8 Corpora

Mètodes empírics 2

30/05/2022

Avui

- Les lleis de Zipf
- Corpora i pre-processament
- Aplicacions
- Language models i més enllà

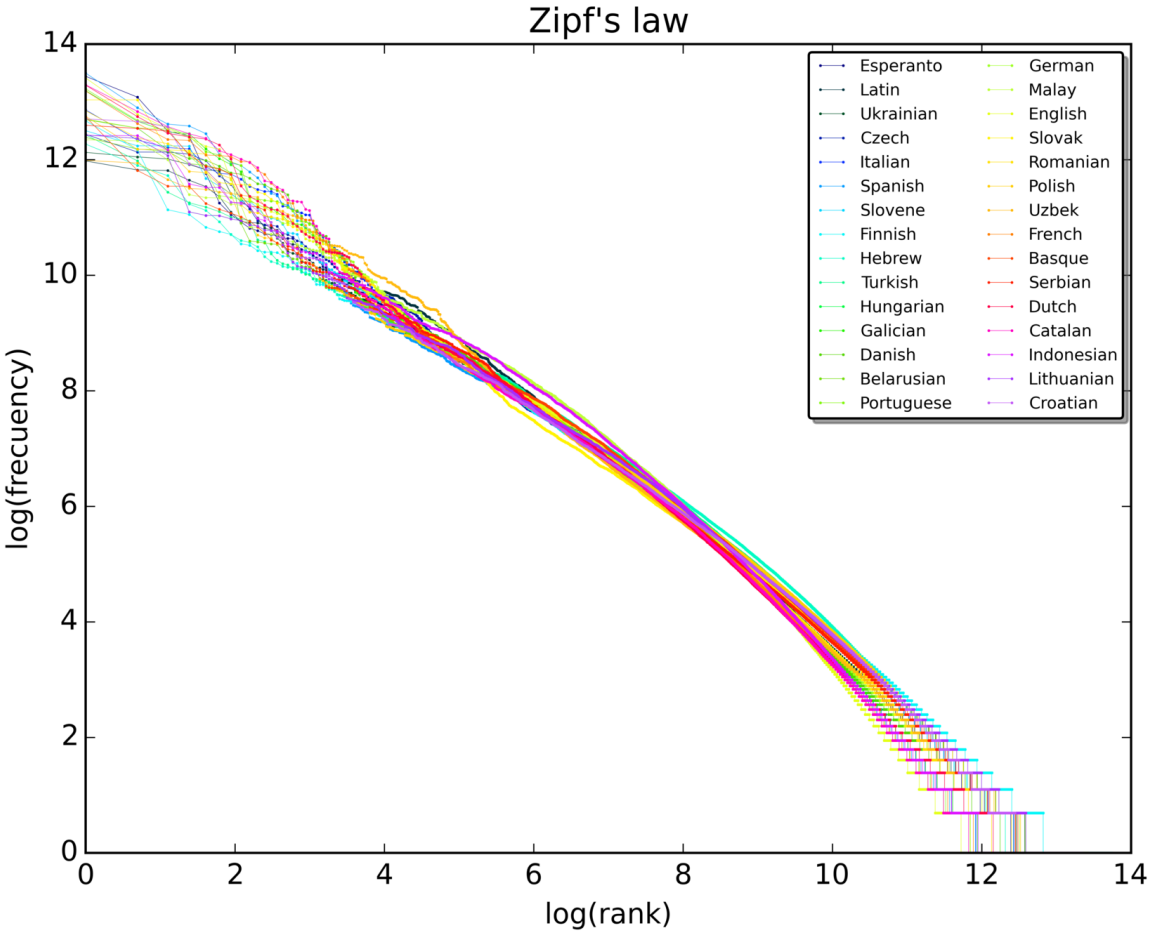
Les lleis de Zipf

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

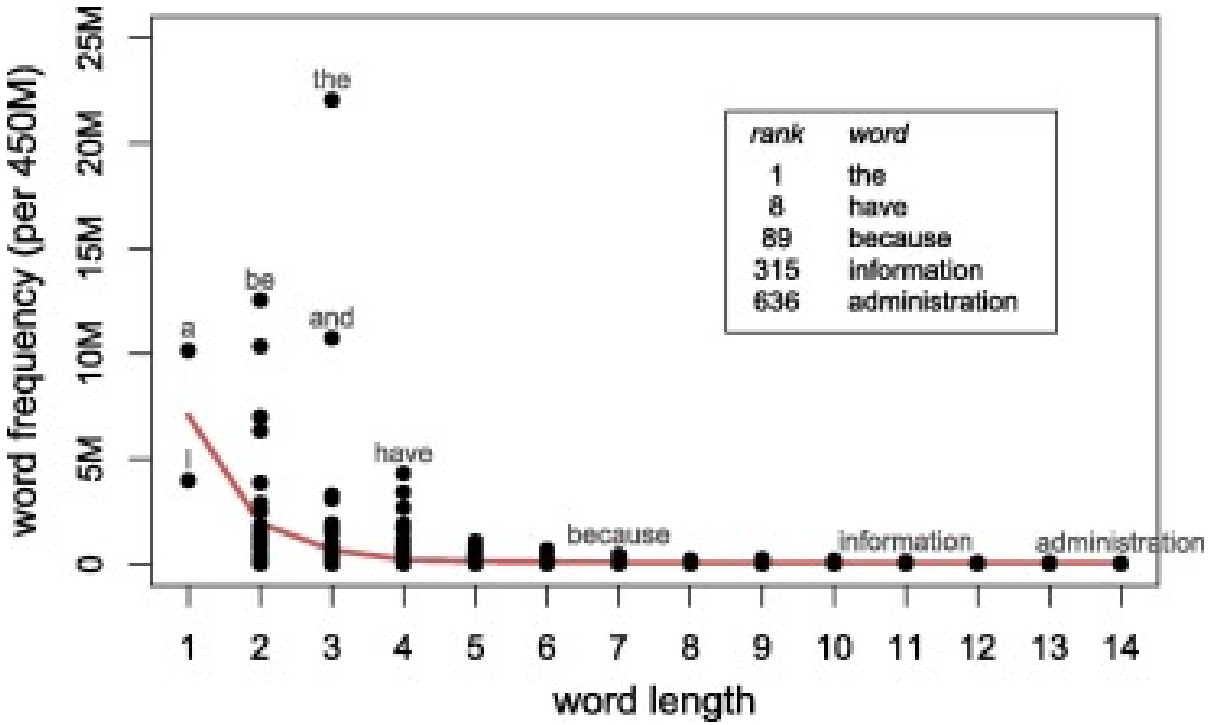
Rank-Frequency Law: La distribució rang-freqüència de paraules és inversa



Per què segueix el llenguatge aquesta llei?

Quines conseqüències té aquesta llei per a (i) l'anàlisi lingüística; (ii) l'aprenentatge de llengües; (iii) la traducció?

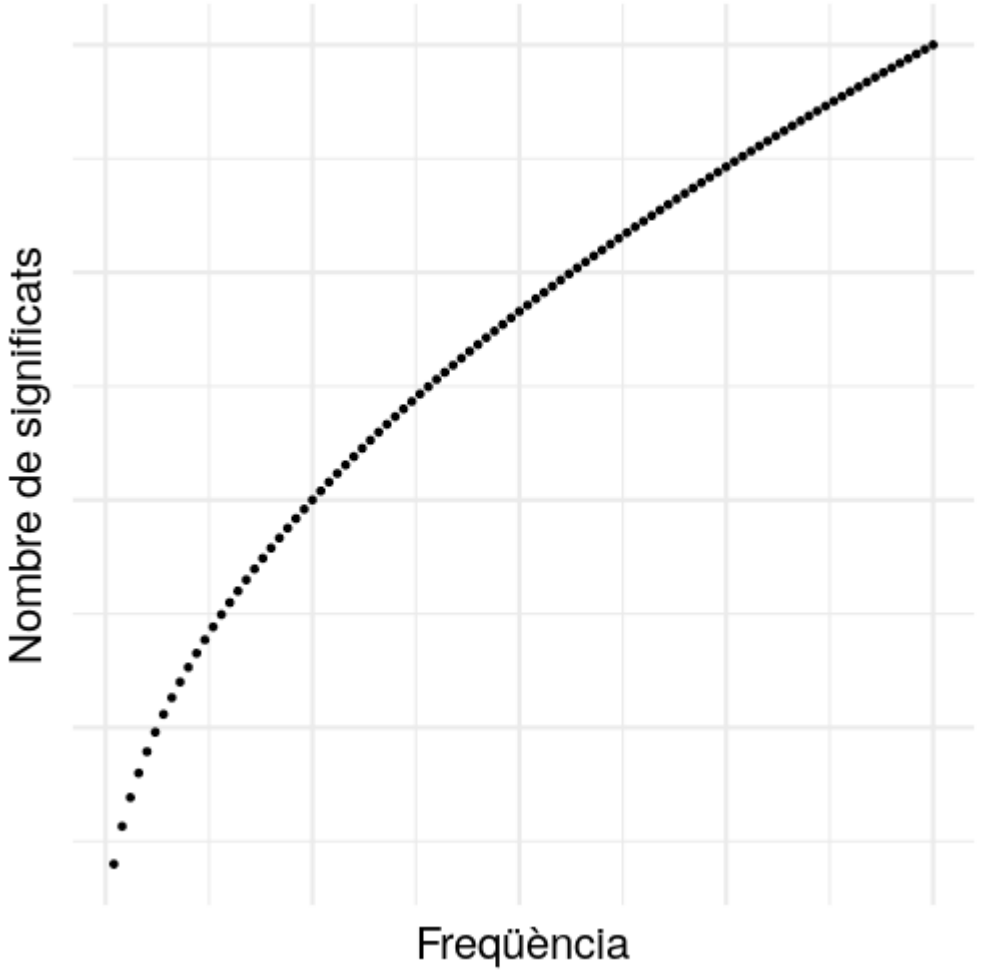
Law of Abbreviation: Formes freqüents tendeixen a ser més curtes



Per què segueix el llenguatge aquesta llei?

Quines conseqüències té aquesta llei per a (i) l'anàlisi lingüística; (ii) l'aprenentatge de llengües; (iii) la traducció?

Meaning-frequency Law: Formes freqüents tendeixen a tenir més significats



Per què segueix el llenguatge aquesta llei?

Quines conseqüències té aquesta llei per a (i) l'anàlisi lingüística; (ii) l'aprenentatge de llengües; (iii) la traducció?

Corpora i pre-processament

Corpora

- Per definició: Qualsevol col·lecció de dades
- Per ús convencional: Col·lecció de dades no estructurades, moltes vegades de gran tamany

---

El que significa gran varia en funció de la naturalesa de les dades, i de quan són / és l'anàlisi.

Associated Press Corpus

Col·lecció de 2246 articles de notícies, la majoria del voltant 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.

Què hauríem de fer per poder analitzar aquestes dades? Per exemple, per comprovar si es compleix la llei de Zipf

Tokenització

Segmentar i transformar el teu corpus perquè representi les unitats de la teva anàlisi.

Per exemple, paraules, morfémes, o caràcters.

Tokenització a nivell de paraules

```
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"
## [16] "pandemic"    "to"          "an"          "endemic"     "stage"
## [31] "recurring"   "disease."
```

Processos de normalització de tokens

Casing

Convertir tot el text a minúscula (o majúscula)

Stemming

Treure material morfològic, quedant-se només amb les arrels

Lematización

Canviar paraules pels respectius lemes.

Casing

```
tolower(tokenized_first_par)
```

```
##  [1] "mexico"      "city"        "(ap)"        "—"           "the"
## [16] "pandemic"    "to"          "an"          "endemic"     "stage"
## [31] "recurring"   "disease."
```

Stemming & lemmatization

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

AP tokenitzada

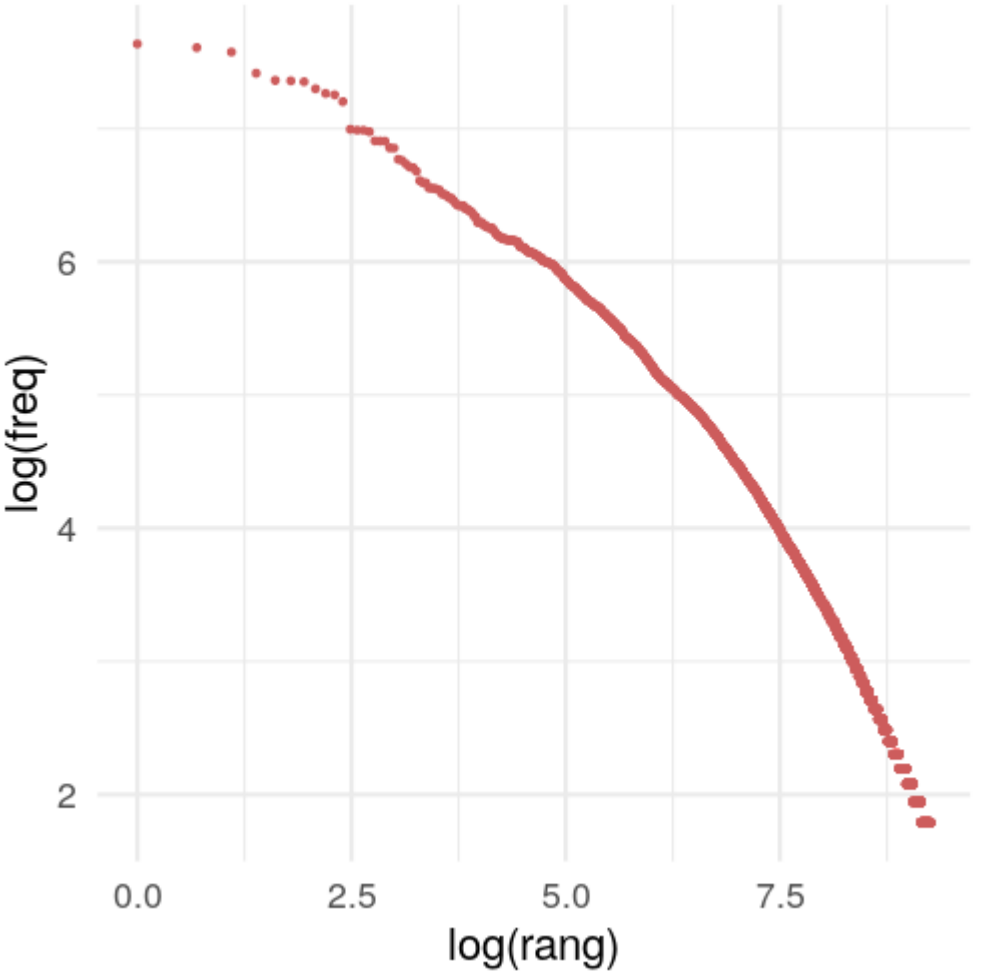
```
library(tidytext)

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
```

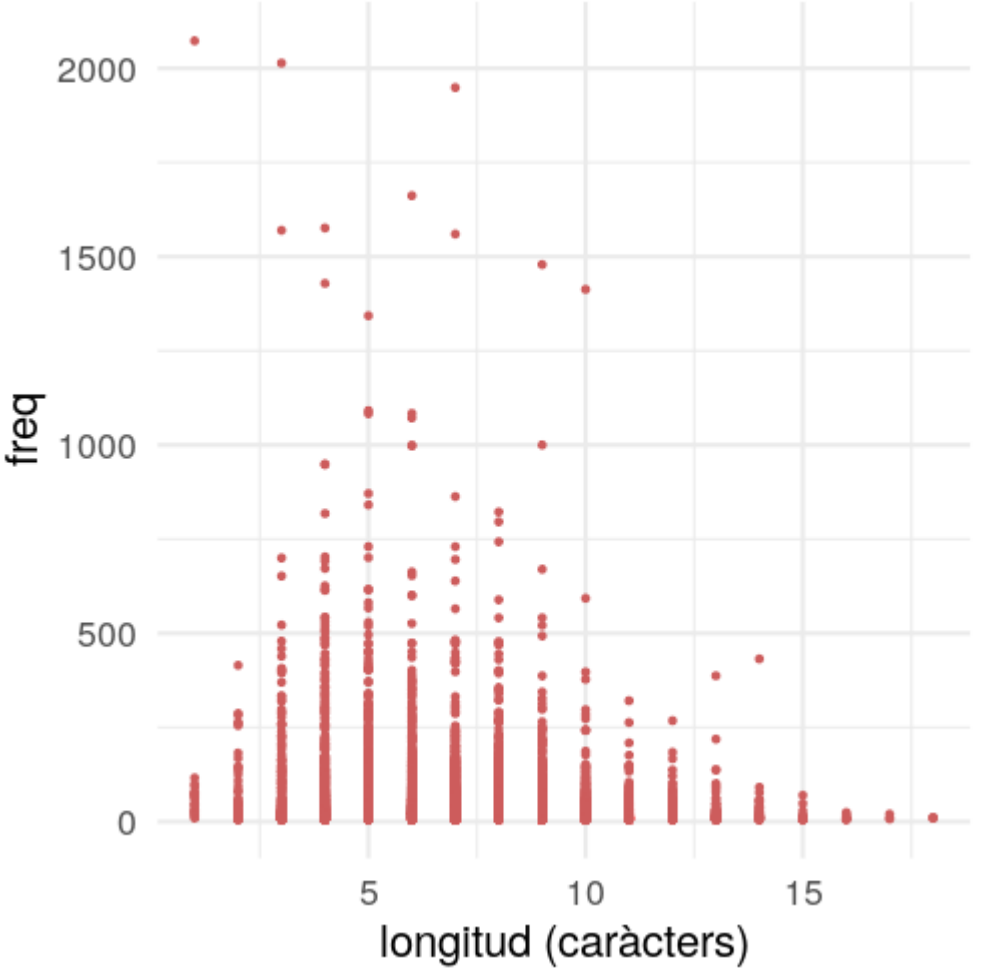
AP i Zipf?

```
## # 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 i Zipf I



AP i 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
```

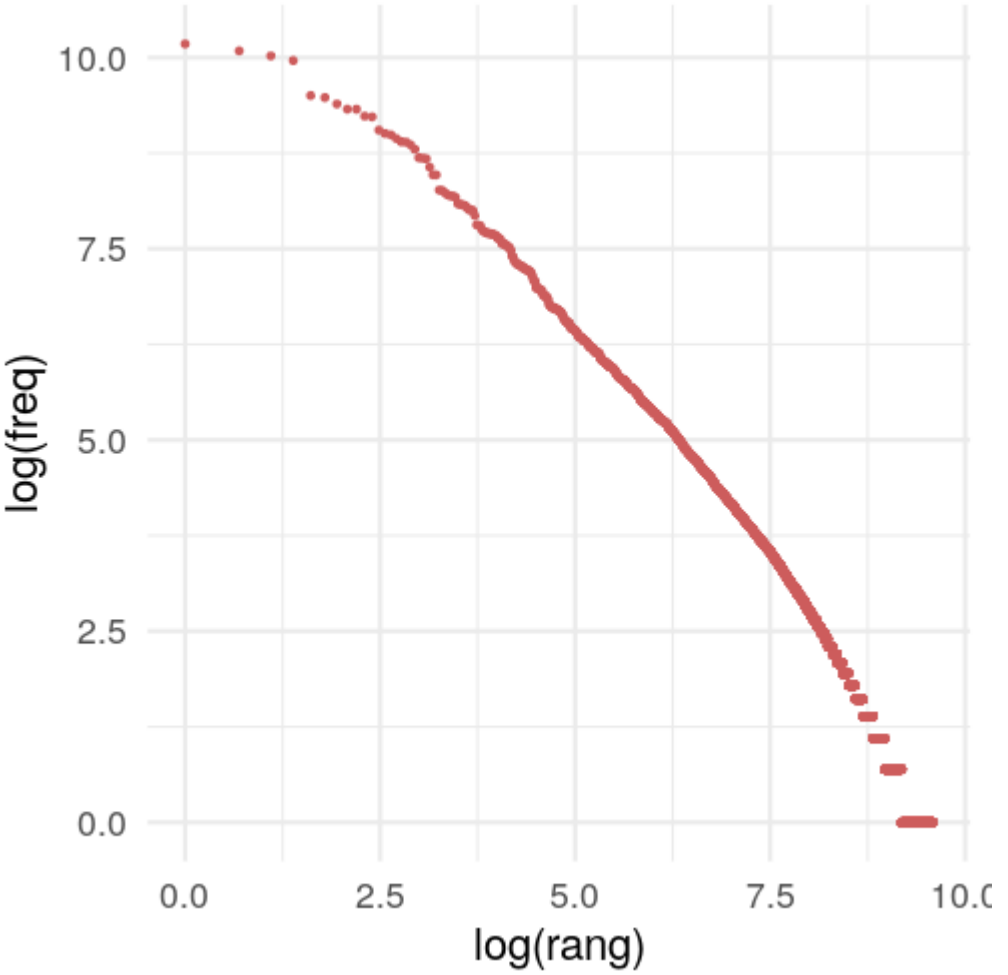
Jane Austen tokenitzada i normalitzada

```
## # 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
```

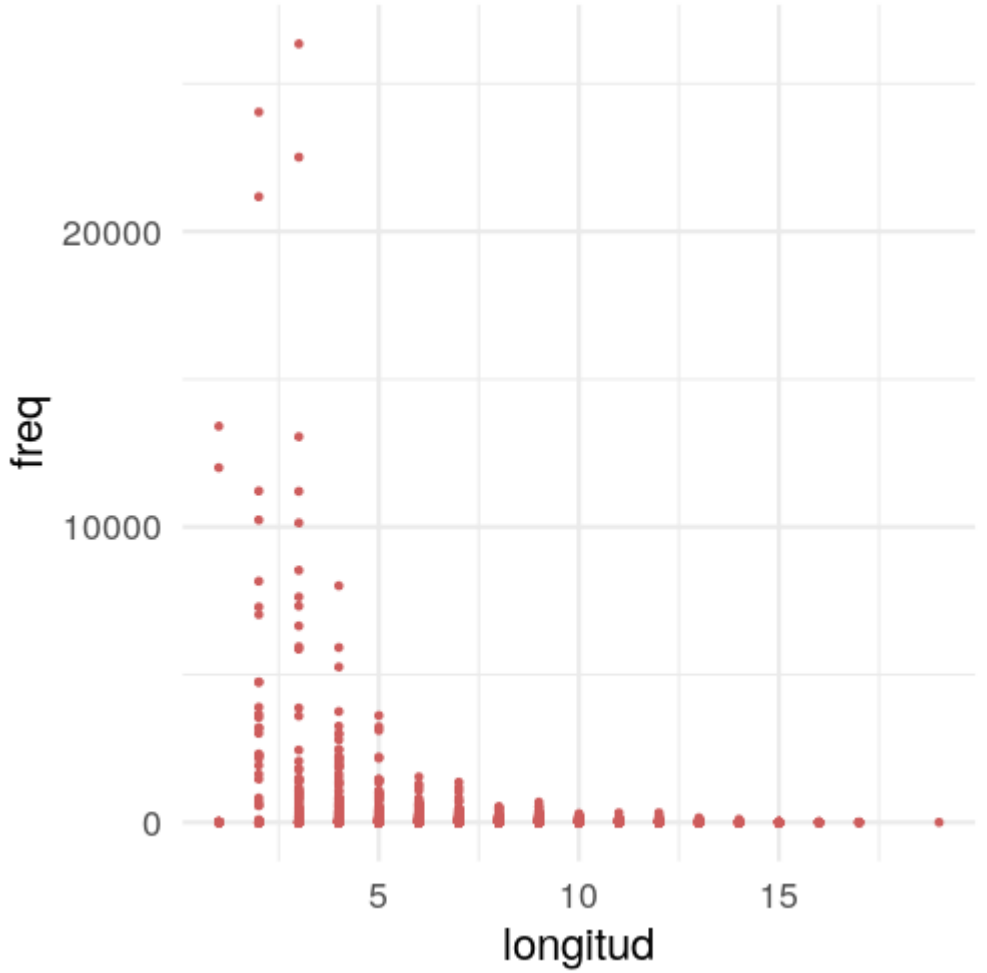
Jane Austen i Zipf?

```
## # 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 i Zipf I



Jane Austen i Zipf II



I l'altra llei de Zipf?

Discussió

- Tot tipus de text segueix les Lleis de Zipf?
- Tota llengua segueix les Lleis de Zipf?

Aplicacions

Investigació

- Indispensable per descobrir o (des)confirmar regularitats en una, o diverses llengües
- Major volum de dades $\Rightarrow$ més sensibilitat per trobar efectes menors (però també més perill de descobrir patrons falsos)
    - https://www.tylervigen.com/spurious-correlations
- Gran potencial --encara per explorar-- per a tipologia i llengües menys descrites

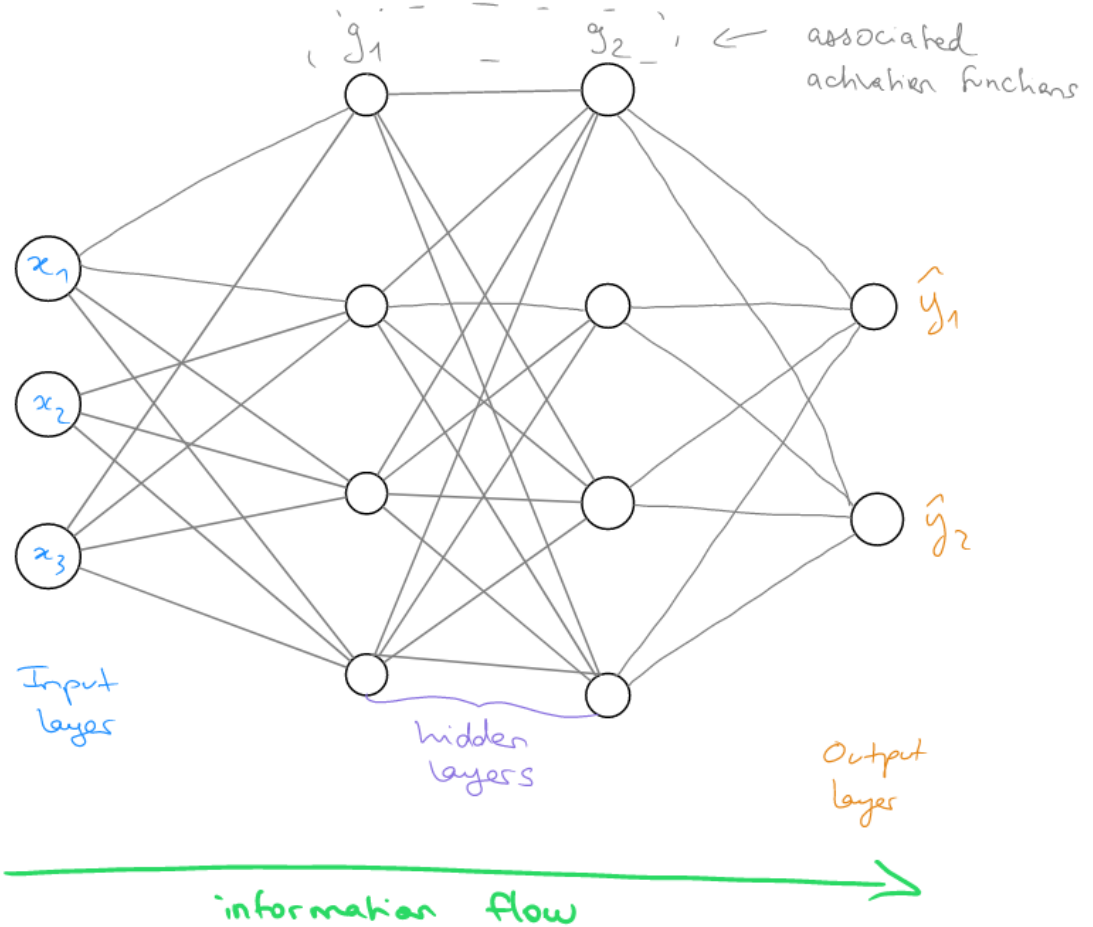
Indústria

- (Pre-)processament de grans volums de dades lingüístiques
- Indispensable per descobrir o (des)confirmar regularitats a nivell d'individus, grups i comunitats
- Enorme mercat que encara s'està obrint

Word embeddings i més enllà

Que es ChatGPT? Com funciona?

Xarxes neuronals



Predicció com a base per a coneixement lingüístic

Les ■■■....■■■

Les tasques ■■■....■■■

Les tasques de ■■■....■■■

Les tasques de remodelació ■■■....■■■

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

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

Predicció com a base per a coneixement lingüístic

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■■■....■■■

Predicció com a base per a coneixement lingüístic

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

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

Predicció com a base per a coneixement lingüístic

Entrenar models amb molts paràmetres a predir informació lingüística en grans volums de dades

$\Rightarrow$ aprenentatge de coneixement lingüístic latent (fins a cert grau)

- Syntàxi ✔
- Morfologia ✔
- Semàntica ✔✘
- Pragmàtica ✘

Language models

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

GPT2, Llama, BERT, ChatGPT, Bard, …

ChatGPT

Després d'aprendre amb self-supervision (predint text per compte propi), aprèn humans amb reinforcement learning en una segona fase

- $\Rightarrow$ millora notable en pragmàtica
- $\Rightarrow$ bona compressió de coneixement lingüístic i coneixement general
- $\Rightarrow$ compressió comporta halucinació i repetició

Preguntes o inquietuds sobre ChatGPT? Canviarà el mercat laboral? El mercat de la traducció? La lingüística? L'educació?

Paquets

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

Següents avenços

- Models multimodals
- Qualitat de dades i mida de model
- Límits d'aprenentatge a base de text
- Black box NLP & llenguatge emergent
- Self-supervised reinforcement learning

Propera sessió

- Practical exercise (06/06)

- **Visualització**

- Informe final: 28/06