

TEMA 2

Statistici status curent WSD Database

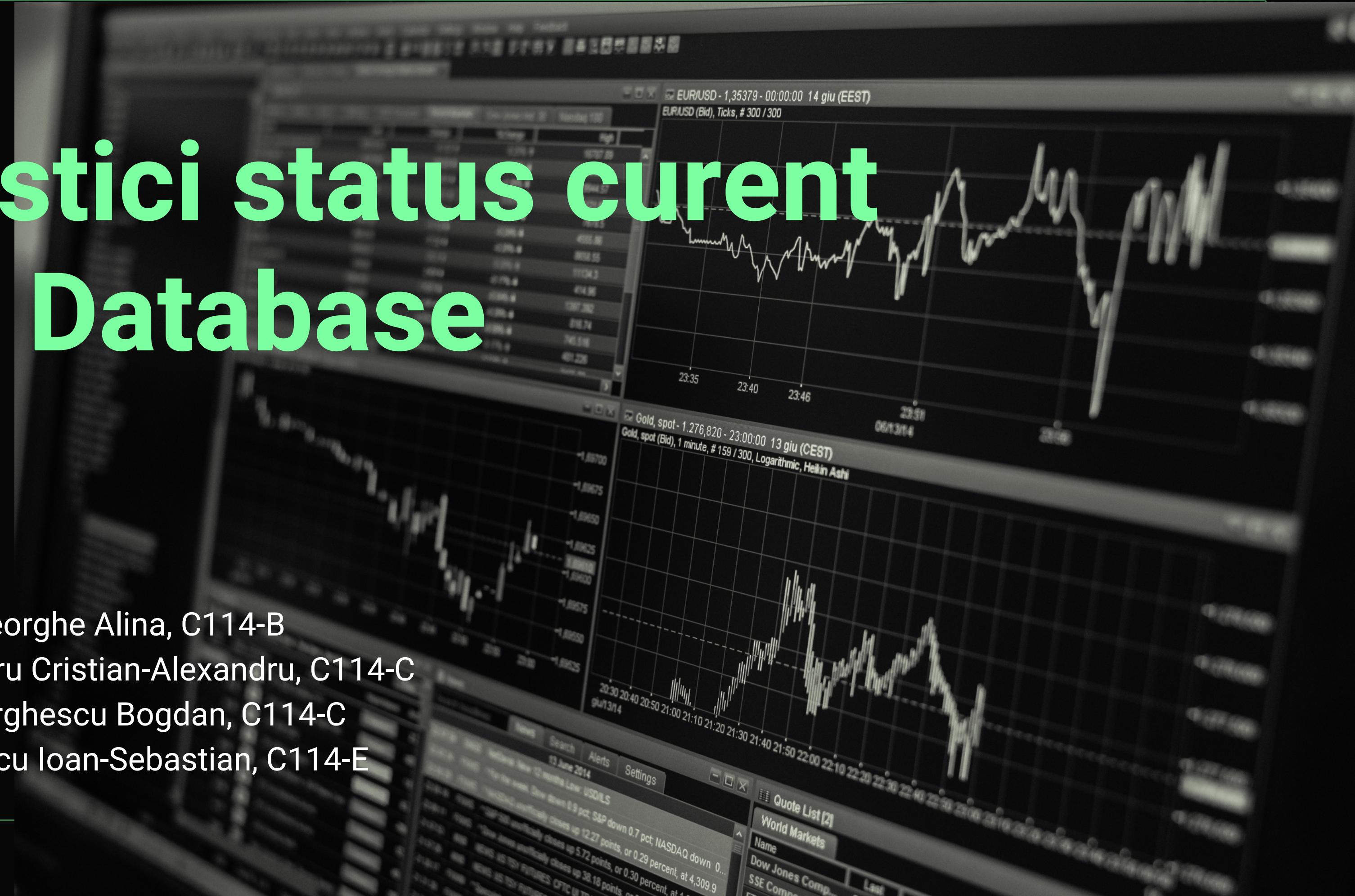
ECHIPA:

Sd. Sg. Maj. Gheorghe Alina, C114-B

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

Sd. Sg. Maj. Marghescu Bogdan, C114-C

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



Agenda

Key topics covered
in this presentation

- Introducere
- Definirea setului de date
- Obiective
- Rezultate
- Interpretări rezultate
- Întrebări

Scopul Proiectului

Scopul nostru este să acoperim baza de date **WSD "dataset.pickle"** la care s-a lucrat până în prezent de către studenții din diferite unități de învățământ, inclusiv Academia Tehnică Militară “Ferdinand I”, din punct de vedere al statisticilor care se pot obține pe acest set de date.

Setul de date

FIŞIER PICKLE

Conține o listă de dicționare

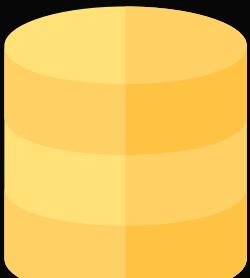
API



RoWordNet - pus la dispoziție de WORDNET;

Lexicul limbii române

Limbaj - Python



SYNSET

Elementul de bază al WORDNET;
Are un ID unic după care poate fi identificat.

Sensul dezambiguizat al unui literal dintr-o anumită propoziție dată.

SYNSETS

Listă de synsets asociate unui literal

Lista cuvintelor cu același sens.

DICȚIONARUL

Cheia - un literal
Valoarea - listă cu dicționare de valori precum : user_id, correct_synset_id, sentence, text,

ROWORDNET

Un graf orientat în care nodurile sunt ID-urile synset-urilor posibile, iar muchiile sunt relațiile dintre acestea.

Synset-urile se păstrează într-un dicționar de tipul {ID: obiect pentru acces};
Acces în O(1).

```
[3] 1 wsd = pd.read_pickle(r'dataset.pickle')
2 rwn = RnWordNet()
3
4 1 wsd['arbore']
5
6 [{"correct_synset_id": "-1",
7     "literal": "arbore",
8     "sentence": "Unul din rolurile sale memorabile a fost cel al Coanei Eleonora Arbore din piesa Micul infern de Mircea Ștefănescu.",
9     "synsets": "ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ",
10    "text": "Arbore",
11    "text_postfix": " din piesa Micul infern de Mircea Ștefănescu.",
12    "text_prefix": "Unul din rolurile sale memorabile a fost cel al Coanei Eleonora ",
13    "user_id": "94"}, {"correct_synset_id": "-1",
14     "literal": "arbore",
15     "sentence": "Piloți Top Gun sunt în partea de sus a arborelui lor, dar chiar și la acest nivel unii piloti sunt în mod constant mai bine la o mai bună anticipare și",
16     "synsets": "ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ",
17     "text": "arborelui",
18     "text_postfix": " lor, dar chiar și la acest nivel unii piloti sunt în mod constant mai bine - mai bine la o mai bună anticipare și",
19     "text_prefix": "Piloți Top Gun sunt în partea de sus a ",
20     "user_id": "94"}, {"correct_synset_id": "ENG30-04111190-n",
21     "literal": "arbore",
22     "sentence": "Mecanismul acestei osii a impus înlăturarea arborelui conector dintre cele două osii principale, lăsând astfel doar osia posterioară motoare.",
23     "synsets": "ENG30-03726760-n ENG30-13912260-n ENG30-12752039-n ENG30-13104059-n ENG30-12662772-n ENG30-04111190-n -1 ",
24     "text": "arborelui",
25     "text_postfix": " conector dintre cele două osii principale, lăsând astfel doar osia posterioară motoare.",
26     "text_prefix": "Mecanismul acestei osii a impus înlăturarea ",
27     "user_id": "94"}]
```

Vizualizarea structurii datelor obținute din
baza de date pentru un literal ales

Objective



STABILIREA STATISTICILOR



Determinarea numărului mediu de poropozitii per literal

Determinarea numărului mediu de synset-uri candidat unui literal

Determinare datelor despre fiecare literal

Determinarea datelor despre fiecare synset

Determinarea numărului de propoziții realizate de fiecare utilizator

Determinarea numărului mediu de propoziții per literal

Pași urmați pentru obținerea
rezultatului final

Creare dicționar de tipul {literal: număr propoziții
asociate}

Aplicarea funcției de medie numpy.mean() pe valorile din
dicționar.

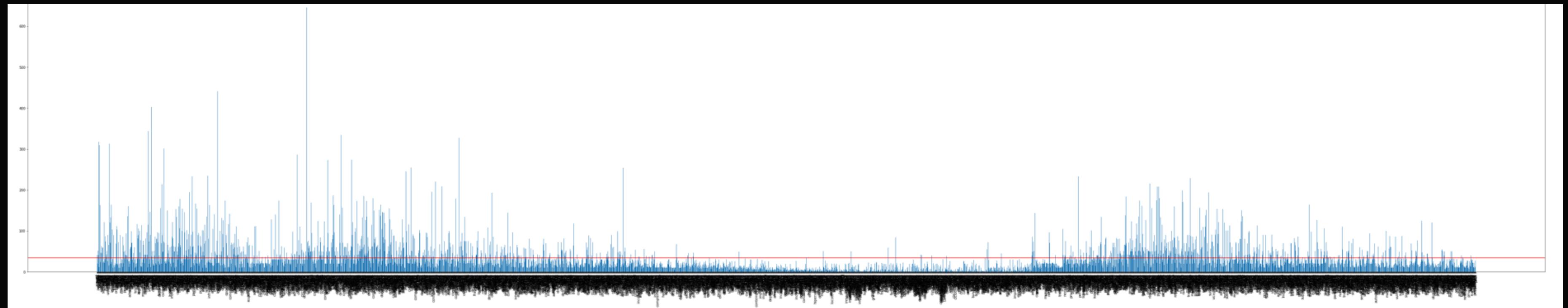
Rezultat obținut

34.463941

În medie, fiecare literal are asociate 34 de propoziții în care
acesta apare cu diferite sensuri care trebuie dezambiguizate.

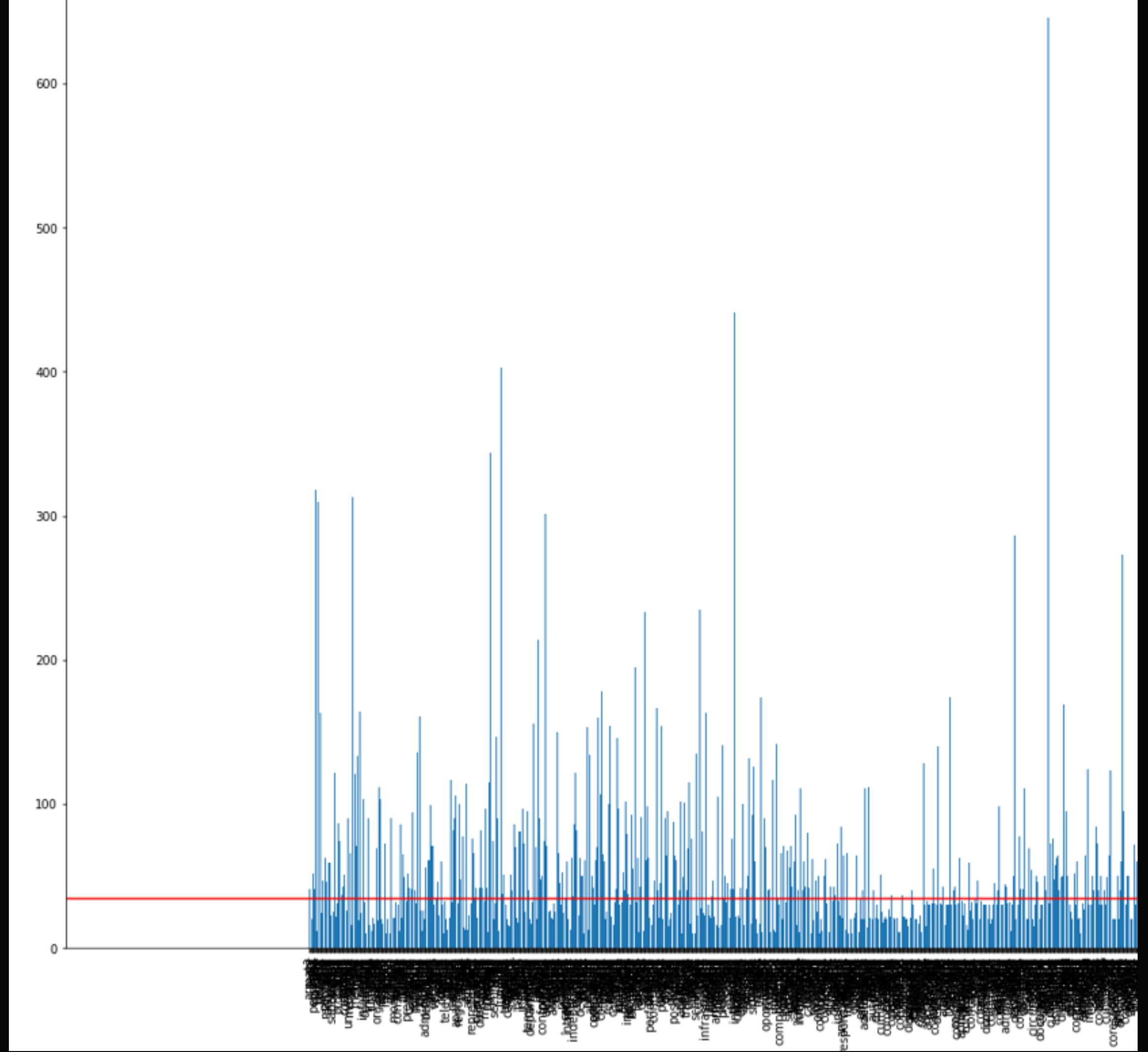
Cum am găsit numărul de propoziții?

Ceea ce returnează baza de date pentru un singur cuvânt este
o listă de dicționare cu structura din slide-ul anterior, având de
fiecare dată altă propoziție asociată literalului respectiv.

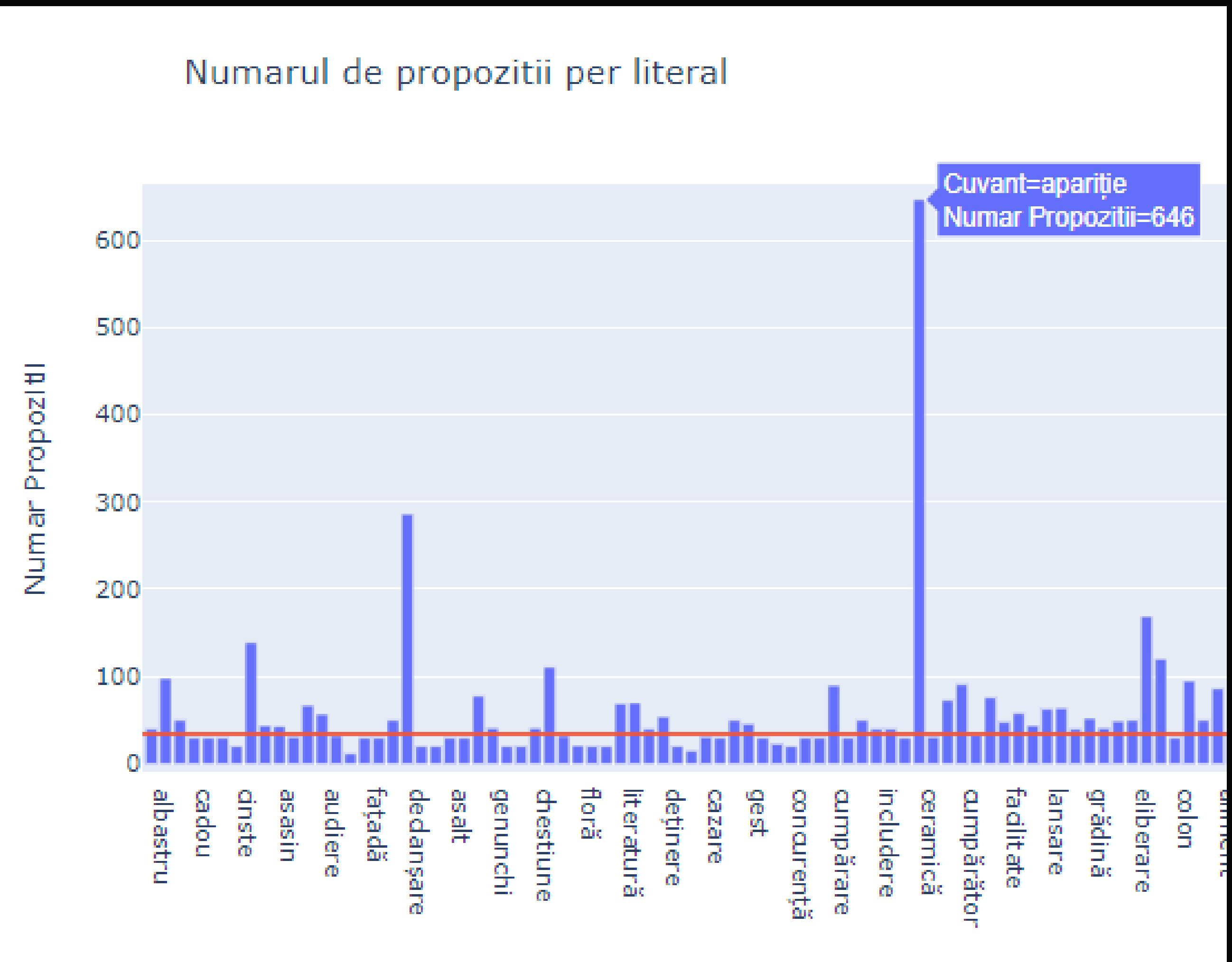


Distribuția numărului mediu de propoziții per literal

Vizualizare mărită



Vizualizare interactivă



Determinarea numarului mediu de synset-uri candidat per literal

Pași urmați pentru obținerea
rezultatului final

Creare dicționar de tipul {literal: număr synsets
candidat fără ultimul (valoare -1) }

Aplicarea funcției de medie numpy.mean() pe lista numărului
de synset-uri asociate fiecărui cuvânt din dicționar.

Rezultat obținut

3.744119

În medie, fiecare literal are asociate 4 sensuri posibile în
cadrul propozițiilor în care acesta apare. Aceste synset-uri
reprezintă sinonimele literalului.

Cum am găsit numărul de synset-uri?

Prin selectarea oricărei propoziții din baza de date pentru un
literal ales, se va obține o listă cu sinonimele posibile asociate
. Pentru oricare intrare din dicționar, lista de sinonime
asociate unui cuvânt este mereu aceeași.

Lista id-urilor sinonimelor asociate cuvântului "rol"

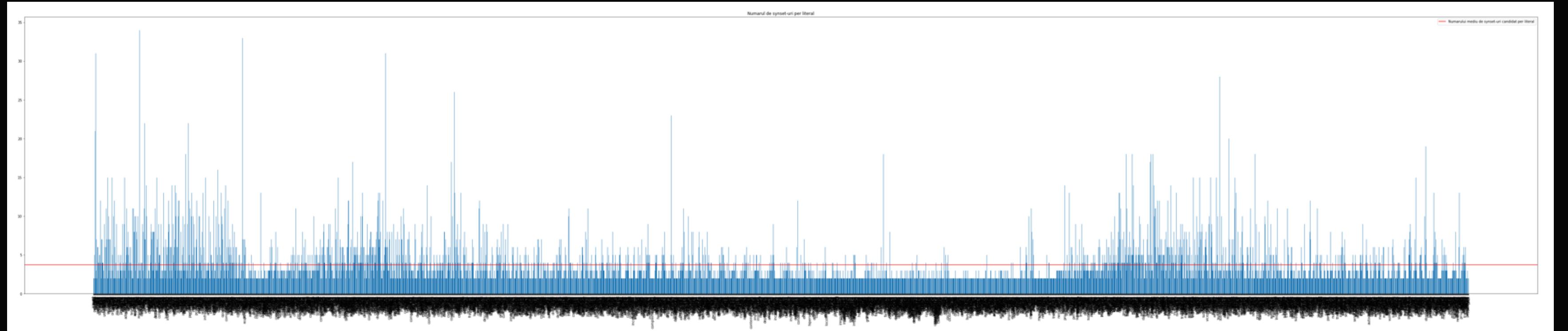
```
[ ] 1 print(len(wsd['rol'])) #nr de propozitii  
2 print(wsd['rol'][0]['synsets'])  
3 print(wsd['rol'][1]['synsets'])  
4 print(wsd['rol'][50]['synsets'])
```

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

Indiferent de propoziția în care apare cuvântul **rol**, lista sensurilor acestui cuvânt rămâne constantă.

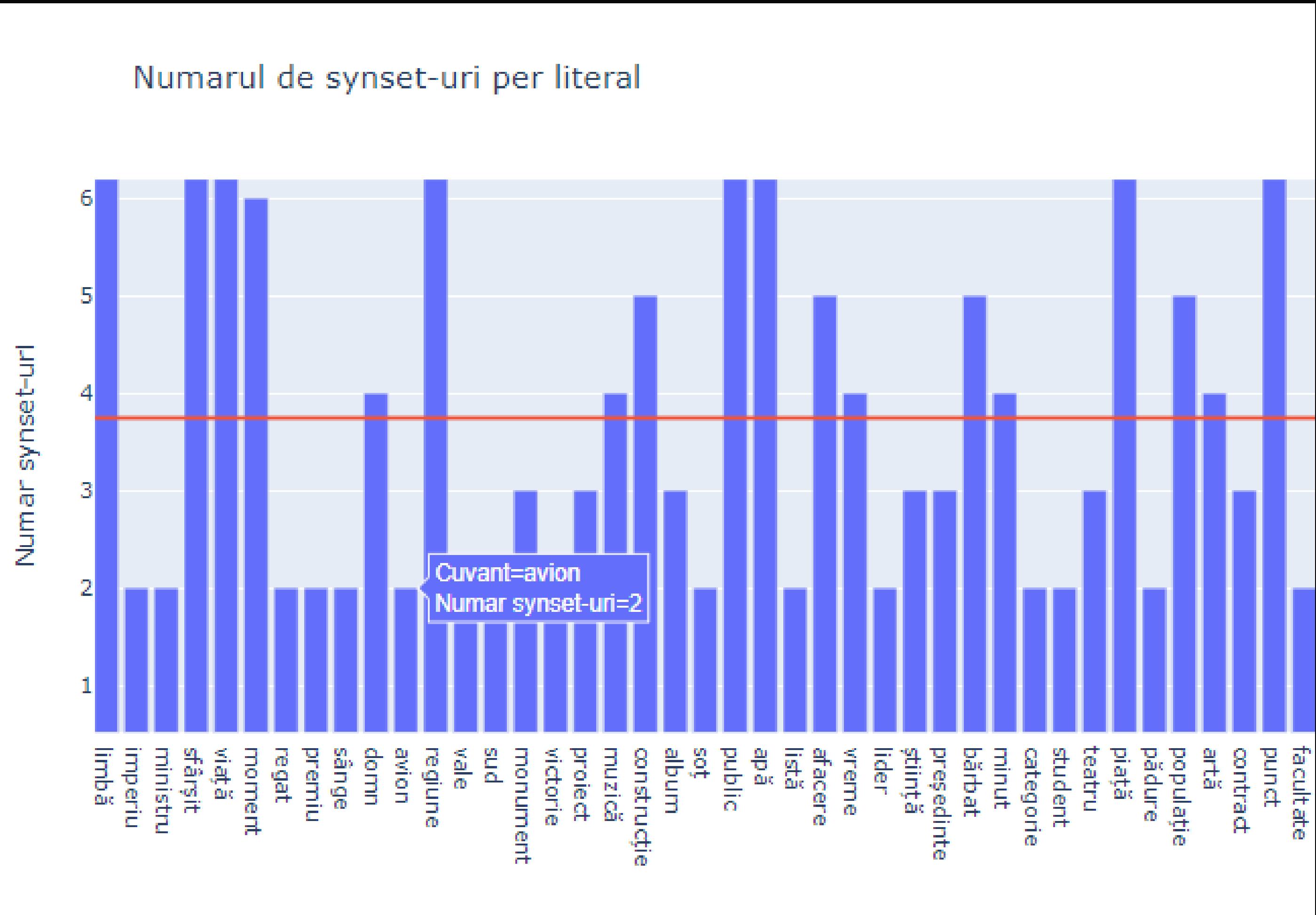
```
1 print(num_synset_candidates_dict['rol'])  
2 print(num_synset_candidates_dict)  
3 print(num_synset_candidates_med)
```

5
{'armată': 4, 'rol': 5, 'secol': 2, 'judet': 2, 'oraș': 5, 'persoană': 3, 'comună': 2,
[3.744118781334362]}



Distribuția numărului de synset-uri per literal

Vizualizare interactivă



Determinarea datelor despre fiecare literal

Pași urmați pentru obținerea
rezultatului final

```
51
[ 'ENG30-09587565-n', 'ENG30-00720565-n', 'ENG30-06331803-n', 'ENG30-05929008-n', 'ENG30-00722061-n']
[ 3. 11. 4. 6. 26.]
```

Creare structură de date care descrie un literal

Am realizat o listă a distribuției numărului de sinonime alese de utilizatori din toate propozițiile asociate cuvântului dat.

Exemplu

Distribuția numărului de variante sinonime alese de utilizatori în cazul cuvântului **rol** este:

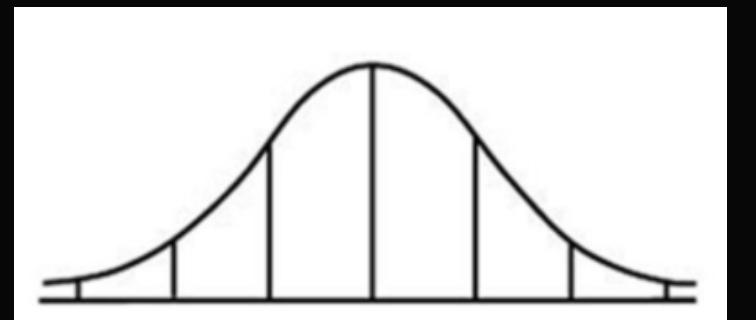
Calcularea statistici pe baza vectorului de distribuție

Calcularea statisticilor asociate literalilor pe baza vectorului de distribuție

```
var = mean (abs (x - x.mean ())2)
```

Varinace

numpy.var()



Mean

numpy.mean()

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

Standard Deviation

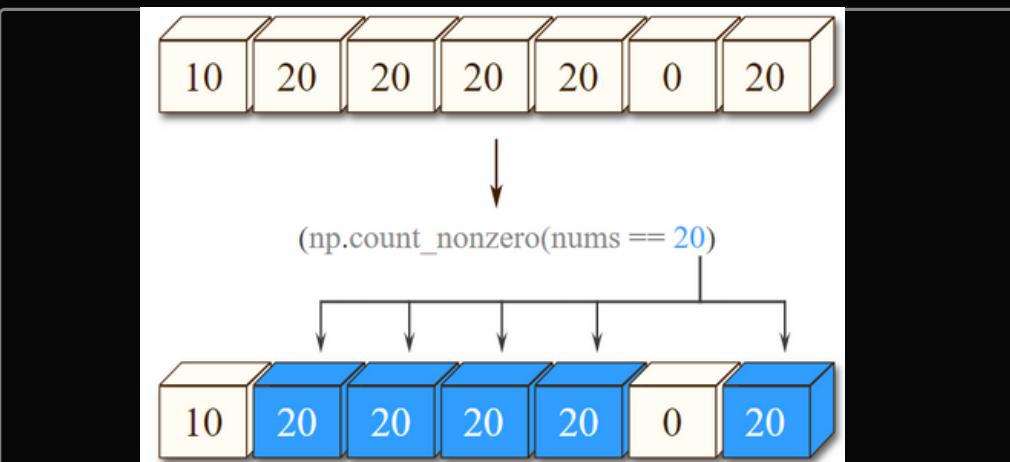
numpy.std()

Calcularea statisticilor asociate synset-urilor pe baza vectorului de distribuție

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

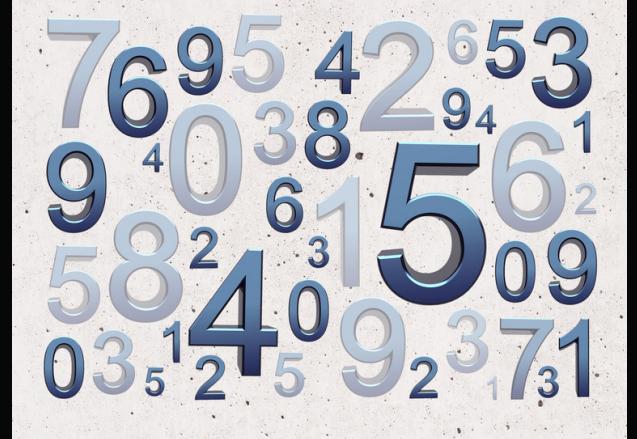
Skewness

scipy.skew()



Number of synsets with 0 sentences

`length(synsets) - np.count_nonzero(synsets)`



Number of sentences

`length(wsd[word])`



```
{"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}
```

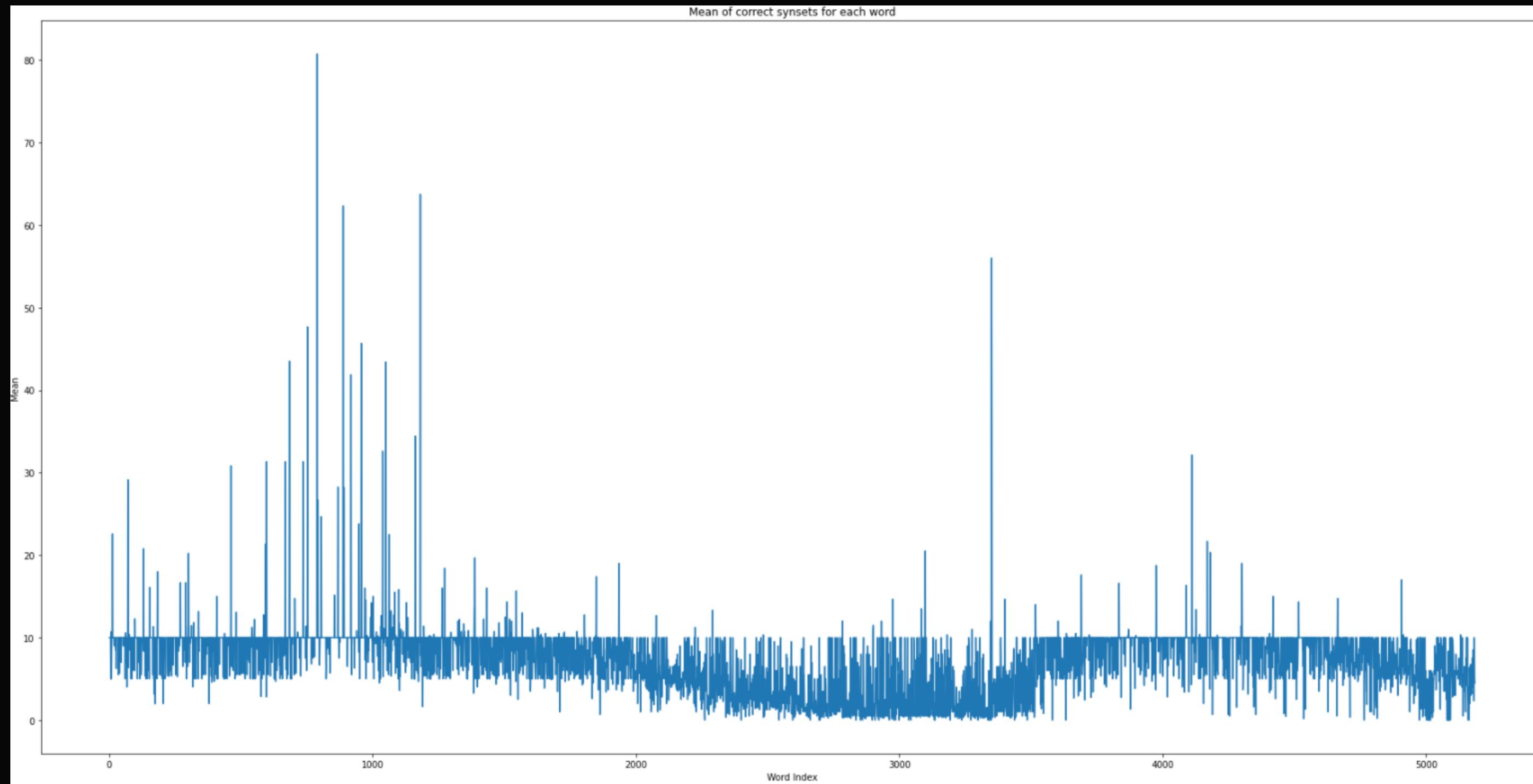


Structura de statistici obținute pentru
cuvântul **rol**

Datasetul cu datele despre fiecare literal

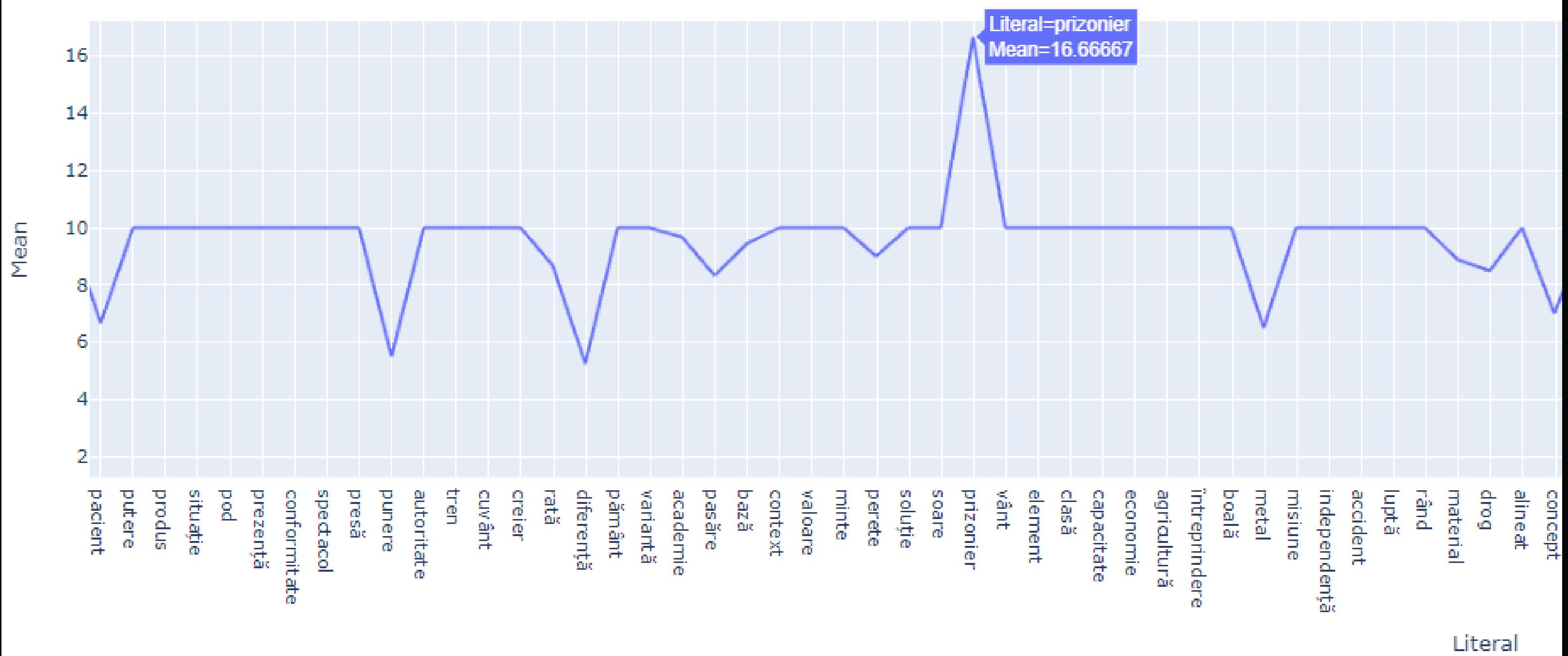
| | Literal | Mean | Variance | Standard Deviation | Skewness | Number of synsets with 0 sentences | Number of sentences |
|------|-------------|-----------|------------|--------------------|-----------|------------------------------------|---------------------|
| 0 | armată | 10.000000 | 31.500000 | 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 | tătă | 5.000000 | 9.000000 | 3.000000 | 0.000000 | 0 | 12 |
| 5185 | pisoii | 4.500000 | 6.250000 | 2.500000 | 0.000000 | 0 | 12 |

5186 rows × 7 columns



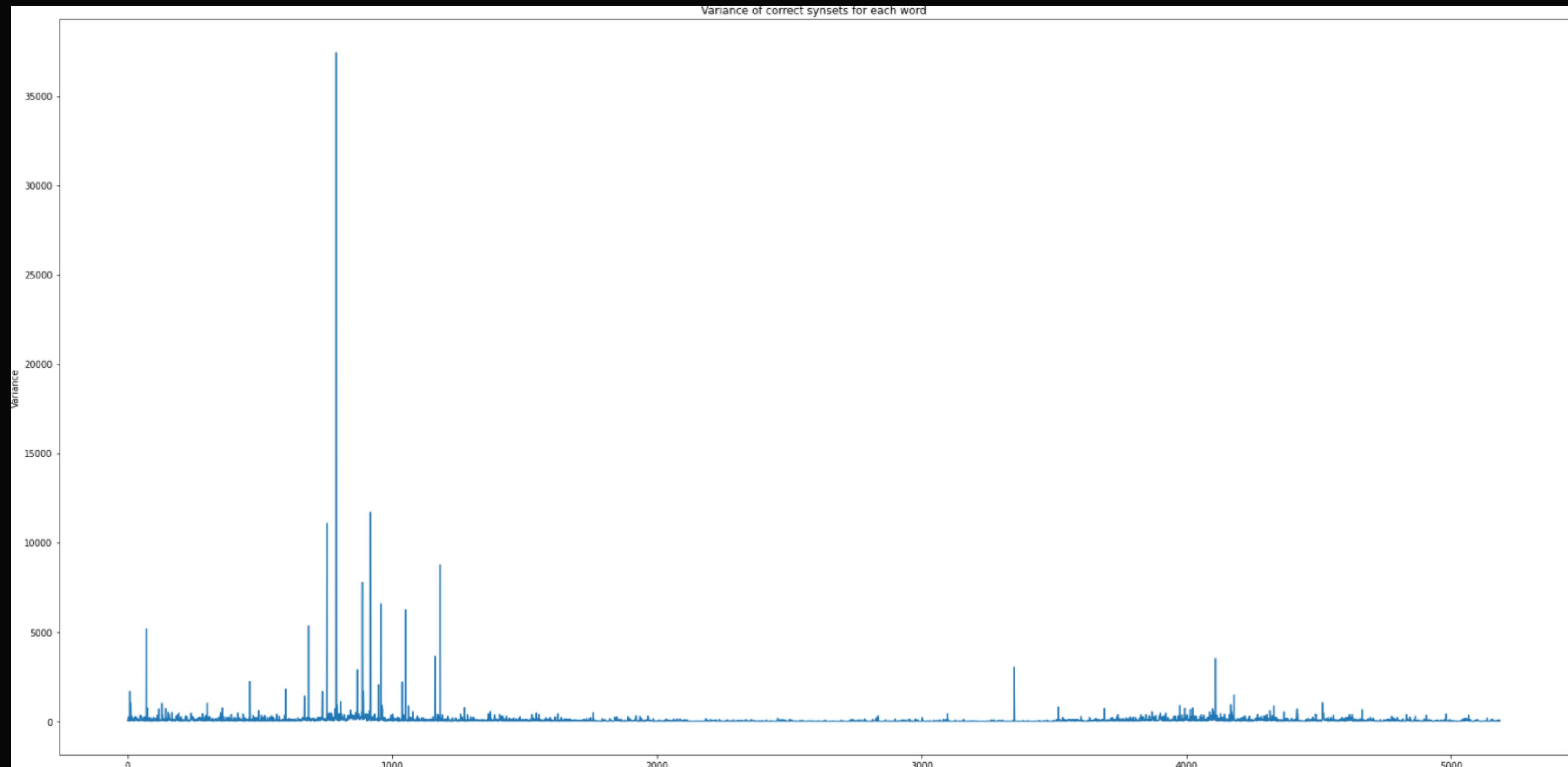
**Mean of correct synsets for each
word**

Mean of correct synsets for each word



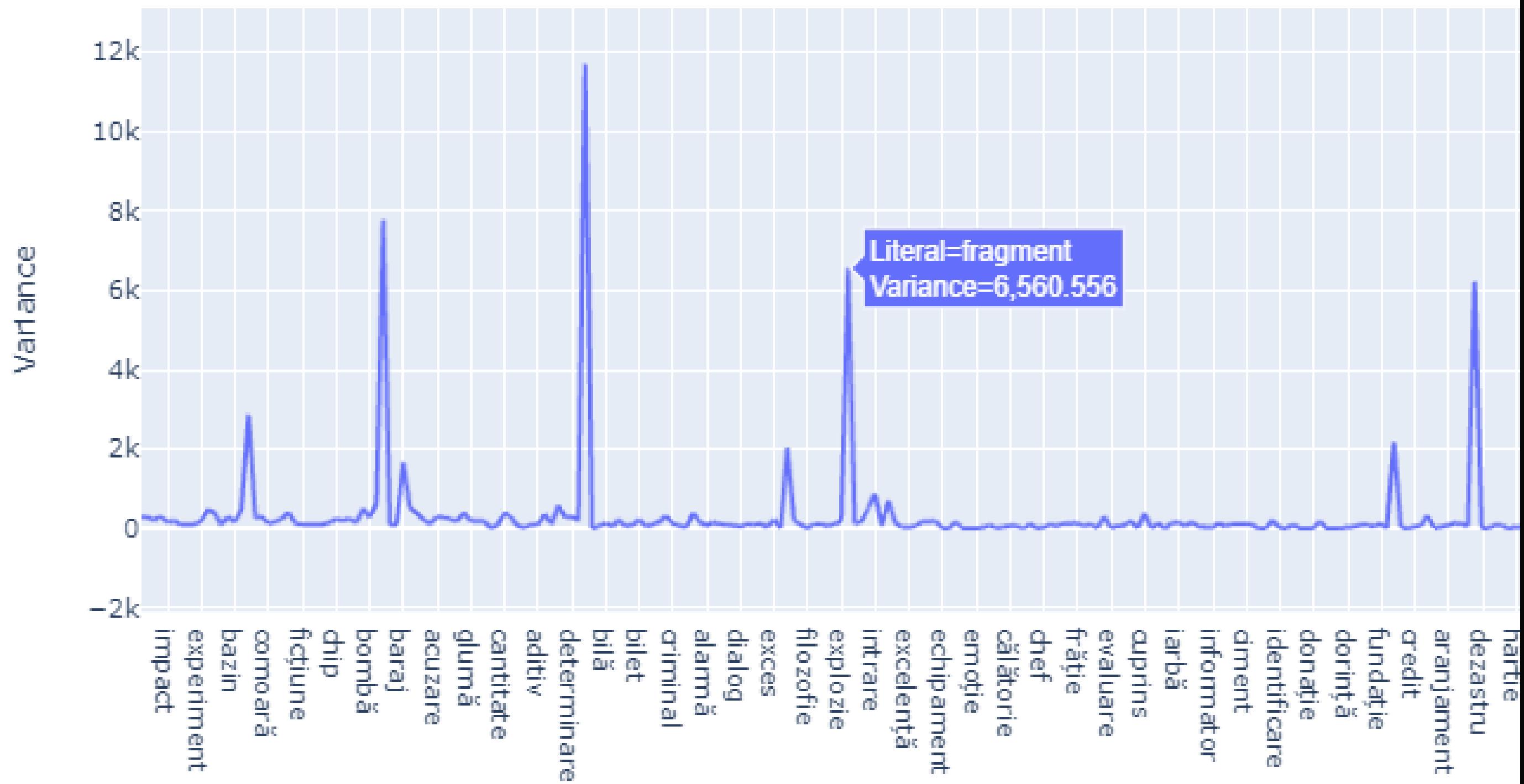
Afișare interactivă

Variance of correct synsets for each word

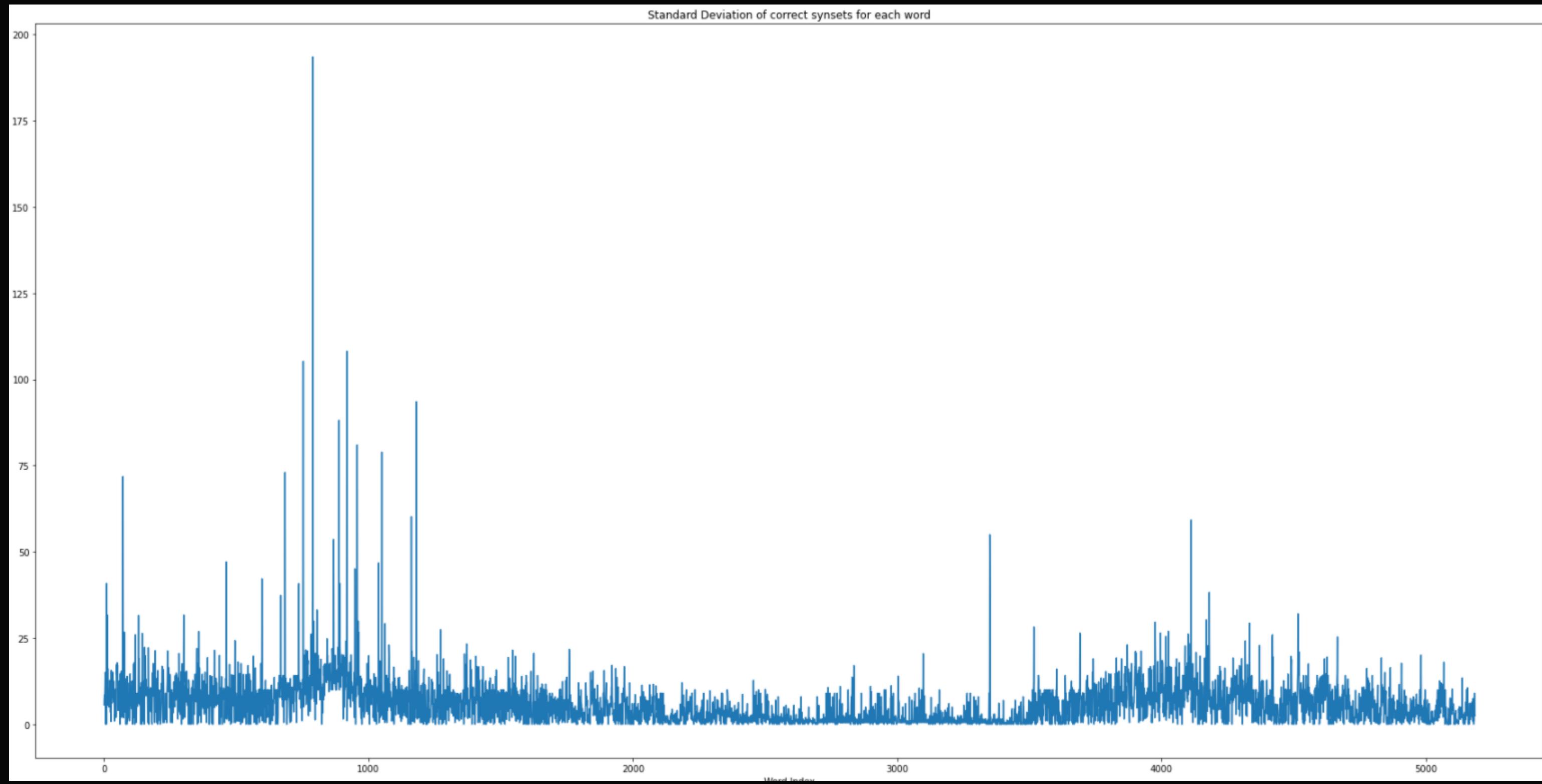


Variance of correct synsets for each word

Variance of correct synsets for each word

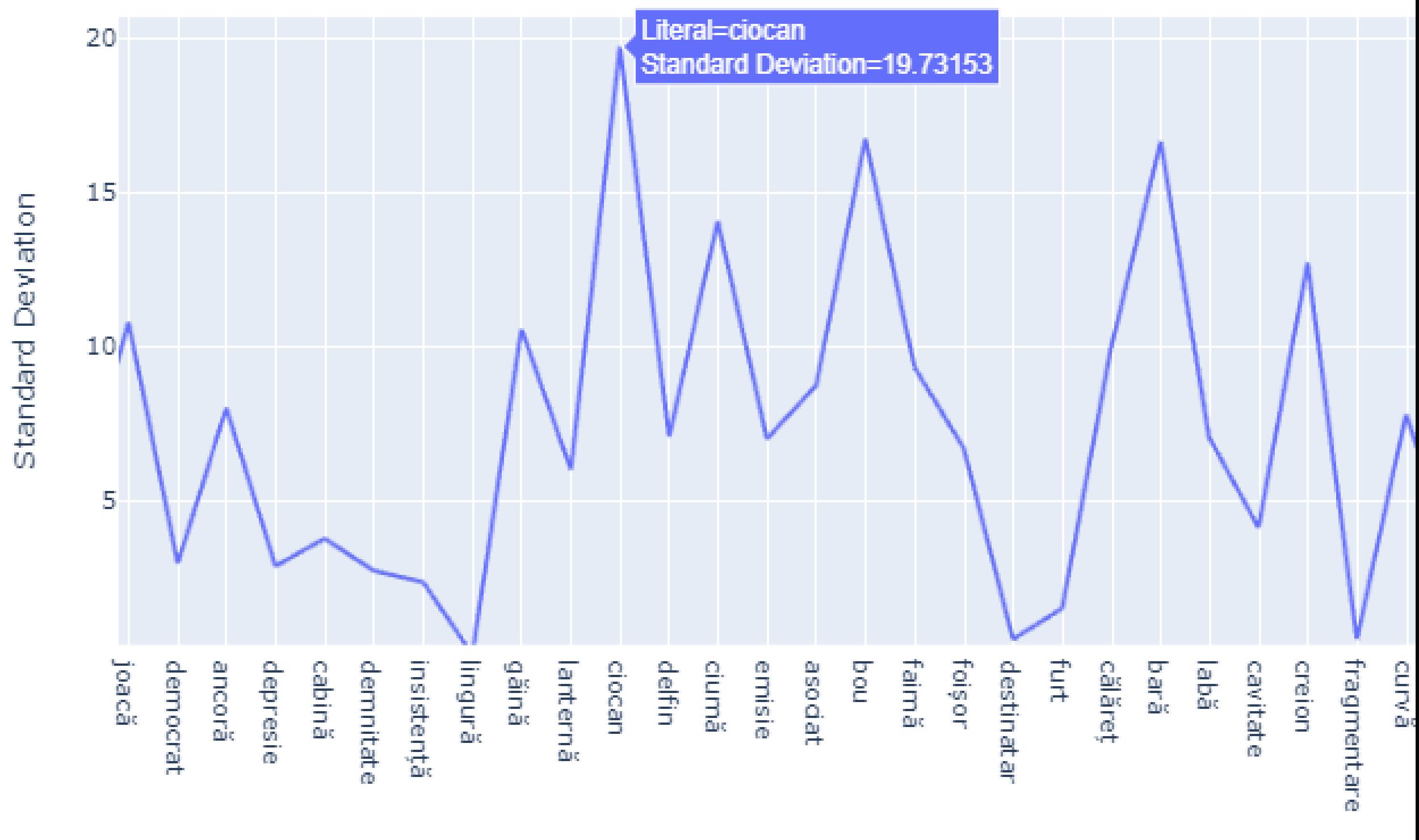


Afișare interactivă

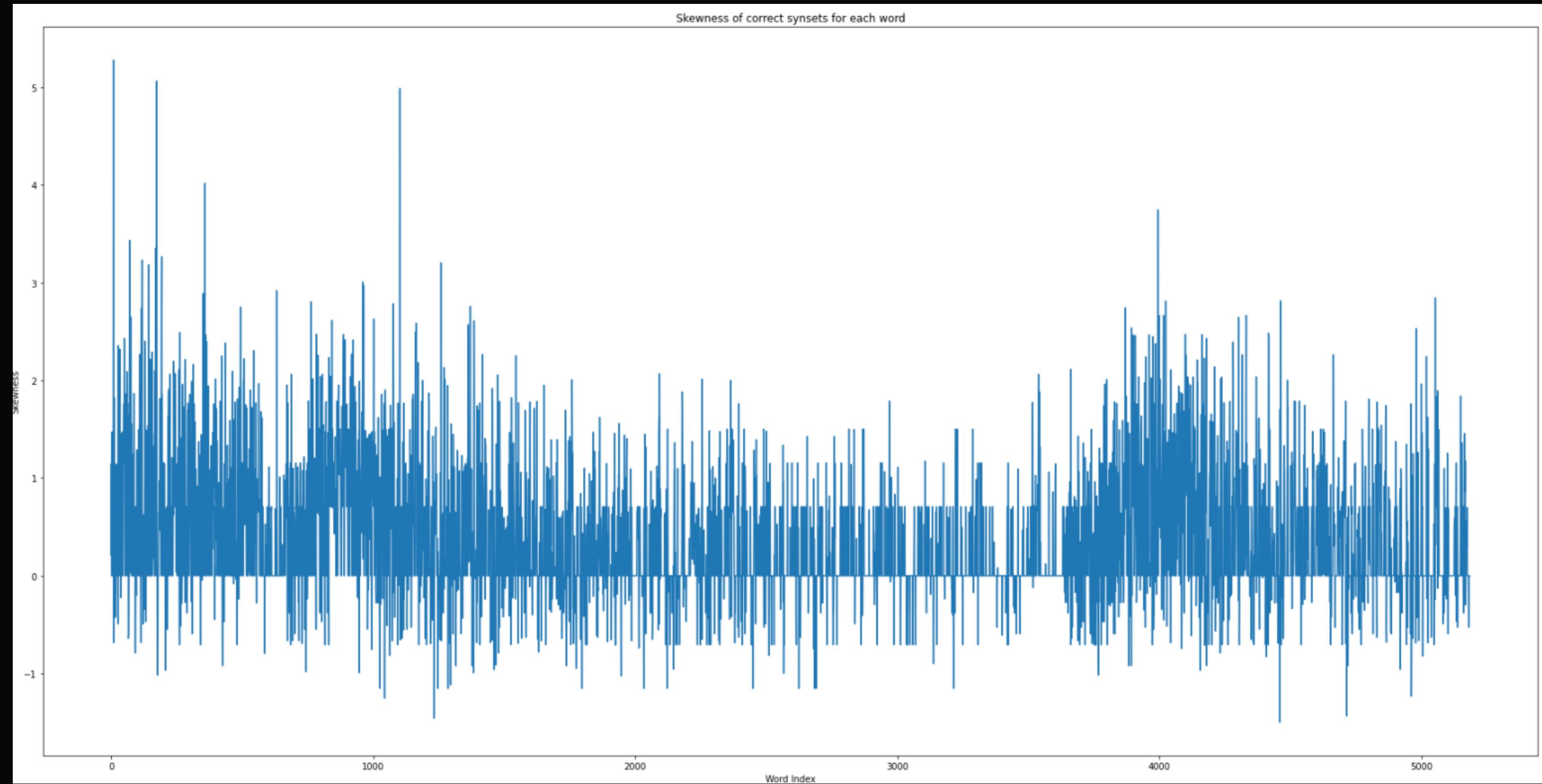


Standard Deviation of correct synsets for
each word

Standard Deviation of correct synsets for each word

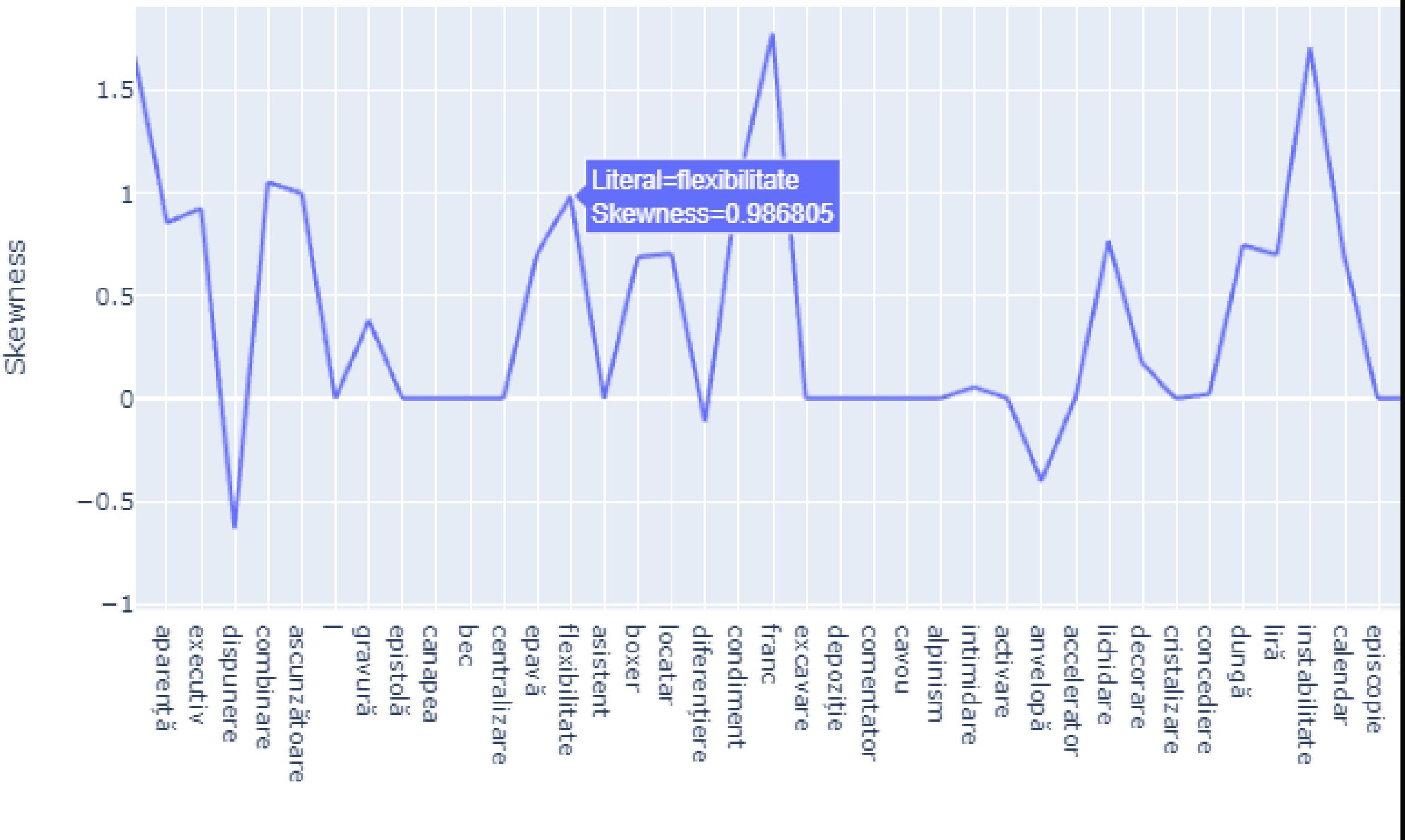


Varianta interactivă

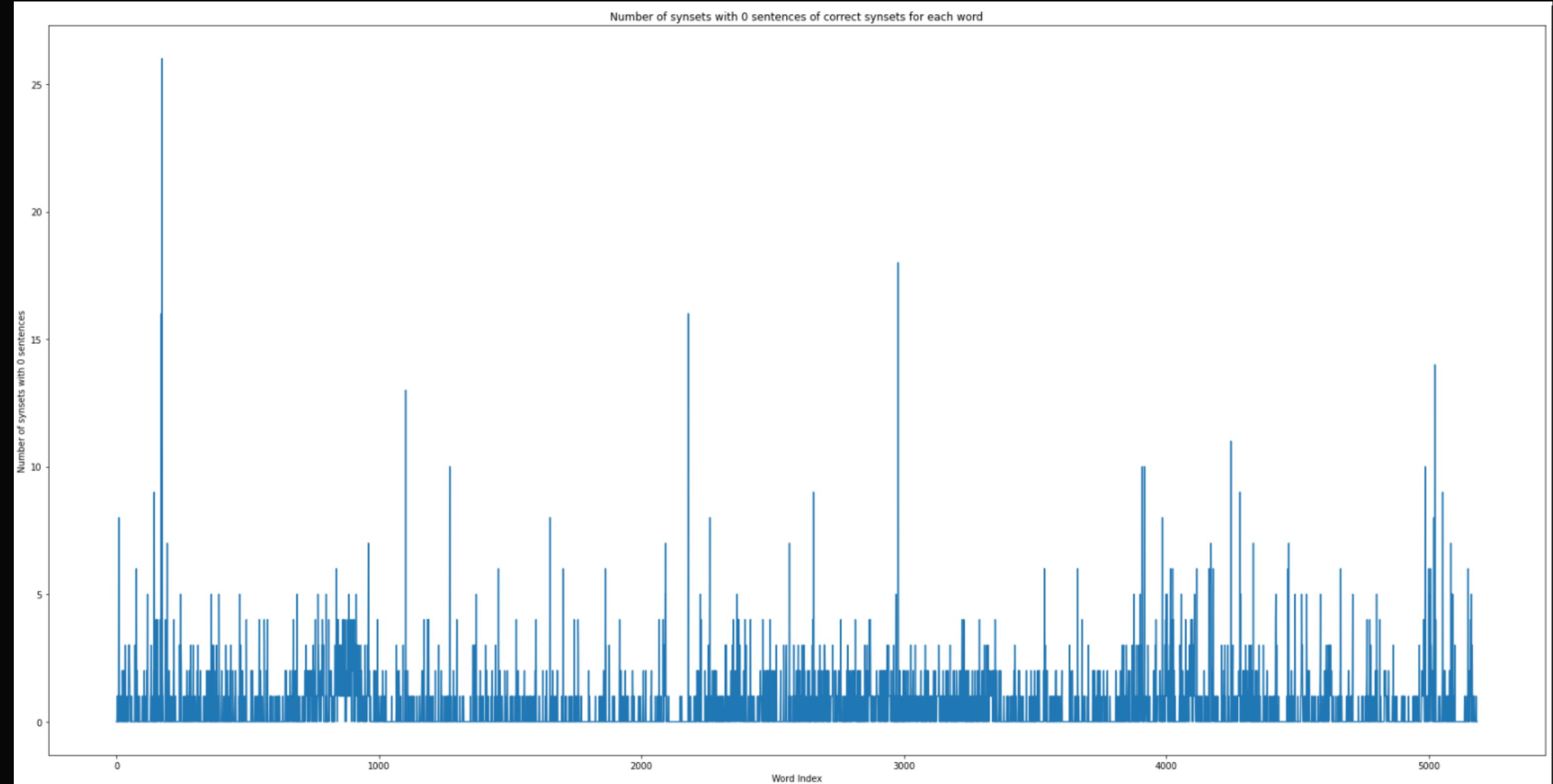


Skewness of correct synsets for each word

Skewness of correct synsets for each word

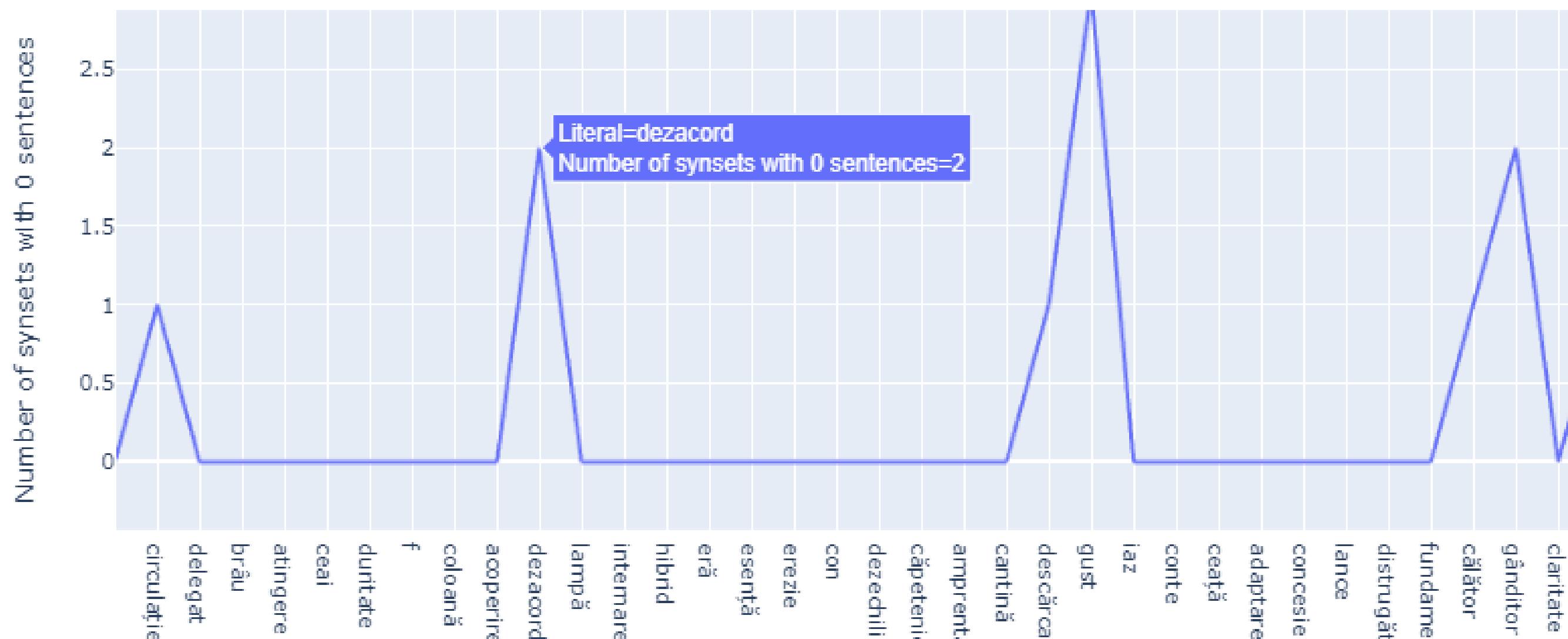


Varianta interactivă

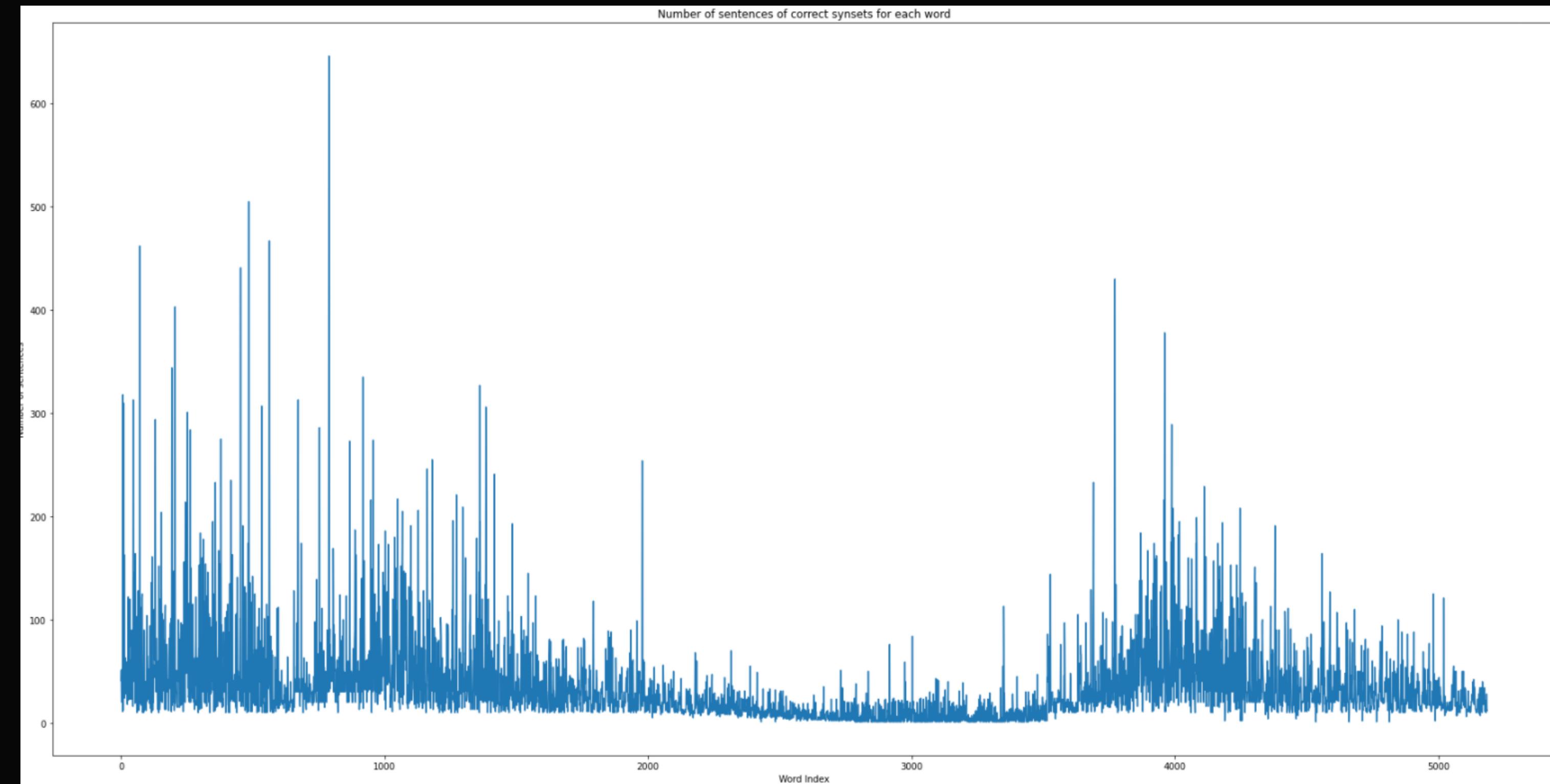


Number of synsets with 0 sentences of correct
synsets for each word

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

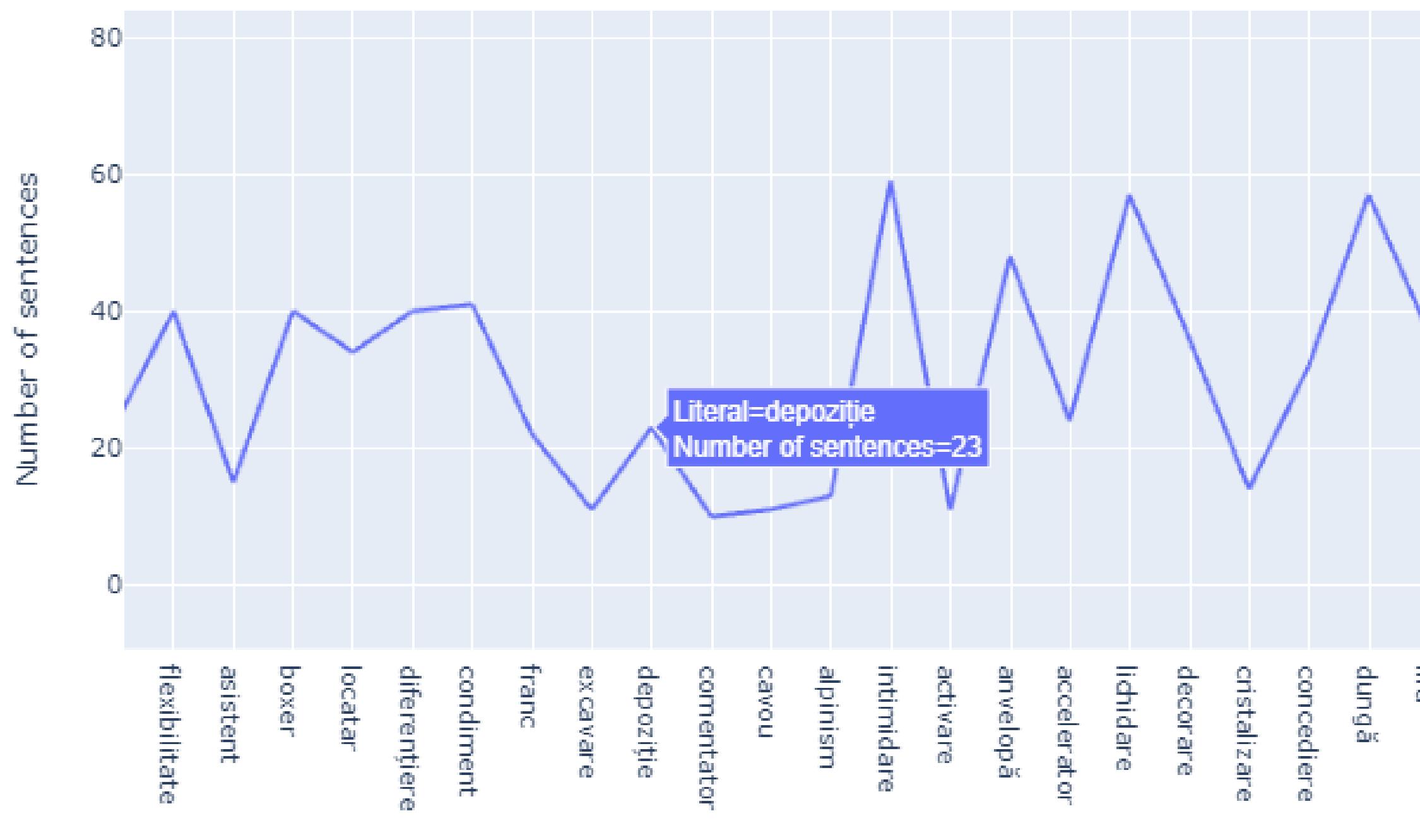


Varianta interactivă



Number of sentences for each word

Number of sentences of correct synsets for each word



Varianta interactivă

Determinarea datelor despre fiecare synset

Pași urmați pentru obținerea rezultatului final

Obținerea unui dicționar cu date ce descriu un synset

Pe baza id-ului unui synset am calculat următoarele date:

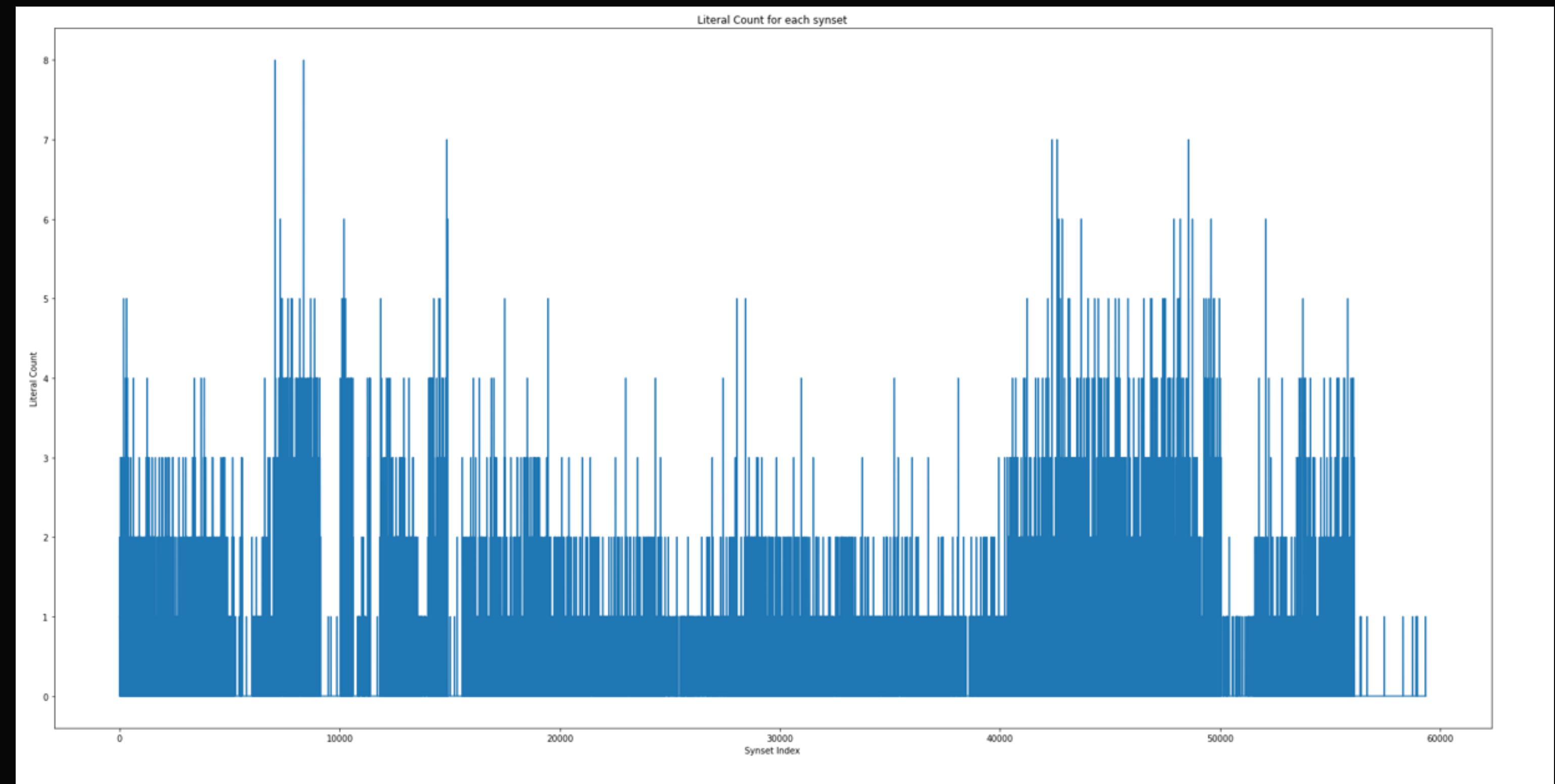
- Listă cu literalii a căror sinonim este synset-ul respectiv
- Numărul de literali găsiți mai sus în lista anterioară
- Numărul total de propoziții în cadrul cărora synset-ul apare drept variantă de răspuns

Rezultat

```
1 print(get_synset_data('ENG30-00017222-n'))  
  
{'ID_Synset': 'ENG30-00017222-n', 'Literal_List': ['plantă'], 'Literal_Count': 1, 'Number_of_sentences': 29}
```

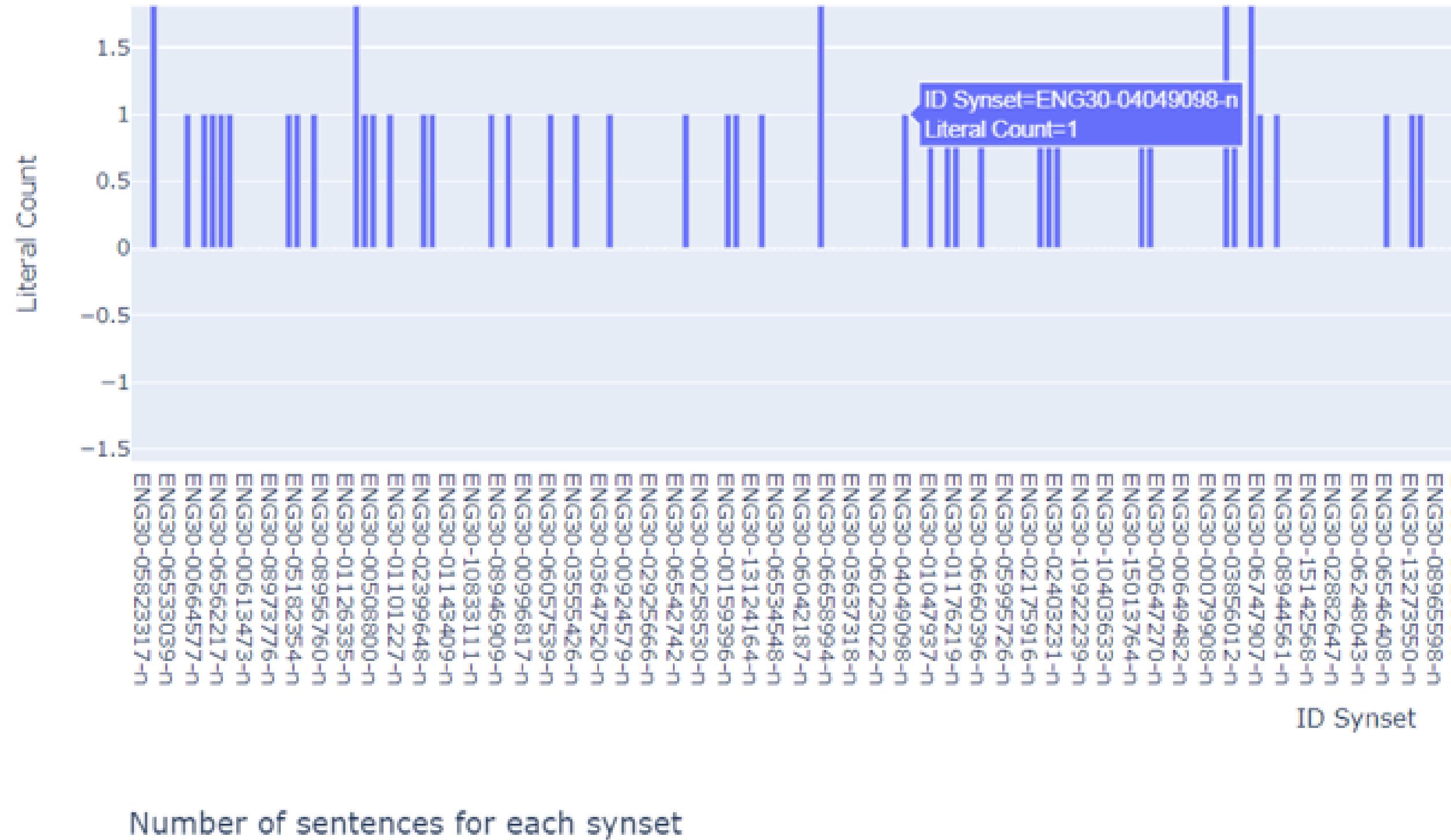
Pe baza acestei structuri de date create, am construit un dicționar cu toate id-urile synset-urilor din RoWordNet și structurile lor asociate.

| | 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 | | | | | |



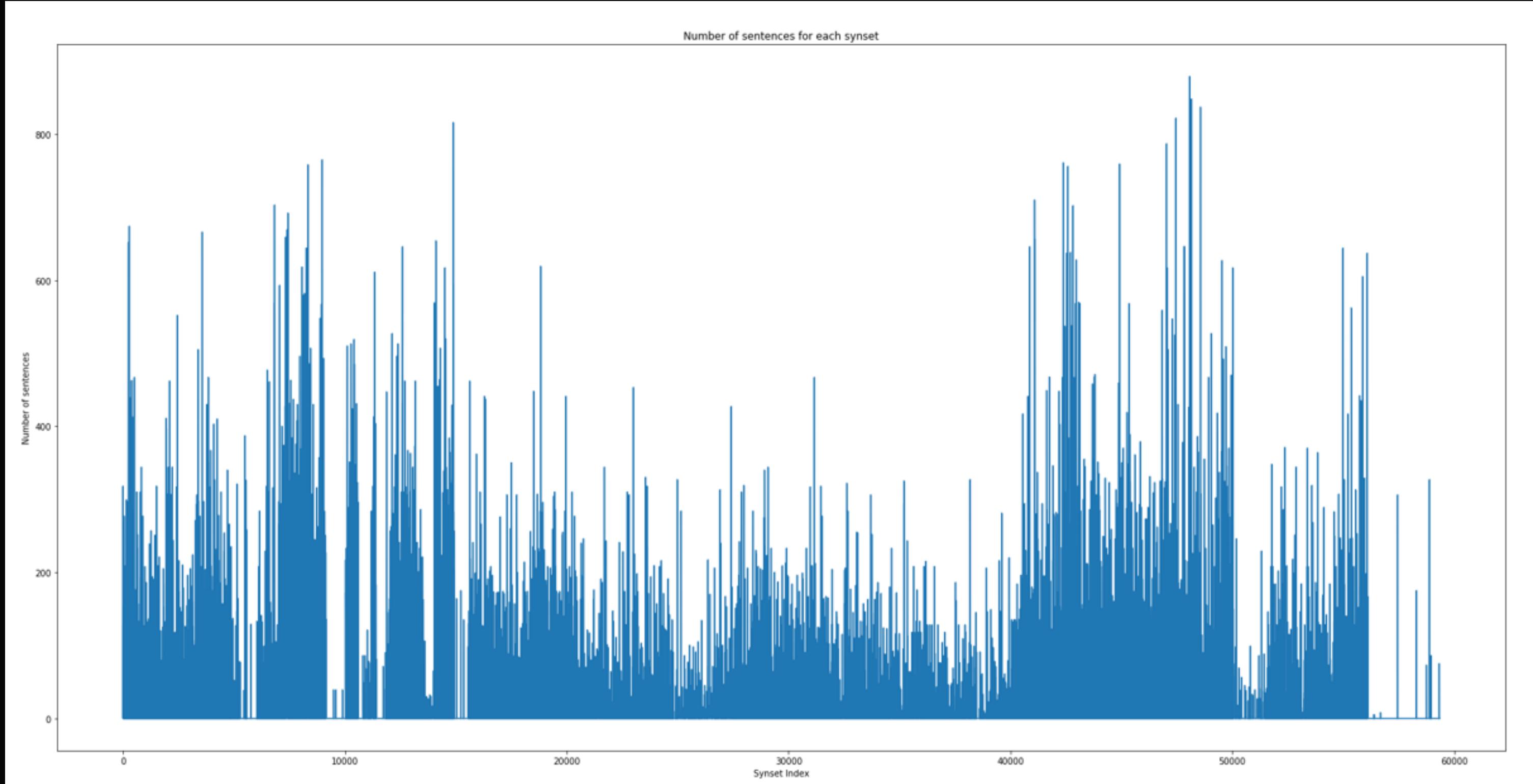
Literal Count for each synset

Literal Count for each synset



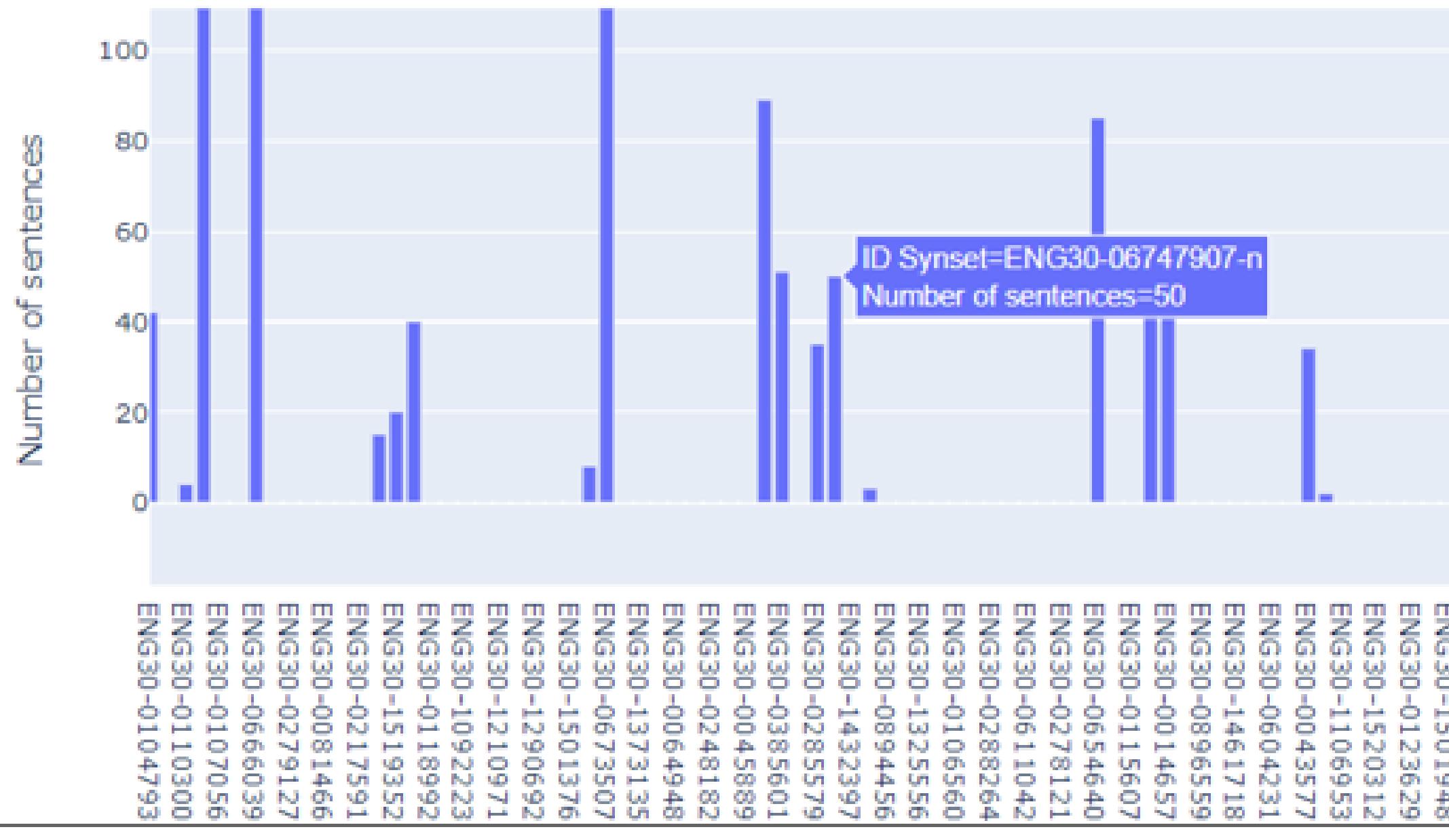
Number of sentences for each synset

Varianta interactivă



Number of sentences for each synset

Number of sentences for each synset



Varianta interactivă

Determinarea numarului de propozitii facut de fiecare utilizator

Pași urmați pentru obținerea
rezultatului final

Obținerea unei liste cu toți utilizatorii

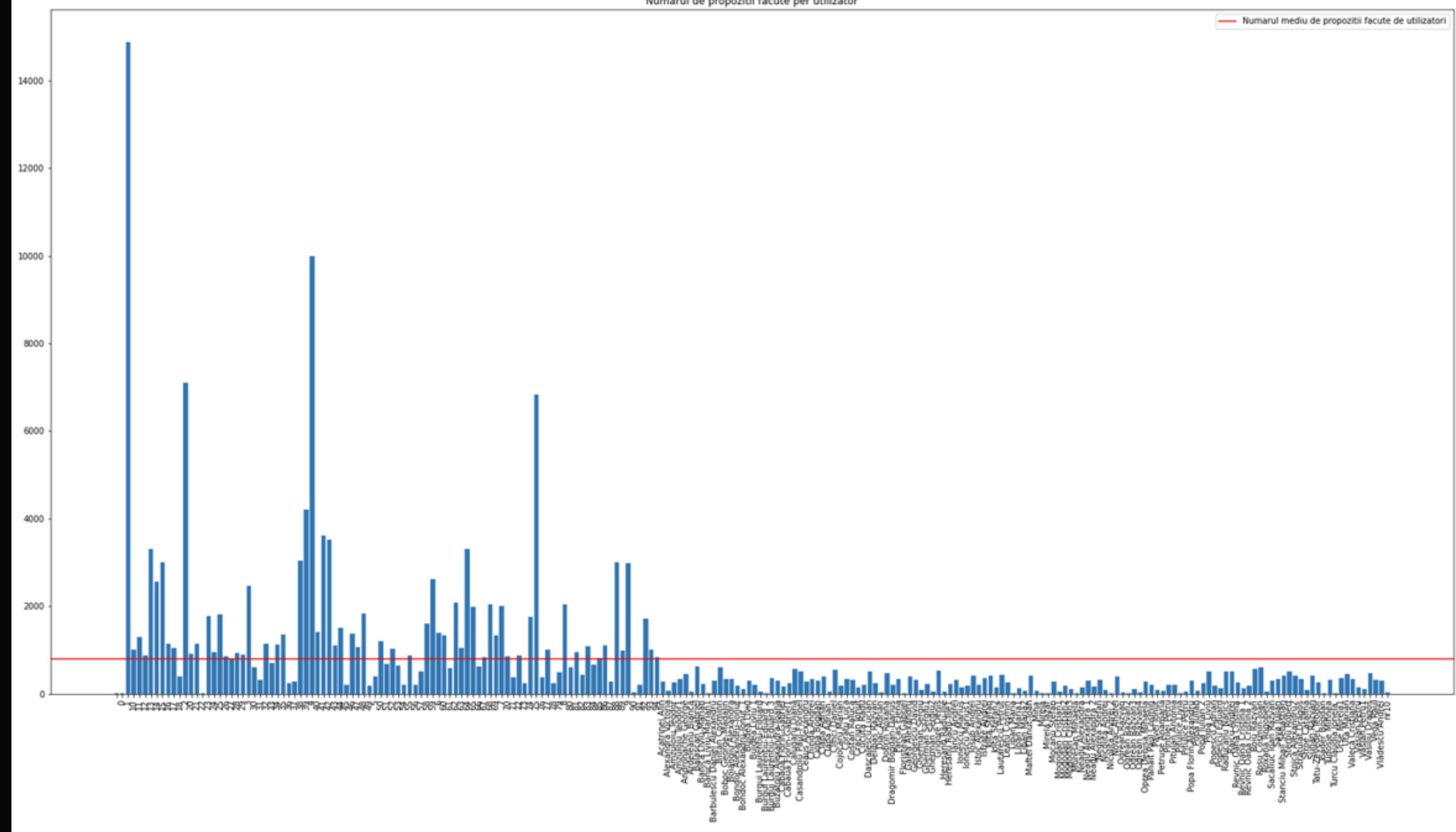
Lista e extrasă pe baza câmpului 'user_id' din fiecare dicționar asociat unei propoziții.

Crearea unui dicționar de tipul {username : număr de propozitii complete}

Se caută în fiecare propoziție asociată fiecărui cuvânt din baza de date daca a fost completată de utilizatorul cu id-ul dat ca argument

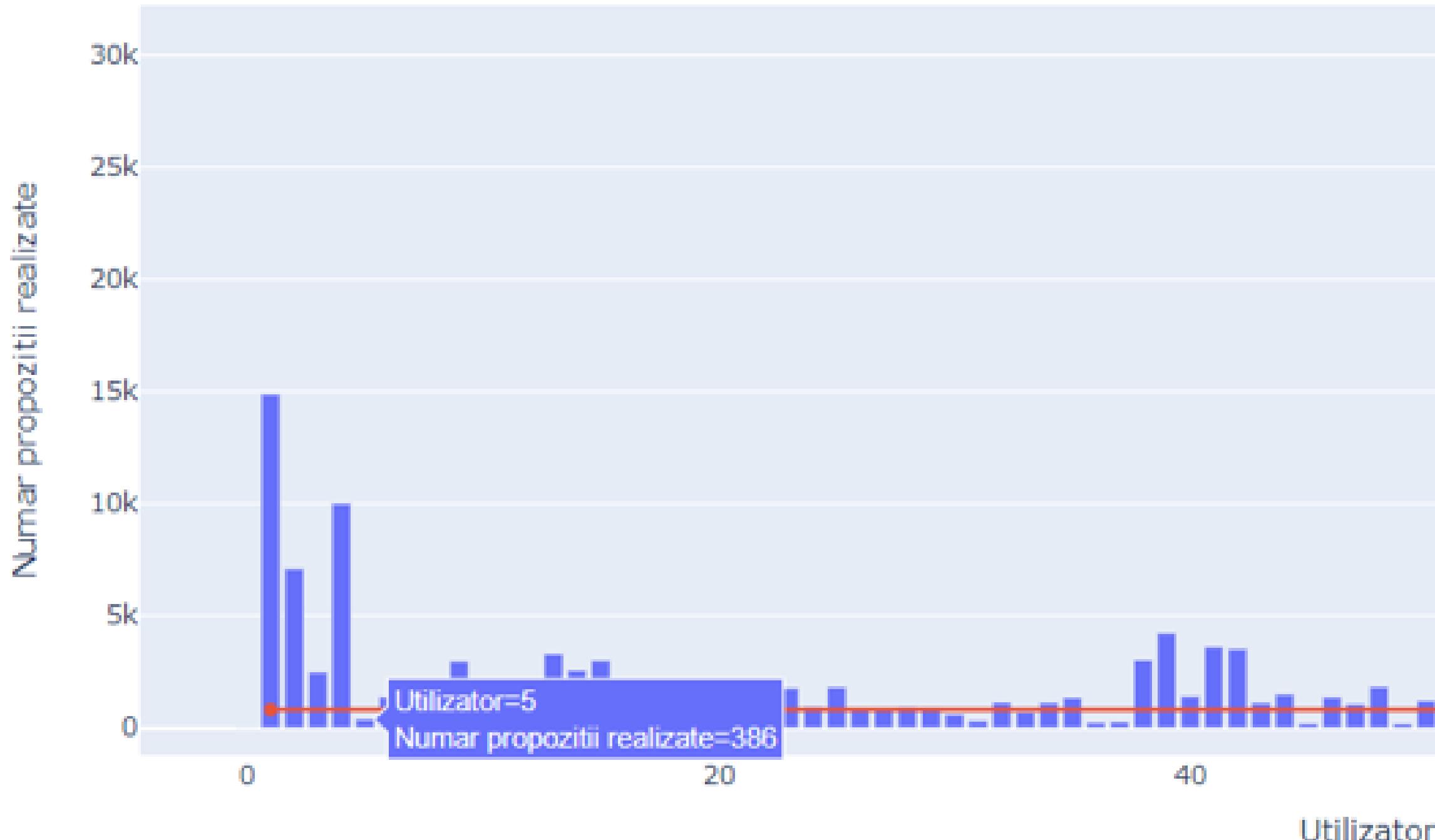
Rezultatul mediei de propozitii per utilizator

804.8243243243244a

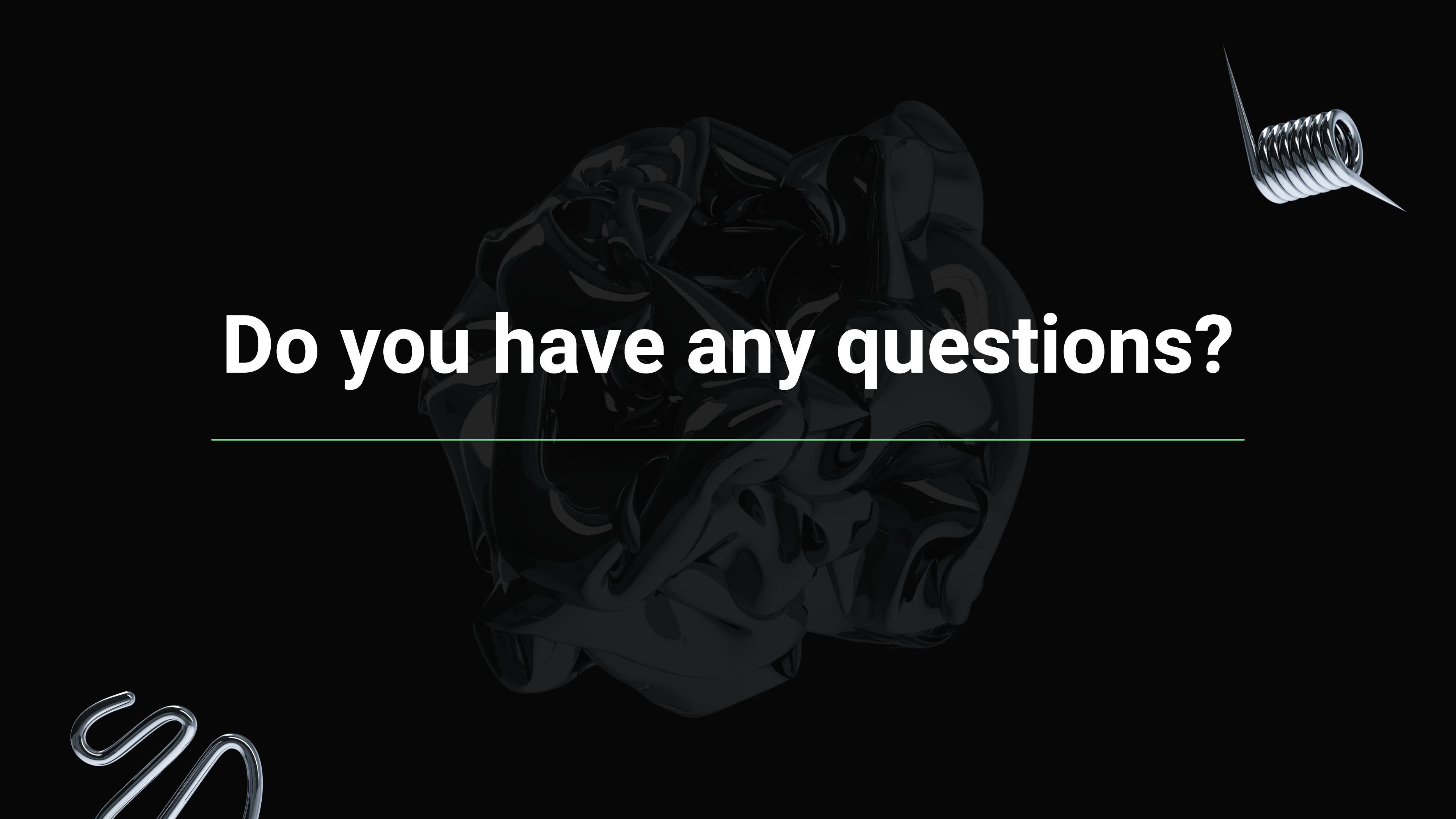


Number of sentences for each user

Numarul de propozitii facute per utilizator



Varianta interactivă



Do you have any questions?

