Tema 2

Std. Plt. Maj. Grecu Gabriela Teodora

Std. Sg. Maj. Vlasceanu Mihnea Constantin

Std. Plt. Zaharia Ciprian Florin

Cuprins

[Scopul temei 2](#_Toc93943565)

[Structura setului de date 2](#_Toc93943566)

[Taskuri 2](#_Toc93943567)

[A. Numărul mediu de propoziții per literal 2](#_Toc93943568)

[B. Numărul mediu de synseturi per literal 2](#_Toc93943569)

[C. Calculul la nivel de literal pentru synseturile posibile 3](#_Toc93943570)

[D. Numărul de propoziții pentru fiecare synset atins 4](#_Toc93943571)

[E. Numărul de propoziții realizat de fiecare utilizator 4](#_Toc93943572)

[Moduri de afișare 4](#_Toc93943573)

[CSV file 4](#_Toc93943574)

[Matplotlib 4](#_Toc93943575)

[Plotly 7](#_Toc93943576)

[Concluzii 9](#_Toc93943577)

# 

# Scopul temei

Scopul acestei teme este dezvoltarea unui cod python in Google Colabs care să facă o analiză asupra conținutului setului de date WSD Database.

# Structura setului de date

Baza de date este stocată sub forma unui fișier pickle ce conține un dicționar în care cheile sunt literali iar valorile sunt o listă de propoziții ce vor fi folosite la antrenarea unui algoritm IA. Fiecare propoziție este codificată printr-un dicționar ce conține următoarele chei:

* user\_id: id-ul utilizatorului care a făcut propoziția
* literal: cuvântul al cărui sens se caută
* synsets: o listă cu id-urile posibilelor întelesuri ale literalului
* correct\_synset\_id: id-ul definiției corecte a literalului în această propoziție
* sentence: întreaga propoziție
* text\_prefix: prima parte a propoziției
* text: literalul așa cum apare în propoziție (poate fi articulat/ la plural)
* test\_postfix: ultima parte a propoziției

# Taskuri

## Numărul mediu de propoziții per literal

Se cere calculul mediei aritmetice a numărului de propoziții al tuturor literalilor din baza de date

## Numărul mediu de synseturi per literal

Se cere calculul mediei aritmetice a numărului de synseturi al tuturor literalilor din baza de date. Fiecare literal are o serie de synseturi candidat (posibile întelesuri ale cuvântului), printre care și “-1”, valoare ce semnalizează faptul că literalul în contextul acelei propoziție nu are sensul niciunuia din synseturile posibile. Se va ignora valoarea “-1” în calcularea mediei numărului de synseturi.

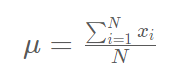
## Calculul la nivel de literal pentru synseturile posibile

Pentru fiecare literal se va calula distribuția de probabilități a synseturilor. Folosind distribuția de probabilități se vor calcula următoarele metrici:

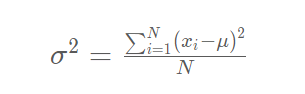
* media
* varianța
* deviația standard
* skewness (coeficientul de asimetrie)

Pe lângă aceste metrici se vor calcula și suma totală de propoziții per literal și numărul de synseturi cu 0 propoziții (posibile definiții care nu au fost alese niciodată).

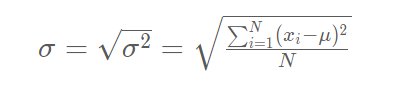
* **Media**arată tendința centrală a valorilor, adică spre care element tind să meargă toate intrările. Se calculează prin suma produselor dintre valorile din vectorul de probabilități și indicii acestor valori.

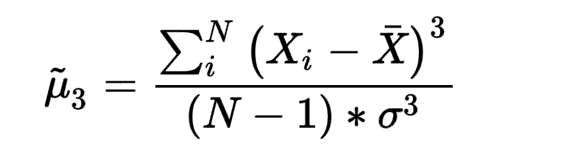


* **Varianța**, denumită și dispersie este o metrică ce arată cât de dispersate sunt valorile în distribuția dată. Se calculează ca suma produselor dintre valorile din vectorul de probabilități și pătratul diferenței dintre indicii valorilor și medie.



* **Deviația standard**este tot o metrică ce indică nivelul de dispersie a valorilor în distribuția dată. Se calculează ca radical din varianță.



* **Skewness**, sau indicele de asimetrie reprezintă o măsură a simetriei distribuției față de mdie. Atunci când skewness este 0, graficul este simetric. Când skewness este mai mic decât 0, distribuția are coada spre stânga, iar când skewness este mai mare decât 0, distribuția are coada spre dreapta. Formula de calcul este următoarea:

## Numărul de propoziții pentru fiecare synset atins

Cerința presupune extragerea tuturor synseturi de pe [RoWordNet](https://github.com/dumitrescustefan/RoWordNet). Se vor afișa toate synseturile atinse in baza de date. Pentru fiecare synset se vor afișa literalii la care sunt conectați și numărul de propoziții ai acestora. De asemenea, se va calcula și numărul total de propoziții pentru un synset, ca suma propozițiilor literalilor la care este conectat.

## Numărul de propoziții realizat de fiecare utilizator

Aceată cerință presupune afișarea numărului de propoziții pe care le a făcut fiecare utilizator

# Moduri de afișare

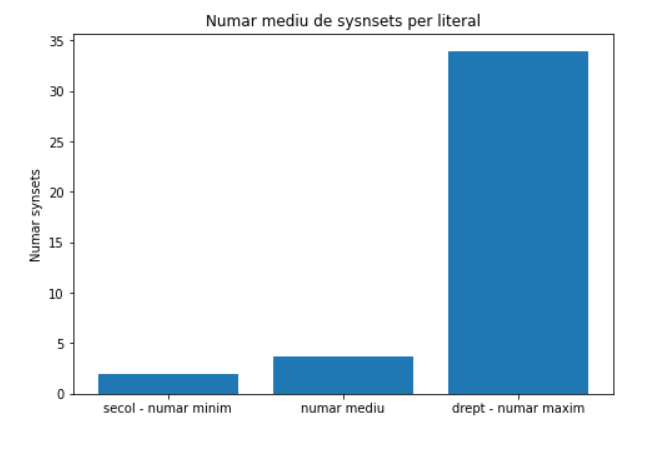
## CSV file

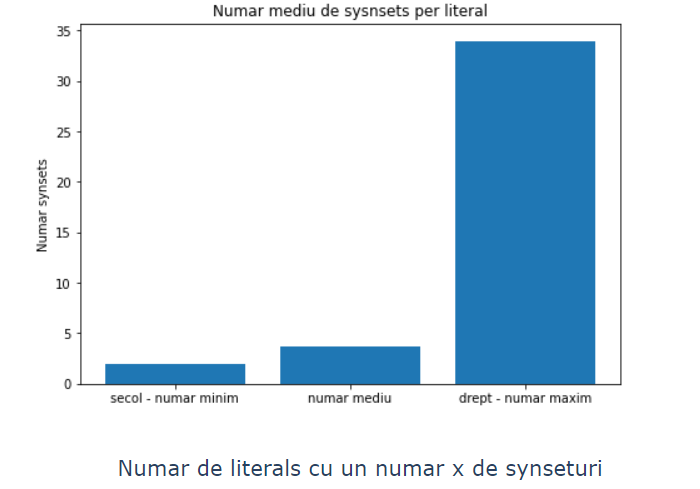
Pentru afișarea în fișiere .csv datele au fost puse în obiecte de tip DataFrame native bibliotecii pandas și pentru fiecare punct în parte rezultatele au fost scrise în fișiere cu denumirea pctX.csv.

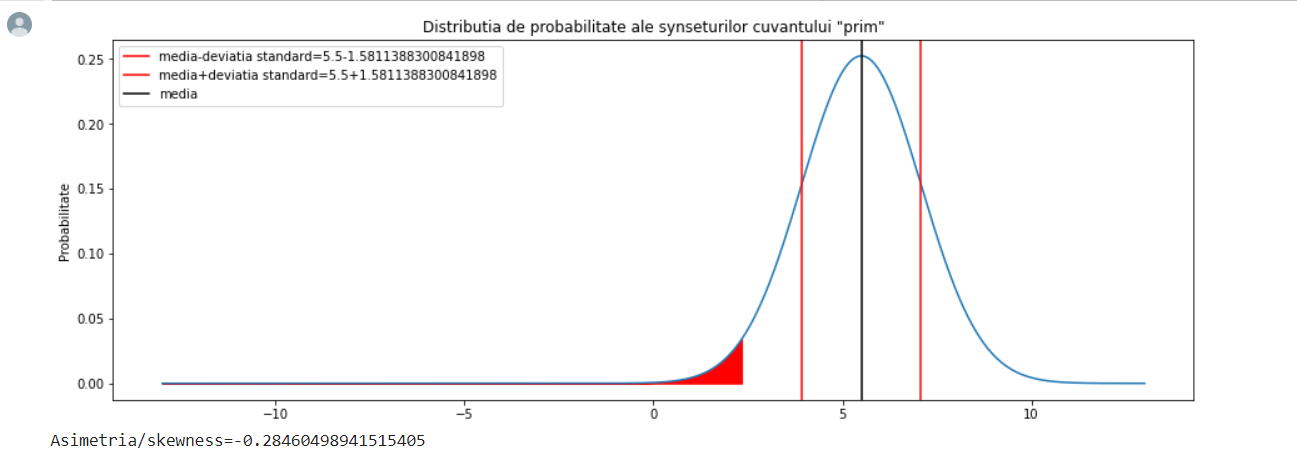
## Matplotlib

Am folosit biblioteca Matplotlib pentru a afișa datele în diferite tipuri de grafice:

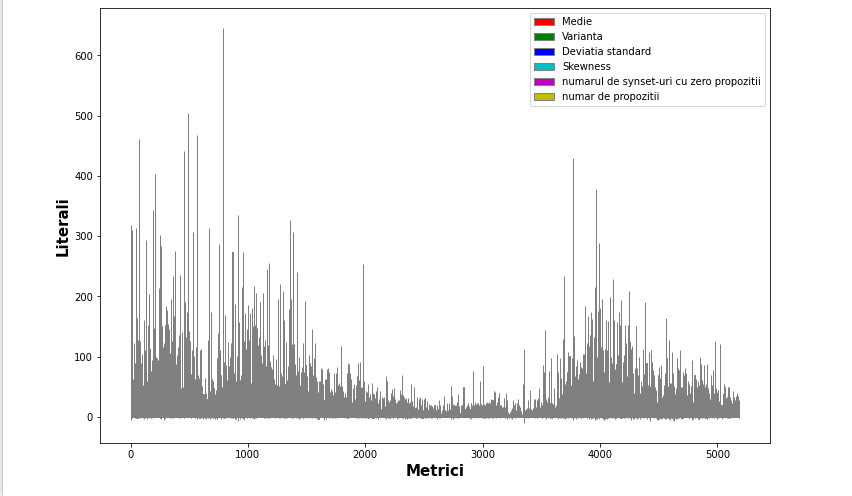
* **Grafic de tip bar pentru numărul de propoziții per literal (A)**



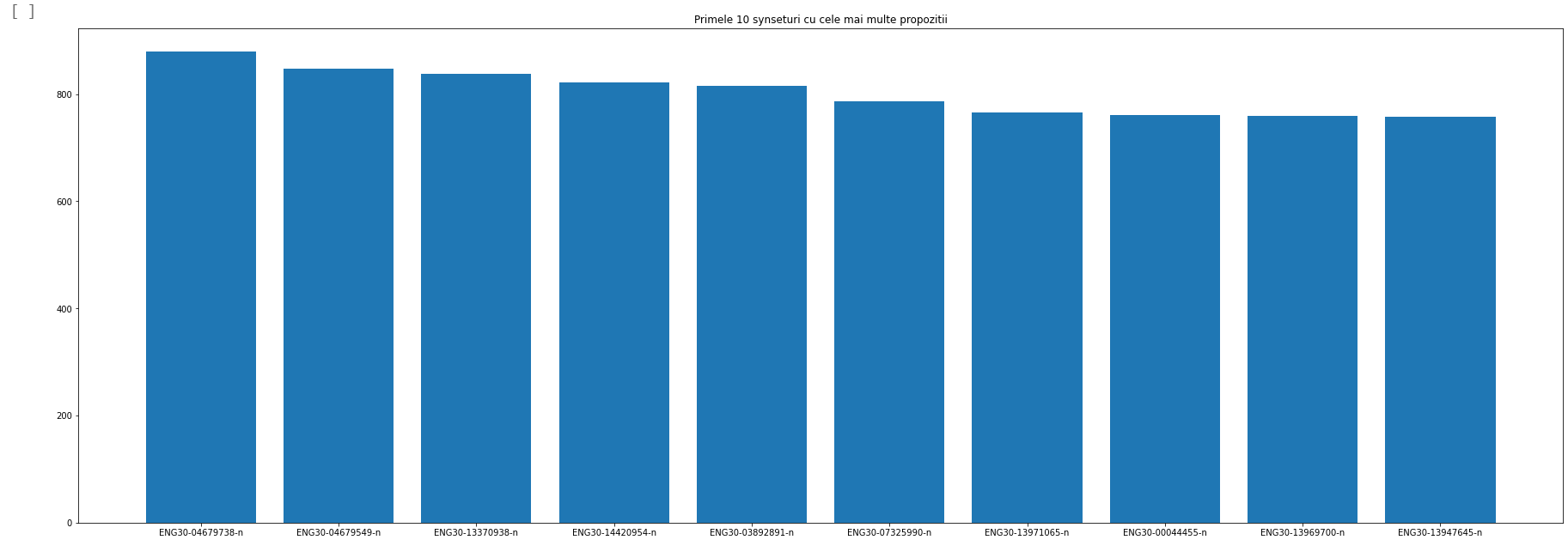
* **Grafic de tip bar pentru numărul de synseturi per literal (B)**
* **Grafic pentru distribuția de probabilități (C)**

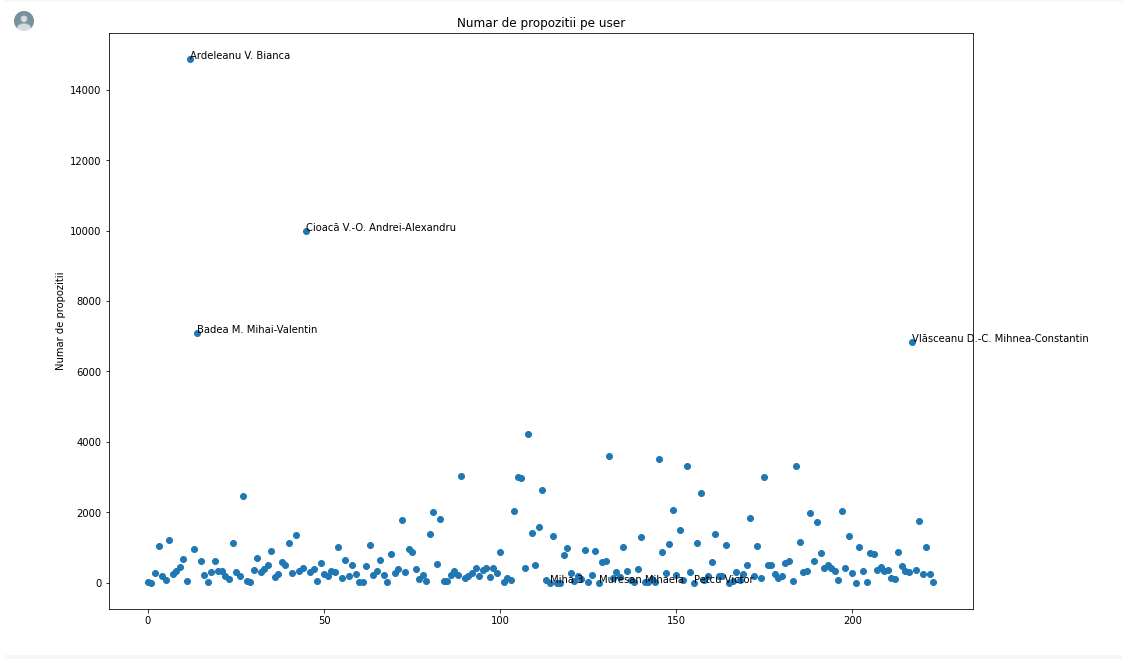


* **Grafic pentru metricile literalilor (C)**



* **Grafic de tip bar pentru numărul de propoziții per synset (D)**

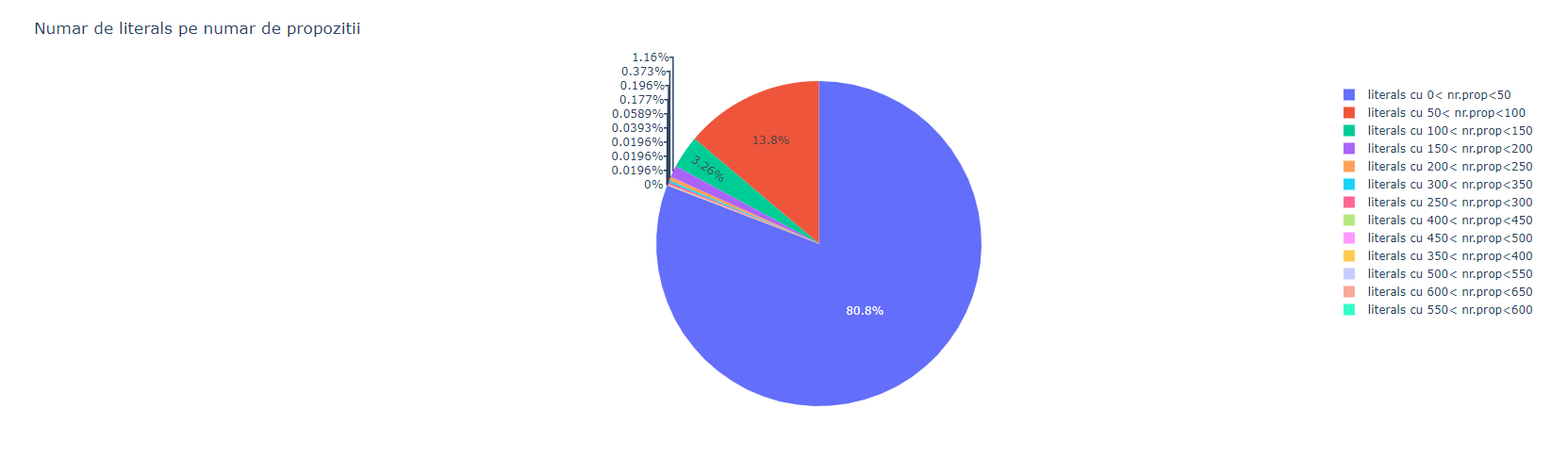


* **Grafic cu numărul de propoziții parcurse de fiecare utilizator (E)**

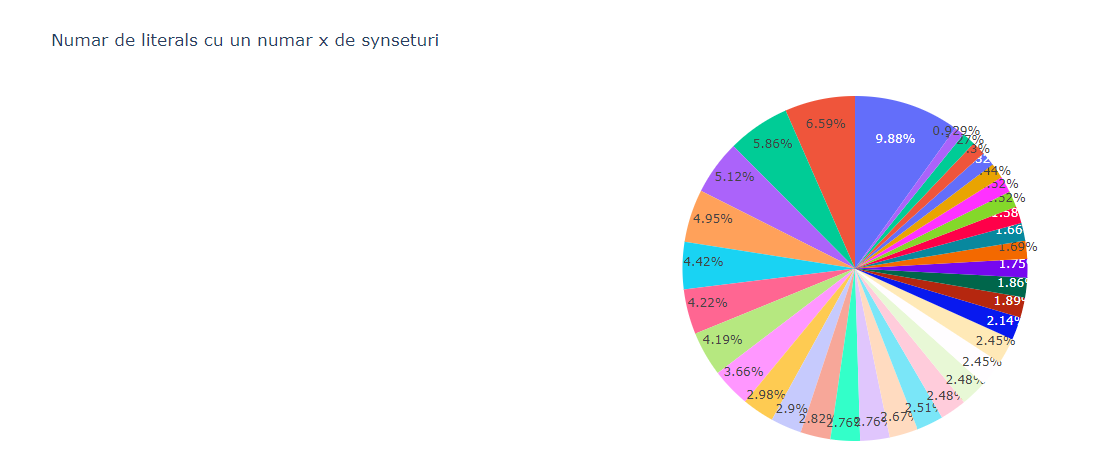
## Plotly

Am folosit plotly pentru a crea grafice interactive ce permit o analiză mai facilă și aprofundată a datelor. Acestea sunt câteva exemple:

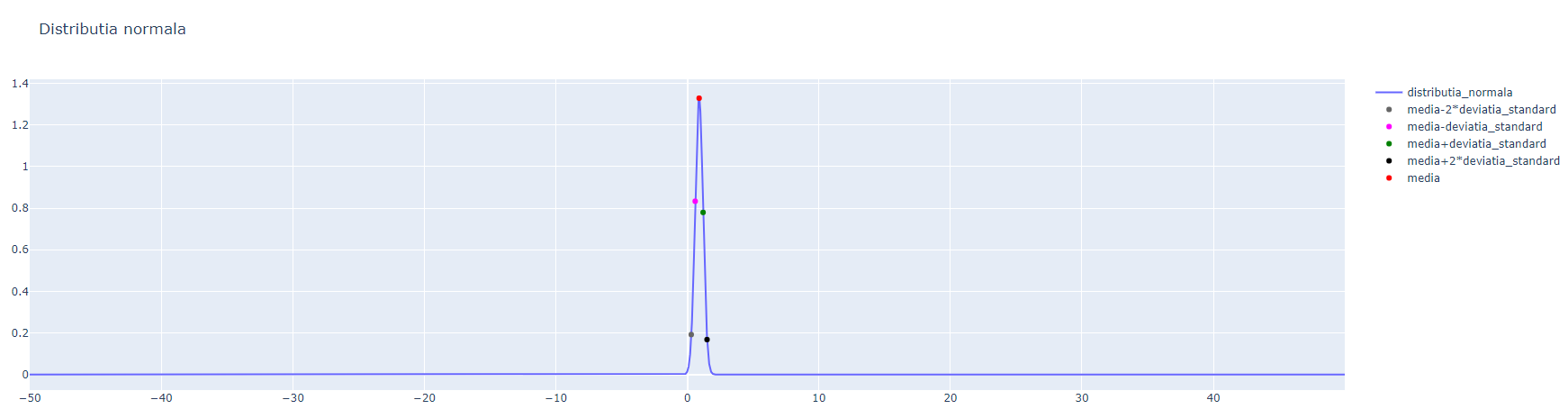
* **Grafic interactiv de tip pie pentru afișarea proporțiile de literali cu numere de propoziții în diverse intervale (A)**



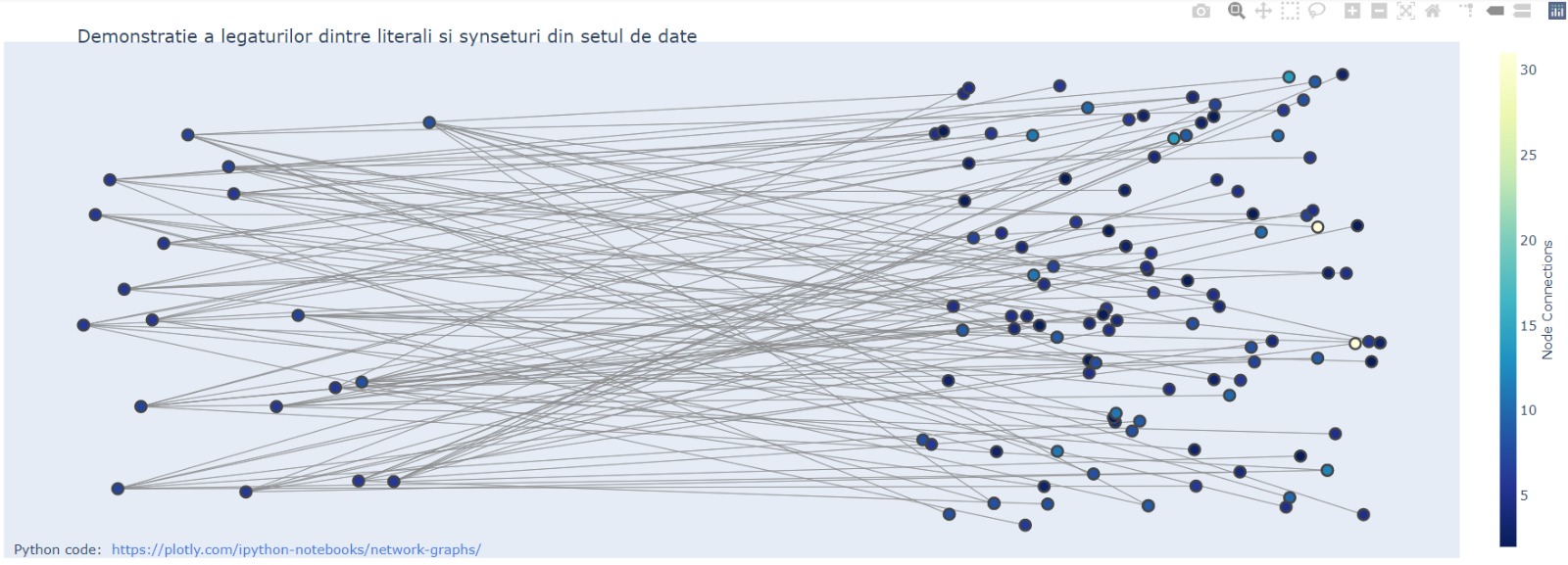
* **Grafic interactiv de tip pie pentru afișarea proporțiile de literali cu diverse numere de synseturi (B)**



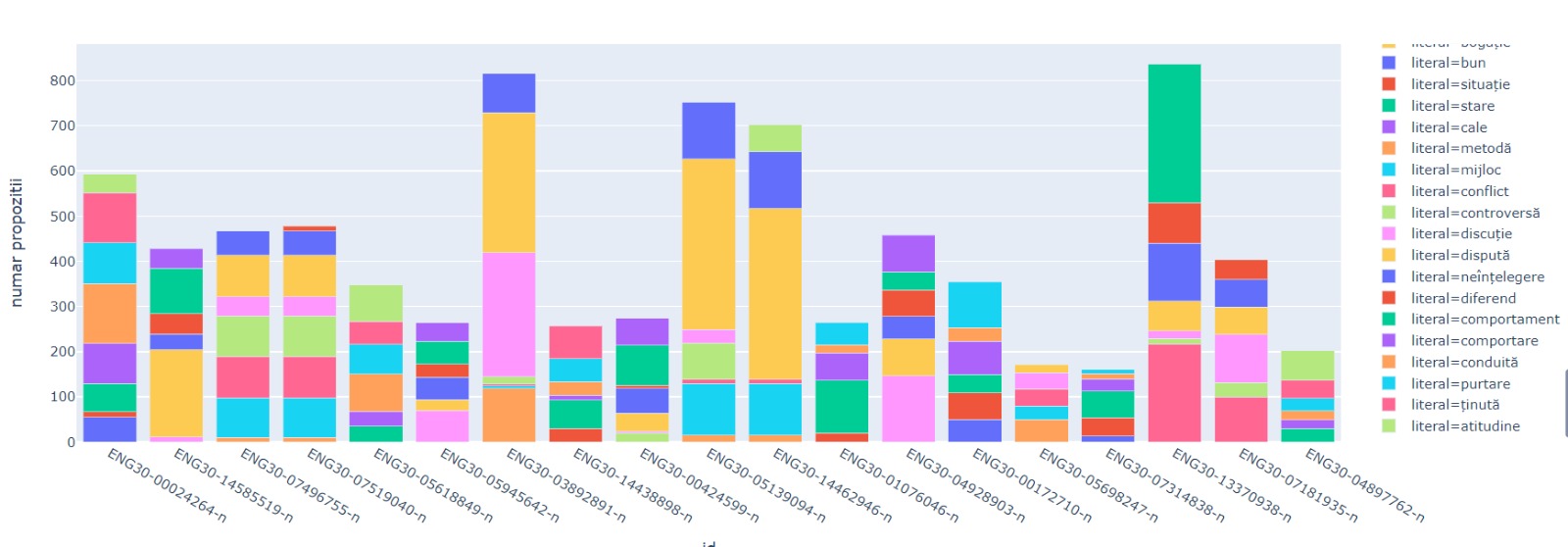
* **Grafic interactiv pentru afișarea distribuției de probabilități și evidențierea skewness-ului (C)**



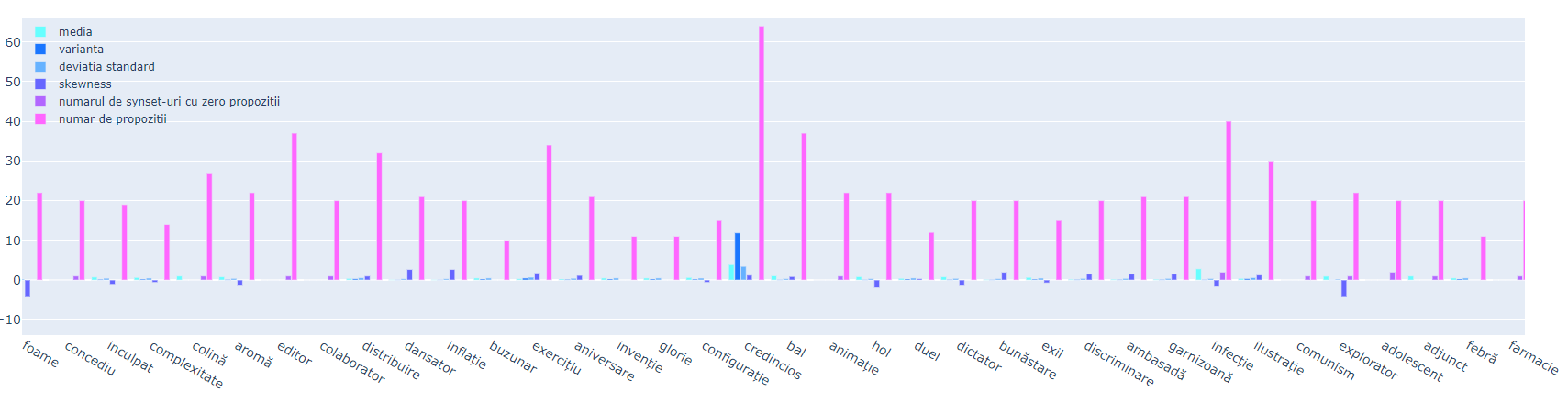
* **Grafic interactiv de tip rețea pentru afișarea legăturilor dintre synseturi și literali (D)**



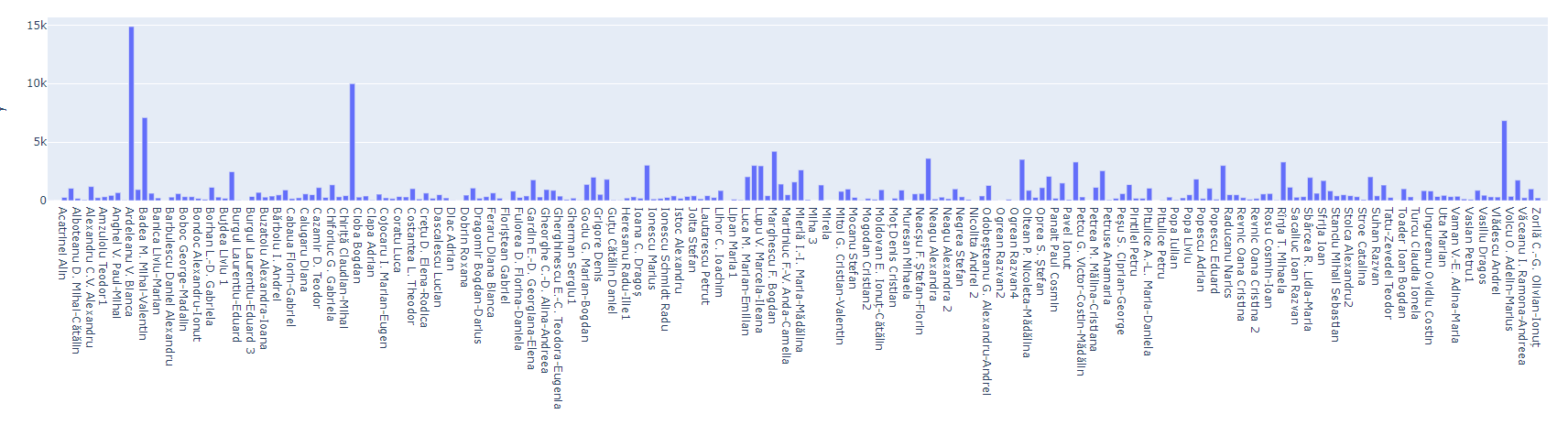
* **Grafic interactiv de tip stacked bar pentru afișarea numărul de propoziții al literalilor conectați la fiecare synset (D)**



* **Grafic interactiv pentru afișarea metricilor pentru fiecare literal (D)**



* **Grafic interactiv pentru afișarea numărului de propoziții parcurse de fiecare utilizator (E)**



# Concluzii

Tema implementeaza diverse instrumente pentru a afișa în moduri cât mai diverse informații de natură statistic despre starea bazei de date și activitatea utilizatorilor care lucrează cu aceasta.