Modelos vectoriales en minería de textos y sus aplicaciones a la elaboración del perfil de autores

María De Arteaga

Universidad Nacional de Colombia
Director: Rodrigo de Castro Korgi

24 de Junio de 2013

Las palabras como dimensiones

- La representación de documentos como vectores permite analizar textos con principios matemáticos.
- Para este caso, cada palabra contribuye en una unidad a la dimensión, de modo que la cardinalidad del espacio será igual al número de palabras en el diccionario construido a partir del corpus.

Construcción binaria

- Solo se toma en cuenta si cada una de las palabras está o no presente en cada documento.
- T: conjunto de los documentos, D: diccionario que se construye a partir de este conjunto, w_i: i-ésima palabra del diccionario D, $|D| = n$, $t \in T$, y D_t el diccionario que se construye a partir de las palabras que ocurren en el documento t.

$$f : T \to \{0,1\}^n,$$

$$f(t) = (v_1, v_2, ..., v_n),$$

$$\text{donde } \begin{cases} v_i = 1 & \text{si} \quad w_i \in D_t, \\ v_i = 0 & \text{si} \quad w_i \notin D_t. \end{cases}$$

Construcción con frecuencia de término e inverso de frecuencia de documento

Si:

- $|T| = N$, $t_j \in T$.
- $tf_{i,j}$: número de veces que aparece la palabra w_i en el documento j.
- df_i: número de documentos en los que aparece w_i.

$f : T \to \mathbb{R}^n$,

$f(t_j) = (v_1, v_2, ...v_n)$ para $j = 1, 2, ..., N$ donde:

$$v_i = \begin{cases} [1 + \log(tf_{i,j})][\log(\frac{N}{df_i})] & \text{si} \quad tf_{i,j}, df_{i,j} \neq 0, \\ 0 & \text{si} \quad tf_{i,j} = 0 \text{ ó } df_i = 0. \end{cases}$$

Construcción con atributos predeterminados

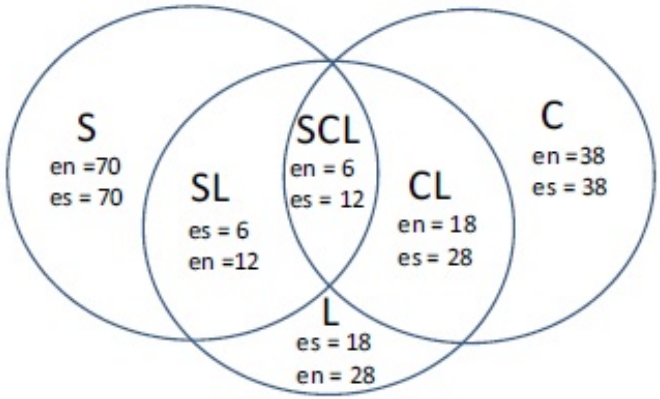
- Se determina una lista de atributos numéricos que contengan información relevante del texto, de modo que cada uno aporte una dimensión al modelo.
- Ventajas:
  - Permite trabajar con corpus y diccionarios de mayor dimensión.
  - Puede optimizar resultados si de antemano se sabe lo que se está buscando.
- Desventaja: Es posible que existan características que la máquina pueda inducir con el primer método y en este caso queden omitidas.

La elaboración de perfil de autores

- En este trabajo se presenta un laboratorio desarrollado por la autora, junto con el grupo de investigación *Informática aplicada a lingüística* de la Universidad Nacional de Colombia, para la conferencia PAN-CLEF 2013.
- El problema de perfil de autores busca ser capaz de determinar características de la persona que escribió un texto, como la edad, el género, su lengua materna y rasgos de su personalidad.
- En este caso en particular, el objetivo es determinar el género y rango de edad según tres opciones: 10's (13-17), 20's (23-27), 30's (33-47).

El modelo vectorial

- Se construyó un modelo a partir de atributos predeterminados.
  - Español: 198 atributos.
  - Inglés: 166 atributos.
- Los atributos pertenecen a tres tipos de categorías: estilísticas (S), lexicones (L) y estadísticas de corpus (C). Los atributos pueden pertenecer a solo una categoría o a más de una a la vez.



Atributos de estilo

- Medidas básicas:
  - Longitud del documento.
  - Número de palabras diferentes utilizadas en el texto.
  - Densidad del vocabulario, es decir, $\frac{v_2}{v_1}$.
  - Largo promedio de las palabras.
  - Número de palabras con 1,2 o 3 caracteres.
  - Palabras con más de 6 caracteres.
- Hapax Legomenon
- Uso de caracteres
  - Densidad de caracteres 'reconocidos'.
  - Mayúsculas.
  - Minúsculas.
  - Ocurrencias de cada caracter.

Atributos de estadísticas de corpus

- **Supervisados** Toman en cuenta las etiquetas de los documentos. Es decir, utilizan estadísticas de cada una de las categorías demográficas.

- **No supervisados** Solo utilizan información del documento y del corpus general.

Puntuación de género

Medida diseñada por el grupo de investigación para este caso en particular, busca dar un criterio que determine si un texto fue escrito por un hombre o una mujer.

$$\text{Si } G_f(t) > 0 \Rightarrow \text{Autor es hombre.}$$
$$\text{Si } G_f(t) < 0 \Rightarrow \text{Autor es mujer.}$$

Puntuación de género de la palabra w

$$g_f(w) = \frac{f_h(w)}{W_h} - \frac{f_m(w)}{W_m} \qquad g_{df}(w) = \frac{df_h(w)}{W_h} - \frac{df_m(w)}{W_m}$$

Puntuación de género del texto t

$$G_f(t) = \sum_{w \in t} [g_f(w) * f_t(w)] \qquad G_{df}(t) = \sum_{w \in t} [g_{df}(w) * f_t(w)]$$

Teorema de Bayes

Aplica el teorema para hallar la probabilidad de una categoría demográfica C dada una palabra w.

Puntuación Bayes del texto t con respecto a la categoría C

$$PB_f^C(t) = \sum_{w \in t} P_f(C|w) * f_t(w)$$

$$PB_{df}^C(t) = \sum_{w \in t} P_{df}(C|w) * f_t(w)$$

Entropía

En el análisis de texto, la entropía es una medida de la riqueza del vocabulario del documento. A mayor entropía, mayor riqueza en la forma de escritura.

Entropía de la palabra w (C puede ser el corpus total)

$$e_f^C(w) = P_f^C(w) * \log_2(P_f^C(w)) \quad e_{df}^C(w) = P_{df}^C(w) * \log_2(P_{df}^C(w))$$

Entropía del texto t

$$E_f^C(t) = \sum_{w \in t} e_f^C(w) * f_t(w) \qquad E_{df}^C(t) = \sum_{w \in t} e_{df}^C(w) * f_t(w)$$

Entropía cruzada

La entropía cruzada mide la predictibilidad de un documento si se toma como modelo el corpus de C.

Funciones utilizadas

$$p_t(w) = \frac{f(w)}{\sum_{w \in t} f(w)} \qquad q_t(w) = \frac{f_t(w)}{|t|}$$

Entropía cruzada de t con respecto a C (C puede ser el corpus general)

$$H^C(t) = \sum_{w \in t} -p_t^C(w) * \log_2(q_t(w))$$

Divergencia Kullback-Leibler

- Mide qué tan representativo es determinado texto del corpus general o de una categoría.
- La divergencia KL mide la pérdida de información si se toma un texto t como representante de la categoría C.

Divergencia KL de t con respecto a C

$$KL^C(t) = \sum_{w \in t} \log\left(\frac{p_t^C(w)}{q_t(w)}\right) * p_t^C(w)$$

Frecuencia de término e inverso de frecuencia de documento

Inverso de frecuencia de documento

$$idf(w) = \log \frac{N}{df(w)}$$

Atributos TF y TF.IDF

$$IDF(t) = \frac{\sum_{w \in t} idf(w)}{W(t)}$$

$$TF.IDF(t) = \frac{\sum_{w \in t} f_t(w) \cdot idf(w)}{W(t)}$$

Atributos de lexicones

- Una serie de lexicones se utiliza para generar cuatro atributos cada uno:
  - Densidad.
  - Densidad ponderada.
  - Entropía según frecuencia de término.
  - Entropía según frecuencia de documento.
- Lexicones utilizados:
  - Temáticos.
  - Indicativos de estilo (groserías, palabras propias de Internet, diccionarios oficiales del idioma, palabras vacías).
  - Diccionarios de emociones de Sidorov.
  - Generados por la prueba T de Student.

Prueba T de Student

La prueba T de Student permite determinar las palabras más representativas de cada categoría demográfica.

Funciones utilizadas

$$\overline{x}_C = \frac{f_C(w)}{W_C} \qquad S_C = \sqrt{\overline{x}_C - \overline{x}_C{}^2}$$

Significancia de género y significancia de edad

$$T_g = \frac{\overline{x}_m - \overline{x}_f}{\sqrt{\frac{S_m{}^2}{N_m} + \frac{S_f{}^2}{N_f}}} \qquad T_e = \frac{\overline{x}_e - \overline{x}}{\sqrt{\frac{S_e{}^2}{N_e} + \frac{S^2}{N}}}$$

Aprendizaje de máquinas

- "El aprendizaje de máquinas proporciona la base técnica de minería de datos. Se utiliza para extraer información de los datos en bruto en bases de datos, información que es expresada en una forma comprensible y puede ser utilizada para una variedad de propósitos. El proceso es uno de abstracción: tomar los datos, con todos sus defectos, e inferir cualquiera que sea la estructura que subyace" (Witten and Frank).
- Se utilizó un modelo de regresión multinomial con estimadores de Ridge de Weka.

Resultados oficiales

El programa obtuvo el 6to lugar entre 20 para el caso de español.

Cuadro : Resultados español

	Total	Género	Edad
Ganador	0.4208	0.6473	0.6430
Resultados	0.3145	0.5627	0.5429
Línea de base	0.1650	0.5000	0.3333

Cuadro : Resultados inglés

	Total	Género	Edad
Ganador	0.3894	0.5921	0.6491
Resultados	0.2450	0.4998	0.4885
Línea de base	0.1650	0.5000	0.3333

Experimentos internos

Cuadro : Precisión promedio de los tres tipos de atributos.

Tipo de atributo	Género *es*	Edad *es*
Estadísticas	0.8038 (0.0007)	0.7866 (0.0004)
Lexicones	0.6261 (0.0007)	0.6446 (0.0006)
Estilísticos	0.5981 (0.0008)	0.6336 (0.0009)
Todos	0.8202 (0.0013)	n/a

Tipo de atributo	Género *en*	Edad *en*
Estadísticas	0.8393 (0.0005)	0.7860 (0.0013)
Lexicones	0.5933 (0.0010)	0.6198 (0.0003)
Estilísticos	0.5502 (0.0012)	0.6048 (0.0003)
Todos	0.8477 (0.0023)	0.7809 (0.0002)

Cuadro : Precisión promedio de las subcategorías estadísticas.

Tipo de atributo	Género *es*	Edad *es*
Bayes	0.7696 (0.0002)	0.7677 (0.0003)
Entropía cruzada	0.5376 (0.0006)	0.5624 (0.0004)
Divergencia KL	0.5896 (0.0005)	0.5952 (0.0007)
Lexicones TT	0,6240 (0.0005)	0.6377 (0.0003)
Todos	0.8202 (0.0013)	n/a

Cuadro : Precisión promedio de las subcategorías estadísticas.

Tipo de atributo	Género *en*	Edad *en*
Bayes	0.7951 (0.0004)	0.7382 (0.0015)
Entropía cruzada	0.5527 (0.0008)	0.5891 (0.0006)
Divergencia KL	0.5485 (0.0005)	0.6034 (0.0003)
Lexicones TT	0.5863 (0.0006)	0.6204 (0.0004)
Todos	0.8477 (0.0023)	0.7809 (0.0002)

Bibliografía

1 María De-Arteaga, Sergio Jimenez, George Dueñas, Sergio Mancera, Julia Baquero. *Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features*. (2013).

2 Manning, Christopher D., and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

3 S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009), *Automatically profiling the author of an anonymous text*, Communications of the ACM 52 (2) :119 – 123.

4 G. Sidorov, S. Miranda-Jiménez, F. Viveros-JimÃ©nez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, J. Gordon. *Empirical Study of Opinion Mining in Spanish Tweets*. LNAI 7629-7630, 2012, 14 p.

Bibliografía

5 J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006.

6 M.Koppel, S. Argamon and A. Shimoni (2003), Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17(4), November 2002, pp. 401-412.