ROMÂNIA

MINISTERUL APĂRĂRII NAŢIONALE

ACADEMIA TEHNICĂ MILITARĂ „FERDINAND I”

Tema 2: Statistici status curent WSD Database și afișare grafică

Sd. Sg. Maj GHEORGHE Alina - Andreea, C114-B

Sd. Sg. Maj MARGHESCU Bogdan, C114-C

Sd. Sg. Maj OLARU Cristian - Alexandru, C114-C

Sd. Sg. Maj TURCU Ioan-Sebastian, C114-E

Table of Contents

Introducere

Tema curentă își propune realizarea statisticilor care se pot obține pe baza de date WSD "dataset.pickle", ce cuprinde lexicul limbii române, și afișarea rezultatelor într-o interfață ce oferă grafice atât statice cât și interactive și user-friendly. Utilitatea acestei teme se raportează la acoperirea bazei de date la care s-a lucrat până în prezent de către studenții din Academia Tehnică Militară "Ferdinand I" și de către alți studenți din alte unități de învățământ din România, în sensul unor interpretări ale datelor înregistrate de platforma aflată la adresa http://138.201.128.20:5000/static/wsd.html.

Componente software

Mediul de dezvoltare: Google Colab.
Limbaj de programare: Python 3.7.12.
Afișare interfață grafică: pandas dataframe, fișiere csv, Matplotlib, Plotly.

Definiții

Table 1 Definirea Termenilor

Termen	**Definiție**
Literal	Cuvânt în forma lemmatizată al cărui sens trebuie dezambiguizat.
Synset	ID-ul corespunzător unei posibile definiții dezambiguizate ale unui literal.
Synsets	Listă de synset-uri candidate pentru un literal.
User	Utilizator logat care a ales cel puțin un synset pentru un literal.
Correct synset	Synset ales de un anumit utilizator pentru o propoziție dată și un literal dat, drept variantă corectă.
Sentence	Propoziție de analizat în cadrul căreia apare literalul.
Text	Cuvânt target ne-lemmatizat, așa cum apare în propoziție (o formă derivată a unui literal).
Text Prefix	Prima parte din propoziție.
Text Postfix	Restul propoziției. (Structura unei propoziții este: Text Prefix – Text – Text Postfix).

Descrierea setului de date

Baza de date pusă la dispoziție, WSD "dataset.pickle", e reprezentată de un fișier pickle (Python) ce conține o listă de dicționare în care cheia e reprezentată de un literal, iar valoarea de o listă în următorul format:

literal: [{„user_id": „valoare_user_id", „literal": „cuvant_target", „synsets": „synset-uri candidat", „correct_synset_id": „valoare_correct_synset_id", „text_prefix": „valoare_text_prefix", „text": „cuvant_de_dezambiguizat", „text_postfix": „valoare_text_postfix", „sentence": „valoare_propozitie_intreaga"} {} …]

```
dict_keys(['armată', 'rol', 'secol', 'județ', 'oraș', 'persoană', 'comună', 'loc', 'uniune', 'parte', 'locuitor', 'biserică', 'fapt', 'muncă',
armată
Print its literals (synonym words): ['armată']
Print its definition: (figurativ) Colectivitate care acționează în vederea unui scop comun
Print its ID: ENG30-08183290-n
```

API-ul pus la dispoziție de Python pentru acest set de date este numit RoWordNet, prescurtat de la Romanian WordNet și derivat din Princeton WordNet, o bază de date cu lexicul limbii engleze, dezvoltată de Universitatea Princeton. Elementul de bază al unui WordNet este synset-ul, lista cuvintelor cu același sens (practic opțiunile pe care un utilizator le poate alege drept definiție a unui literal dat, în cadrul platformei de lucru al cărei link e atașat în descrierea proiectului).

Forma sub care se găsește un literal în cadrul API-ului și al bazei de date este cea din dicționar, forma lemmatizată.

RoWordNet este, în esență, un graf orientat (derivat din networkx) având drept noduri ID-urile synset-urilor posibile, iar ca muchii relațiile dintre acestea. Synset-urile (obiectele) sunt păstrate într-un dicționar de tipul {ID : obiect pentru acces} în O(1).

Întrucât cuvintele sunt polisemantice, căutarea unui cuvânt va produce probabil mai multe synset-uri. Un cuvânt este cunoscut ca un literal în RoWordNet și fiecare synset are unul sau mai multe literale care sunt sinonime. Acest aspect poate fi vizualizat în captura de ecran de mai jos:

Figure 1 Obținerea listei de synset-uri pentru un literal

În figura 1 se poate observa lista id-urilor synset-urilor associate literalului 'arbore' și detalierea primului synset din această listă. Synsetul cu id-ul ENG30-08102402-n are ca definiție "*Urmărire*

sistematică a filiației existente între membrii unei familii (marcante), făcută pentru a stabili originea și gradul lor de înrudire.”.

Fiecare synset are un ID unic prin care poate fi identificat. De asemenea, WordNet-ul românesc conține și cuvinte care sunt de fapt expresii precum „tren_de_marfă”, iar căutarea „tren” va găsi și acest synset.

Obiectivele Proiectului

În cadrul acestui proiect ne propunem să calculăm următoarele statistici:

A. Numărul mediu de propoziții per literal.
B. Numărul mediu de synset-uri candidat unui literal (fără synset-ul “-1” care corespunde variantei pentru niciun răspuns corect)
C. Calculul distribuției la nivel de literal pentru synset-urile posibile:
    a. Listă cu toți literalii
    b. Pentru fiecare literal se vor calcula: media, varianța, deviația standard, skewness (indice/coeficient de asimetrie), numărul de synset-uri cu zero propoziții și suma totală de propoziții per literal.
D. Lista tuturor synset-urilor atinse și numarul de propozitii pentru fiecare.
E. Lista cu numarul de propozitii realizate de fiecare utilizator.

Pentru fiecare subpunct vom afișa rezultatele în toate cele 3 formate: pandas dataframe scris într-un fișier csv, grafice statice realizate cu Matplotlib, grafice interactive realizate cu Plotly.

Cerința A

Pentru găsirea numărului mediu de propoziții per literal am pornit de la crearea unui dicționar în care:

- Cheia e reprezentată de fiecare literal (cuvânt din limba română) din baza de date anterior citită din formatul pickle.
- Valoarea e reprezentată de numărul de propoziții în care se regăsește acel literal cheie.

Mai jos se pot vedea primele intrari din dicționar:

{'armată': 41, 'rol': 51, 'secol': 23, 'județ': 20, 'oraș': 52, 'persoană': 41,…}

Numărul de propoziții în care apare fiecare literal l-am determinat prin numărarea elementelor listei de dicționare întoarse de baza de date pentru cuvântul respectiv, întrucât fiecare propoziție este unică:

```
[3]  1 wsd = pd.read_pickle(r'dataset.pickle')
     2 wn = rwn.RoWordNet()

     1 wsd['arbore']

[{'correct_synset_id': '-1',
  'literal': 'arbore',
  'sentence': 'Unul din rolurile sale memorabile a fost cel al Coanei Eleonora Arbore din piesa Micul infern de Mircea Ștefănescu.',
  'synsets': 'ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ',
  'text': 'Arbore',
  'text_postfix': ' din piesa Micul infern de Mircea Ștefănescu.',
  'text_prefix': 'Unul din rolurile sale memorabile a fost cel al Coanei Eleonora ',
  'user_id': '94'},
 {'correct_synset_id': '-1',
  'literal': 'arbore',
  'sentence': 'Piloți Top Gun sunt în partea de sus a arborelui lor, dar chiar și la acest nivel unii piloti sunt în mod constant ma
  'synsets': 'ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ',
  'text': 'arborelui',
  'text_postfix': ' lor, dar chiar și la acest nivel unii piloti sunt în mod constant mai bine - mai bine la o mai bună anticipare ș
  'text_prefix': 'Piloți Top Gun sunt în partea de sus a ',
  'user_id': '94'},
 {'correct_synset_id': 'ENG30-04111190-n',
  'literal': 'arbore',
  'sentence': 'Mecanismul acestei osii a impus înlăturarea arborelui conector dintre cele două osii principale, lăsând astfel doar o
  'synsets': 'ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ',
  'text': 'arborelui',
  'text_postfix': ' conector dintre cele două osii principale, lăsând astfel doar osia posterioară motoare.',
  'text_prefix': 'Mecanismul acestei osii a impus înlăturarea ',
  'user_id': '94'},
```

În această captură se poate observa faptul că pentru literalul “arbore” s-a obținut o listă de dicționare ce conțin informații precum:

- Propoziția în care apare literalul, contextul de dezambiguizat
- Varianta aleasă de utilizator
- ID-ul utilizatorului care a lucrat la dezambiguizarea literalului din aceasta propoziție
- Synset-urile posibile, adica id-urile sinonimelor literalului “arbore” ce reprezintă variante de răspuns
- Literalul
- Textul prefix literalului
- Textul postfix literalului

Numărul mediu al acestor propozitii per literal l-am calculat cu ajutorul functiei mean() din librăria numpy, obținând rezultatul: **34.463941380640186** .

Afișare Matplotlib

Acest rezultat poate fi vizualizat în graficul de mai jos realizat cu Matplotlib:

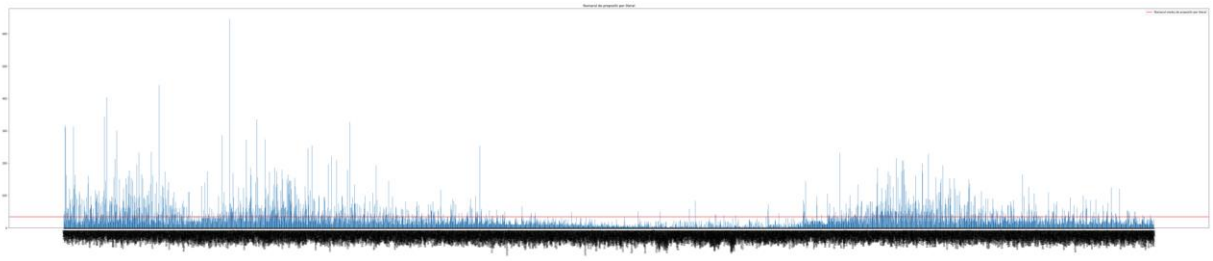


Figure 2 Numărul de propoziții per literal [Matplotlib]

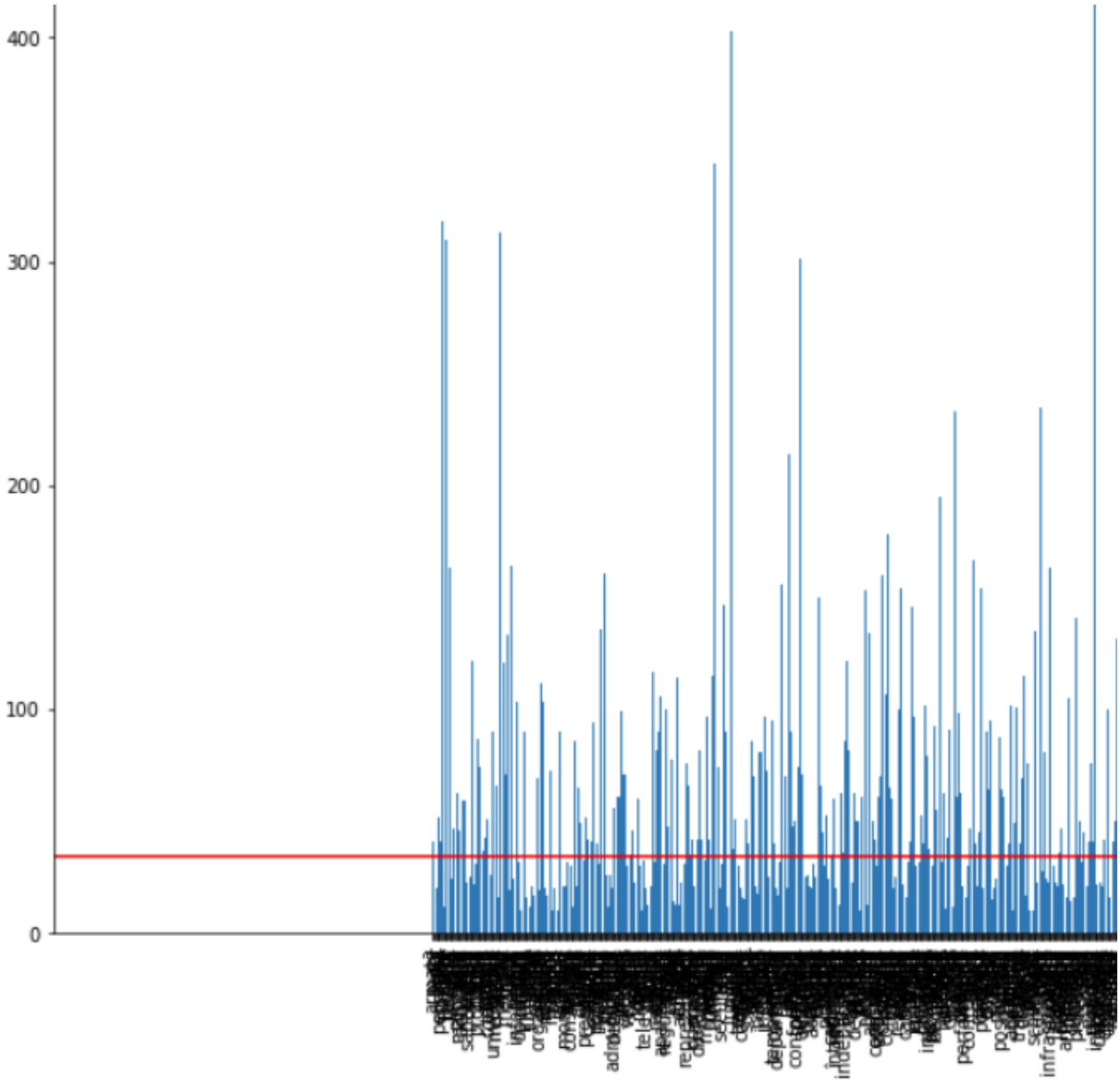


Figure 3 Zoom In – numărul de propozitii per literal

Interpretare: în medie, fiecărui literal i se asociază 34 de propoziții în cadrul cărora trebuie dezambiguizat sensul lui.

Afișare Plotly Interactiv

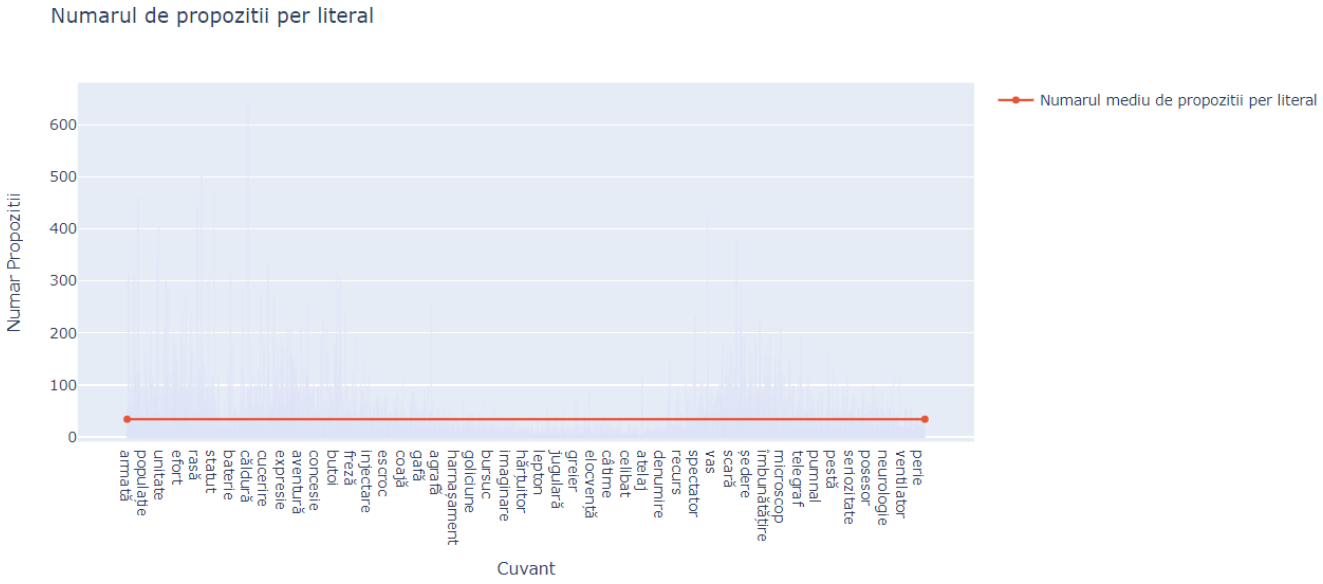


Figure 4 Numărul de propoziții per literal

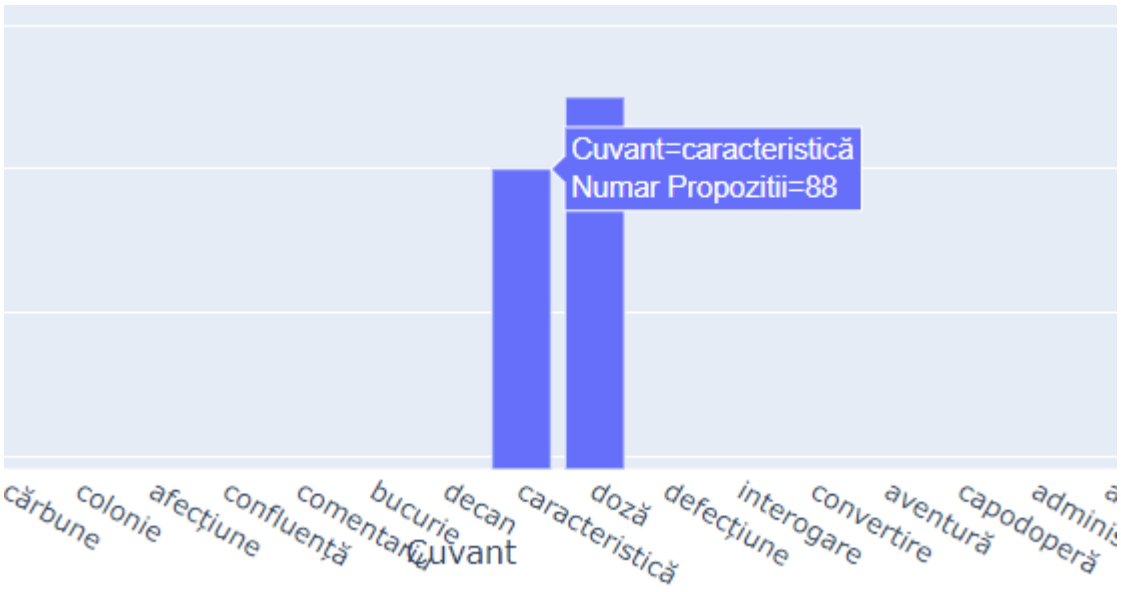


Figure 5 Zoom In - numărul de propoziții per literal Plotly

Interpretare: pe axa OX se regăsesc cuvintele, iar pe axa OY este numărul de propoziții asociat. Linia roșie vizibilă în primul grafic reprezintă media valorilor, adică **34.463941380640186** .

Cerința B

Pentru găsirea numărului mediu de synset-uri candidat unui literal am procedat în mod analog primei cerințe, plecând de la crearea unui dicționar care reține literalii și numărul de synset-uri asociate fiecăruia.

Dicționarul arată astfel:

{'armată': 4, 'rol': 5, 'secol': 2, 'județ': 2, 'oraș': 5, 'persoană': 3, 'comună': 2,…}

Numărul de synset-uri asociate unui literal a fost determinat prin numărarea elementelor listei ‘synsets’ din cadrul oricărui dicționar asociat unui literal. Se poate observa raționamentul în cadrul rezultatului de mai jos:

```
1 print(len(wsd['rol'])) #nr de propozitii
2 print(wsd['rol'][0]['synsets'])
3 print(wsd['rol'][1]['synsets'])
4 print(wsd['rol'][50]['synsets'])
```

```
51
ENG30-09587565-n ENG30-00720565-n ENG30-06331803-n ENG30-05929008-n ENG30-00722061-n -1
ENG30-09587565-n ENG30-00720565-n ENG30-06331803-n ENG30-05929008-n ENG30-00722061-n -1
ENG30-09587565-n ENG30-00720565-n ENG30-06331803-n ENG30-05929008-n ENG30-00722061-n -1
```

De exemplu, cuvântul “rol” apare în 51 de propoziții, iar pentru fiecare propoziție există 5 sinonime cu care poate fi înlocuit și o varintă “-1” asociată unui răspuns general valabil prin care utilizatorul selectează “Nicio opțiune validă”.

Numărul mediu de synset-uri candidat per literal este: **3.744119**.

Afișare Matplotlib

Valoarea medie, alături de cuvintele din baza de date și numărul de sinonime asociate fiecăruia se poate vizualiza în figura de mai jos.

Linia roșie evidențiază valoarea medie: **3.744119**.

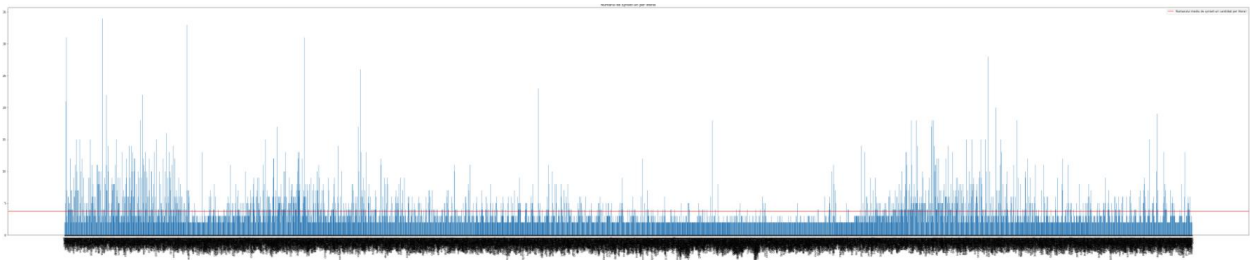


Figure 6 Numărul de synset-uri per literal - Matplotlib

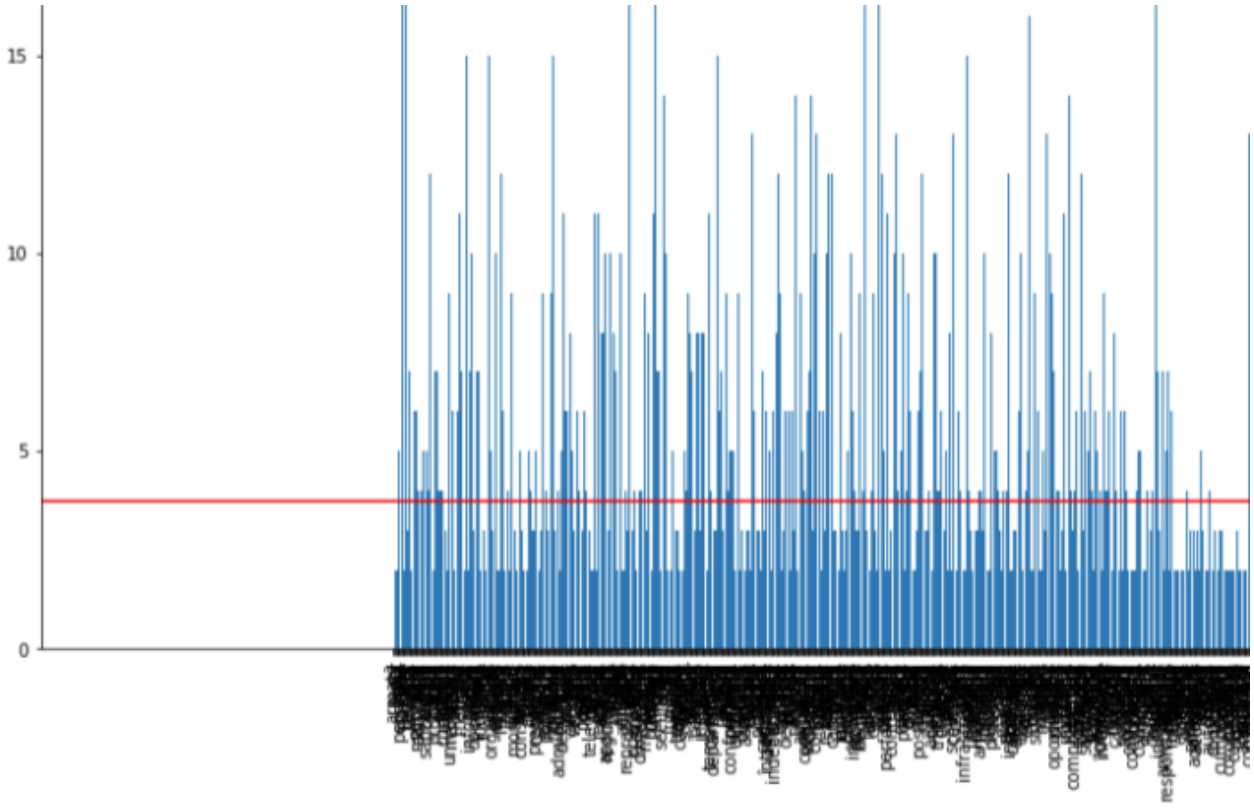


Figure 7 Zoom In

Interpretare: fiecare cuvânt din dicționar are asociate în medie 4 sinonime prin care poate fi înlocuit în diferite contexte.

Afișare Plotly interactiv

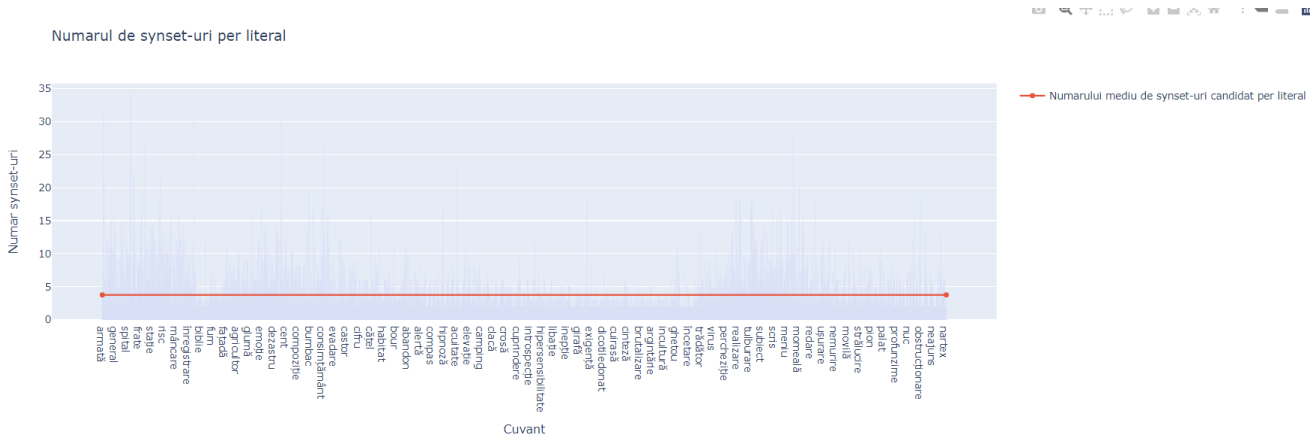


Figure 8 Numărul de synset-uri per literal Plotly

Valoarea medie dintre numerele de synset-uri per literal este afișată prin dreapta roșie orizontală.

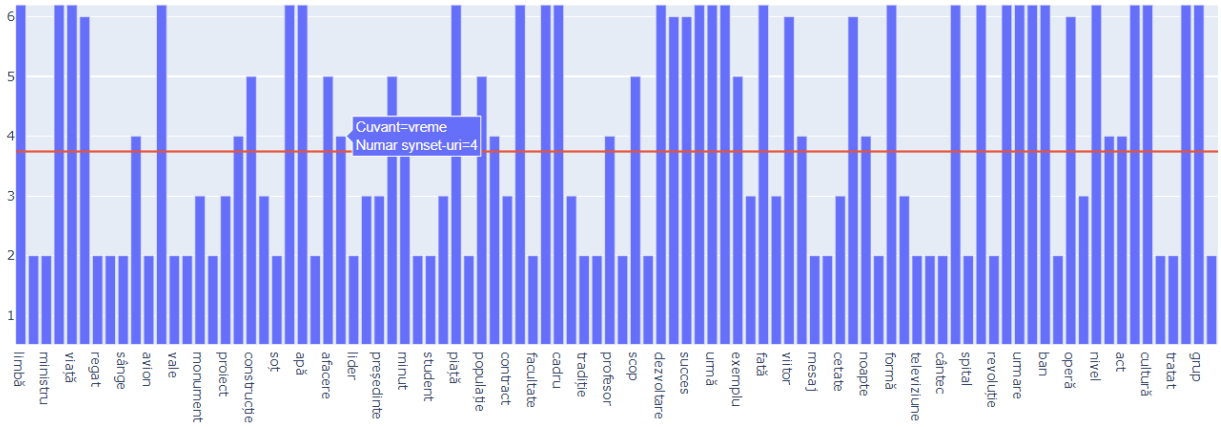


Figure 9 Zoom In

Printr-un zoom in al graficului realizat cu librăria Plotly se pot observa detaliile despre fiecare cuvânt în mod interactiv.

Cerința C

Calculul distribuției la nivel de literal pentru synset-urile posibile: listă cu toți literalii, media, varianța, deviația standard, skewness (indice/coeficient de asimetrie), numărul de synset-uri cu zero propoziții și suma totală de propoziții per literal.

Pentru obținerea acestor statistici, am pornit de la crearea unei liste ce reprezintă o distribuție a synset-urilor alese de utilizatori, în raport cu un literal dat. Mai exact, această listă o să conțină pe poziția synset-ului cu id-ul x valoarea dată de numărul de alegeri ale opțiunii x.

Vom exemplifica în continuare rezultatele obținute pentru cuvântul “rol”:

```
['ENG30-09587565-n', 'ENG30-00720565-n', 'ENG30-06331803-n', 'ENG30-05929008-n', 'ENG30-00722061-n']
[ 3. 11.  4.  6. 26.]
```

De exemplu, synset-ul cu id-ul *ENG-39-09587565-n* a fost ales de 3 ori .

Plecând de la această distribuție, am calculat :

- **Valoarea medie a numărului de synseturi alese per literal** : ***'Mean': 10.0***.
  - Modalitatea de calcul : funcția mean() din librăria numpy.
  - Formulă: media aritmetica a valorilor din lista de distribuție de mai sus.
- **Variația valorilor distribuției de synset-uri per literal** :***'Variance': 71.6*** .
  - Modalitate de calcul : funcția var() din librăria numpy.
  - Formulă :

    $$var = mean(abs(x - x.mean())^2)$$

  - Unde x reprezintă vectorul distribuției synset-urilor alese de utilizatori per literal.
- **Deviația standard**: ***'Standard Deviation': 8.461678320522472***.
  - Modalitate de calcul: funcția std() din librăria numpy.

- Formulă:

$$Standard\ Deviation\ = \sqrt{\text{mean}\ (\text{abs}(\text{x}\ -\ \text{x.mean}())^2)}$$

- **Valoarea skewness**, adică indicele/coeficient de asimetrie: ***'Skewness': 1.146806844570046.***
  - Modalitate de calcul: funcția skew() din librăria scipy.stats .
  - Formulă:

$$\textbf{Skewness} = \frac{\textbf{3}\ (\boldsymbol{Mean - Median})}{\boldsymbol{Standard\ Deviation}}$$

  - Interpretare:
    - **skewness = 0 :** normally distributed.
    - **skewness > 0 :** more weight in the left tail of the distribution.
    - **skewness < 0 :** more weight in the right tail of the distribution.
- **Numărul de synset-uri cu zero propozitii asociate: *'Number of synsets with 0 sentences': 0***
  - Modalitate de calcul: funcția len() și funcția count_nonzero() din librăria numpy.
  - Formula : diferența dintre numărul total de synset-uri asociate unui literal și numărul de zerouri găsite în lista cu distribuțiile synset-urilor considerate corecte de utilizatori.

Aceste statistici sunt returnate într-un dicționar, care se poate vizualiza mai jos:

```
{'Literal': 'rol',
 'Mean': 10.0,
 'Number of sentences': 51,
 'Number of synsets with 0 sentences': 0,
 'Skewness': 1.146806844570046,
 'Standard Deviation': 8.461678320522472,
 'Variance': 71.6}
```

Figure 10 Dicționar cu statistici despre literalul "rol"

Pentru vizualizarea statisticilor de mai sus la nivelul întregului corpus, am creat un dicționar ce contine:

- Cheia : literalul din lexic
- Valoarea : dicționarul de mai sus cu statisticile despre un literal.

Dataframe-ul creat din acest dicționar arată astfel:

```
1 df_wsd_literal_data
```

	Literal	Mean	Variance	Standard Deviation	Skewness	Number of synsets with 0 sentences	Number of sentences
0	armată	10.000000	31.500000	5.612486	0.212112	0	41
1	rol	10.000000	71.600000	8.461678	1.146807	0	51
2	secol	10.000000	49.000000	7.000000	0.000000	0	23
3	județ	10.000000	100.000000	10.000000	0.000000	1	20
4	oraș	10.000000	226.800000	15.059880	1.471024	0	52
...	...	...	...	...	...	...	...
5181	persuasiune	2.333333	2.888889	1.699673	-0.528005	1	10
5182	repeziciune	6.000000	1.000000	1.000000	0.000000	0	13
5183	păsărică	10.000000	81.000000	9.000000	0.000000	0	28
5184	țâță	5.000000	9.000000	3.000000	0.000000	0	12
5185	pisoi	4.500000	6.250000	2.500000	0.000000	0	12

5186 rows × 7 columns

Figure 11 Dataframe cu statisticile despre literali

Afișarea cu Matplotlib

Pentru fiecare coloană din dataframe-ul de mai sus am realizat câte un grafic în care pe axa OX e reprezentat indexul literalului, iar pe OY valorile statisticilor din dataframe.
Astfel, pentru afișarea valorii medie a synset-urilor corecte pentru fiecare literal, am obținut graficul de mai jos:

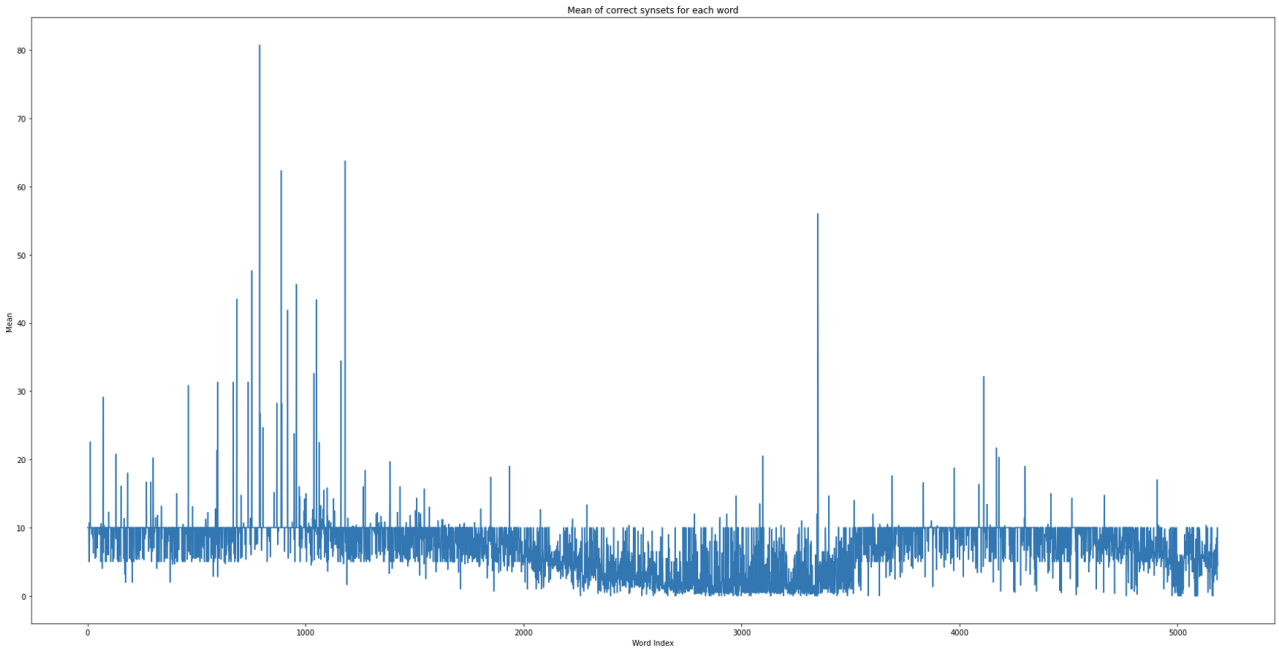


Figure 12 Mean of the correct synsets for each word

Pentru afișarea variației synset-urilor corecte pentru fiecare literal am obținut graficul:

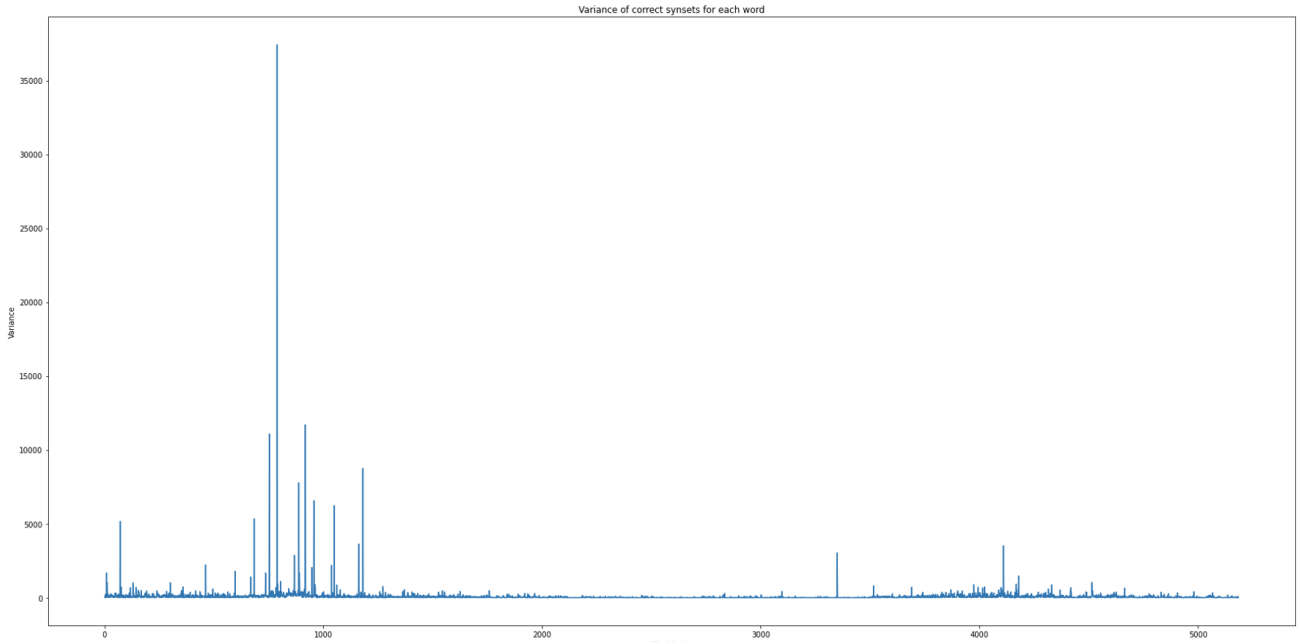


Figure 13 Variance of correct synsets for each word

Variația standard a synset-urilor alese drept corecte de către utilizatori:

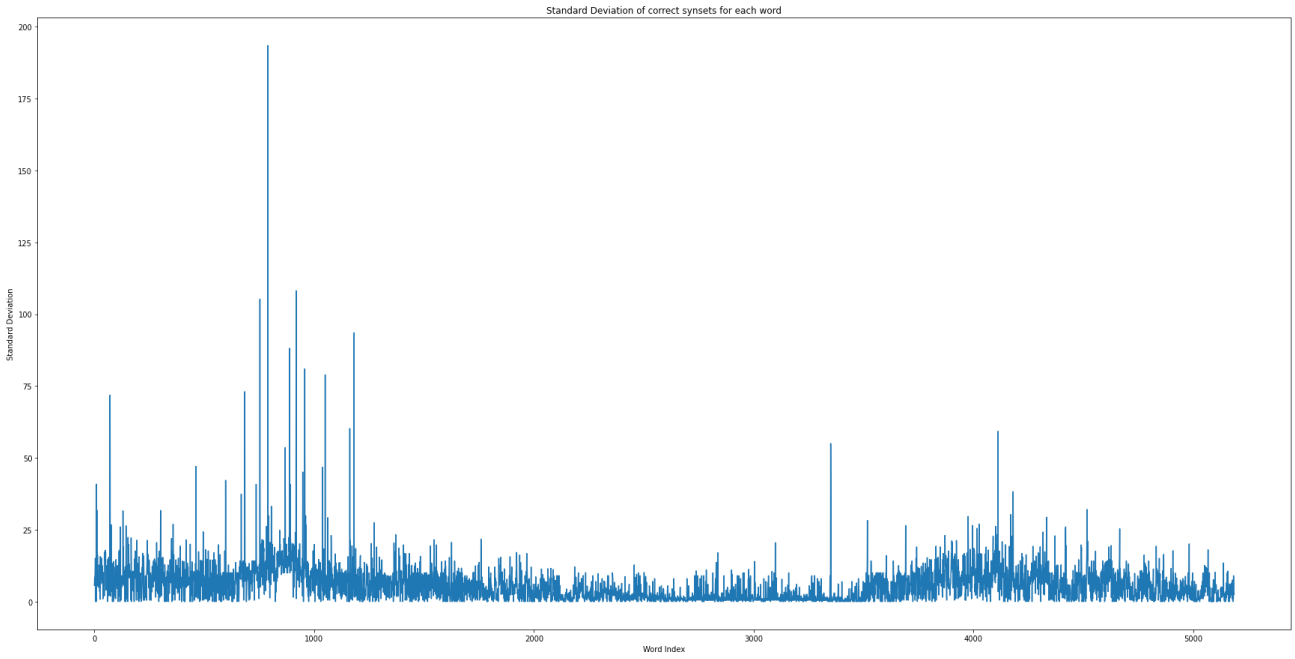


Figure 14 Standard Deviation of correct synsets for each word

Valoarea skewness pentru synset-urile alese drept corecte de către utilizatori:

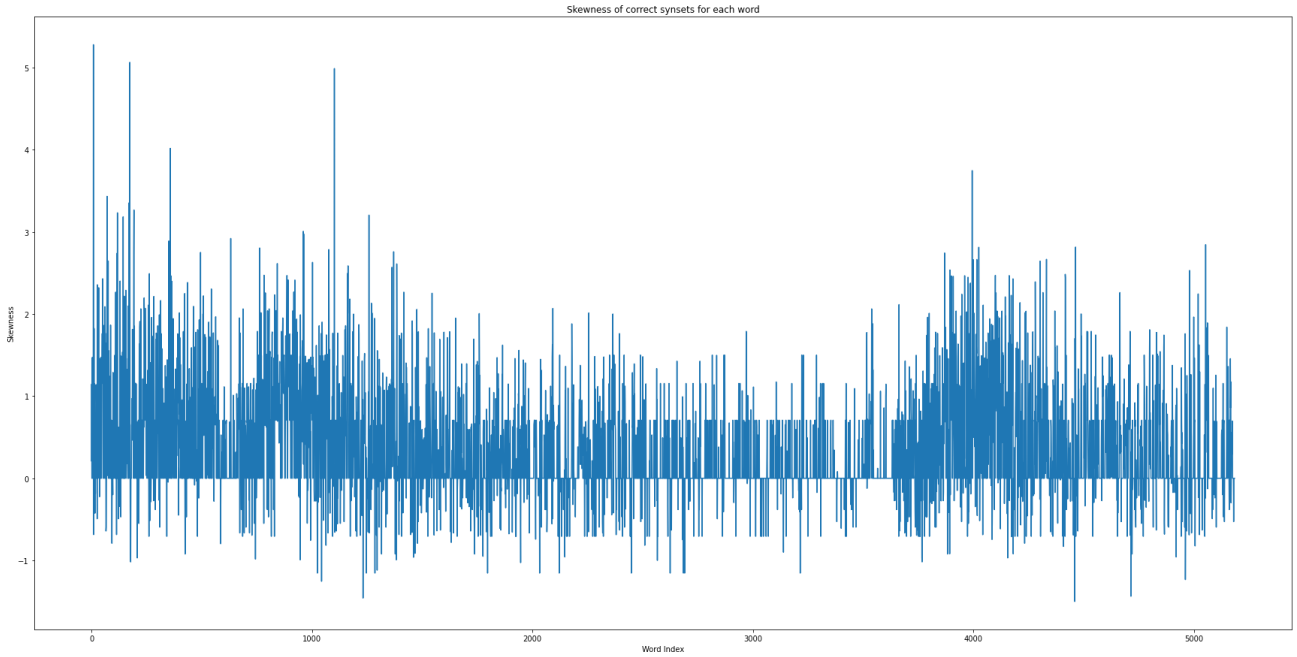


Figure 15 Skewness of correct synsets for each word

Numărul de synsets cu zero propoziții asociate:

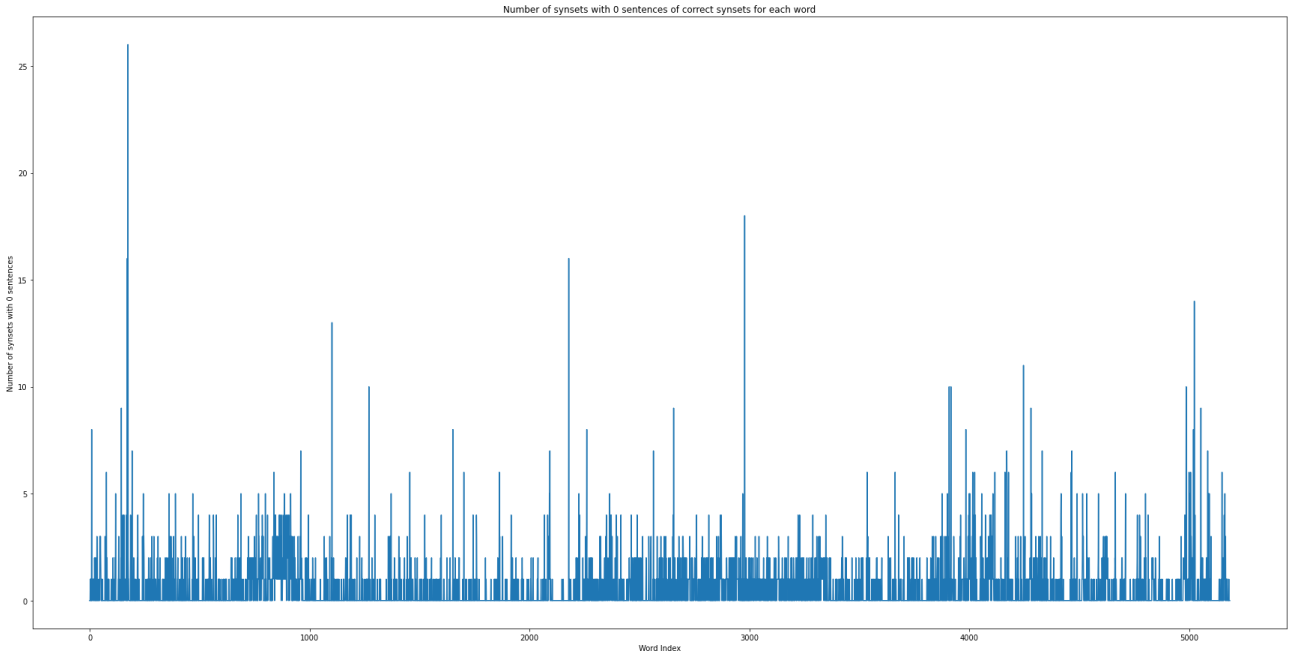


Figure 16 Number of synsets with 0 sentences of correct synsets for each word

Numărul de propoziții asociate synset-urilor alese drept corecte de către utilizatori:

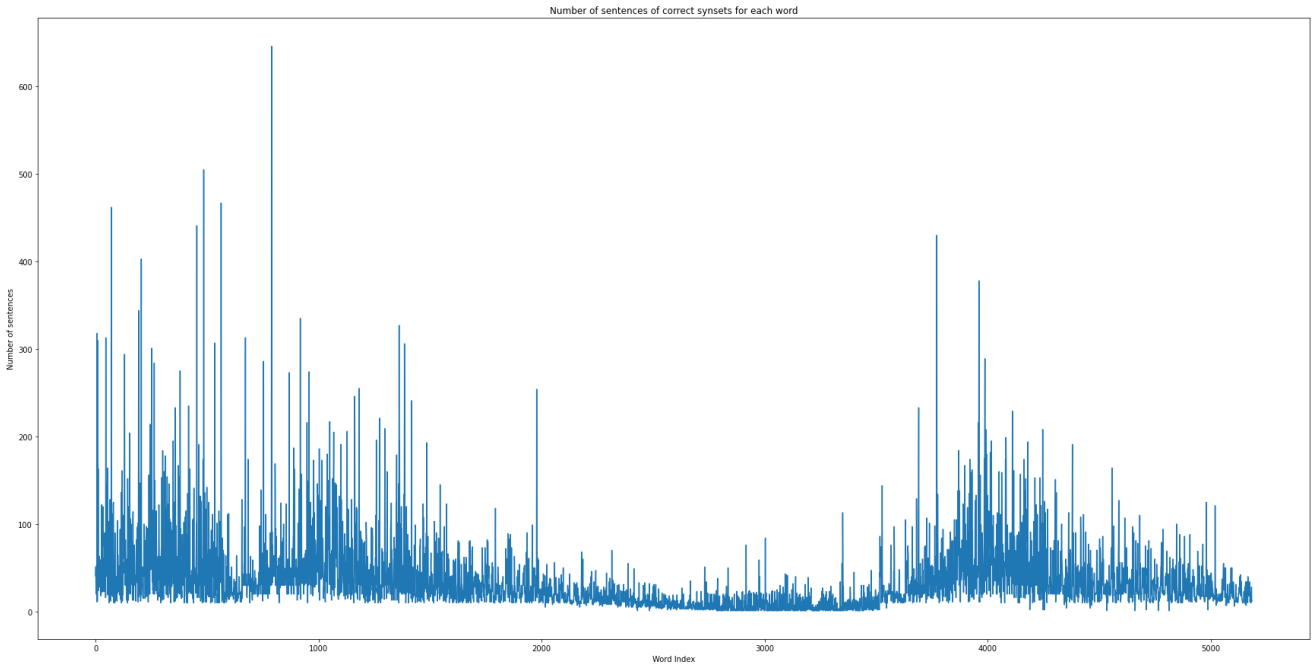


Figure 17 Number of sentences of correct synsets for each word

Afișarea cu Plotly

Pentru exemplificarea vizualizării rezultatelor de mai sus în mod interactiv, cu Plotly, am captat un zoom-in al graficelor obținute și valoarea asociată unui cuvânt ales la întâmplare din fiecare grafic. Pe axa OX sunt reprezentate cuvintele, literalii, iar pe OY valoarea asociată literalului conform coloanelor din dataframe-ul de statistici obținut anterior:

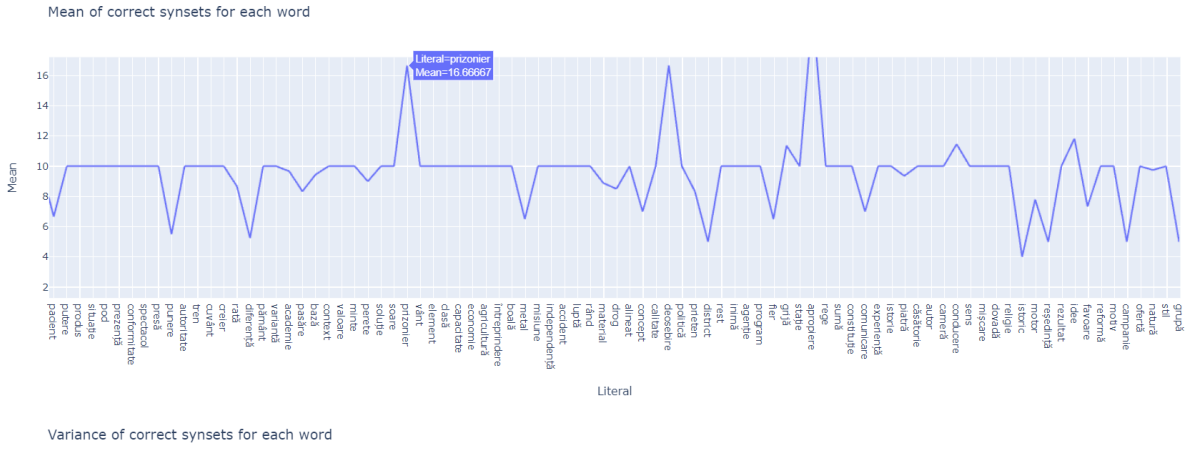


Figure 18 Mean of correct synsets for each word

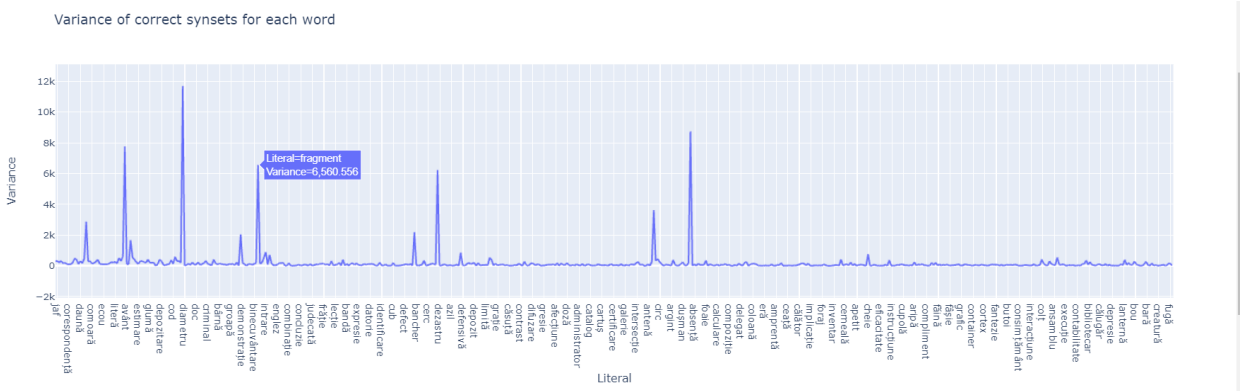


Figure 19 Variance of correct synsets for each word

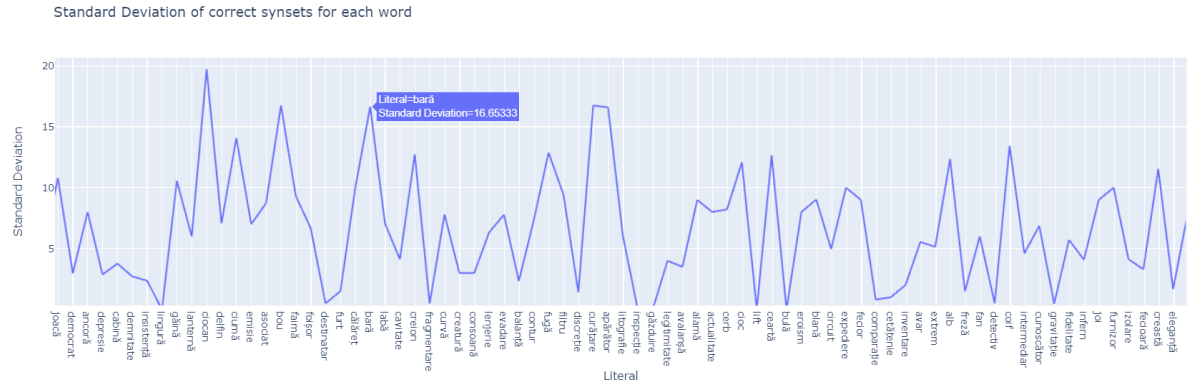


Figure 20 Standard Deviation of correct synsets for each word

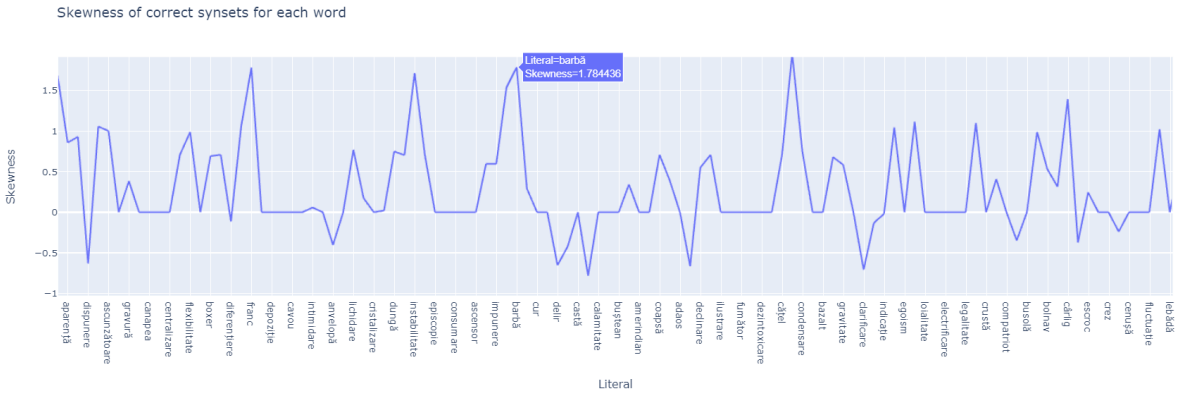


Figure 21 Skewness of correct synsets for each word

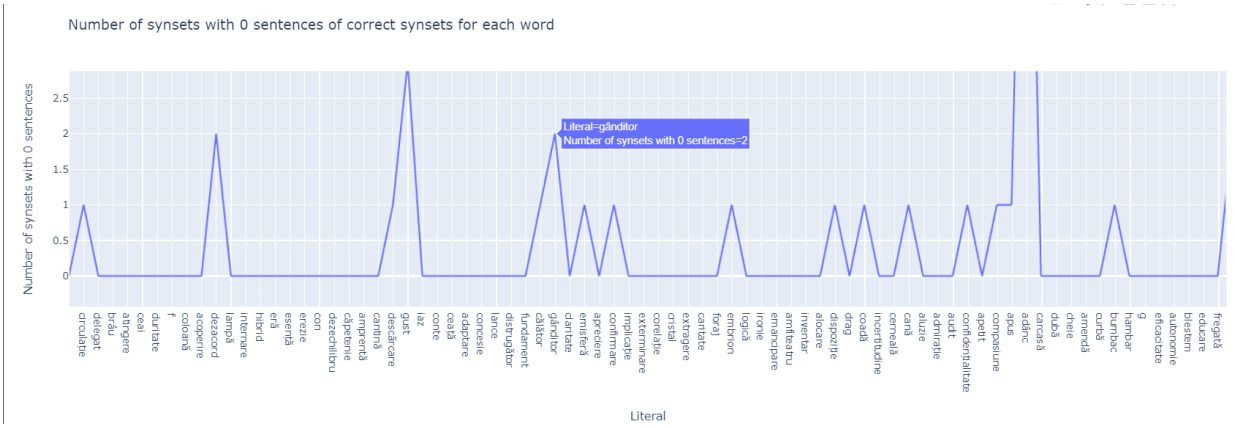


Figure 22 Number of synsets with 0 sentences of correct synsets for each word

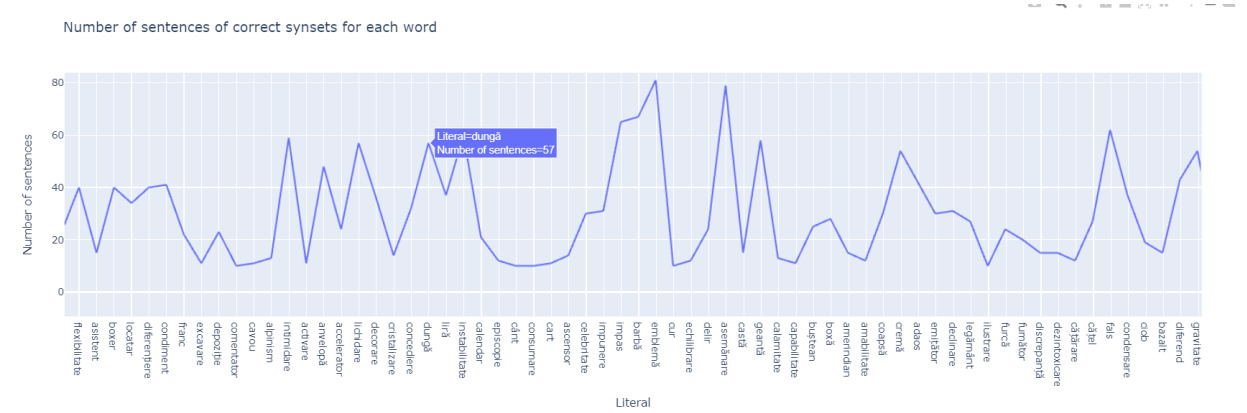


Figure 23 Number of sentences of correct synsets for each word

Cerința D

Pentru determinarea datelor despre fiecare synset am creat un dicționar care cuprinde următoarele date pentru un synset identificat după ID:

- ID-ul asociat synset-ului ales
- Listă cu literalii a căror sinonim este synset-ul, practic sunt literalii care au ca variantă de dezambiguizare synset-ul respectiv
- Numărul de literali găsiți mai sus
- Numărul total de propoziții în cadrul cărora synset-ul apare drept variantă de răspuns

Pentru aflarea numărului de literali care au synset-ul cu ID-ul dat, am căutat prin listele de sinonime asociate fiecarui cuvant din baza de date.

Dicționarul obținut are structura din captura de mai jos:

```
1 print(get_synset_data('ENG30-00017222-n'))
{'ID Synset': 'ENG30-00017222-n', 'Literal List': ['plantă'], 'Literal Count': 1, 'Number of sentences': 29}
```

Figure 24 Synset Data

Pe baza acestei structuri de date create, am construit un dicționar cu toate id-urile synset-urilor din RoWordNet și structurile lor asociate. Acest dicționar de dicționare l-am transformat într-un dataframe în care coloanele sunt reprezentate de cheile dicționarului de mai sus.

	ID Synset	Literal List	Literal Count	Number of sentences
0	ENG30-00006269-n	[viață]	1	125
1	ENG30-00006484-n	[celulă]	1	26
2	ENG30-00017222-n	[plantă]	1	29
3	ENG30-00023100-n	[]	0	0
4	ENG30-00027167-n	[loc]	1	318
...	...	...	...	...
59343	ENG30-05622076-n	[]	0	0
59344	ENG30-01353670-v	[]	0	0
59345	ENG30-01457710-v	[]	0	0
59346	ENG30-07269552-n	[]	0	0
59347	ENG30-01244178-v	[]	0	0

59348 rows × 4 columns

Figure 25 Dataframe cu datele despre fiecare synset din RoWordNet

Afișare Matplotlib

Pentru prezentarea acestor statistici, am realizat două grafice în care pe axa OX se regăsește indexul fiecărui synset din RoWordNet, iar pe OY se regasesc valorile *Literal Count* (Numărul de cuvinte care au synset-ul respectiv drept sinonim), respectiv *Number of Sentences* (Numărul de propoziții în care apare un synset drept variantă de dezambiguizare a literalului).

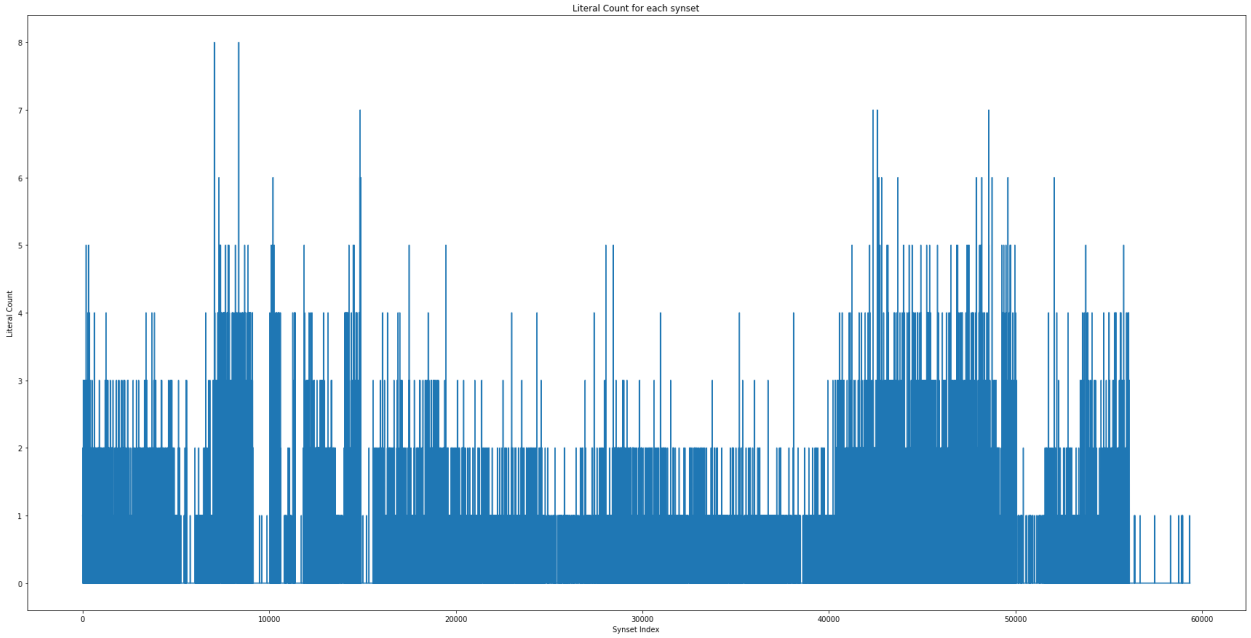


Figure 26 Numărul de literali pentru care reprezintă un sinonim fiecare synset

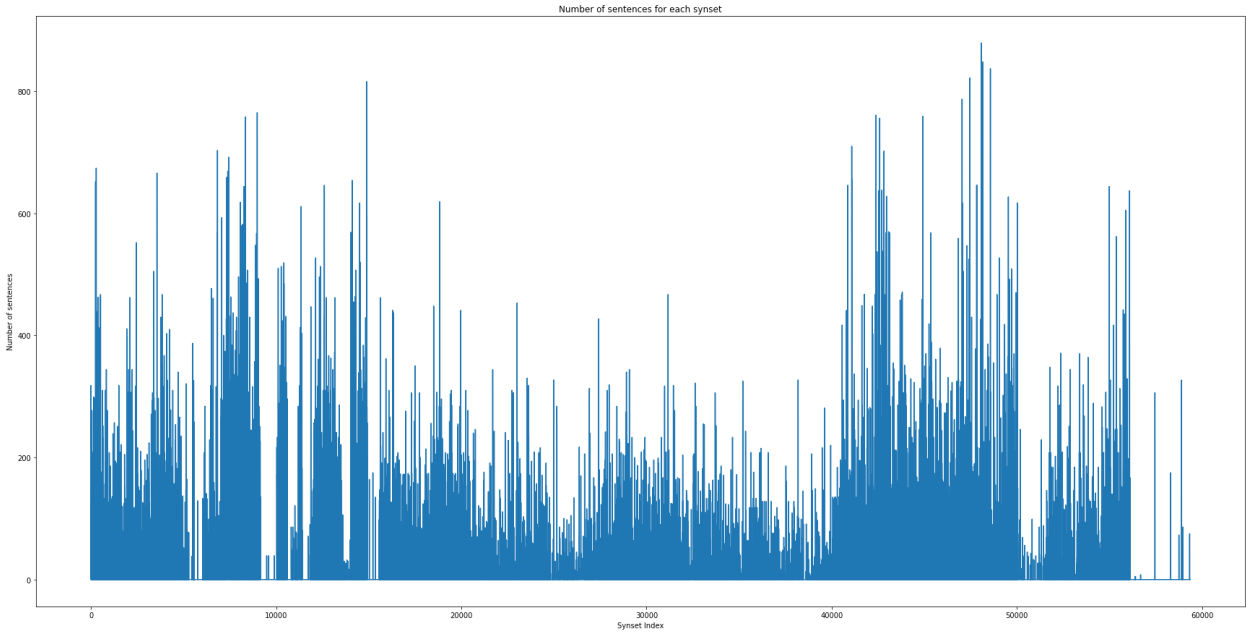


Figure 27 Numărul de propoziții în care apare fiecare synset

Afișare Plotly

Cele două grafice explicate mai sus pot fi vizualizate și în mod interactiv, cu Plotly, în varianta zoom-in într-o zonă aleasă aleator:

Figure 28 Literal Count for Each Synset

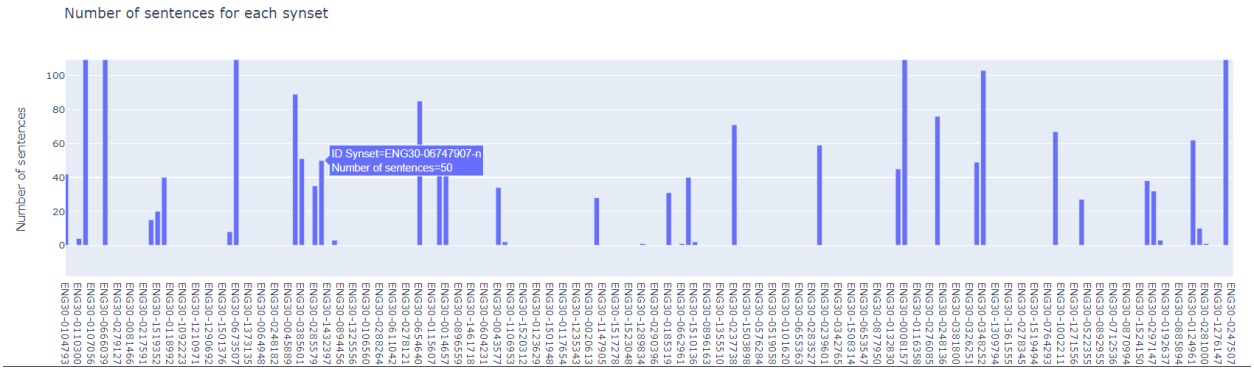


Figure 29 Number of Sentences for Each Synset

Cerința E

Pentru determinarea listei cu numărul de propoziții realizate de un utilizator am iterat prin toate propozițiile asociate unui literal din baza de date și am verificat daca în câmpul “user_id” se identifică username-ul utilizatorului respectiv.

Totodată, am creat o listă sortată alfabetic, cu intrări unice, ce conține id-urile tuturor utilizatorilor care apar în baza de date.

```
1 print(user_list)
```

```
[' ' '0' '1' '10' '11' '12' '13' '14' '15' '16' '17' '18' '2' '20' '21'
 '22' '23' '24' '25' '26' '27' '28' '29' '3' '30' '31' '32' '33' '34' '35'
 '36' '37' '38' '39' '4' '40' '41' '42' '43' '44' '45' '46' '47' '48' '49'
 '5' '50' '51' '52' '53' '54' '55' '56' '57' '58' '59' '6' '60' '61' '62'
 '63' '64' '65' '66' '67' '68' '69' '7' '70' '71' '72' '73' '74' '75' '76'
 '77' '78' '79' '8' '80' '81' '82' '83' '84' '85' '86' '87' '88' '89' '9'
 '90' '91' '92' '93' '94' 'Acatrinei Alin' 'Alexandra Victoria'
 'Amzuloiu Teodor' 'Amzuloiu Teodor1' 'Andronache George'
 'Ardeleanu Bianca' 'Balascan Gabriel' 'Banica Liviu-Marian'
 'Banica Liviu-Marian1' 'Barbulescu Daniel Alexandru' 'Biltan Cosminn'
 'Boboc George-Madalin' 'Boiangiu Alexandra' 'Bondoc Alexandru-Ionut'
 'Bondoc Alexandru-Ionut_1' 'Bujdea Liviu' 'Bujdea Liviu 1'
 'Burgui Laurentiu-Eduard' 'Burgui Laurentiu-Eduard 2'
 'Burgui Laurentiu-Eduard 3' 'Buzatoiu Alexandra-Ioana'
 'Cabaua Florin-Gabriel' 'Cabaua Florin-Gabriel1' 'Calugaru Diana'
 'Casandroiu Paul-Florinel' 'Ceaus Alexandru' 'Chinda Andrei'
 'Cioba Bogdan' 'Clapa Adrian' 'Clapa Adrian ' 'Clinci Daniel'
 'Cojocaru Marian' 'Coratu Luca' 'Costan Raluca' 'Craciun Ionut'
 'Cucuta Radu' 'Dascalescu Lucian' 'Delibas Razvan' 'Diac Adrian'
 'Dobrin Roxana' 'Dragomir Bogdan-Darius' 'Farcas Elena'
 'Floriștean Gabriel' 'Galea Alexandru' 'Georgescu David' 'Gherman Sergiu'
 'Gherman Sergiu1' 'Gherman Sergiu2' 'Grigore Denis' 'Heresanu Radu-Ilie'
```

Figure 30 Lista cu utilizatori

În fina, am creat un dicționar cu următoarea structura:

- Cheia : reprezentată de username-ul fiecărui utilizator
- Valoare : reprezentată de numărul de propoziții asociate fiecărui utilizator

Acest dicționar transpus într-un pandas dataframe, arată astfel sortat:

	Utilizator	Numar propozitii realizate
167	Muresan Mihaela	1
1	0	2
161	Mirela	2
162	Mirela lazar	2
210	Toader Bogdan	3
...	...	...
33	39	4215
73	75	6836
12	2	7100
34	4	9999
2	1	14876

Figure 31 Dataframe cu utilizatorii și numărul de propoziții asociate

Din acest tabel putem observa faptul că utilizatorul Muresan Mihaela a completat cele mai putine propozitii (o propozitie), iar utilizatorul cu id-ul 2 a completat cele mai multe propoziții (14876).

Numărul mediu de propoziții realizate de fiecare utilizatorul l-am determinat prin aplicarea funcției mean() din librăria numpy pe dicționarul cu utilizatorii și numărul lor de propoziții completate. Astfel, am obținut o medie de **804.8243243243244** propoziții per utilizator.

Afișarea Matplotlib

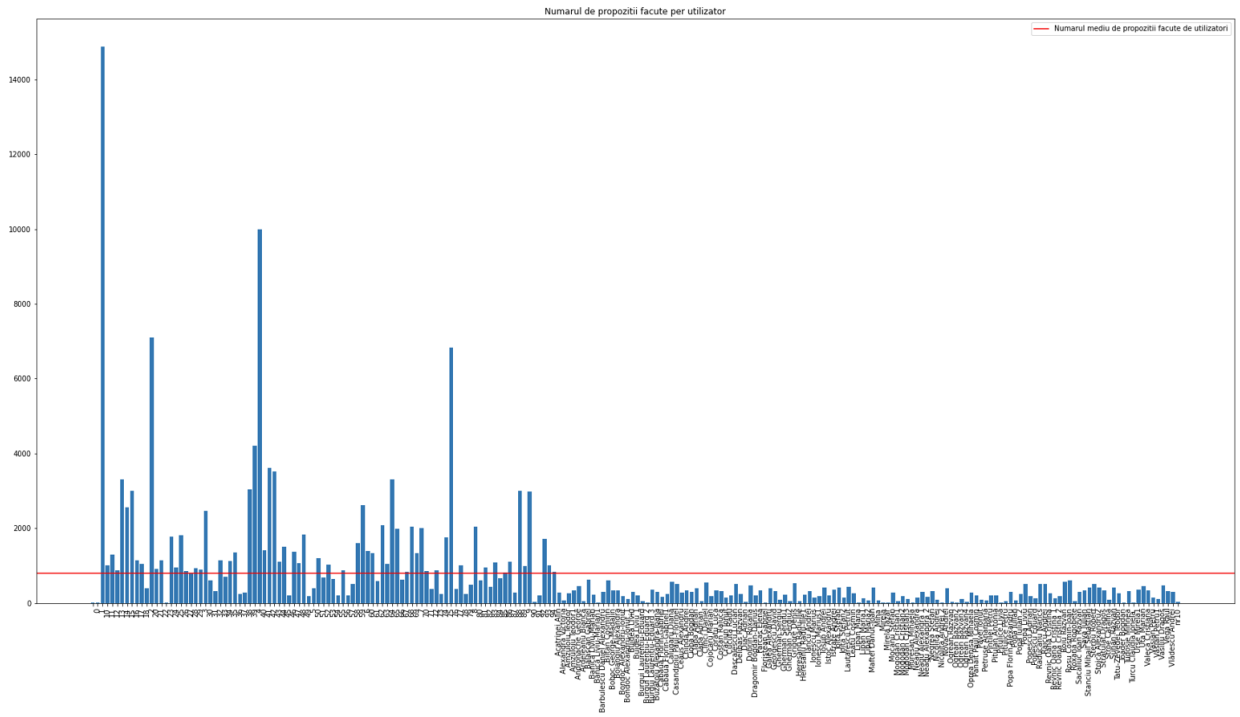
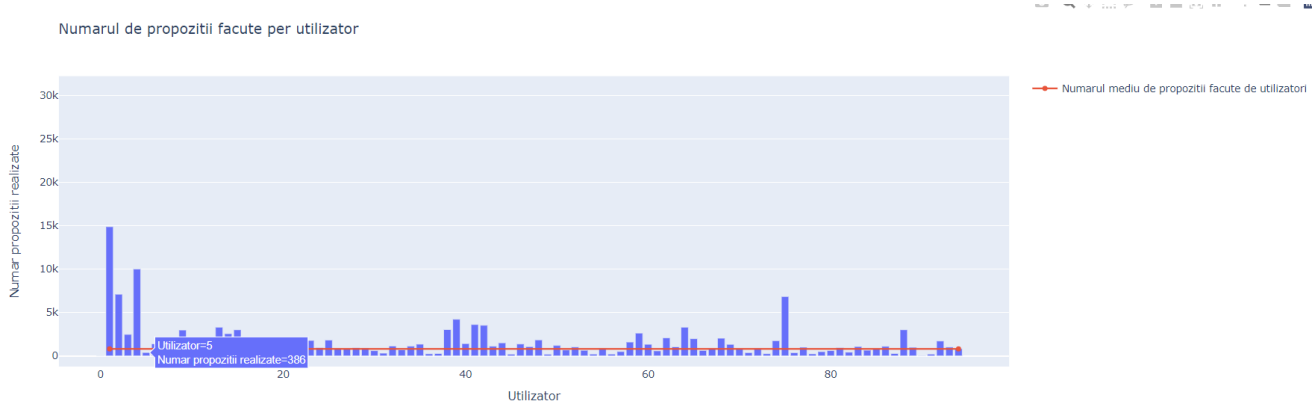


Figure 32 Numărul de propoziții făcute per utilizator

În cadrul acestui grafic am reprezentat pe OX id-urile utilizatorilor (fie numerice fie literale, în funcție de cum s-au înregistrat), iar pe OY numarul de propozitii corespunzător fiecăruia. Se poate observa media numărului de propozitii realizate prin dreapta orizontală roșie poziționată la valoarea aproximativă 805 pe OY.

Afișarea Plotly

Același grafic l-am reprezentat și interactiv, obținând următorul rezultat pentru un utilizator ales random:



De exemplu, utilizatorul cu id-ul 5 a completat 386 de propoziții.

Bibliography

Dan Tufiş, Verginica Barbu Mititelu, The Lexical Ontology for Romanian, in Nuria Gala, Reinhard Rapp, Nuria Bel-Enguix (Ed.), Language Production, Cognition, and the Lexicon, series Text, Speech and Language Technology, vol. 48, Springer, 2014, p. 491-504.

https://wiki.mta.ro/c/4/ia/hw/2021/t2

https://wiki.mta.ro/c/4/ia/hw/2021/t2/wsd

https://wordnet.princeton.edu

https://www.geeksforgeeks.org/numpy-var-in-python