

WCES, PARIS – FRANCE, 03-05 FEBRUARY 2022

## Analizimi i tekstit në gjuhën shqipe

Ilir Perolli

*Kosovar, 9 Nentori, Gjakove 50000, Kosove*

---

### Abstrakti

Një gjuhë është një sistem i strukturuar i komunikimit i përdorur nga njerëzit, duke përfshirë shkrimin, të folurit (gjuhë e folur) dhe gjeste (gjuhë shenjash) [1].

Gjuhët evoluojnë dhe diversifikohen me kalimin e kohës, dhe historia e evolucionit të tyre mund të rindërtohet duke krahasuar gjuhët moderne për të përcaktuar cilat tipare duhet të kishin gjuhët e tyre paraardhëse në mënyrë që të ndodhin fazat e mëvonshme të zhvillimit. Një grup gjuhësh që vijnë nga një paraardhës i përbashkët njihet si një familje gjuhësore; në të kundërt, një gjuhë që është demonstruar se nuk ka ndonjë marrëdhënie të gjallë ose jo të gjallë me një gjuhë tjetër quhet izolim i gjuhës. Ka edhe shumë gjuhë të paklasifikuara, marrëdhëniet e të cilave nuk janë krijuar, dhe gjuhët false mund të mos kenë ekzistuar fare. Konsensusi akademik mban se midis 50% dhe 90% të gjuhëve të folura në fillim të shekullit 21 do të jenë zhdukur deri në vitin 2100 [1].

*Fjalët kyqe:* Python, Tekst, Bigram, Zanore;

---

## **1. Hyrje**

Në fushën e sotme të internetit dhe shërbimeve online, të dhënat po gjenerojnë me shpejtësi dhe sasi të jashtëzakonshme. Në përgjithësi, analisti i të dhënave, inxhinieri dhe shkencëtarët janë duke trajtuar të dhëna relacionale ose tabelare. Këto kolona të të dhënave tabelare kanë të dhëna numerike ose kategorike. Të dhënat e gjeneruara kanë një larmi strukturash si teksti, imazhi, audio dhe video. Aktivitetet në internet si artikuj, teksti i faqes në internet, postimet në blog, postimet në mediat sociale po gjenerojnë të dhëna tekstuale të pastrukturuara. Korporata dhe biznesi duhet të analizojnë të dhënat tekstuale për të kuptuar aktivitetet, mendimet dhe reagimet e klientëve për të nxjerrë me sukses biznesin e tyre. Për të konkurruar me të dhëna të mëdha tekstuale, analiza e tekstit po zhvillohet me një ritëm më të shpejtë se kurrë më parë [2].

## **2. Word Count**

Word Count është numri i fjalëve në një dokument. Numërimi i fjalëve mund të jetë i nevojshëm kur një tekst kërkohet për të qëndruar brenda një numri të caktuar të fjalëve. Kjo mund të jetë veçanërisht rasti në akademinë, procedurat ligjore, gazetarinë dhe reklamën. Numri i fjalëve përdoret zakonisht nga përkthyesit për të përcaktuar çmimin e një pune përkthimi. Pikat e fjalëve mund të përdoren gjithashtu për të llogaritur masat e lexueshmërisë dhe për të matur shpejtimin dhe shpejtësinë e leximit (zakonisht në fjalë për minutë). Kur konvertimi i karakterit llogaritet me fjalë, një masë prej 5 ose 6 karakteresh në një fjalë përdoret përgjithësisht për anglisht [3].

## **3. Numërimi i shkronjave në tekst**

Numërimi i shkronjave është specifik për çdo gjuhë. Gjuha angleze përdor 26 letra latine. Alfabeti shqiptar përbëhet nga 36 shkronja. Në këtë rast, përveç letrave latine, përdoren shkronja të dyfishta si bigrame. Përderisa gjuha shqipe përdor bigramet, nevojitet një dizajnim i veçantë i algoritmit për numërimin e shkronjave. Gjithashtu gjuha shqipe shkronja që nuk i përkasin alfabetit latin si: "Ç ç" dhe "Ë ë".

## **4. Aplikacioni për analizën tekstuale**

Për realizimin e këtij procesimi të tekstit është përdorur gjuha programuese Python. Python realisht ka librari të gatshme për numërimin e fjalëve, shkronjave etj., por pasi ne në fokus e kemi gjuhën shqipe, atëherë është realizuar një algoritëm i ri për këtë qëllim.

Së pari inputi i dhënë nga file i hiqen shenjat e pikësimit si: . , ? ‘ : ; etj. dhe zëvendësohen me hapësira. Pas zëvendësimit me hapësira, ndahen fjalët nga hapësirat dhe futen në një varg. Nga madhësia e vargut tregojmë realisht se sa fjalë ka teksti. Për analizën e fjalive, nga file i hiqen shenjat e pikësimit si: pikat (.) ku gjithashtu vendosen në varg dhe shikohet madhësia e vargut.

Për numërimin e shkronjave në tekst është marrur një fjalor (dictionary) me të gjitha shkronjat e gjuhës shqipe në të dhe me frekuencën e përsëritjes së tyre. Shembull: shkronjat = {'a':0,'b':0,'c':0,'ç':0,'d':0,'dh':0}

Më pas i qasemi tekstit shkronjë për shkronjë dhe për çdo shkronjë e rrisim numrin e përsëritjeve në fjalor. Pas numërimit të fjalëve, fjalive etj. këto rezultate i ruajmë në një tekst file dhe i shfaqim grafikisht ato.

#### 4.1 Çfarë është Python?



Python është një gjuhë programuese e interpretuar, e orientuar drejt objektit, e nivelit të lartë me semantikë dinamike. Sintaksa e thjeshtë, e lehtë për të mësuar e Python thekson lexueshmërinë dhe për këtë arsye ul koston e mirëmbajtjes së programit.

Python mbështet module dhe paketa, të cilat inkurajojnë modularitetin e programit dhe ripërdorimin e kodit. Përkthyesi Python dhe biblioteka e gjerë standarde janë në dispozicion në formë burimore ose binare pa pagesë për të gjitha platformat kryesore dhe mund të shpërndahen lirisht.

Versioni i përdorur për zhvillimin e eksperimentit është: 3.7.4

#### 4.2 Si të ekzekutojmë programin?

Së pari hapim filen `article.txt` për të shënuar tekstin. Programi automatikisht kërkon për këtë tekst file. Për këtë shembull kemi marrur romanin Gjakftoftësia nga Ismail Kadare.

Name	Date modified	Type	Size
 <code>text_analysis</code>	3/29/2021 00:24	Python File	5 KB
 <code>article</code>	3/29/2021 00:21	Text Document	484 KB

*Foto 1. Tekst file.*

Pas shënimit të tekstit, hapim filen `text_analysis.py` për të ekzekutuar programin.

```

Numri i fjaleve: 84450
Numri i karaktereve: 466772
Numri i fjaltive: 7169

Fjala me e shpeshte eshte: e -> 4570 here

Shkronjat e perseritura:
a -> 25705
b -> 3838
c -> 1034
d -> 101
e -> 8998
dh -> 2758
e -> 49926
o -> 13062
f -> 3270
g -> 3194
gj -> 2232
h -> 6691
i -> 27665
j -> 11826
k -> 11681
l -> 5326
ll -> 1803
m -> 11241
n -> 19520
nj -> 2243
o -> 10928
p -> 10169
q -> 3576
r -> 22108
w -> 1836
s -> 11327
sh -> 9211
t -> 28946
th -> 2535
u -> 14692
v -> 4381
x -> 214
xh -> 670
w -> 61
y -> 3456
z -> 1883
zh -> 307
. -> 6546
, -> 6343
? -> 463
! -> 159

```

*Foto 2. Hapja e files për ekzekutim dhe paraqitja e rezultateve.*

```

v -> 4381
x -> 214
xh -> 670
e -> 61
y -> 3456
z -> 1883
zh -> 307
. -> 6546
, -> 6343
? -> 463
! -> 159

Shkronja me e shpeshte eshte: e -> 49926 here.

Perseritja e zanoreve:
a -> 25705
e -> 49926
s -> 13062
i -> 27665
o -> 10928
u -> 14692
y -> 3456

Perseritja e bigrameve:
dh -> 2758
gj -> 2232
ll -> 1803
nj -> 2243
rr -> 1836
sh -> 9211
th -> 2535
xh -> 670
zh -> 307

```

Foto 3. Hapja e files për ekzekutim dhe paraqitja e rezultateve.

Nga foto 2 dhe 3 kemi paraqitjen totale të frekuencës së paraqitjes së fjalëve, fjalive, shkronjave etj. Duhet marrur parasysh mënyrën se si janë llogaritur bigramet në tekst. Së pari është marrë një varg i të gjitha bigrameve në gjuhën shqipe. Pastaj në momentin që janë numëruar shkronjat, kemi pyetur për secilin element të vargut se mos gjinden në tekst. Në rast se gjinden, atëherë numëron frekuencën e paraqitjeve në tekst dhe afishon rezultatin. Por nëse ka bigrame, atëherë programi do t'i llogaritë ato si shkronja të vetme. Si p.sh nëse mirret teksti: **Shkova në punë**, në paraqitjen e rezultateve Sh paraqitet 1 herë kurse në llogaritje mirret edhe S me H. Për eliminimin e këtij problemi llogaritet se sa herë paraqitet ai bigram, dhe në shumën totale të paraqitjes së shkronjës S dhe H zbritet totali i paraqitjes së atij bigrami.

```

Perseritja e bigrameve:
dh -> 0
gj -> 0
ll -> 1
nj -> 0
rr -> 0
sh -> 2
th -> 0
xh -> 0
zh -> 0

```

Foto 4. Përsëritja e bigrameve.

```

h -> 0
i -> 0
j -> 0
k -> 2
l -> 0
ll -> 1
m -> 0
n -> 1
nj -> 0
o -> 2
p -> 0
q -> 0
r -> 0
rr -> 0
s -> 0

```

Foto 5. Përsëritja e shkronjave.

Përveq paraqitjes së frekuencës së fjalëve dhe shkronjave, paraqiten edhe grafikë të ndryshëm për vizualizim të analizimit tekstit.

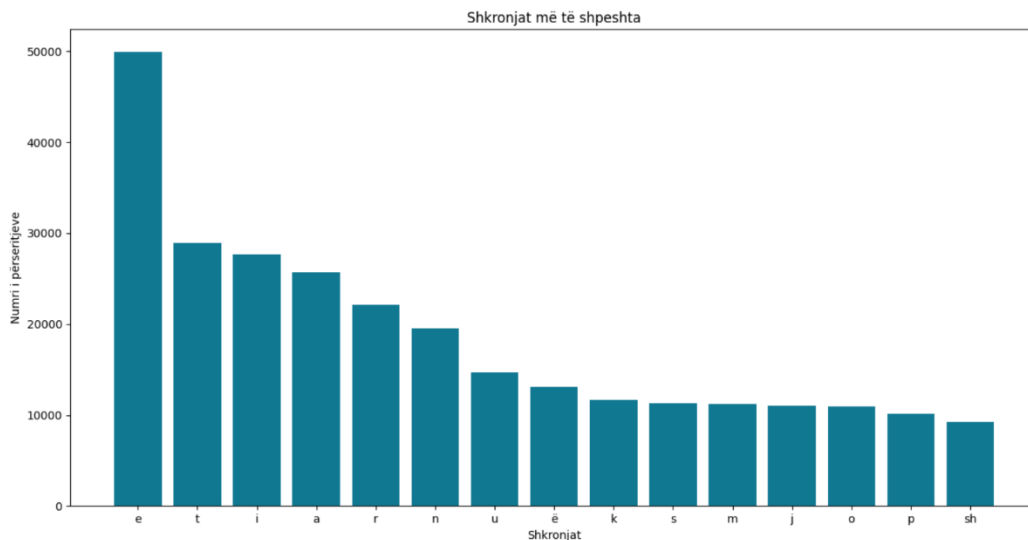


Foto 6. Grafiku për shfaqjen e frekuencës së paraqitjes së shkronjave më të shpeshta.

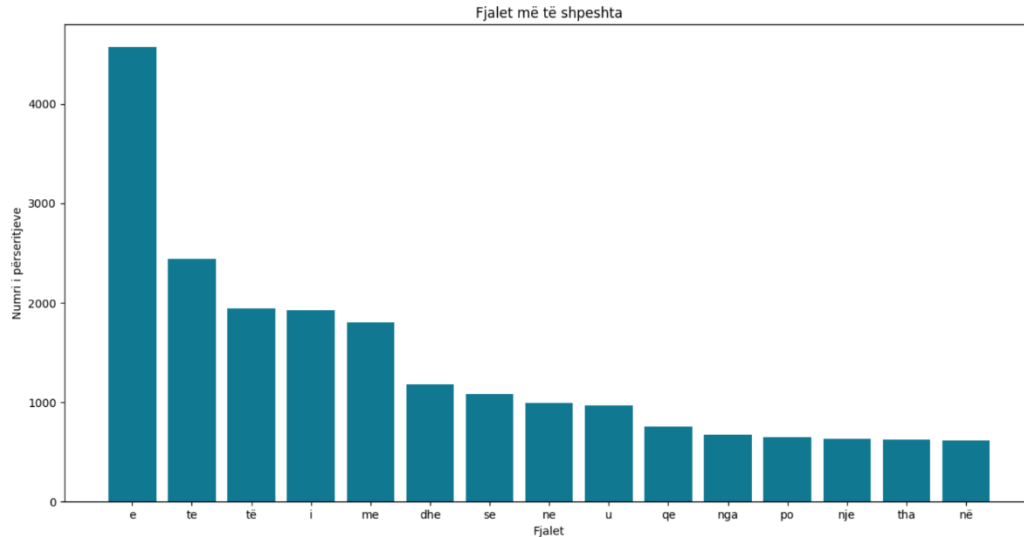


Foto 7. Grafiku për shfaqjen e frekuencës së paraqitjes së fjalëve më të shpeshta.

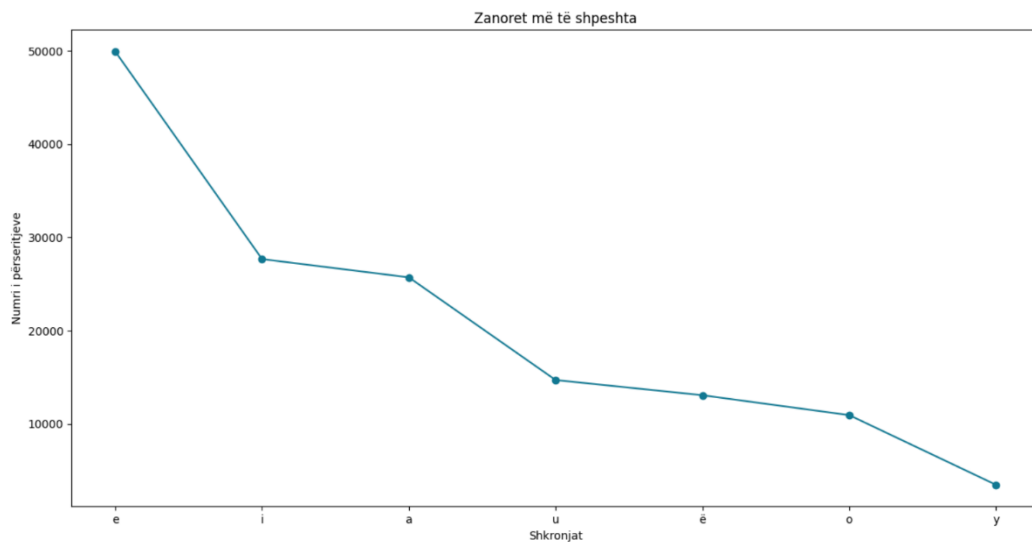


Foto 8. Grafiku për shfaqjen e frekuencës së paraqitjes së zanoreve.

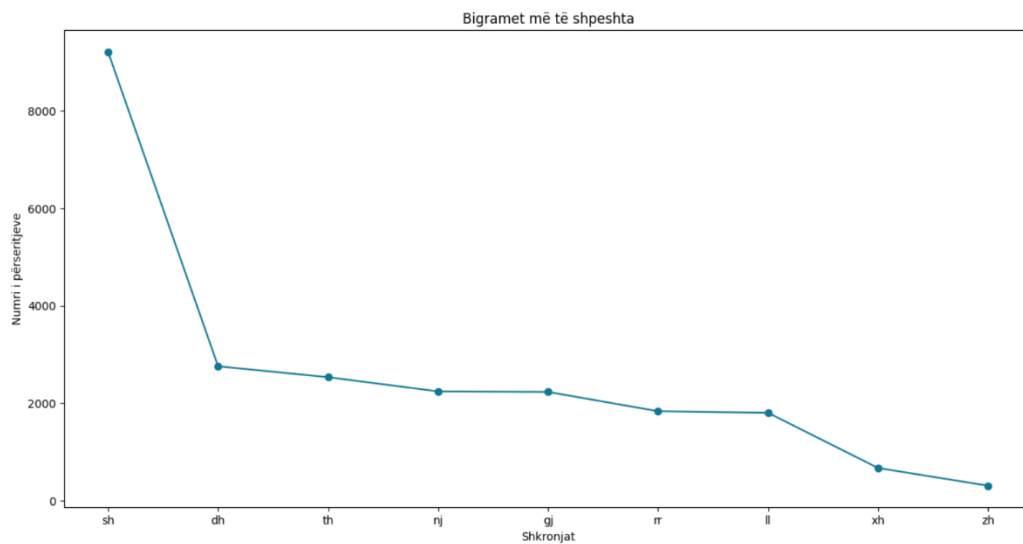


Foto 9. Grafiku për shfaqjen e frekuencës së paraqitjes së bigrameve.

Në foton 6 kemi paraqitjen e 15 shkronjave më të shpeshta me rënditje nga më e përsëritura deri tek më e rralla. Në foton 7 kemi të njëjtin funksion por kemi paraqitjen e 15 fjalëve më të shpeshta. Tek foto 8 dhe 9 kemi frekuencën e paraqitjeve të zanoreve dhe bigrameve. Nga këto rezultate, shohim se shkronja ‘e’ është shkronja më e përdorur në alfabetin e gjuhës shqipe.

```

output - Notepad
File Edit Format View Help
Numri i fjaleve: 87074
Numri i karaktereve: 469702
Numri i fjalive: 7169

Fjala me e shpeshte eshte: e -> 4570 here.

Shkronjat e perseritura:
a -> 25705
b -> 3838
c -> 1834
d -> 181
dh -> 8998
dh -> 2758
e -> 49926
ë -> 13062
f -> 3270
g -> 3194
gj -> 2232
h -> 6691
i -> 27665
j -> 11026
k -> 11681
l -> 5326
ll -> 1803
m -> 11241
n -> 19520
nj -> 2243
o -> 10928
p -> 10169
q -> 3576
r -> 22108
rr -> 1836
s -> 11327
sh -> 9211
t -> 28946
th -> 2535
u -> 14692
v -> 4381
x -> 214
xh -> 670
w -> 61
y -> 3456
z -> 1883
zh -> 307
. -> 6546
, -> 6343
? -> 463
! -> 159

```

*Foto 10. Paraqitja e frekuencës së shkronjave dhe fjalëve*

```

output - Notepad
File Edit Format View Help
m -> 11241
n -> 19520
nj -> 2243
o -> 10928
p -> 10169
q -> 3576
r -> 22108
rr -> 1836
s -> 11327
sh -> 9211
t -> 28946
th -> 2535
u -> 14692
v -> 4381
x -> 214
xh -> 670
w -> 61
y -> 3456
z -> 1883
zh -> 307
. -> 6546
, -> 6343
? -> 463
! -> 159

Shkronja me e shpeshte eshte: e -> 49926 here.

Perseritja e zanoreve:
a -> 25705
e -> 49926
ë -> 13062
i -> 27665
o -> 10928
u -> 14692
y -> 3456

Perseritja e bigrameve:
dh -> 2758
gj -> 2232
ll -> 1803
nj -> 2243
rr -> 1836
sh -> 9211
th -> 2535
xh -> 670
zh -> 307

```

*Foto 11. Paraqitja e frekuencës së shkronjave dhe fjalëve*

Si përfundim të procesit të analizës së tekstit, në foton 10 dhe 11 kemi output filen në formatin .txt për të gjithë frekuencën e paraqitjeve të shkronave, fjalëve etj.

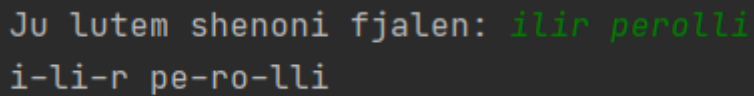
## **Ndarja e tekstit në rrokje**

Një shembull tjetër të analizimit të tekstit kemi edhe rrokëzimin. Në këtë shembull kemi marrë tekstin si input. E dimë që rrokjet gjinden pas zanoreve. Për këtë arsye, kemi marrur një varg të zanoreve:

zanoret = ['a','e','ë','i', 'o', 'u','y']

Pas inicializimit të vargut i jemi qasur çdo shkronje të inputit. Në rast se ekziston zanorja atëherë shfaqim në output një vizë (-).

Shembull:



```
Ju lutem shenoni fjalen: ilir perolli
i-li-r pe-ro-lli
```

I gjithë programi i realizuar për analizimin e tekstit mund të ketë çasje në linkun:  
<https://github.com/IlirPerolli/Albanian-letters-count>



## **Përfundimi**

Këshilla e sotme e shkrimit ka të bëjë me rëndësinë e numërimit të fjalëve. Numërimi i mërzitshëm i fjalëve. Si mund të ketë rëndësi? Shumica e njerëzve as nuk e dinë se çfarë është, dhe ata që e dinë, mendojnë se është thjesht një punë që duhet lënë mënjane. Nuk ka rëndësi.

Apo mos ndoshta ka rëndësi?

Le ta heqim atë pyetje. Po, numri i fjalëve ka rëndësi. Është një punë që duhet lënë mënjane, por kjo nuk do të thotë se nuk është e rëndësishme. Çdo shkrimtar në një moment apo në një tjetër ka luftuar me numërimin e fjalëve. Pavarësisht nëse po i përmbahet një gjatësie të kërkuar ose po shkruan romanin tënd personal, numërimi i fjalëve ka qenë një qëllim i pakapshëm.

Në këtë hulumtim kemi gjeneruar një frekuencë të përsëritjes së shkronjave, fjalëve, fjalive dhe gjenerimin e rezultateve në grafike.

Gjithashtu nga ky punim kemi parë që gjuha programore Python plotëson plotësisht kërkesat për analizimin e tekstit dhe që është shumë e lehtë për modifikim. Gjithashtu kodi burimor është i hapur nga të gjithë për kontribute.

## **Bibliografia**

[1] Language definition and meaning, Collins English Dictionary

Available: [www.collinsdictionary.com](http://www.collinsdictionary.com).

[Accessed 29 March 2021].

[2] Text Analytics for Beginners using NLTK, Datacamp.

Available: <https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk>

[Accessed 28 March 2021].

[3] The Science Fiction and Fantasy Writers of America suggest 6 chars to a word

Available: <https://www.sfwaw.org/2005/01/04/what-is-a-word/>

[Accessed 29 March 2021].