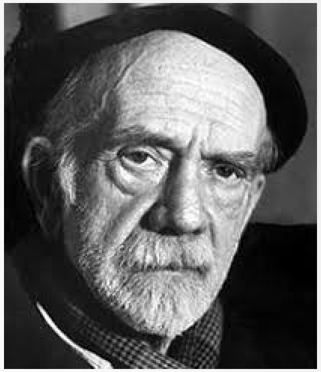
Galdós

Bazán

Ibañez

Baroja



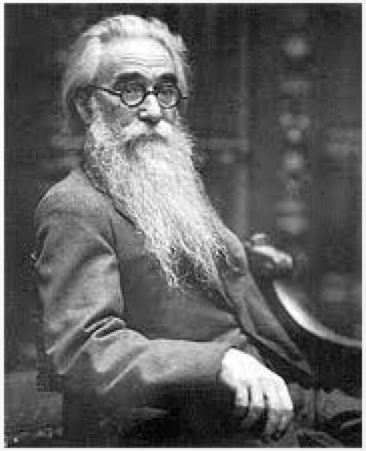
The Librarian “Amigo”

Valera

Clarín

Unamuno

Valle-Inclán



The librarian hopes to

- Identify author

?

The librarian hopes to

- Identify author

?

- Identify main topics

  (clustering)

Resources

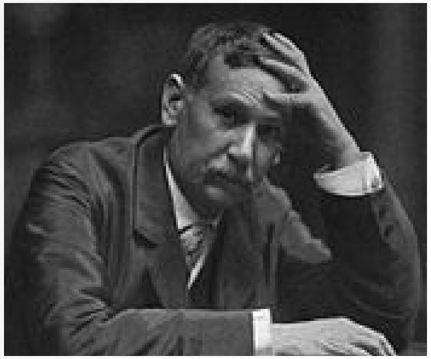
- Public Domain spanish literature
  - gutenberg.org: text files
  - ataun.net: pdf files

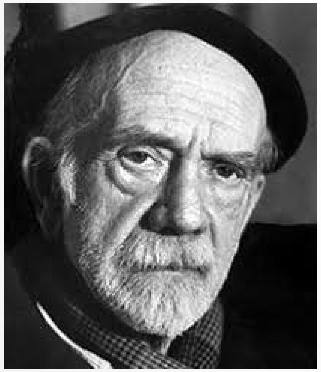
Resources

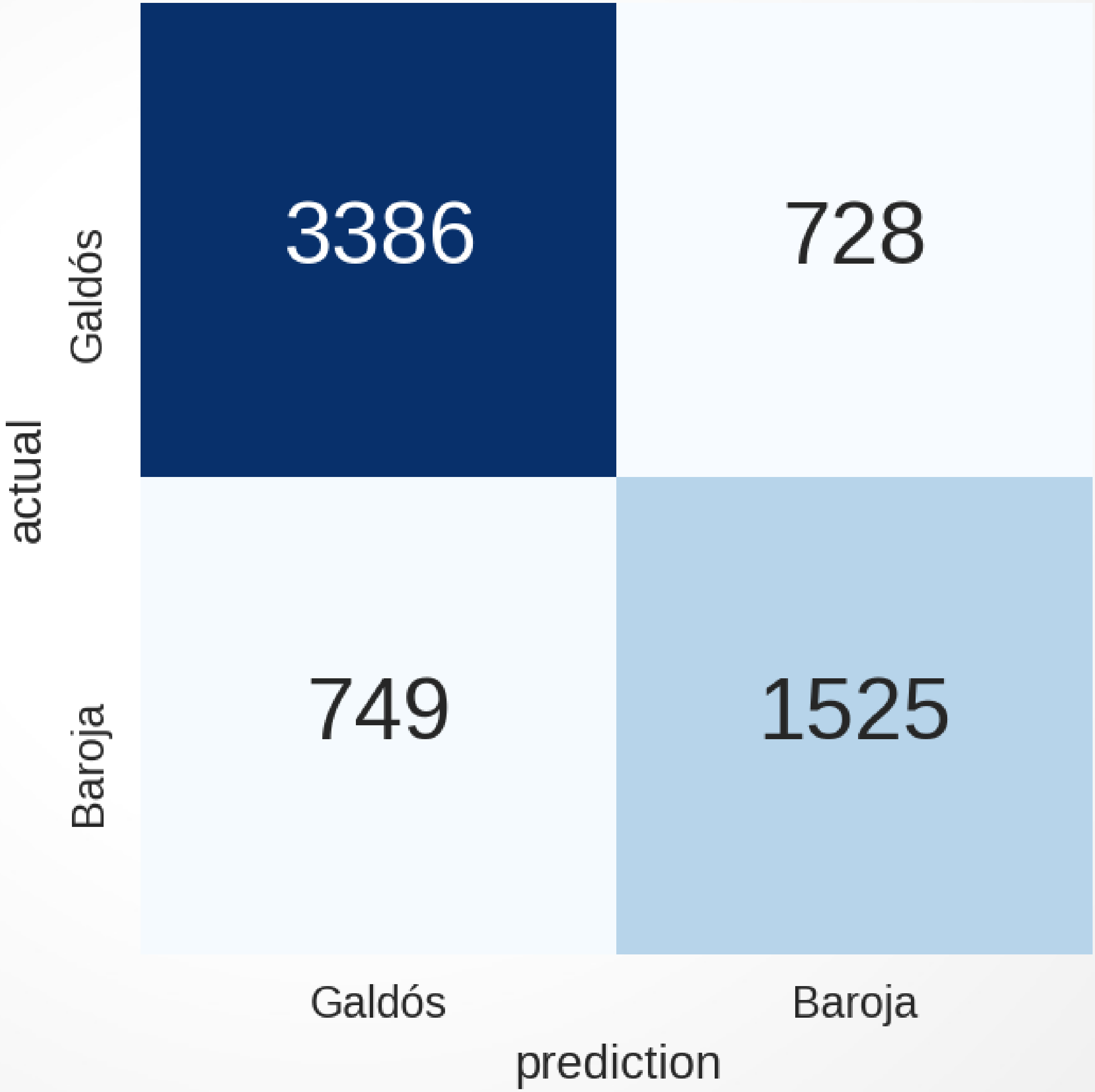
- Public Domain spanish literature
  - gutenberg.org: text files
  - ataun.net: pdf files
- 214 Books ~ 45000 fragments

Identify author (Naive Bayes(tfidf/LSA))

Acc. train: 77.2 %
Acc. test: 76.9 %

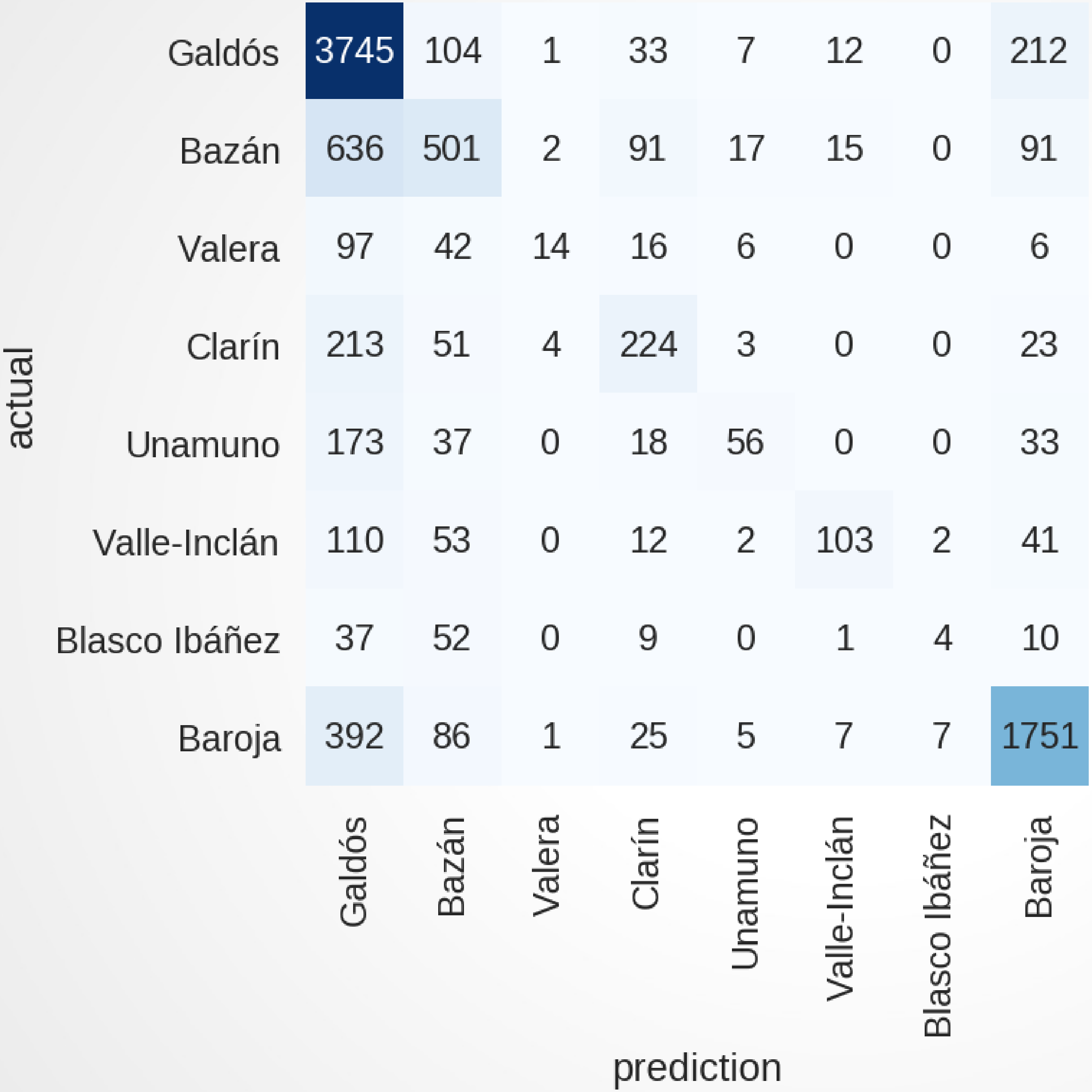






actual \ prediction	Galdós	Baroja
Galdós	3386	728
Baroja	749	1525

Identify author (SVC(tfidf/LSA))

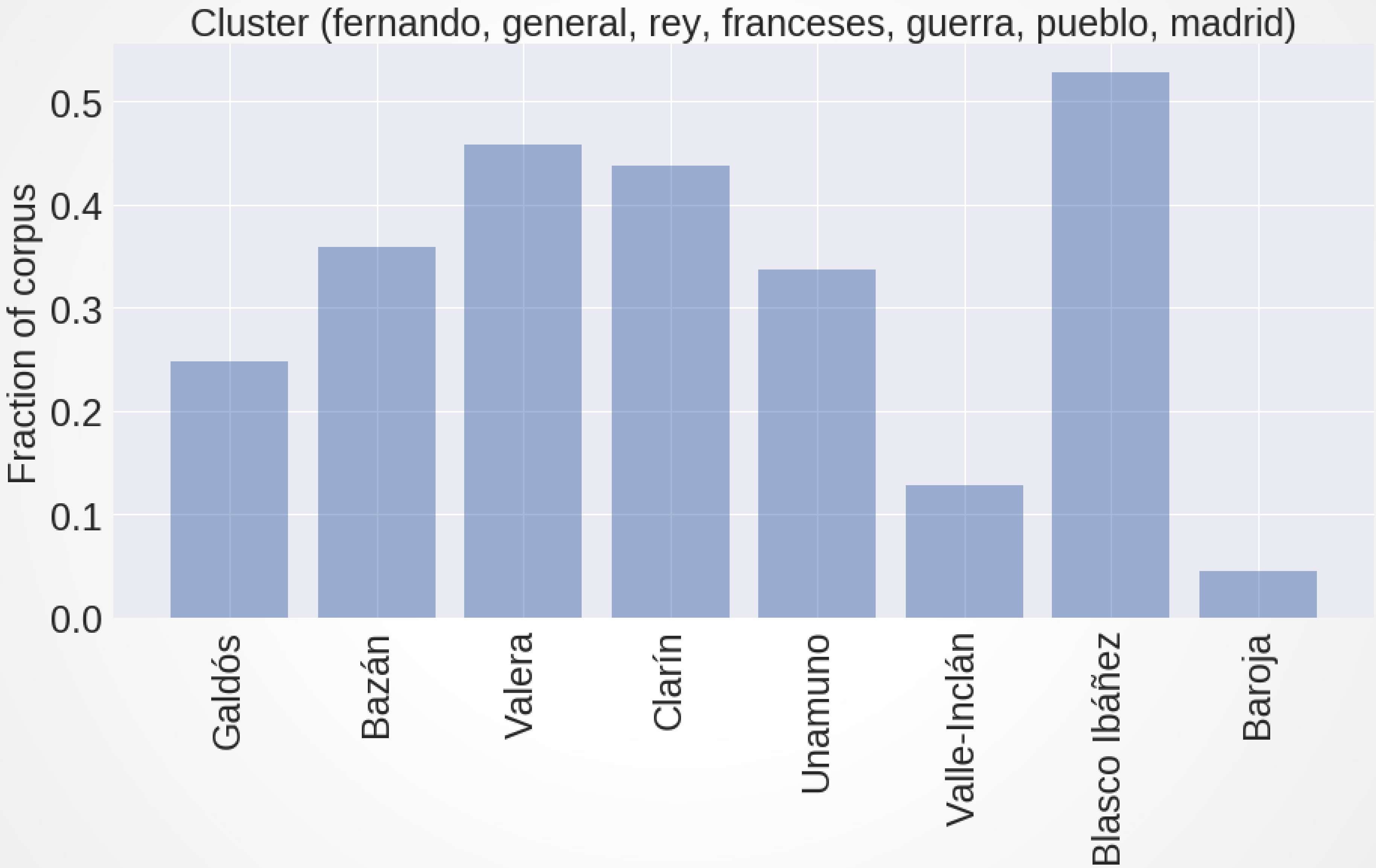


	Galdós	Bazán	Valera	Clarín	Unamuno	Valle-Inclán	Blasco Ibáñez	Baroja
Galdós	3745	104	1	33	7	12	0	212
Bazán	636	501	2	91	17	15	0	91
Valera	97	42	14	16	6	0	0	6
Clarín	213	51	4	224	3	0	0	23
Unamuno	173	37	0	18	56	0	0	33
Valle-Inclán	110	53	0	12	2	103	2	41
Blasco Ibáñez	37	52	0	9	0	1	4	10
Baroja	392	86	1	25	5	7	7	1751

actual
prediction

Acc. train: 70.4 %
Acc. test: 69.6 %

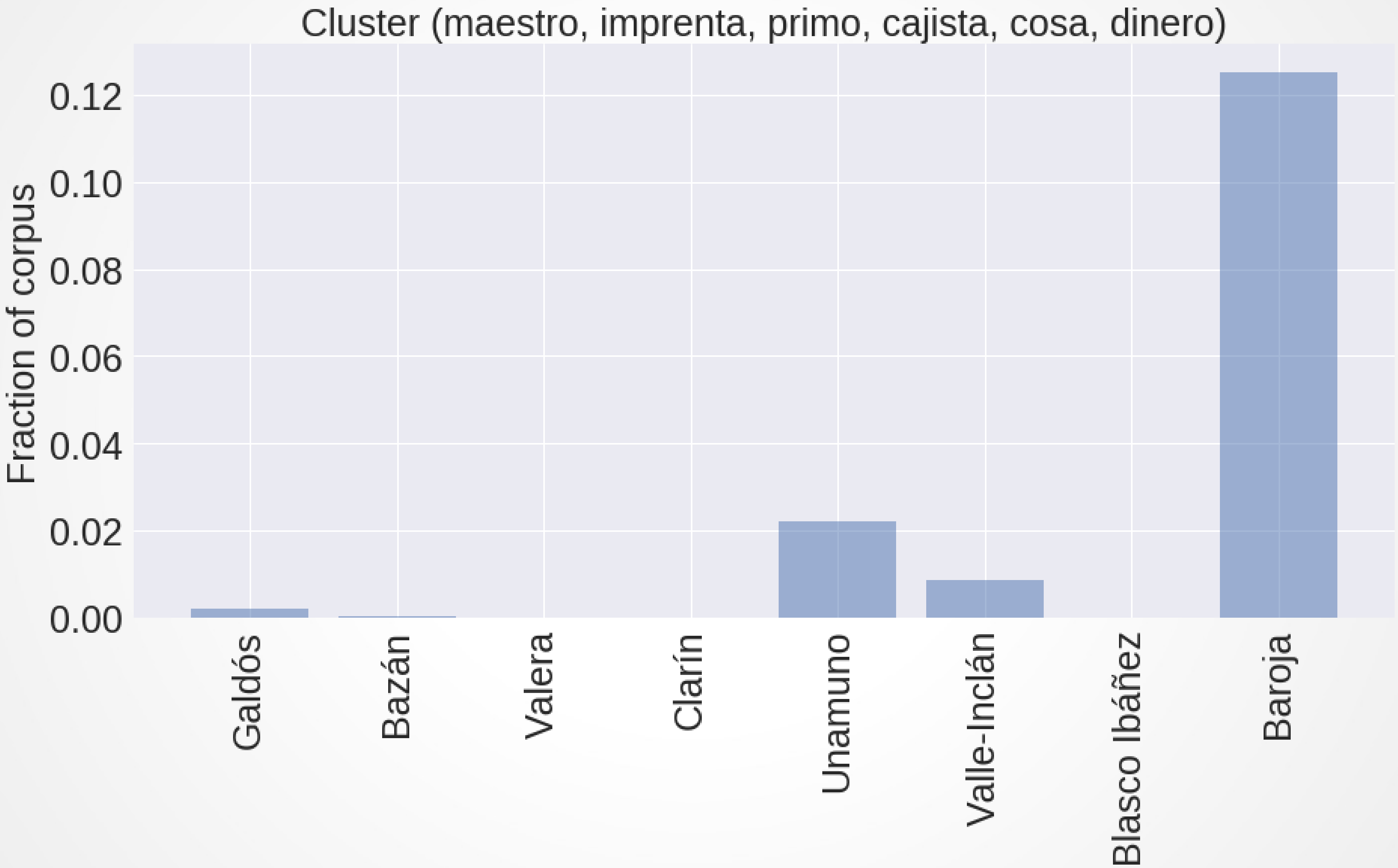
Main Topics (k-means(tfidf/LSA))



Main Topics (k-means(tfidf/LSA))

C
drid)
Fraction of corpus
0.5
0.4
0.3
0.2
0.1
0.0
Ur
Vall
Blasco
Baroja

Main Topics (k-means(tfidf/LSA))



Main Topics (k-means(tfidf/LSA))

Thank you for your attention!

Questions?

Word use for different authors

“Atacar” = to attack

Galdos

```
('villarreal', 0.9676589965820312),
 ('unirse', 0.9671500325202942),
 ('castaños,', 0.9619759321212769),
 ('dupont', 0.9614170789718628),
 ('andújar', 0.9606595039367676),
 ('tropas,', 0.9606088995933533),
 ('armar', 0.9605693817138672),
 ('guipúzcoa', 0.9598258137702942),
 ('castaños', 0.9596730470657349),
 ('iturralde', 0.9587283730506897)
```

Baroja

```
('lúzaro.', 0.9962713718414307),
 ('costumbre,', 0.9960745573043823),
 ('advertir', 0.9957879185676575),
 ('meter', 0.9956990480422974),
 ('gibraltar.', 0.9956982135772705),
 ('quedé', 0.9956833124160767),
 ('sótano.', 0.9956726431846619),
 ('dediqué', 0.9955557584762573),
 ('decidimos', 0.9954980611801147),
 ('siete,', 0.9954259395599365)
```