WORD ASSOCIATION NORMS, MUTUAL INFORMATION, AND LEXICOGRAPHY

(Journal of Computational Linguistics, Volume 16 Issue 1, March 1990, pp 22-29, MIT Press Cambridge, MA, USA)

Kenneth Ward Church
Bell Laboratories Murray Hill, N.J.

PatrickHanks
Collins Publishers Glasgow, Scotland

Word association

"Generally speaking, subjects respond quicker than normal to the word nurse if it follows a highly associated word such as doctor."

Empirical estimates of word association

Palermo, D. and Jenkins, J. 1964 "Word AssociationNorms." University of Minnesota Press, Minneapolis, MN.

An information theoretic measure

Association ratio, based on the information theoretic concept of **mutual information**.

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- If there is a **genuine association** between *x* and *y*, then *P(x,y)* will be much larger than *P(x)* . *P(y)*, and ***I(x,y) >> 0***.
- If there is **no interesting relationship** between *x* and *y*, then *P(x,y) = P(x) . P(y)*, and thus, ***I(x,y) ~ 0***.
- If x and y are in **complementary distribution**, then P(x,y) will be much less than P(x) P(y), forcing ***I(x,y) <<0***.

Probability estimation

P(x) and P(y) are estimated by counting the number of observations of *x* and *y* in a corpus, *f(x)* and *f(y)*, and normalizing by N, the size of the corpus.

P(x,y), are estimated by counting the number of times that *x* is followed by *y* in a window of *w* words, *f_w(x,y)*, and normalizing by N.

(...the window size, w, will be set to five words as a compromise)

Since the association ratio becomes unstable when the counts are very small, we will not discuss word pairs with f(x,y) < 6.

Measure properties

- I (x, y) = I(y, x). But, f(x, y) ~= f(y, x)
- Expected: f (x, y) <= f(x) and f(x, y) <= f(y)
  - "Library workers were prohibited from saving books from this heap of ruins,"
  - f(prohibited) = 1 and f(prohibited, from) = 2
  - This problem can be fixed by dividing f(x, y) by (w-1)

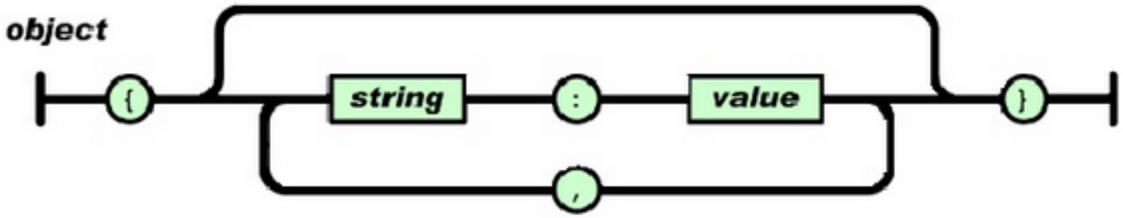
Corpus

- This paper: AP
- Other corpus: TASA, Pagina12, La Nación, Google

Más formatos... JSON

JSON: JavaScript Object Notation.

Formato para serializar objetos (listas, diccionarios)



Ejemplo:

d = {'mensaje': 'Hola a todos'}

json(d) --> "{'mensaje': 'Hola a todos'}"

Ejemplo de JSON

```
import json

d = {'mensaje': 'Hola a todos'}

# Convert dict to JSON
mi_json = json.dumps(d)

# Convert JSON to dict
d2 = json.loads(mi_json)
```

Ejercicios

1) Levantar el corpus AP, separando cada noticia como un elemento distinto en un diccionario (<DOCNO> : <TEXT>)

2) Calcular el tamaño del vocabulario.

3) Para las 500 palabras con más apariciones, calcular el par más asociado según la medida presentada.

4) Repetir los ejercicios con los artículos de La Nación, levantando los textos usando JSON.