Perspectives on Data Access And Use at Scale: Lessons from the Field

Topic Modeling on the Kadi Registers

Dr. Sümeyye Akça
Assistant Professor
Marmara Univresity
Information and Records Management
Turkey/Istanbul
sumeyyesakca@gmail.com
sumeyye.akca@marmara.edu.tr

Kadi Registers?

- One of the most important sources
- Turkish, Arabic, and Persian
- Maintained in book format
- 15th - 20th centuries
- Judgement for every unique event
- Transcription??

Challenges

- 32 letters Persian and Turkish,
- Writing order is from left to right,
- Vowel point!
- Difficult to distinguish words!
- Separation of lines!
- Horizontal plane!
- Changing writing styles!

Problem?

- Need for authority
- Develop community belonging
- Understand present problems!!

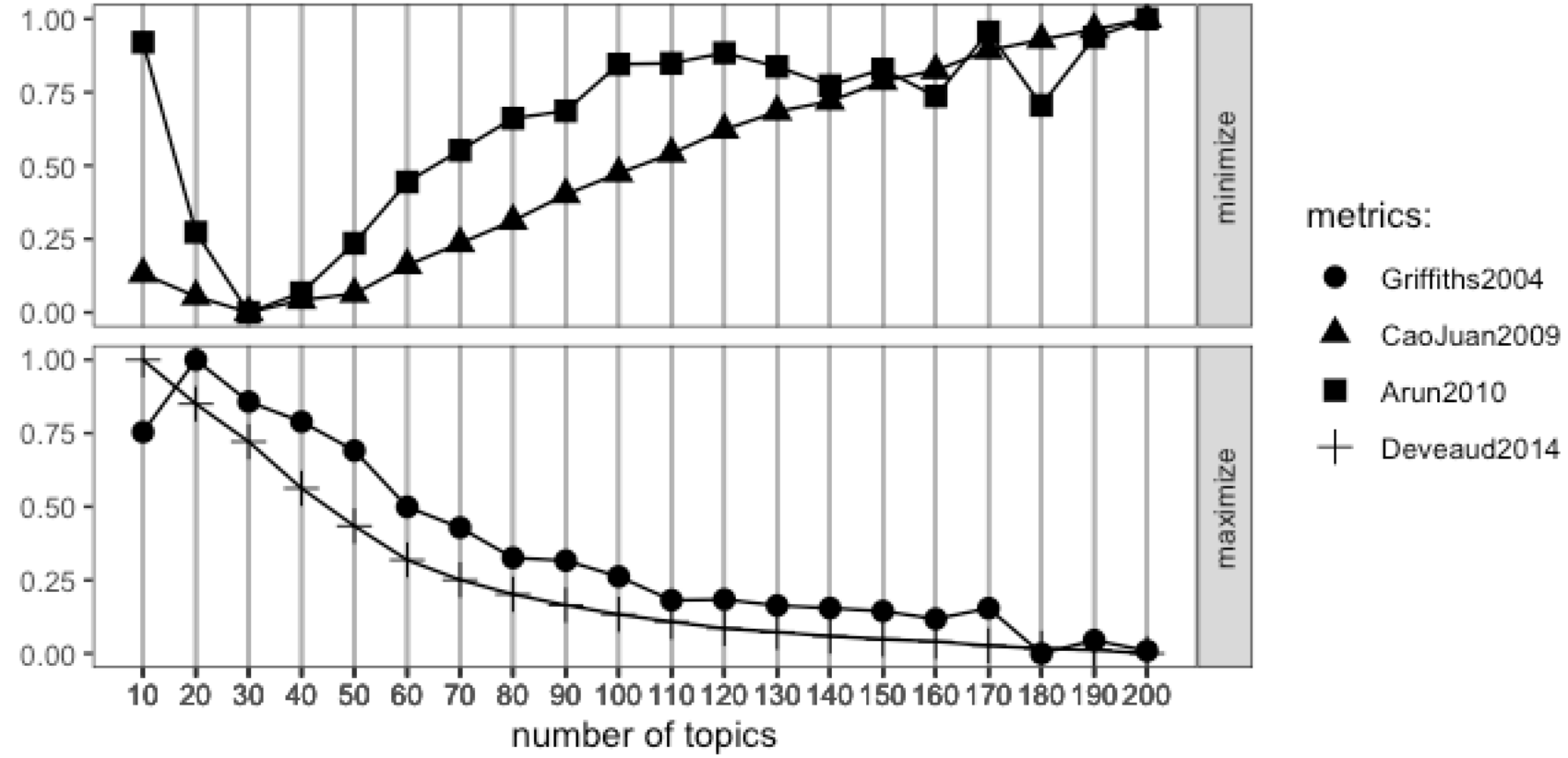
Latent Dirichlet Allocation

- We need algorithmic tools!
- Probability distribution and Topic Modeling!
- LDA algorithm!

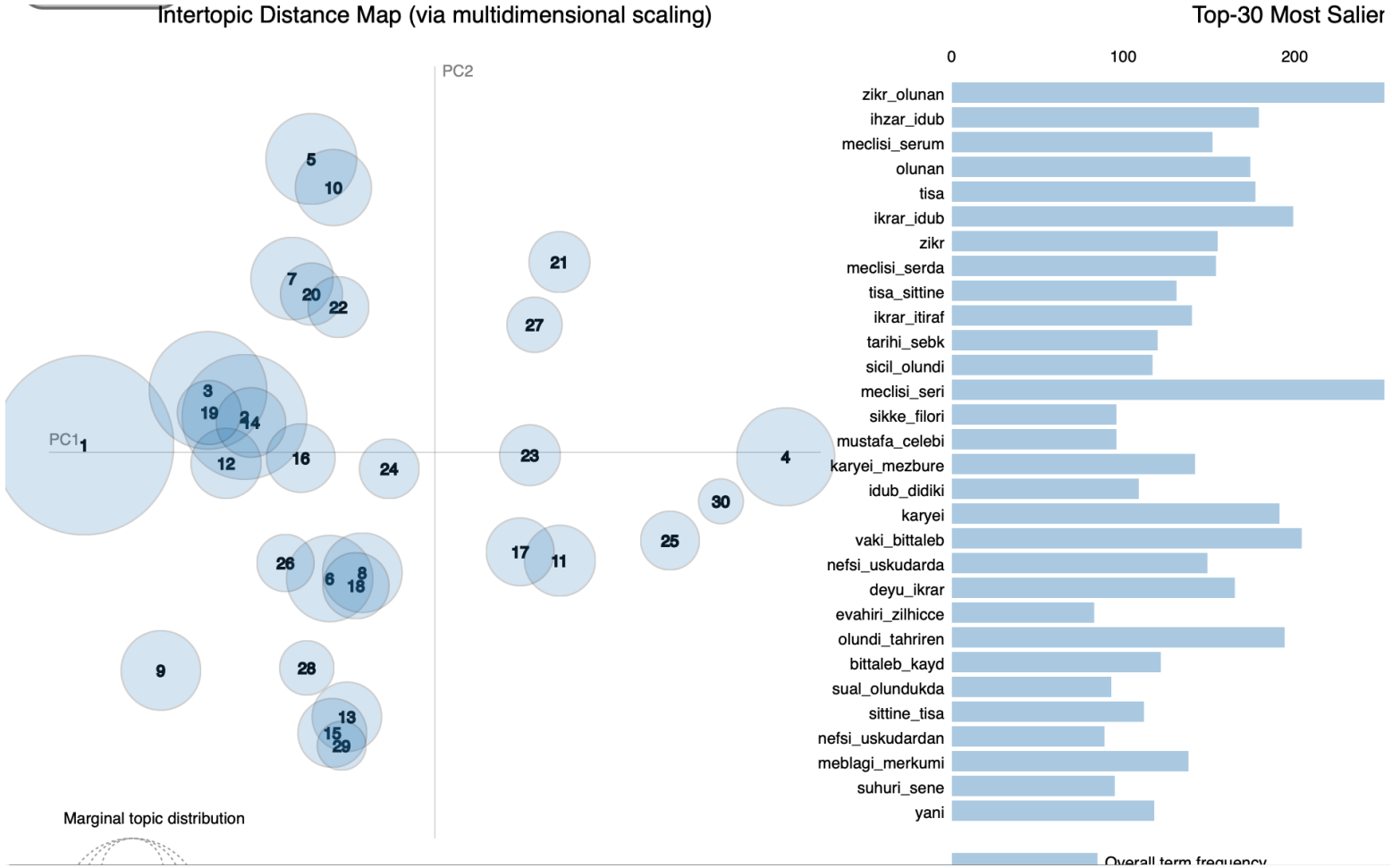
The Dataset

- 23. Kadi Register book
- Consists of 648 single items (problem or case)
- I had transcribed it as my master's thesis
- To determine the topics of the each case =>

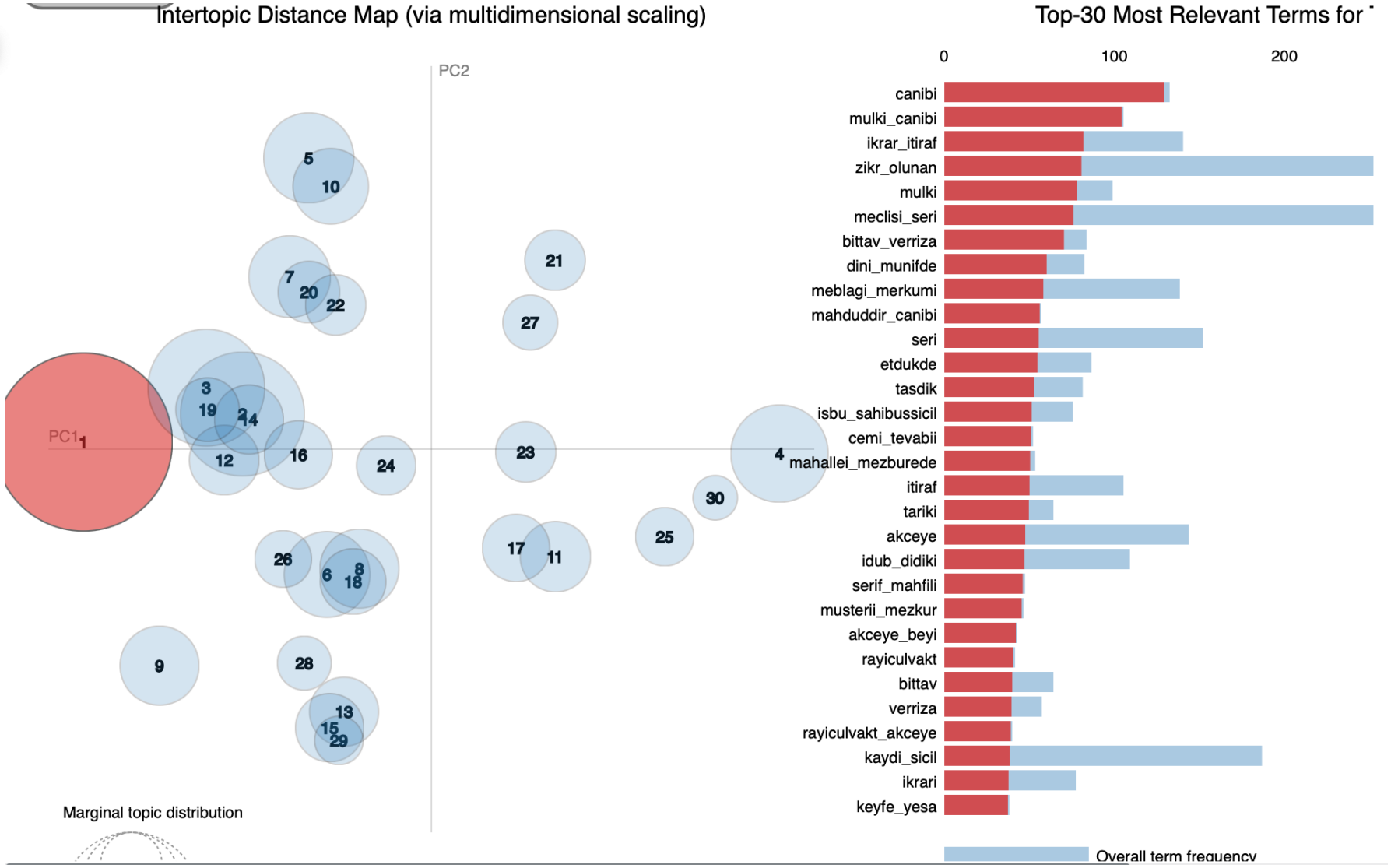
Select # of The Topics



Results

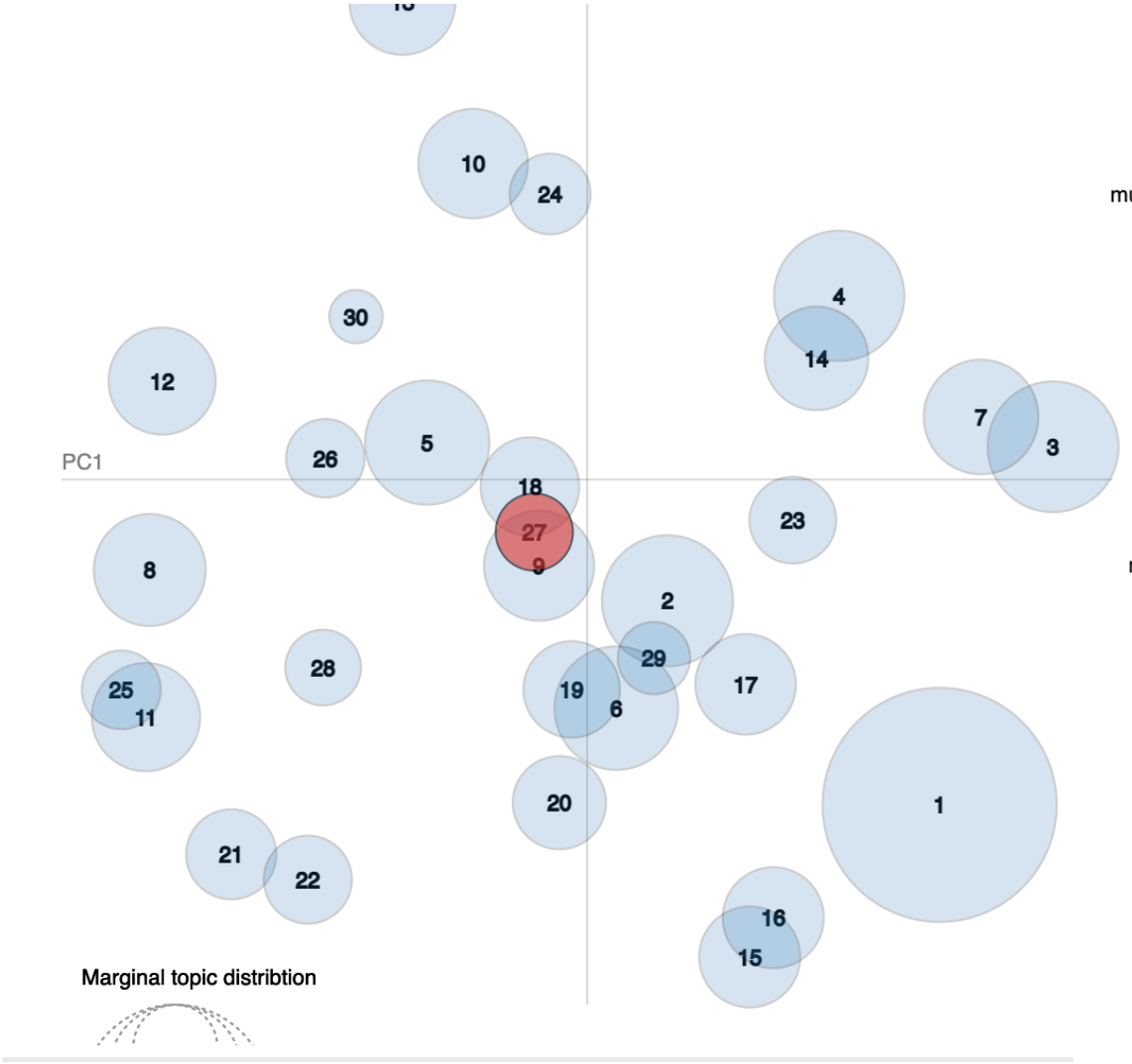

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Marginal topic distribution
Top-30 Most Salier
0
100
200
zikr_olunan
ihzar_idub
meclisi_serum
olunan
tisa
ikrar_idub
zikr
meclisi_serda
tisa_sittine
ikrar_itiraf
tarihi_sebk
sicil_olundi
meclisi_seri
sikke_filori
mustafa_celebi
karyei_mezbure
idub_didiki
karyei
vaki_bittaleb
nefsi_uskudarda
deyu_ikrar
evahiri_zilhicce
olundi_tahriren
bittaleb_kayd
sual_olundukda
sittine_tisa
nefsi_uskudardan
meblagi_merkumi
suhuri_sene
yani
Overall term frequency

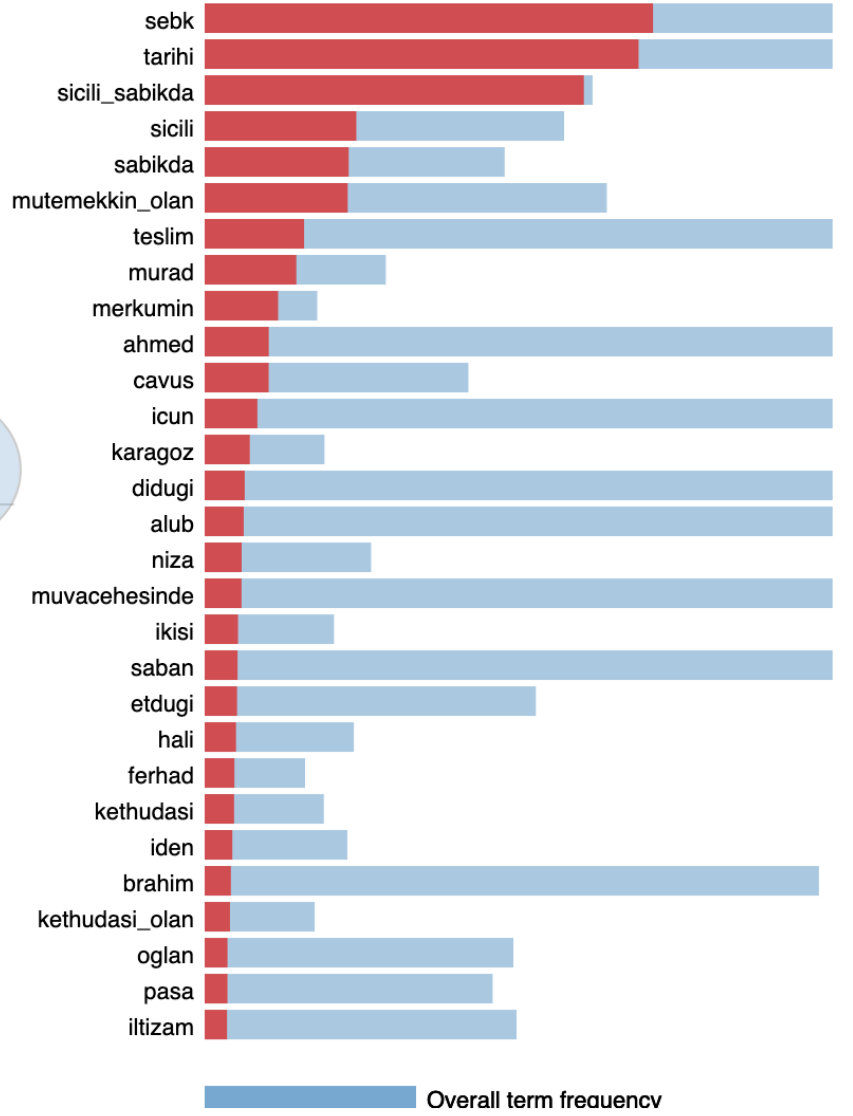
[12]:

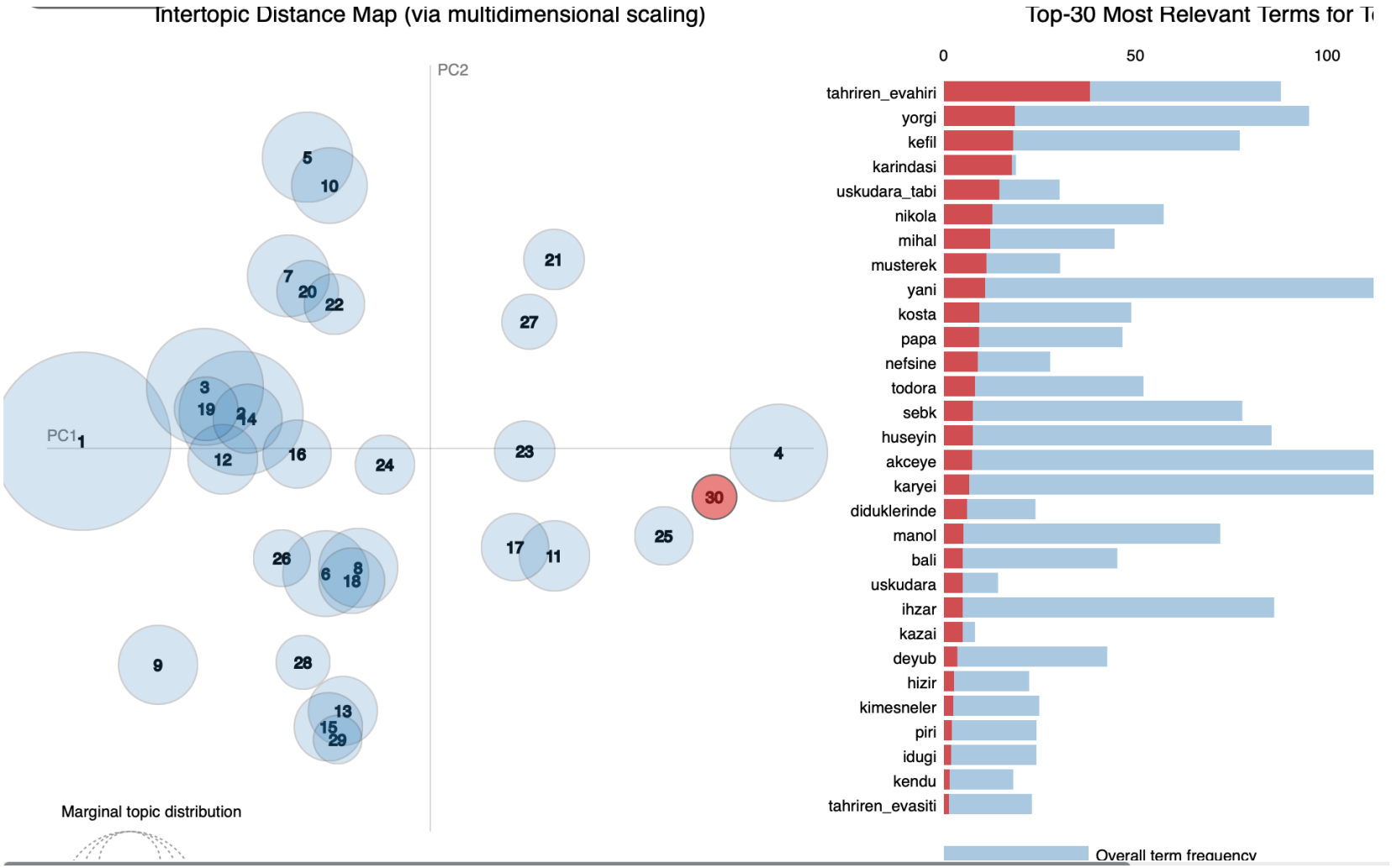

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Marginal topic distribution
Top-30 Most Relevant Terms for
0
100
200
canibi
mulki_canibi
ikrar_itiraf
zikr_olunan
mulki
meclisi_seri
bittav_verriza
dini_munifde
meblagi_merkumi
mahduddir_canibi
seri
etdukde
tasdik
isbu_sahibussicil
cemi_tevabii
mahallei_mezburede
itiraf
tariki
akceye
idub_didiki
serif_mahfili
musterii_mezkur
akceye_beyi
rayiculvakt
bittav
verriza
rayiculvakt_akceye
kaydi_sicil
ikrari
keyfe_yesa
Overall term frequency

Most Used Words..

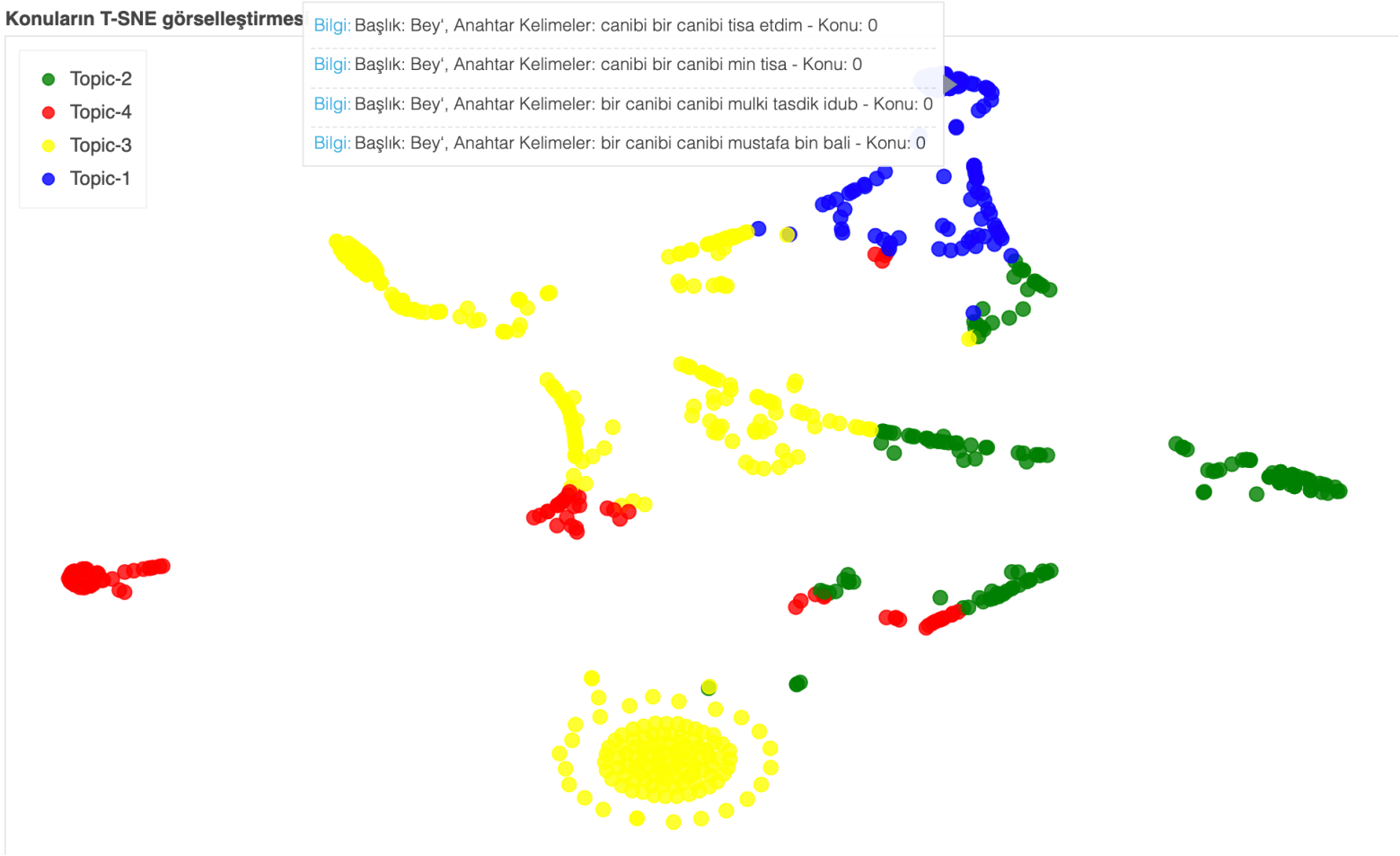
- Vechi tahriri huruf oldur ki
- Vechi tahriri kaziyye budur ki
- Ve tarihi ma sebk
- Ve tarihi ma zikru filevvel
- Binti-Bt
- Bin
- Suhud Ma sebk
- Tahriren fi


PC1
10
24
4
14
30
12
26
5
18
27
9
7
3
23
8
2
29
28
25
11
19
6
17
1
20
21
22
16
15
Marginal topic distribtion

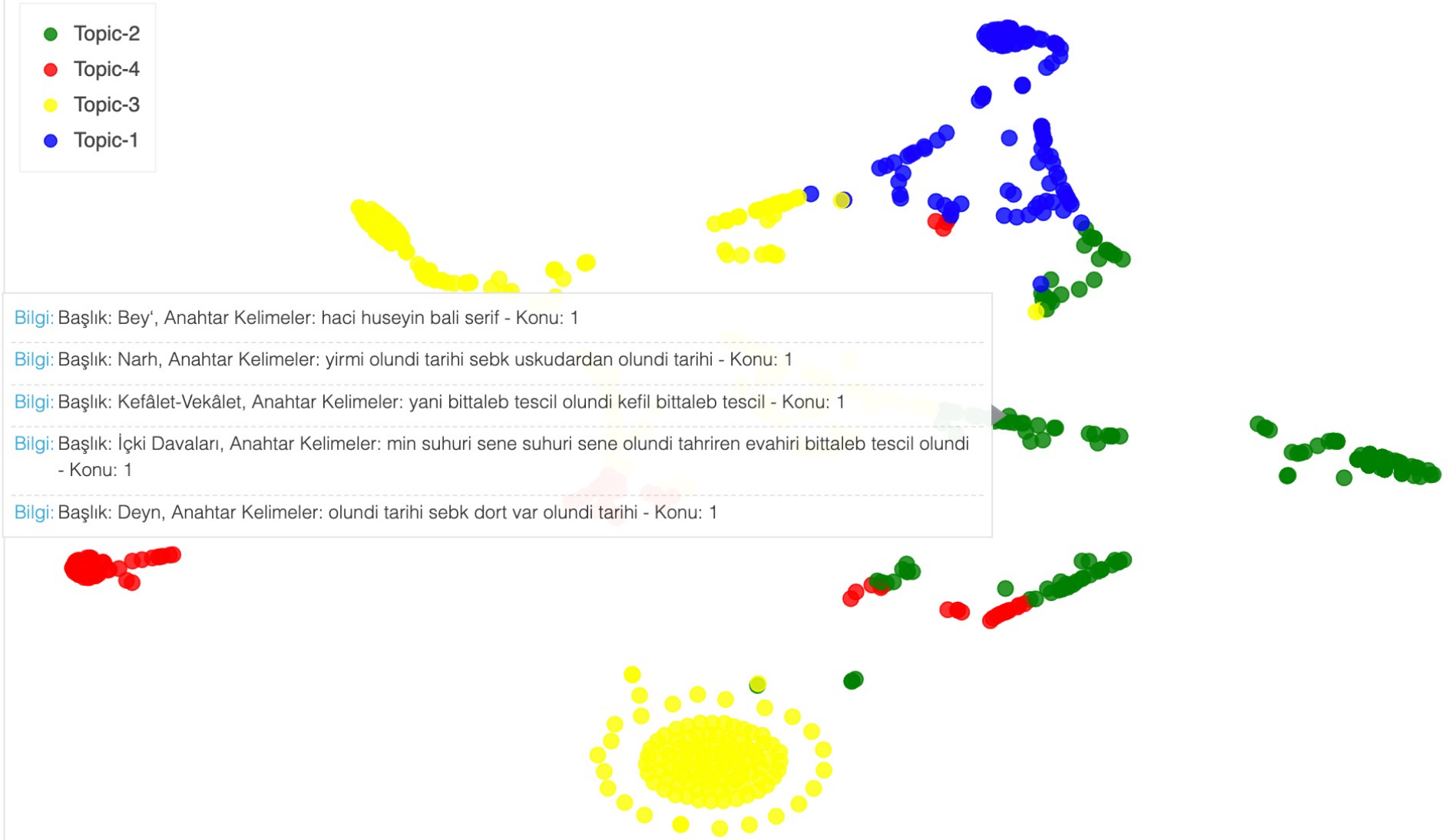

sebk
tarihi
sicili_sabikda
sicili
sabikda
mutemekkin_olan
teslim
murad
merkumin
ahmed
cavus
icun
karagoz
didugi
alub
niza
muvacehesinde
ikisi
saban
etdugi
hali
ferhad
kethudasi
iden
brahim
kethudasi_olan
oglan
pasa
iltizam
Overall term frequency

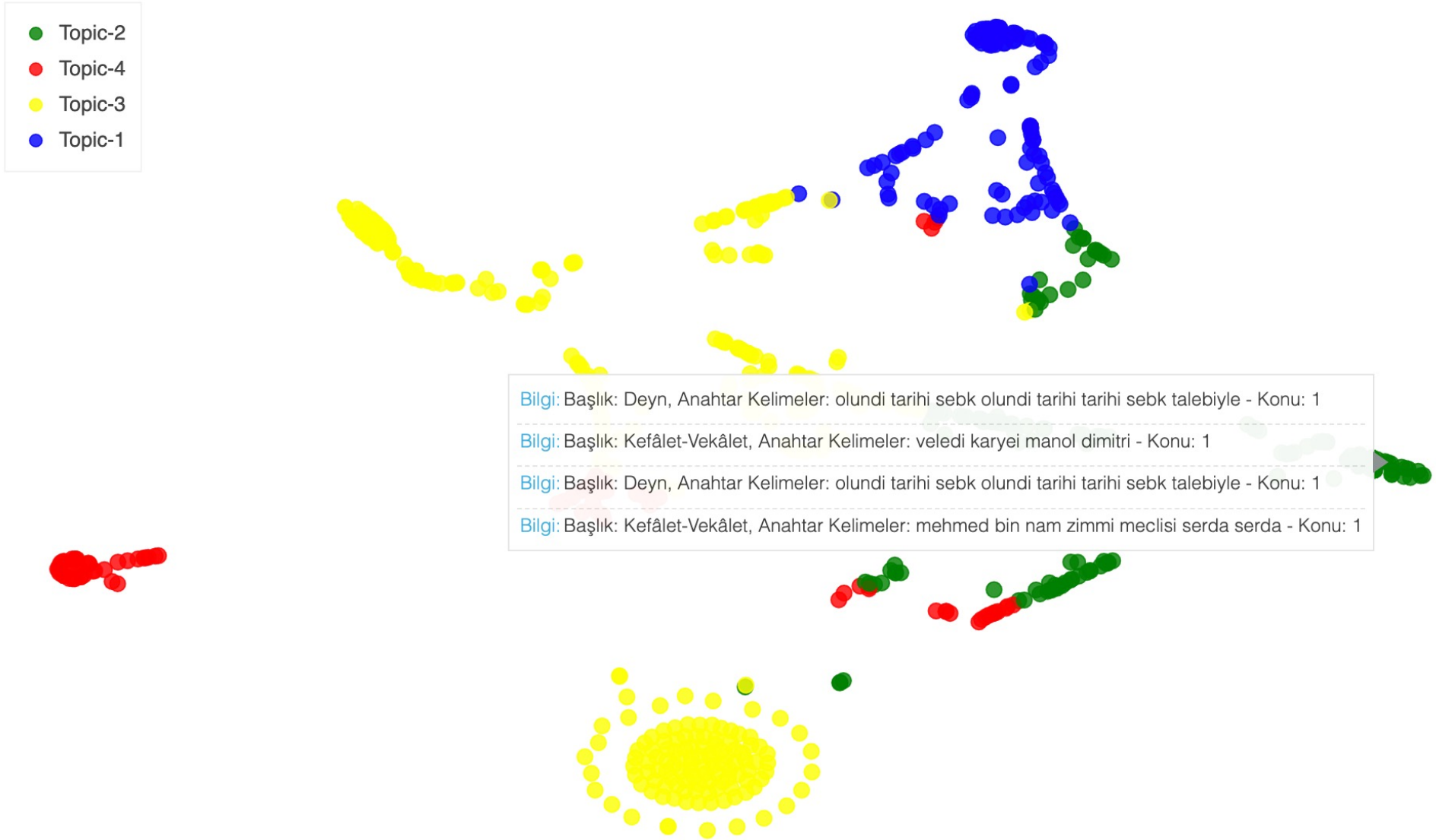

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
10
7
20
22
21
27
3
19
2
14
1
12
16
24
23
4
30
25
17
11
26
6
8
18
9
28
13
15
29
Marginal topic distribution
Top-30 Most Relevant Terms for T
0
50
100
tahriren_evahiri
yorgi
kefil
karindasi
uskudara_tabi
nikola
mihal
musterek
yani
kosta
papa
nefsine
todora
sebk
huseyin
akceye
karyei
diduklerinde
manol
bali
uskudara
ihzar
kazai
deyub
hizir
kimesneler
piri
idugi
kendu
tahriren_evasiti
Overall term frequency

Dimension Reduction



Konuların T-SNE görselleştirmesi




Topic-2
Topic-4
Topic-3
Topic-1
Bilgi: Başlık: Deyn, Anahtar Kelimeler: olundi tarihi sebk olundi tarihi tarihi sebk talebiyle - Konu: 1
Bilgi: Başlık: Kefâlet-Vekâlet, Anahtar Kelimeler: veledi karyei manol dimitri - Konu: 1
Bilgi: Başlık: Deyn, Anahtar Kelimeler: olundi tarihi sebk olundi tarihi tarihi sebk talebiyle - Konu: 1
Bilgi: Başlık: Kefâlet-Vekâlet, Anahtar Kelimeler: mehmed bin nam zimmi meclisi serda serda - Konu: 1

Conclusion

- With this model, it has been seen more clearly which topics in the Kadi Registers are included with which words.
- This facilitates the identification of documents and automatic classification.
- In the transcription of documents, each case is assigned to a single subject.
- However, in the LDA model, it is possible to have different topics in each case at the same time.
- At the point of access to these documents, the user is offered a wider choice of topics.
- the LDA algorithm is based on the bag of words approach!! (Limit)

Perspectives on Data Access And Use at Scale: Lessons from the Field

Topic Modeling on the Kadi Registers

Dr. Sümeyye Akça
Assistant Professor
Marmara Univresity
Information and Records Management
Turkey/Istanbul
sumeyyesakca@gmail.com
sumeyye.akca@marmara.edu.tr