Analizando el Español de Argentina en Twitter

Damian Aleman - Santiago Kalinowski

Santiago Kalinowski

Doctor en Estudios Hispánicos de la University of Western Ontario

Director del Departamento de Investigaciones Lingüísticas y Filológicas (DILyF), Academia Argentina de Letras (AAL).

Damián Aleman

Licenciado en Ciencias de la Computación (Universidad de Buenos Aires)

Data Scientist en Grandata

¿Qué hacemos con la AAL?

- Estudiar uso del Español de Argentina en Internet
- Aplicado principalmente a Lexicografía (construcción de diccionarios)

Diccionario Del Habla De Los Argentinos

Una pequeña inspiración

Berta Vidal de Battini (1964) en “El Español de Argentina” demarcó cinco regiones dialectales en Argentina:

1. Norteño
2. Guaranítico
3. Cuyano
4. Cordobés/Central
5. Rioplatense

Detección de regionalismos

- Regionalismo: palabra/expresión utilizada en un región geográfica particular de una lengua
- La detección y registro de estos se suele hacer de manera muy manual (encuestas, cuestionarios)

Ejemplos:

culeado, patrá, chomaso, waso, teres, locasa, mavale, chombi, carnasas, cequin, visitarle, angá

Elección de Corpus

- Creado por RAE
- Textos orales y escritos de España, América, Filipinas y Guinea Ecuatorial (subrepresentados!)
- Sólo se puede acceder a través de una interfaz web

Mejor Twitter

- Buscamos usuarios con location encendida
- Agrupamos por provincia y no por ciudad
- Normalización muy básica

	Total	Per province
Words (in millions of words)	647	28.14 +- 6.64
Tweets (in millions of tweets)	80.9	3.51 +- 0.91
Users (in thousands of users)	56.2	2.44 +- 0.04
Vocabulary (in millions of words)	7.5	0.32 +- 0.04

Métrica para regionalismos

Dada una palabra, tenemos ahora la cantidad de ocurrencias y usuarios que la utilizan por provincia.

Para calcular su “dispersión”, usamos la entropía (tanto de word counter como user count)

$$H(w) = -\sum_{i=1} p(w_i) \log(p(w_i))$$

Si H(w) = 0 => palabra totalmente concentrada en una provincia

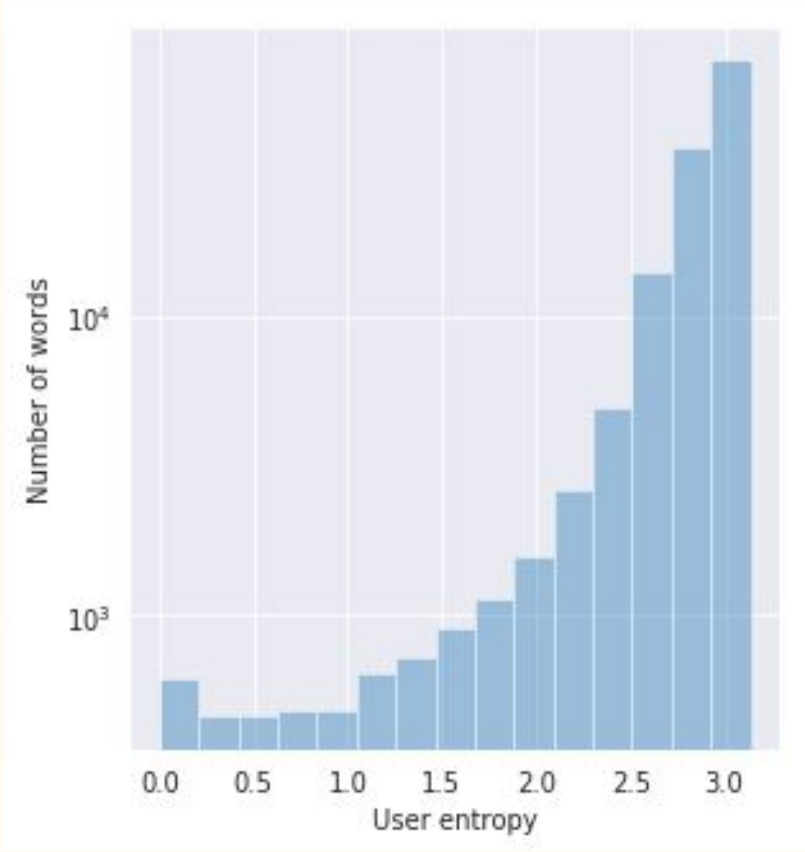
Si H(w) = log(23) (máximo) => palabra usada uniformemente en todas las provincias

Métrica para regionalismos (II)

Definimos la información de w como

$$I(w) = n(w)(M - H(w))$$

Donde n(w) es el logaritmo de la cantidad de ocurrencias normalizado y M la entropía máxima.



Montemurro, Marcelo A., and Damián H. Zanette. "Entropic analysis of the role of words in literary texts." *Advances in complex systems* 5.01 (2002): 7-17.

Métrica para regionalismos (III)

Definimos entonces tres métricas

1. I(w) usando la cantidad de ocurrencias
2. I(w) usando la cantidad de usuarios
3. La multiplicación de 1. y 2.

Con esto, generamos listados de palabras de posible relevancia para los lexicógrafos.

$$H(w) = -\sum_{i=1} p(w_i) \log(p(w_i))$$

$$I(w) = n(w)(M - H(w))$$

Validación lexicográfica

Una vez generados los listados con las métricas, tomamos las 1000 primeras palabras de cada una.

Los lexicógrafos etiquetaron si dichas palabras tenían alguna relevancia para su tarea.

Ranking	Palabra	Usuario	Ambas
1	ushuaia	chivil	chivilcoy
2	rioja	ush	ush
3	chivilcoy	poec	tolhuin
4	bragado	**malpegue**	blv
5	viedma	**aijue**	chivil
6	logroño	tolhuin	logroño
7	chepes	vallerga	bragado
8	oberá	yarca	vallerga
9	**cldo**	blv	breñas
10	tdf	portho	**malpegue**
11	riojanos	jumeal	**aijue**
12	breñas	**sinf**	choele
13	choele	plottier	oberá
14	gallegos	kraka	obera
15	tiemposur	fsa	**cldo**
16	fueguinos	**bombola**	plottier
17	chilecito	**yarco**	kraka
18	blv	sanagasta	**sinf**
19	ush	wika	poec
20	merlo	obera	nqn

Resultados

- Mejor métrica: contando usuarios
- Elimina palabras “raras” (de bots/spammers)

Metric	% of interesting words
Word Information Value	21.9%
User Information Value	30.2%
Mixed Information Value	25.3%

Palabras relevantes

Coloquialismos: chombi, culeado, carnasas, chuño, pororó, mimosear, abuelear

Gentilicios: casildense, concordiense, obereño

Acepciones: Atinar, mansa, asada (Cuyo)

Indigenismos: angá, mitaí (Formosa y Corrientes), ura(Tucumán), angaú

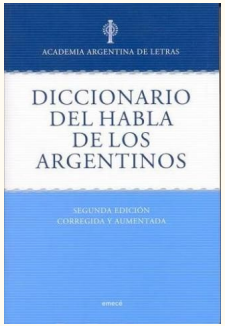
Leísmos: saludarle, visitarle

Morfológicas: Malaso, chomaso, locaso

Variantes ortográficas: pasao, qliaw, teres

Conclusiones y futuro

1. Se añadieron palabras al Diccionario del Habla de los Argentinos(de la AAL)
2. Trabajo multidisciplinario en área no muy explorada (al menos en Español)
3. A futuro: replicación en Iberoamérica, clusterización de regiones dialectales, geolocalización mediante palabras de dialectos.



¡Gracias!

daleman@dc.uba.ar

s.kalinowski@aal.edu.ar

LIAA
Laboratorio de Inteligencia Artificial Aplicada

ACADEMIA ARGENTINA DE LETRAS