8 Corpora

Métodos empíricos 2

31/05/2022

Hoy

- Las leyes de Zipf
- Corpora y pre-procesamiento
- Aplicaciones
- Word embeddings y más allá

Las leyes de Zipf

G.K. Zipf (1935) *The psycho-biology of language*

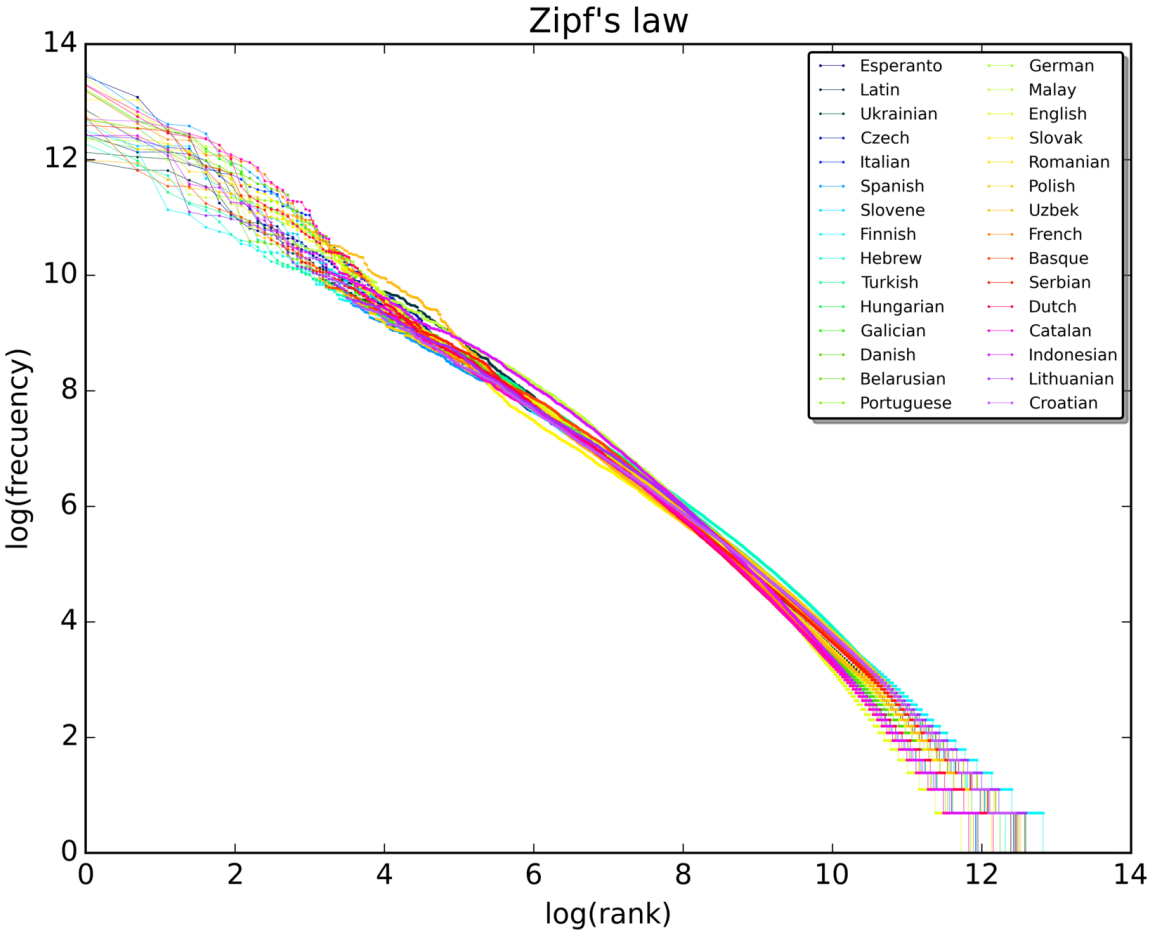
G.K. Zipf (1949) *Human behavior and the principle of least effort*

1. Zipf's (Rank-Frequency) Law

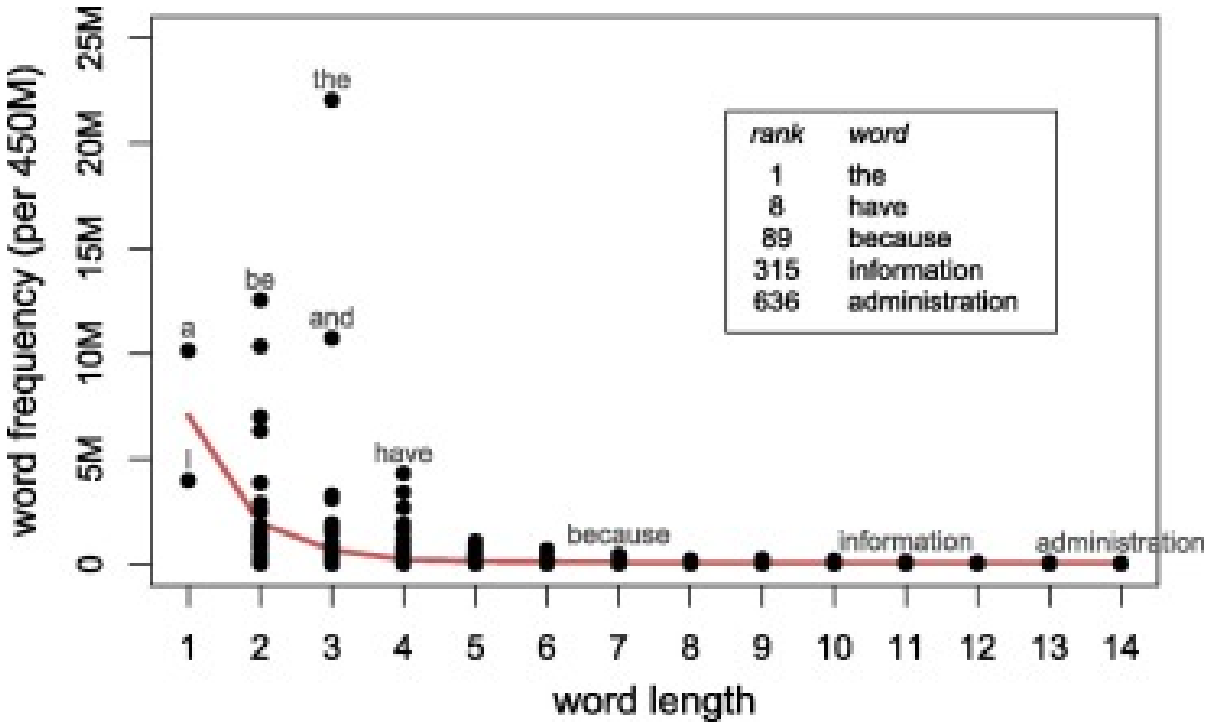
2. Zipf's Law of Abbreviation

3. Zipf's Meaning-Frequency Law

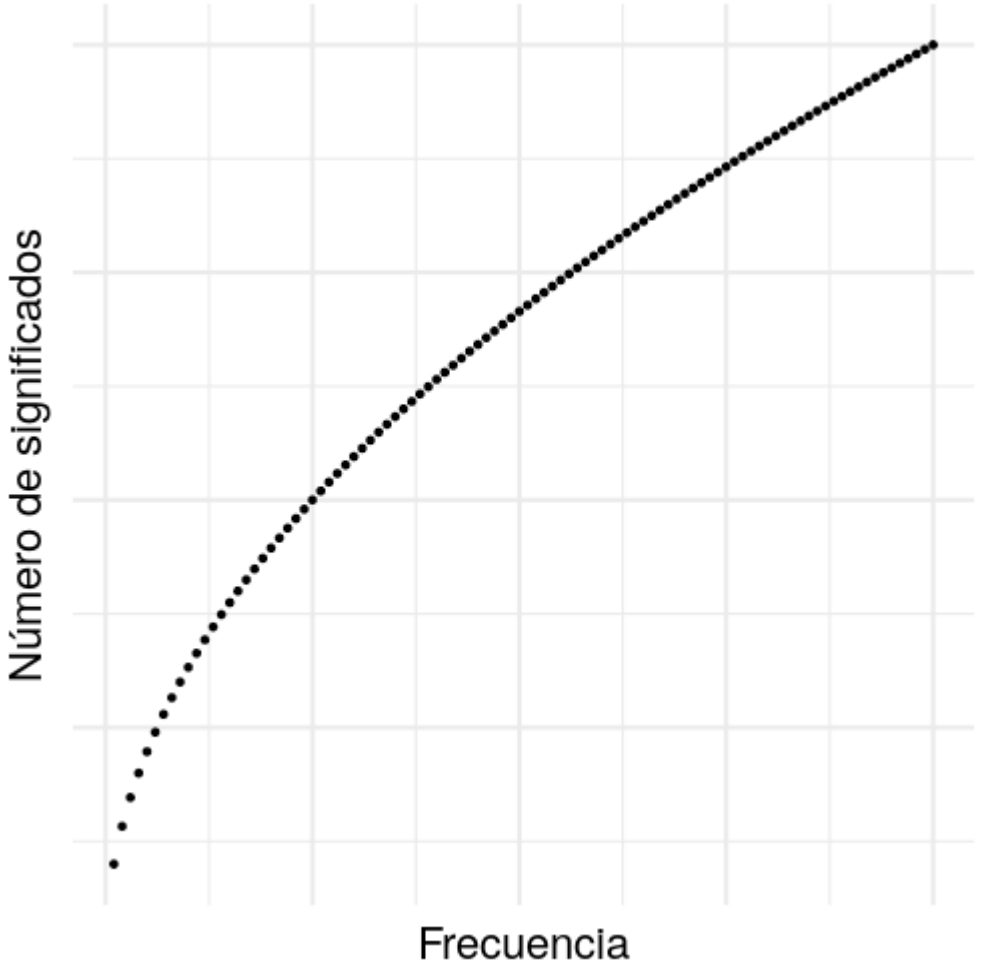
Rank-Frequency Law: La distribución rango-frecuencia de palabras es inversa



Law of Abbreviation: Formas frequentes tienden a ser más cortas



Meaning-frequency Law: Formas frequentes tienden a tener más significados



Corpora y pre-procesamiento

Corpora

- Por definición: Cualquier colección de datos
- Por uso convencional: Colección de datos no estructurados, muchas veces de gran tamaño

---

Lo que significa gran varía en función a la naturaleza de los datos, y de cuándo son / es el análisis.

Associated Press Corpus

Colección de 2246 artículos de noticias, la mayoría de alrededor de 1988

MEXICO CITY (AP) — The Mexican government said Tuesday that COVID-19 has passed from a pandemic to an endemic stage in Mexico, meaning authorities will treat it as a seasonally recurring disease.

Mexico never enforced face mask requirements, and the few partial shutdowns of businesses and activities were lifted weeks ago.

“It is now retreating almost completely,” said President Andrés Manuel López Obrador.

New case numbers have declined. But that may be because Mexico, which never did much testing, is now offering even fewer tests.

Daily death rates have also dropped sharply.

Mexico has recorded almost 325,000 test-confirmed deaths, but government reviews of death certificates suggest the real toll is almost 490,000.

About 90% of adult Mexicans have recieved at least one dose of the coronavirus vaccine.

Tokenización

Segmentar y transformar tu corpus para que represente las unidades de tu análisis.

Por ejemplo, palabras, morfémas, o caracteres.

Tokenización a nivel de palabras

```
library(stringr)

first_par <- 'MEXICO CITY (AP) — The Mexican government said Tuesda

tokenized_first_par <- str_split(first_par, pattern = " ")[[1]]
tokenized_first_par
```

```
##  [1] "MEXICO"      "CITY"        "(AP)"        "—"           "The"
##  [6] "Mexican"     "government"  "said"        "Tuesday"     "that"
## [11] "COVID-19"    "has"         "passed"      "from"        "a"
## [16] "pandemic"    "to"          "an"          "endemic"     "stage"
## [21] "in"          "Mexico,"     "meaning"     "authorities" "will"
## [26] "treat"       "it"          "as"          "a"           "seasonally
## [31] "recurring"   "disease."
```

Procesos de normalización de token(e)s

Casing

Convertir todo el texto a minúscula (o mayúscula)

Stemming

Quitar material morfológico, quedandose sólamente con las raíces

Lematización

Cambiar palabras por sus respectivos lemas.

Casing

```
tolower(tokenized_first_par)
```

```
##  [1] "mexico"      "city"        "(ap)"        "—"           "the"
##  [6] "mexican"     "government"  "said"        "tuesday"     "that"
## [11] "covid-19"    "has"         "passed"      "from"        "a"
## [16] "pandemic"    "to"          "an"          "endemic"     "stage"
## [21] "in"          "mexico,"     "meaning"     "authorities" "will"
## [26] "treat"       "it"          "as"          "a"           "seasonally"
## [31] "recurring"   "disease."
```

Stemming & lemmatization

- cat, cats, cat's, cats'; ...
- to be; am; are; were; ...

AP tokenizada

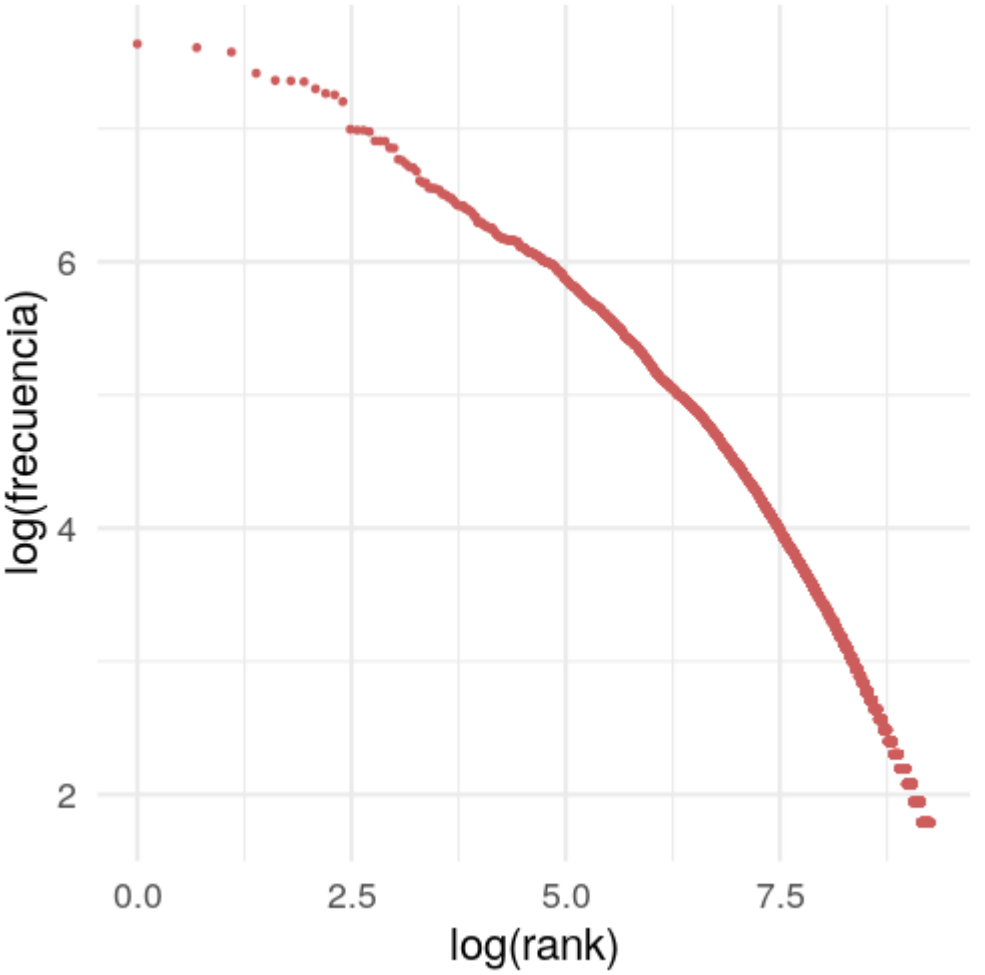
```
library(tidytext)

data("AssociatedPress", package = "topicmodels")
tidy(AssociatedPress)
```

```
## # A tibble: 6 × 3
##   document term      count
##      <int> <chr>     <dbl>
## 1        1 adding        1
## 2        1 adult         2
## 3        1 ago           1
## 4        1 alcohol       1
## 5        1 allegedly     1
## 6        1 allen         1
```

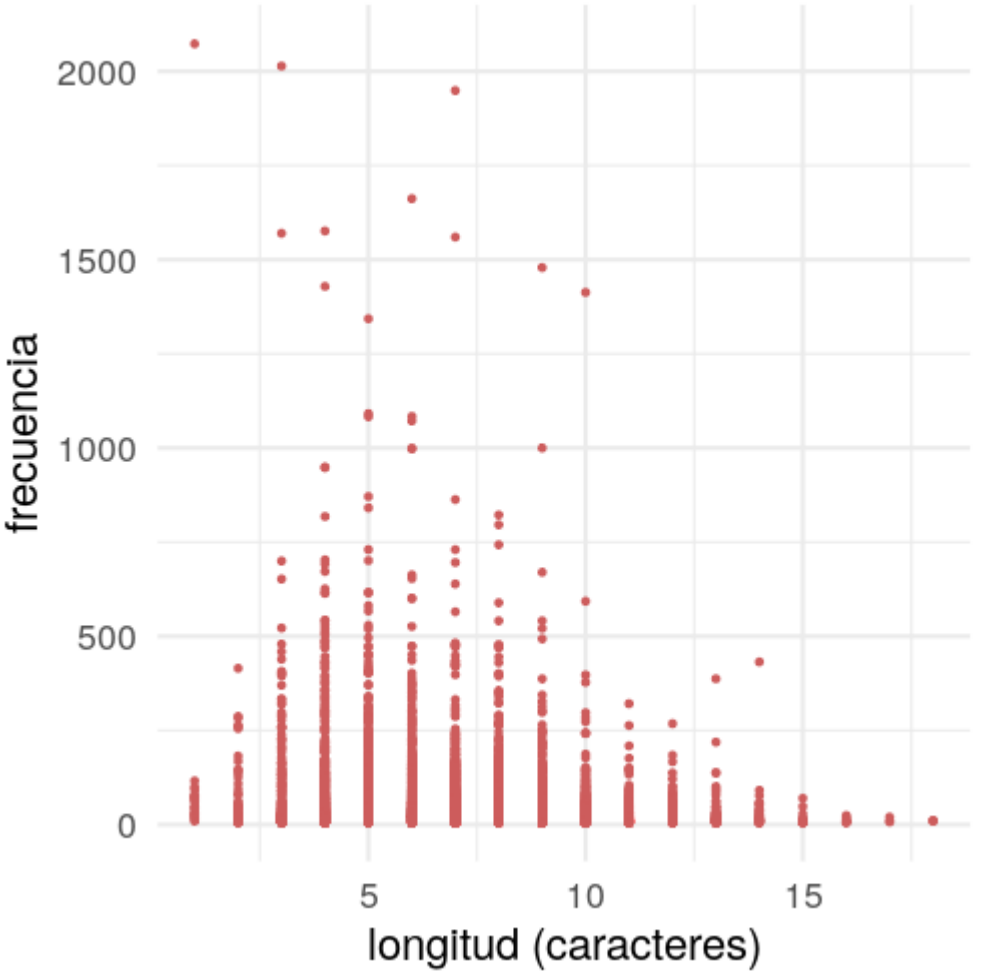
AP y Zipf?

```
## # A tibble: 6 × 6
##   term    count length  rank log.count log.rank
##   <chr>   <dbl>  <int> <int>     <dbl>    <dbl>
## 1 i        2073      1     1      7.64    0
## 2 new      2014      3     2      7.61    0.693
## 3 percent  1949      7     3      7.58    1.10
## 4 people   1662      6     4      7.42    1.39
## 5 year     1576      4     5      7.36    1.61
## 6 two      1570      3     6      7.36    1.79
```

AP y Zipf I



AP y Zipf II



Jane Austen

```
## # A tibble: 73,422 × 4
##    text                  book                 line chapter
##    <chr>                 <fct>               <int>   <int>
##  1 "SENSE AND SENSIBILITY" Sense & Sensibility     1       0
##  2 ""                    Sense & Sensibility     2       0
##  3 "by Jane Austen"      Sense & Sensibility     3       0
##  4 ""                    Sense & Sensibility     4       0
##  5 "(1811)"              Sense & Sensibility     5       0
##  6 ""                    Sense & Sensibility     6       0
##  7 ""                    Sense & Sensibility     7       0
##  8 ""                    Sense & Sensibility     8       0
##  9 ""                    Sense & Sensibility     9       0
## 10 "CHAPTER 1"           Sense & Sensibility    10       1
## # … with 73,412 more rows
```

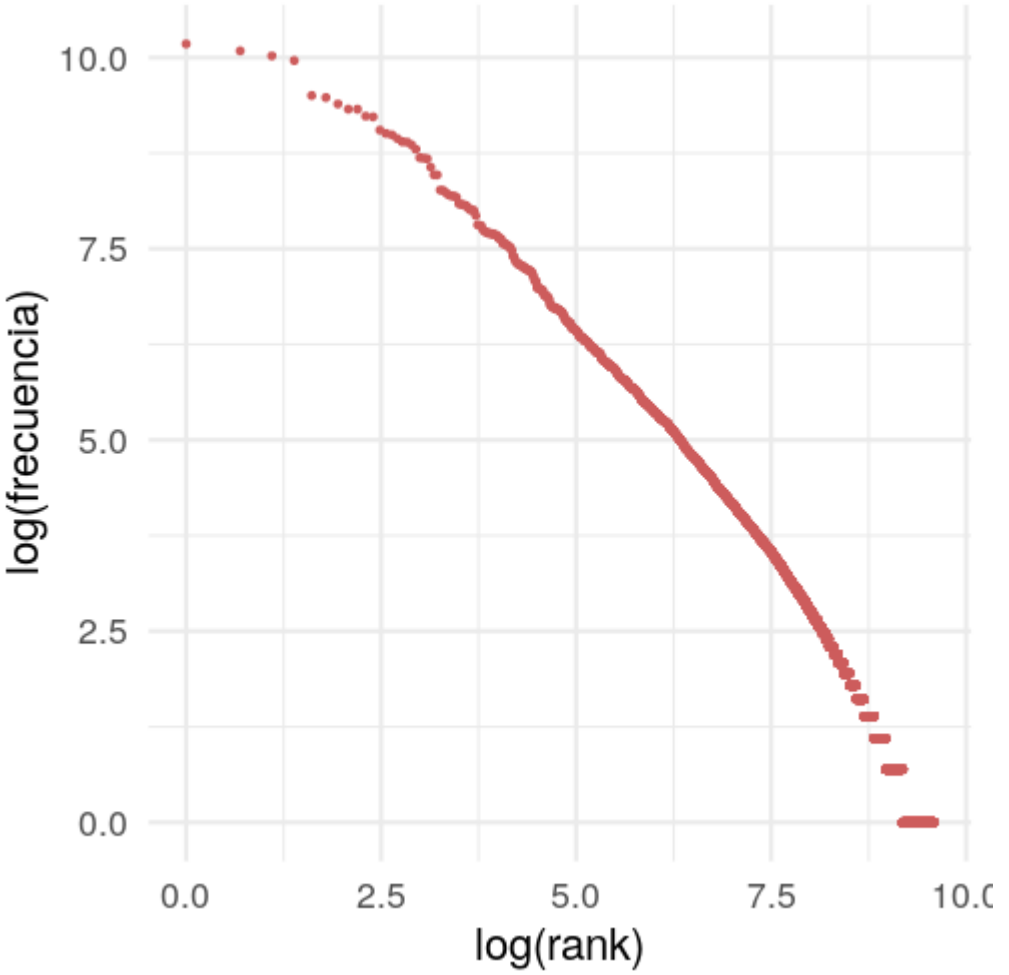
Jane Austen tokenizada y normalizada

```
## # A tibble: 725,055 × 4
##    book                line chapter word
##    <fct>              <int>   <int> <chr>
##  1 Sense & Sensibility     1       0 sense
##  2 Sense & Sensibility     1       0 and
##  3 Sense & Sensibility     1       0 sensibility
##  4 Sense & Sensibility     3       0 by
##  5 Sense & Sensibility     3       0 jane
##  6 Sense & Sensibility     3       0 austen
##  7 Sense & Sensibility     5       0 1811
##  8 Sense & Sensibility    10       1 chapter
##  9 Sense & Sensibility    10       1 1
## 10 Sense & Sensibility    13       1 the
## # … with 725,045 more rows
```

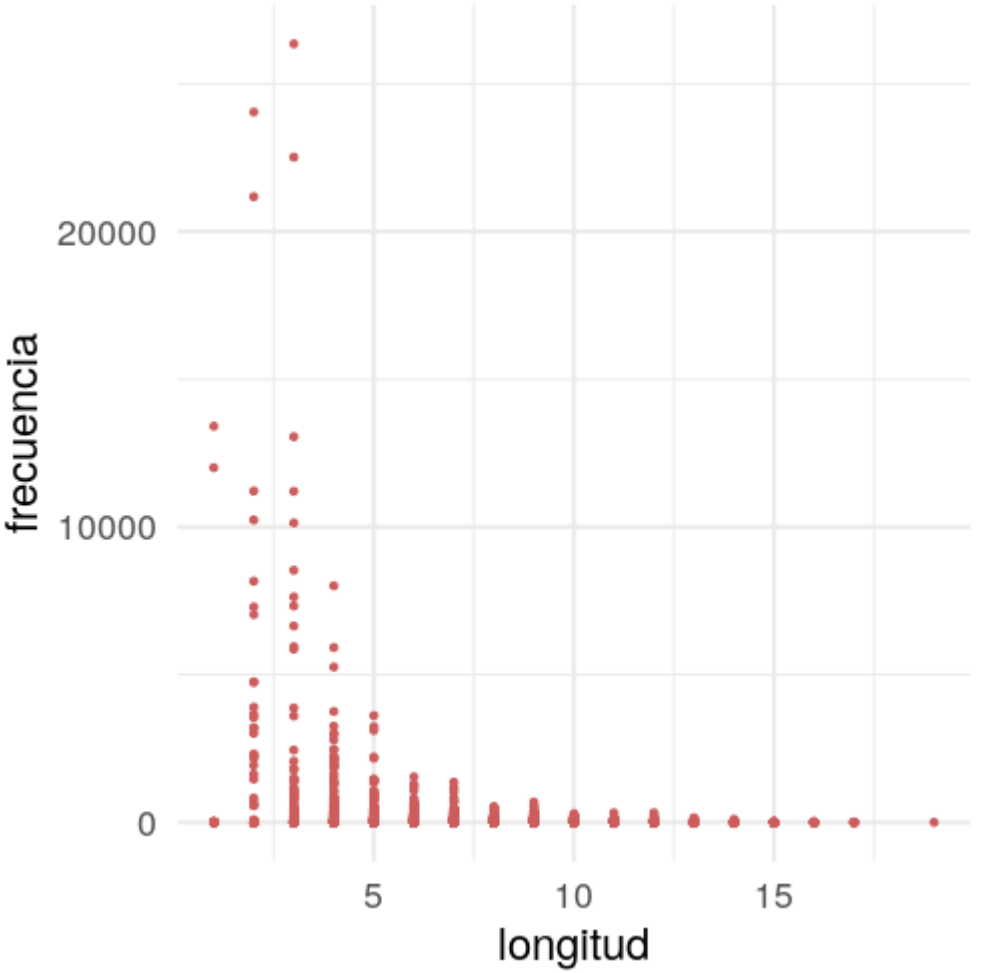
Jane Austen y Zipf?

```
## # A tibble: 6 × 6
##   word  count length log.count  rank log.rank
##   <chr> <int>  <int>     <dbl> <int>    <dbl>
## 1 the   26351      3     10.2      1    0
## 2 to    24044      2     10.1      2    0.693
## 3 and   22515      3     10.0      3    1.10
## 4 of    21178      2      9.96     4    1.39
## 5 a     13408      1      9.50     5    1.61
## 6 her   13055      3      9.48     6    1.79
```

Jane Austen y Zipf I



Jane Austen y Zipf II



Y la otra ley de Zipf?

Discusión

- Todo tipo de texto sigue las Leyes de Zipf?
- Toda lengua sigue las Leyes de Zipf?

Aplicaciones

Investigación

- Indispensable para descubrir o (des)confirmar regularidades en una, o varias lenguas
- Mayor volúmen de datos $\Rightarrow$ mayor sensibilidad para encontrar efectos menores (pero también más peligro de descubrir patrones falsos)
    - https://www.tylervigen.com/spurious-correlations
- Gran potencial --aún por explorar-- para tipología y lenguas menos descritas

Industria

- (Pre-)procesamiento de grandes volúmenes de datos lingüísticos
- Indispensable para descubrir o (des)confirmar regularidades a nivel de individuos, grupos y comunidades
- Enorme mercado que todavía se está abriendo

Word embeddings y más allá

Predicción como base para conocimiento lingüístico

Les ■■■....■■■

Les tasques ■■■....■■■

Les tasques de ■■■....■■■

Les tasques de remodelació ■■■....■■■

Les tasques de remodelació i ■■■....■■■

Les tasques de remodelació i ampliació ■■■....■■■

Predicción como base para conocimiento lingüístico

Les tasques de remodelació i ampliació de ■■■....■■■

Les tasques de remodelació i ampliació de l'
■■■....■■■

Les tasques de remodelació i ampliació de l'estadi
■■■....■■■

Les tasques de remodelació i ampliació de l'estadi començaran
■■■....■■■

Predicción como base para conocimiento lingüístico

Les tasques de remodelació i ampliació de l'estadi començaran al

Les tasques de remodelació i ampliació de l'estadi començaran al juny

Predicción como base para conocimiento lingüístico

Entrenar modelos con muchos parámetros a predecir información lingüística en grandes volúmenes de datos

$\Rightarrow$ aprendizaje de conocimiento lingüístico latente (hasta cierto grado)

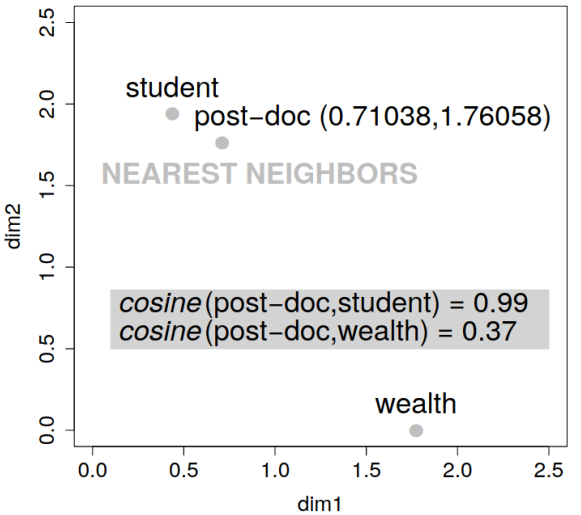
- Syntáxis ✔
- Morfología ✔
- Semántica ✔✘
- Pragmática ✘

Word embeddings

Any grad student or **post-doc** he'd have would be a clonal copy of himself.
During that **post-doc**, I didn't publish much.
. . .

⇓

	dim1	dim2
post-doc	0.71038	1.76058
student	0.43679	1.93841
wealth	1.77337	0.00012



Word embeddings

Language models

https://transformer.huggingface.co/doc/distil-gpt2

Paquetes

- python: spaCy, (py)torch
- R: tidytext, stringr

Siguentes avances

- Modelos multi-modales
- Calidad de datos vs. tamaño de modelo
- Límites de aprendizaje a base de texto
- Black box NLP & lenguaje emergente

Próxima sesión

- No hay entrega

- **Visualización**

- Informe final: 28/06