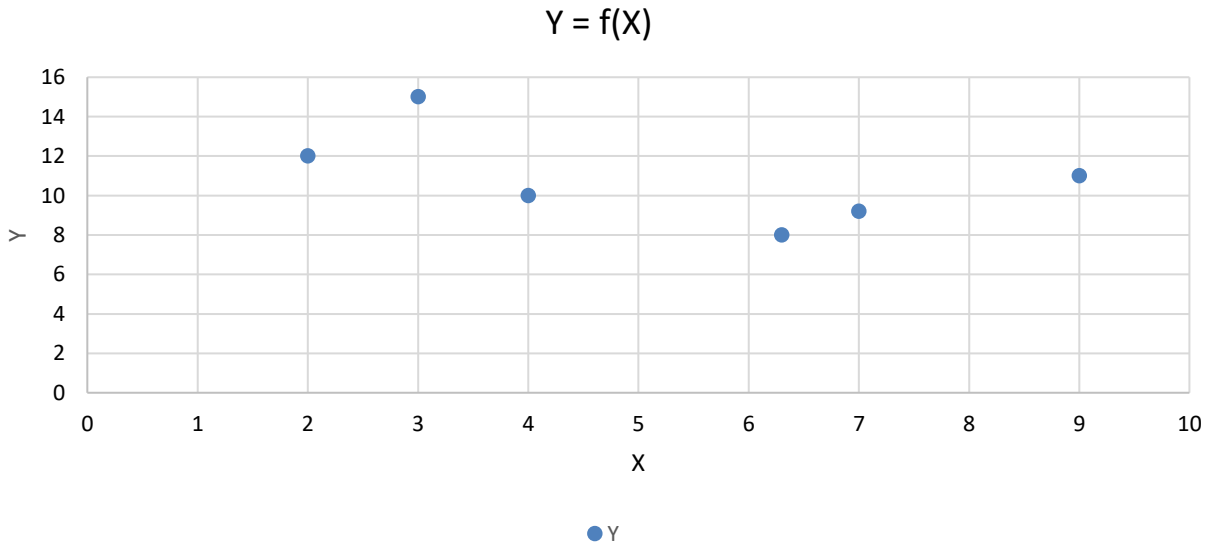


Exerciții

1. Pentru modelare prin regresie liniară a ecuației y = f(x), parametrii sunt:

a = - 0.5, b = 6.6

a) Care este ecuația de regresie liniară?
b) Reprezentați pe grafic dreapta de regresie.
c) Calculați valorile estimate pentru X = [2, 3, 4, 6.3, 7, 9]
d) Calculați eroarea și eroarea medie pătratică.



Exerciții

2. Pentru setul de date reprezentat în figură, se definesc:

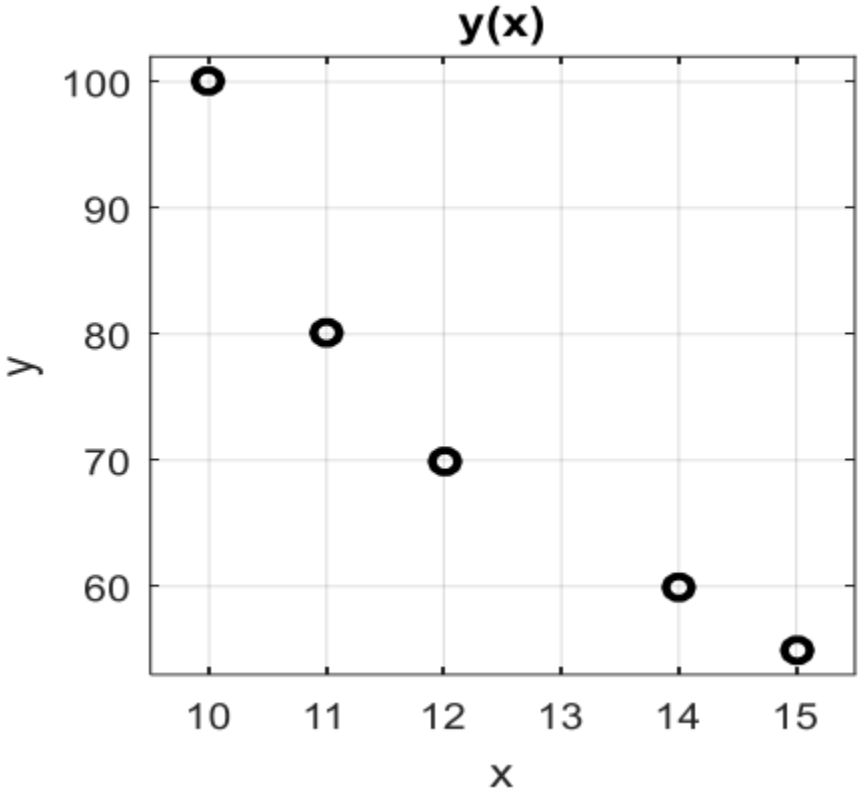
- model de regresie liniară

$\widehat{y_{lin}} = ax + b$, cu b = 175, a = −8

- model de regresie pătratică

$\widehat{y_{quad}} = a_1 x + a_2 x^2 + b$, cu b = 470; $a1$ = −56; $a2$ = 1.9

a) Care este ecuația modelului liniar? Reprezentați pe grafic dreapta de regresie.
b) Care este ecuația modelului pătratic? Reprezentați pe grafic curba de regresie.
c) Calculați valoarea estimată pentru x = 14, utilizând ambele modele. Care dintre cele două modele este mai precis? De ce?



Big Data

❑ Definiție:

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." (Gartner)

Big Data ≠ 3 V (https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#4195b17442f6)

1. 3 V – volume, velocity, variety
2. Cost-effective, innovative forms of information processing
3. Enhanced insight and decision making

Big Data

- Volume – volumul de date generate zilnic de platforme online, social media, aparatură, rețele, etc. Datele sunt stocate în "data warehouses".

- Velocity – viteza cu care datele sunt generate, în timp real
  - dinamica legăturii/raportului dintre date
  - legături între date care sunt generate cu viteze diferite
  - more data becomes less data

- Variety – date provenite din diferite surse, sub diferite forme
  - În trecut: baze de date, fișiere Excel (format tabelar)
  - În prezent: email, documente PDF, imagini, video, mesaje, postări social media, etc

Big Data

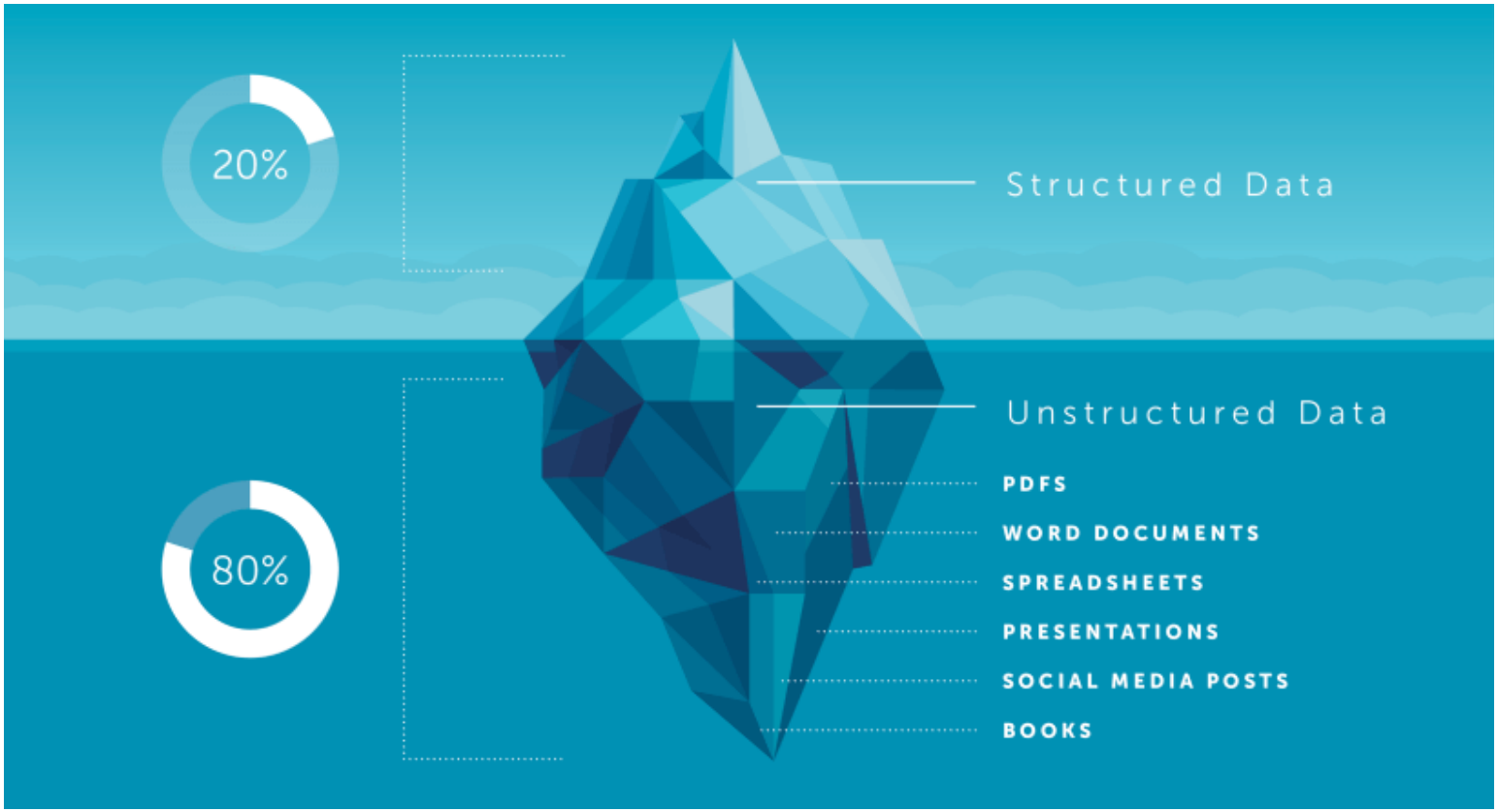
- Date structurate: date care pot fi procesate, stocate și accesate în format fix
  - informație organizată (tabelar, bază de date)
  - accesare prin algoritmi de căutare simpli

Ex: tabel cu informații despre candidații înscriși la admitere, cu nume, prenume, medie Bac, opțiuni

- Date nestructurate: date care nu sunt organizate după o formă sau structură
  - interpretare extrem de dificilă

Ex: email

Big Data



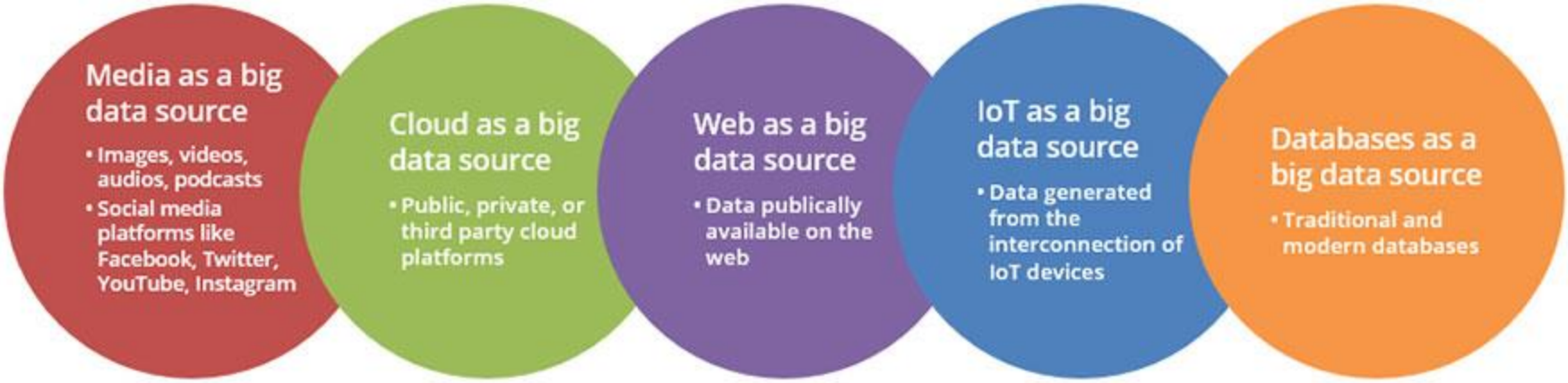
https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

- Date semi-structurate: date care nu sunt complet organizate în format fix, dar permit o căutare prin cuvinte cheie (tag)

Ex: date personale într-un fișier XML

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Big Data



https://www.allerin.com/blog/top-5-sources-of-big-data

Generarea limbajului natural

"Natural language generation (NLG) is the natural language processing task of generating natural language from a machine representation system such as a knowledge base or a logical form." (Wikipedia)

Funcție principală: transformarea datelor în reprezentare ușor de înțeles pentru oameni

Generarea limbajului natural

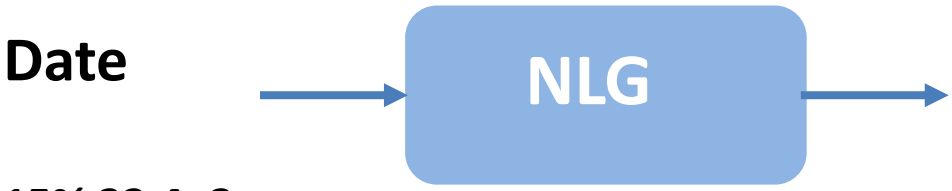
NLG + NLU = NLP

NLG/U/P – natural language generation/understanding/processing

NLU – înțelegere și interpretare a limbajului natural
NLG – transformarea datelor în limbaj natural

***NLG** can write, but cannot read. **NLU** can read, but cannot write.*

Generarea limbajului natural



Exemplu: generare scrisori pentru consumatori (de la bănci, companii de utilități, companii de telefonie mobilă, etc)

Generarea limbajului natural

❑ Tipuri de NLG

Basic NLG – propoziții simple, în care se înlocuiesc valori pe poziții predefinite

Ex. Temperatura în _____ este de ___ grade Celsius.

Template NLG – paragrafe mai complexe, generate pe bază de reguli, poziții predefinite ale unor valori (*placeholders*); analizează un set de date

Ex. generare rapoarte financiare, știri bazate pe rezultate sportive, etc

❑ Tipuri de NLG

Advanced NLG – produce texte complexe în limbaj natural, utilizând reguli sintactice și morfologice; cel mai complex tip de NLG

Ex. generare rapoarte de final de an detaliate pe baza situației financiare.

ChatGPT este un motor cu NLG/NLU/NLP?

Generarea limbajului natural

- Etape ale NLG

1. Analiza conținutului

   Care sunt principalele subiecte? Care este legătura dintre ele?

2. Înțelegerea datelor

   Analiza (inclusiv statistică) a datelor

   Implică tehnici de data mining/machine learning (identificare șabloane, tendințe)

3. Structurarea documentului

   Similar cu “scheletul” unei povestiri

❑ Etape ale NLG

4. Agregarea propozițiilor/frazelor

Mariei îi este foame. Maria are un măr. Maria mănâncă mărul.
Mariei îi este foame, așa că își mănâncă mărul.

5. Structurare gramaticală/sintactică – textul generat trebuie să fie corect

6. Rafinare limbaj – utilizare sinonime, cuvinte care să fie cât mai apropiate de limbajul natural

- Generați un paragraf pe baza tabelului:

ValidDate	AreaID	Value
27/06/2005	1 (North)	6
27/06/2005	2 (North West)	5
27/06/2005	3 (Central)	5
27/06/2005	4 (North East)	6
27/06/2005	5 (South West)	8
27/06/2005	6 (South East)	8

- Rezultat:

Grass pollen levels for Monday remain at the moderate to high levels of recent days with values of around 5 to 6 across most parts of the country. However, in southern areas, pollen levels will be very high with values of 8.

Monday looks set to bring another day of relatively high pollen counts, with values up to a very high eight in the Central Belt. Further North, levels will be a little better at a moderate to high five to six. However, even at these lower levels it will probably be uncomfortable for Hay fever sufferers.

Care dintre ele "sună" mai bine, mai natural?

please generate a paragraph using the following data, where on each line the first value is the date, the second is the region, and the last is the pollen level
27/06/2005 1 (North) 6
27/06/2005 2 (North West) 5
27/06/2005 3 (Central) 5
27/06/2005 4 (North East) 6
27/06/2005 5 (South West) 8
27/06/2005 6 (South East) 8

On June 27, 2005, various regions across the country experienced different pollen levels. In the northern region, the pollen level was moderate, measuring at 6 on the scale. Moving towards the northwestern area, residents enjoyed slightly better conditions with a pollen level of 5. In the central region, the pollen level mirrored that of the northwestern area at 5. As we ventured into the northeastern part, the pollen count increased to 6, signifying a moderate pollen presence. Meanwhile, in the southwestern and southeastern corners of the country, individuals had to contend with higher pollen levels, as both regions recorded an 8 on the pollen scale, indicating a significant pollen presence. This data underscores the regional variability in pollen levels on that particular summer day.

regenerate the previous paragraph, but using simpler words

On June 27, 2005, different parts of the country had varying amounts of pollen in the air. In the northern region, there was a moderate amount of pollen, which was a 6 on the scale. Moving to the northwest, there was a bit less pollen with a level of 5. In the central area, it was the same as the northwest with a level of 5. As we went to the northeast, there was a bit more pollen with a level of 6. But in the southwest and southeast areas, there was even more pollen, with a level of 8. This shows that the amount of pollen in the air was different in various regions on that summer day.

Exerciții

❑ Exercițiu:

Generați un paragraf cu tema:

1. *Raport despre activitatea copilului la grădiniță*
2. *Prognoza meteo pentru mâine*
3. *Știre despre deschiderea unei noi companii în oraș*
4. *Profil de utilizator pe o platformă social media*
5. *Descriere a unui model nou de telefon mobil*

utilizând și evidențiind etapele NLG.