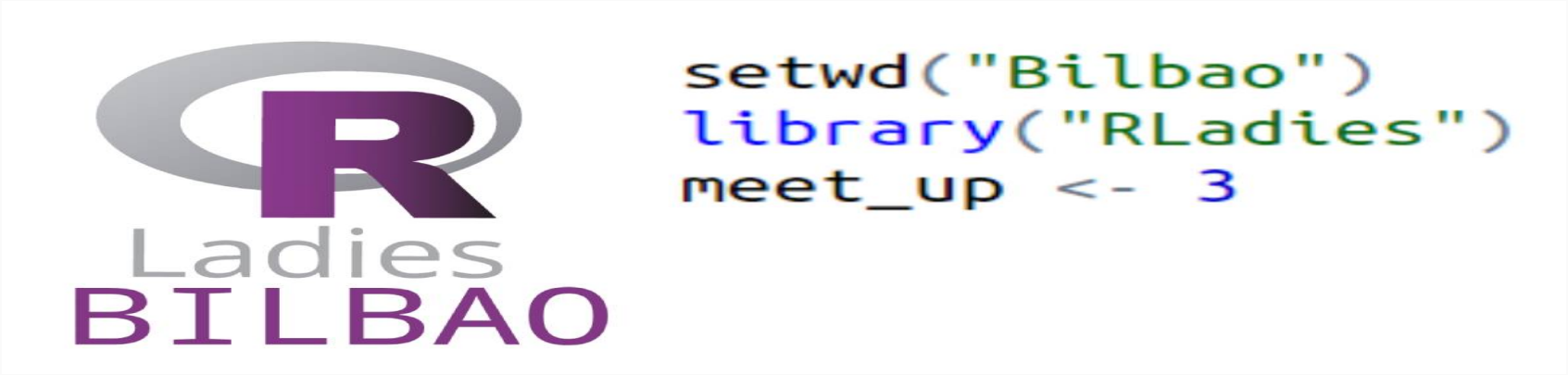


Testuak prozesatzen eta bisualizatzen R-rekin

Itziar Gonzalez-Dios

Bilbo, 2010/01/14

Aurkibidea

Nor naiz?

- Aleman Filologian lizentziatua (2010), Hizkuntzaren Azterketan eta Prozesamenduan masterra (2011) eta Hizkuntzalaritza Konputazionalean doktoretza (2016)
- Ixa taldeko ikertzaile 2010etik
- Bilboko Ingeniaritza Eskolako irakasle, Euskal Hizkuntza eta Komunikazioa sailekoa
- Ikerketa-lerro nagusiak: testuen konplexutasunaren azterketa, sinplifikazio automatikoa, lexikografia, semantika eta terminologia konputazionala
- @ItziarGD

Zer egingo dugu?

- Testuak prozesatzen (analizatzen) ikasi
- Testuen bisualizazioak egin
  - *barchart*
  - *wordcloud*
  - *wordnetwork*
- Hizkuntzalaritza konputazionaleko eta hizkuntzaren prozesamenduko kontzeptu garrantzitsu batzuk ikasi

Hasi baino lehen...

- Beharrezko softwarea
  - R `https://cran.r-project.org/`
  - R studio `https://rstudio.com/`
  - R bertsioa eguneratzeko `https://www.datatechnotes.com/2017/07/updating-r-in-rstudion-and-solving-slam.html`
- Paketeak instalatu eta kargatu!
- plyr, dplyr, readr, topicmodels, tidytext, ggplot2, tidyr, tm, wordcloud, udpipe, lattice, igraph, ggraph, textrank

Paketeak instalatu

```
install.packages("PAKETEA")
```

Paketeak kargatu

```
library("PAKETEA")
```

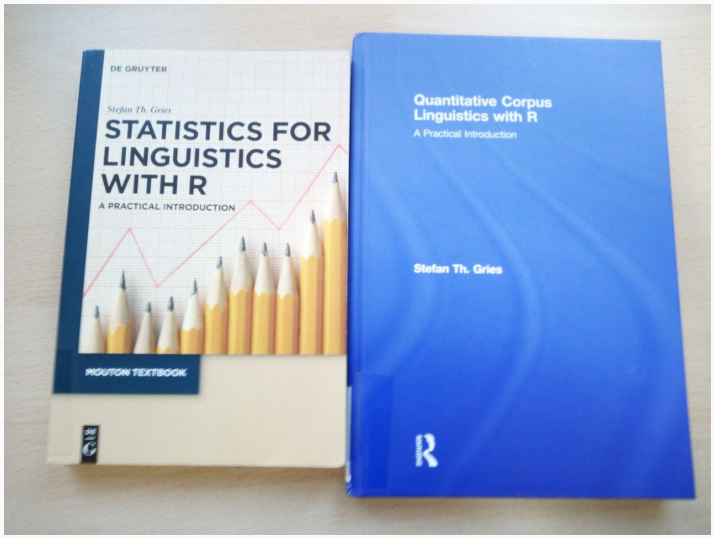
Interneten

- Wordcloudak egiteko hainbat orri
- Voyant tools tresna

- Baina
  - Ez daude hizkuntza batzuetarako prestatuta (gehianak ingeleserako)
  - Ez daukagu kontrola! *Gauzak egiten dituzte*

Literaturan

- Stefan Th. Gries-en lanak



- Baina
  - Ingeleserako!

Aurkibidea

Testuen analisi automatikoa

- Analizatzaile sintaktikoak edo parserrak
- *Universal Dependencies* edo dependentzia unibertsaletan oinarritutako udpipe parserra erabili
  `https://universaldependencies.org/`
- Eleaniztuna, euskarako egokia
  `https://zientziakaiera.eus/2019/06/28/maria-jesus-aranzabe-hizkuntza-teknologiak-ezinbesteko-tresna-dira-euskara`

Analizatzailearen irteera

	doc_id	paragraph_id	sentence_id	sentence	start	end	term_id	token_id	token	lemma	upos	xpos
1	doc1	1	1	Gaur euria ari du.	1	4	1	1	Gaur	gaur	ADV	<NA>
2	doc1	1	1	Gaur euria ari du.	6	10	2	2	euria	euri	NOUN	<NA>
3	doc1	1	1	Gaur euria ari du.	12	14	3	3	ari	ari	AUX	<NA>
4	doc1	1	1	Gaur euria ari du.	16	17	4	4	du	ukan	AUX	<NA>
5	doc1	1	1	Gaur euria ari du.	18	18	5	5	.	.	PUNCT	<NA>

	feats	head_token_id	dep_rel	deps	misc
1	<NA>	2	advmod	<NA>	<NA>
2	Animacy=Inan|Case=Abs|Definite=Def|Number=Sing	0	root	<NA>	<NA>
3	<NA>	2	cop	<NA>	<NA>
4	Mood=Ind|Number[abs]=Sing|Number[erg]=Sing|Person[abs]=3|Person[erg]=3	3	aux	<NA>	SpaceAfter=No
5	<NA>	2	punct	<NA>	SpacesAfter=\\n

Aurreprozesua

- Fitxategia irakurri (normalean txt formatuan) `text <- readLines("FITX.txt")`
- Hizkuntza aukeratu `language = "basque"`, modeloa jaitsi eta kargatu
- Testua analizatu!

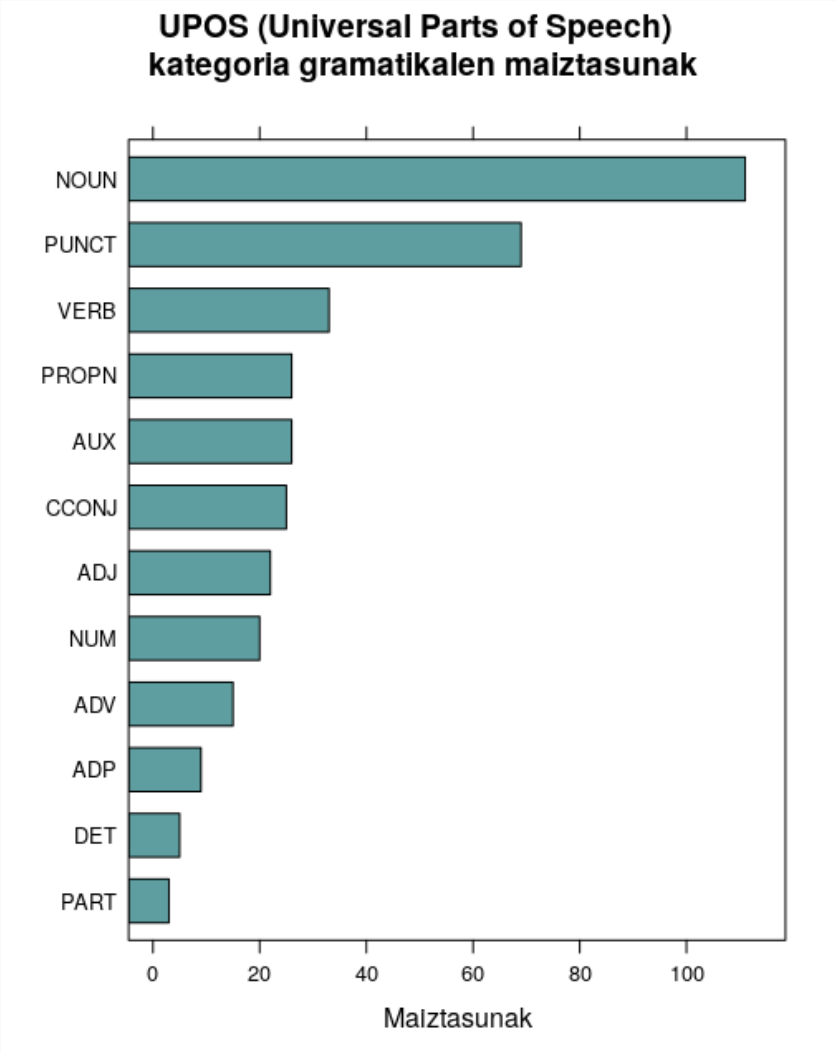
Analisia ikusteko

```
head(x)
```

Oinarrizko estatistikak atera I

- Kategoria gramatikalen distribuzioak

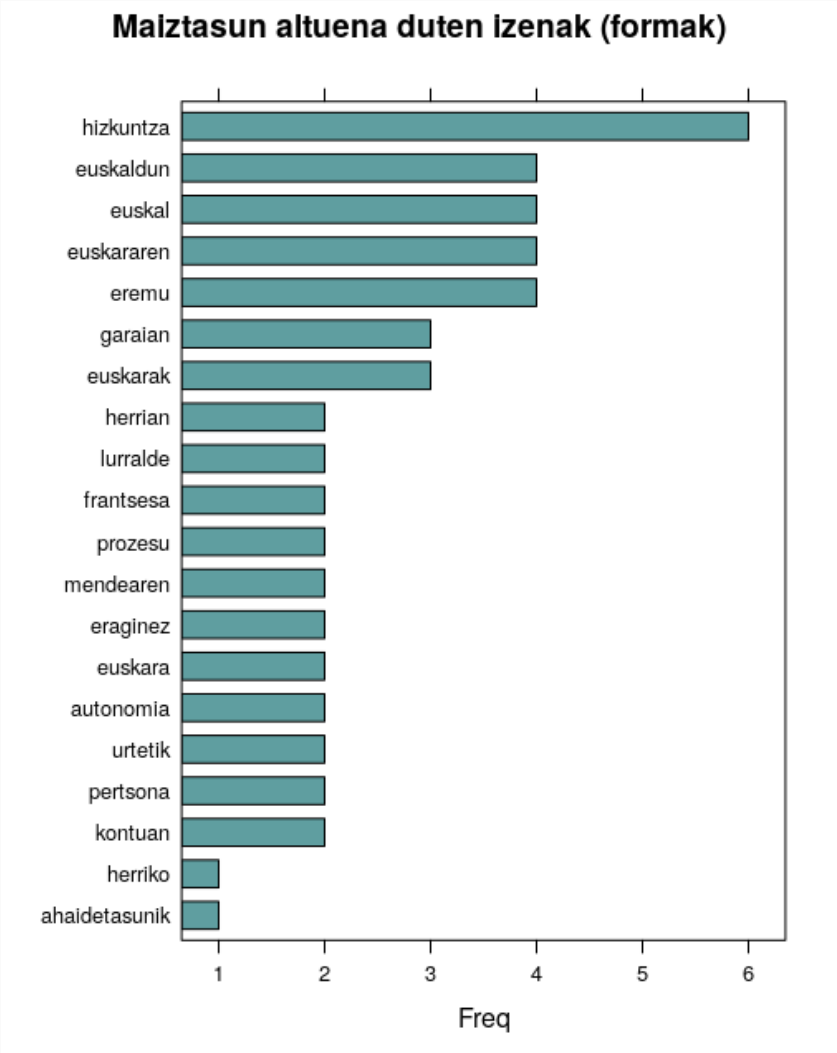
Kategoria gramatikalen distribuzioak (grafikoa)



Oinarrizko estatistikak atera II

- Gehien erabilitako izenak adibidez

Maiztasun handiena duten izenak (grafikoa)

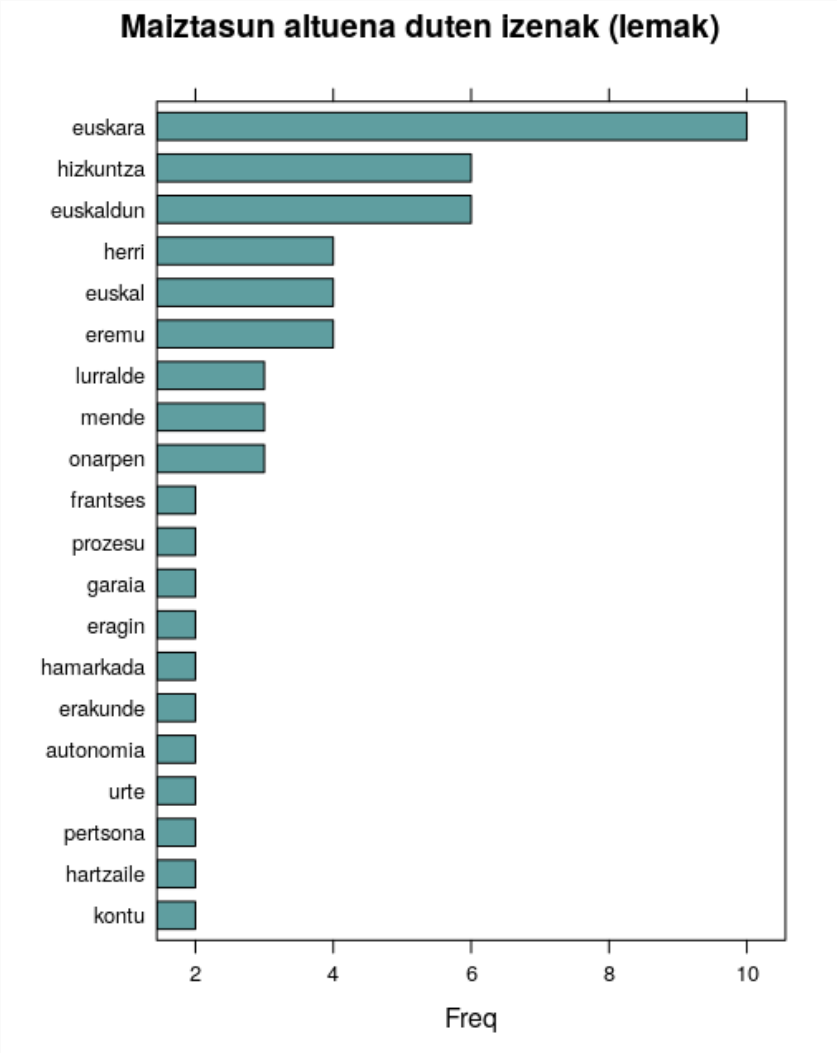


Kontzeptu garrantzitsua I: forma vs. lema

- **Forma:** hitza bere horretan adib. etxearen, etxeko, etxetik... dadin, dezan...
- **Lema:** hiztegian dagoen forma, flexiorik ez duena adib. etxe

Lematizazioa: formei dagokien lema ematearen prozesu automatikoa adib. etxearen -> etxe

Maiztasun handiena duten izenak II (grafikoa)



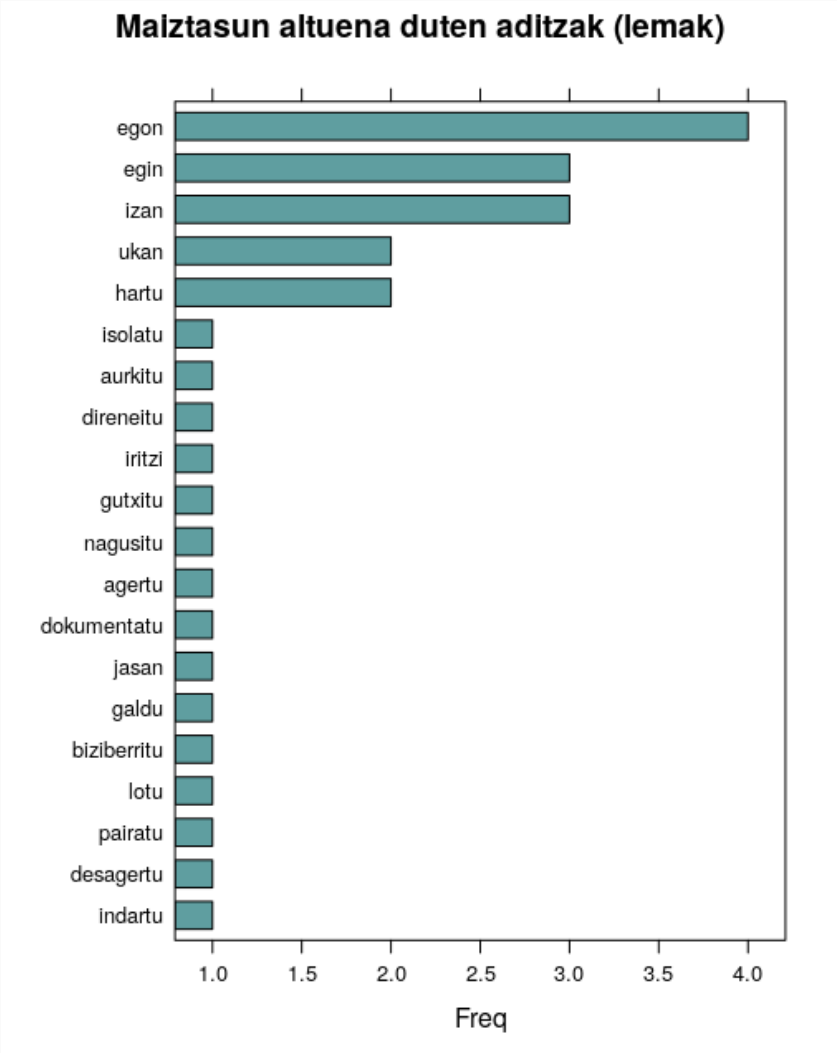
1. ariketa

- Atera testuan gehien erabili diren aditzen grafikoa

1. ariketa

- Atera testuan gehien erabili diren aditzen grafikoa
- Aditzen UPOSa VERB
- Lemak atera behar dira

Maiztasun handiena duten aditzak (grafikoa)



Parserraren akatsak

- *direneitu* aditza. Zer da?
- Lematizazio akats bat
- Normala tresna automatikoetan
- Emaitzak: `https://github.com/jwijffels/udpipe.models.ud.2.4/blob/master/inst/udpipe-ud-2.4-190531/README`

Aurkibidea

Zer da *wordcloud* bat?

- Testu baten agertzen diren hitzak irudikatzea, maiztasunaren arabera antolatuta
- Hezkuntzan baliagarria testuak irakurtzen hasi aurretik hiztegia lantzeko (*testuaren laburpena*)
- Sarean orri ugari, baina desegokiak morfologia aberatsa duten hizkuntzentzat; stopwords zerrendarik ez...

Kontzeptu garrantzitsua II: stopwords-ak

- **Stopwords:** Maiz agertzen diren hitzak, semantikoki arinak (adib. eman, alde...) edo hitz funtzionalak (adib. baina, hau...)
- R-en tm paketeak ingelesezkoa baditu, baina beste hizkuntzetakoak ez
- Zerrenda bezala sartzean arazoak

Eduki semantikoa duten hitzak erabili

Testuaren analisitik eduki semantikoa duten hitzak kategoriaren arabera filtratu

Kategoria gramatikalak aukeratzeko

```
x <- filter(x, upos %in% c("PROPN", "NOUN", "ADJ",
"VERB"))
```

Dataframetik lemak jaso

Lemen zutabea aukeratu

```
lemak <- as.vector(x$lemma)
```

Corpus txukuna sortu

“SimpleCorpus” eta “Corpus” datu-egiturak

Corpusa begiratu

```
inspect(docs)
```

Corpusa txukundu: karaktere arraroak kendu, minuskula bihurtu, zenbakiak kendu, maiz erabiltzen diren hitzak (stopwordak) ezabatu

Stopwordak kendu

```
docs <- tm_map(docs, removeWords, c( "egin",
"egon", "izan", "ukan", "eman", "aurre", "nahi",
"alde", "behar"))
```

Wordclouda sortu

Matrizea sortu hitzen maiztasunarekin

Matrizearen x lerroak ikusi

```
head(d, 10)
```

Wordcloud funtzioan parametroekin jolastu: min.freq., max.words, colors...

Kolorea aukeratu

```
display.brewer.all()
```

Wordclouda


euskara
euskaldun
euskal
hizkuntza
lurralde
eremu
nafarroa
hamarkada
erakunde
eragin
mende
handi
herri
onarpen
autonomia
frantses
urte
oso
prozesu
pertsona
hartzaile
ofizial
garaia
hartu
arana
gaztelania
kontu
ipar-mendebalde
arrotz
erdialde

2. ariketa: wordclouda egin

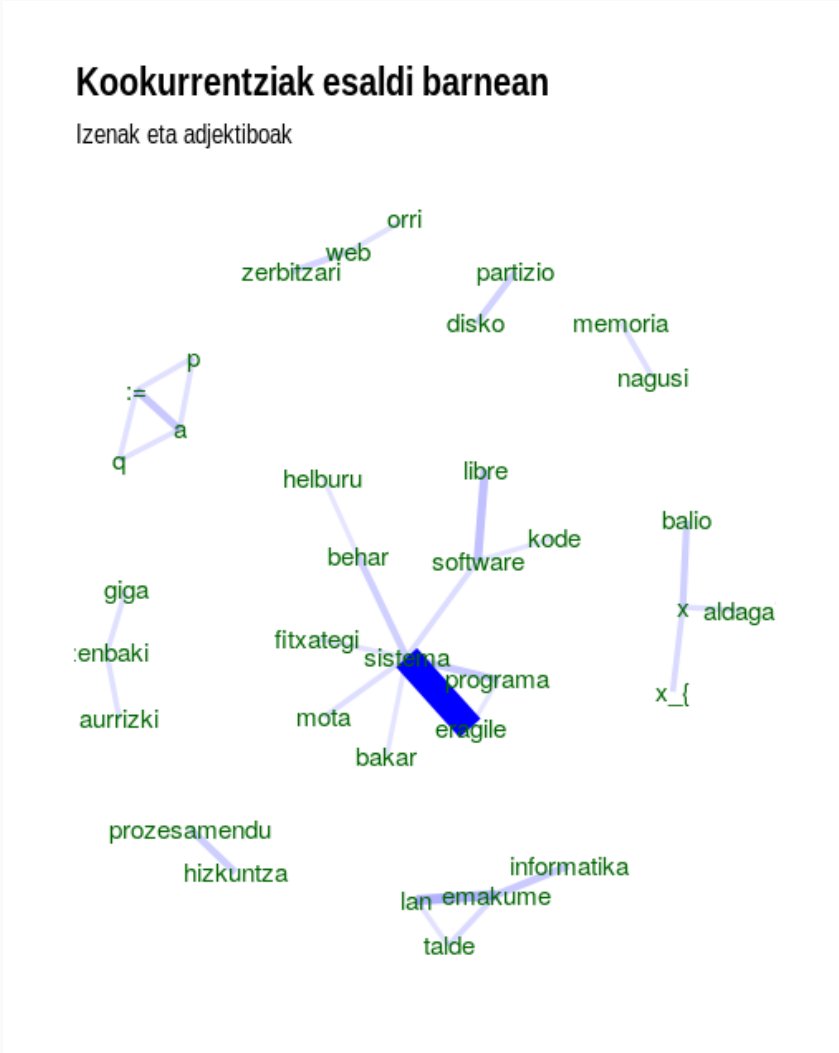
- Koloreak aldatu
- Parametroekin jolastu
- Hitz gehiago stopwords zerrendan sartu

Aurkibidea

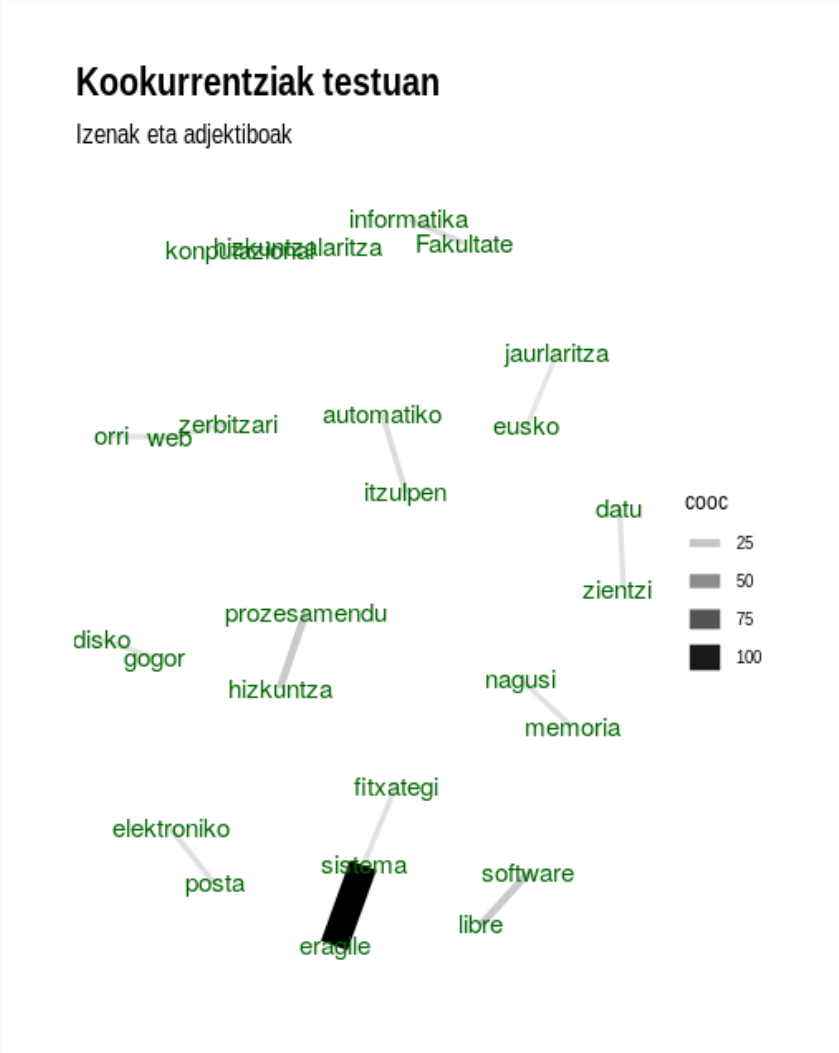
Zeinekin zabiltza???

- Testu luzeago bat erabili
- Kookurrentziak: maiz jarraian dauden hitzak

Kookurrentziak esaldi barnean (izenak eta adjektiboak)



Kookurrentziak testuan (izenak eta adjektiboak)



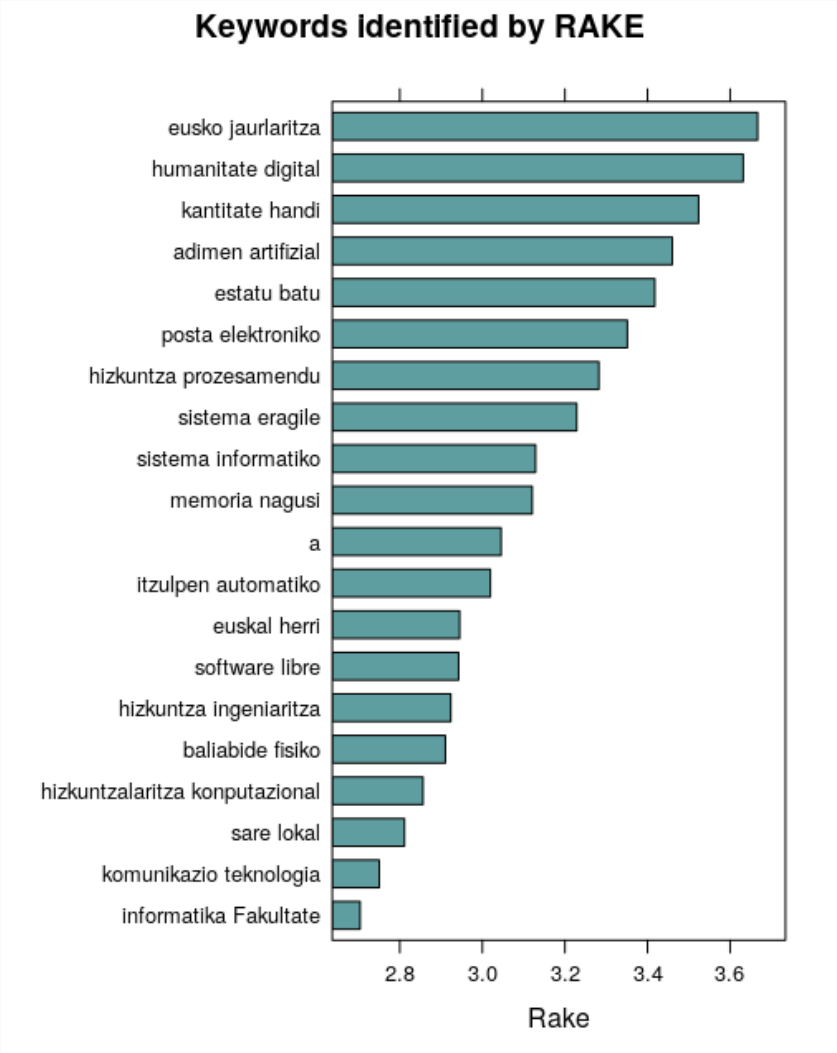
Aurkibidea

Zer da garrantzitsuena?

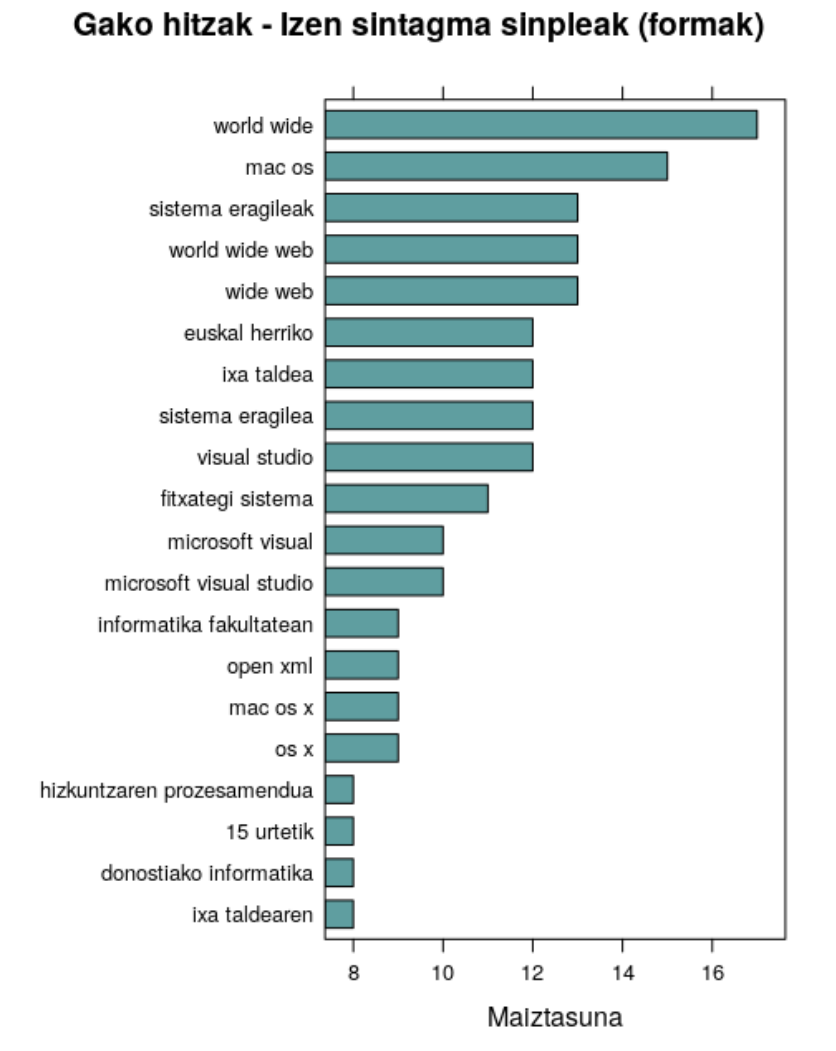
- **Gako hitzak (keywords)**: testu batean garrantzitsuenak diren hitzak
- TextRank erabilita: hitzen sare bat osatu jarraian dauden hitzekin eta ondoren *Google Pagerank* algoritmoa aplikatu
- Rapid Automatic Keyword Extraction (RAKE) algoritmoa erabilita: domeninuarekiko independientea, maiztasunak eta kookurrentziak kontuan hartu
- Gramatika kategorien sekuentziak erabilita

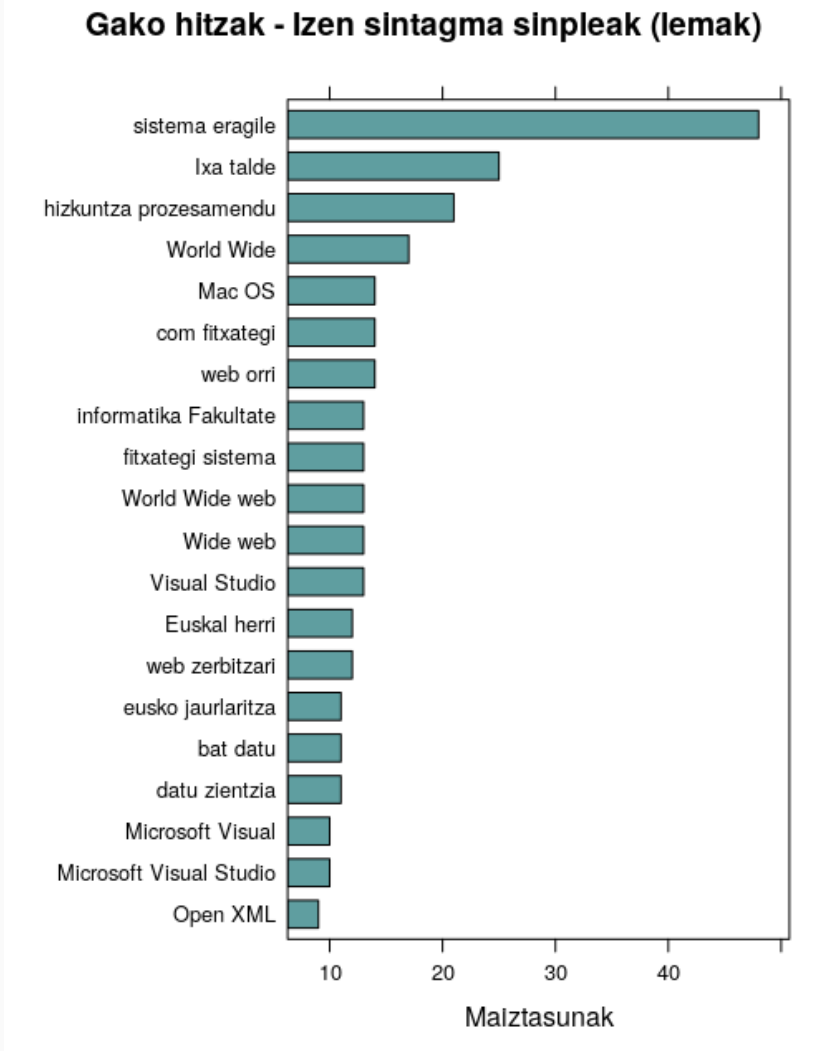
TextRank-en araberako gako hitzak

RAKE-ren araberako gako hitzak



Informazio linguistikoan oinarritutako gako hitzak (lemak vs. formak)





Aurkibidea

Eta gero, zer?

- Ehunka analisi linguistiko
- Topic modelling-ak
- Informazio erauzketa
- ...

Lizentzia

Pakete garrantzitsuen dokumentazioa

- udpipe `https://cran.r-project.org/web/packages/udpipe/udpipe.pdf`
- tm `https://cran.r-project.org/web/packages/tm/tm.pdf`

Web orri garrantzitsuak

- Universal dependencies
  `https://universaldependencies.org/`
- udpipe lantzeko
  `https://bnosac.github.io/udpipe/en/index.html`
- Wordcloudak egiteko
  `http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simpl`
- Testuen meatzaritza `https://www.tidytextmining.com/`

Mila esker!!!