Descubrimiento de relaciones sintagmáticas: entropía

Salvador López Mendoza

Mayo de 2018

Relación sintagmática

Aparición de palabras correlacionadas.

Mi gato come pez los sábados.
Su gato come pavo los martes.
Mi perro come carne los domingos.
Su perro come pavo los martes.

¿Qué palabras suelen aparecer a la izquierda de *come*?

¿Qué palabras suelen aparecer a la derecha de *come*?

PEUVI

Predicción de palabras

¿La palabra *w* está presente o ausente en este segmento?

Un *segmento de texto* es cualquier unidad dentro del texto.

Puede ser una oración, párrafo o documento.

¿Algunas palabras son más fáciles de predecir que otras?

- *carne*.
- *una*.
- *unicornio*.

PEUVI

Definición formal

Se define X_w como una variable aleatoria: $X_w \in \{0, 1\}$

La variable $X_w = 1$ indica que la palabra w está presente.

La variable $X_w = 0$ indica que la palabra w está ausente.

$$p(X_w = 1) + p(X_w = 0) = 1$$

Mientras que la variable X_w sea más aleatoria, la dificultad en la predicción aumenta.

¿Cómo se mide cuantitativamente la *aleatoriedad* de una variable aleatoria como X_w?

PEUVI

Entropía

La entropía ($H(X)$) mide la aleatoriedad de la variable aleatoria X.

$$H(X_w) = \sum_{v \in \{0,1\}} -p(X_w = v) log_2 p(X_w = v)$$

$$H(X_w) = -p(X_w = 0) log_2 p(X_w = 0) - p(X_w = 1) log_2 p(X_w = 1)$$

¿Para cuáles X_w $H(X_w)$ alcanza su máximo/mínimo?

Lanzar una moneda

La entropía $H(X)$ es como lanzar una moneda.

$$H(X_{moneda}) = -p(X_{moneda} = 0)log_2 p(X_{moneda} = 0) - p(X_{moneda} = 1)log_2 p(X_{moneda} = 1)$$

$X_{moneda} = 1$, indica que el resultado es *águila*.

$X_{moneda} = 0$, indica que el resultado es *sol*.

Si la moneda es equitativa, $p(X = 1) = p(X = 0) = 1/2$

$$H(X) = -1/2log_2 1/2 - 1/2log_2 1/2 = 1$$

Si la moneda está cargada a un lado, $p(X = 1) = 1$

$$H(X) = -0 * log_2 0 - 1 * log_2 1 = 0$$

PEUVI

Entropía para la predicción de palabras

La palabra w, ¿está presente (o ausente) en este segmento?

Si w es *carne*, o *el* o *unicornio*, ¿cuál tiene un valor mayor o menor?

¿$H(X_{carne})$, $H(X_{el})$ o $H(X_{unicornio})$?

$H(X_{el}) \approx 0$, no hay *incertidumbre* porque $p(X_{el}) \approx 1$

Las palabras con entropía alta son difíciles de predecir.

PEUVI