Análisis de Datos con Twitter

Javier Castro M.

Universidad De Chile

July 9, 2021

LDA: "Latent Dirichlet allocation"

LDA es un modelo tipo **topic modelling**; trata de capturar temas latentes en el texto mediante la calibración de dos parámetros, $\alpha \in [0,1]^K$ y $\beta \in [0,1]^{K \times V}$. Donde K es la cantidad de tópicos y V el tamaño del vocabulario.

- Se samplea una distribución sobre tópicos $\theta \sim Dir(\alpha, K)$. Notar que $\sum \theta_i = 1$.
- Dado θ, se samplea un tópico $t \sim \theta$.
- Dado el tópico t, se samplea una palabra $w \sim \beta_t$..

LDA: "Latent Dirichlet allocation"

Para aplicar el LDA podemos usar:

- `gensim.models.LdaMulticore` (`broken pipe`)
- `gensim.models.LdaModel`

Se necesita transformar los textos en "bag of words". Los tweets procesados quedan como se ven a continuación:

```
tweet 0:  [(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 2), (6, 1), (7, 1),
(8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1)]
tweet 1:  [(15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1),
(22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1)]
```

El algoritmo entrega, para cada tópico, una distribución sobre las palabras y una medida llamada "coherencia".

Resultados

Se aplicó el algoritmo a tweets de la cuenta `@gabrielboric`, a estos se les realizo la limpieza correspondiente.

- Resultan 20902 tweets.
- La cantidad de palabras resultantes resultó ser 1853.
- Se buscarán 5 tópicos.

Resultados

```
Topic: 0
Words: 0.014*"magallanes" + 0.012*"buena" + 0.009*"mañana" + 0.008*"arenas" +
0.008*"punta" + 0.007*"gente" + 0.006*"@valenzuelalevi" + 0.006*"reunión" +
0.006*"cabros" + 0.005*"trabajadores"
Topic: 1
Words: 0.012*"ahora" + 0.012*"educación" + 0.012*"chile" + 0.010*"política" +
0.008*"vamos" + 0.008*"acuerdo" + 0.006*"gobierno" + 0.006*"izquierda" +
0.006*"fech" + 0.006*"derecha"
Topic: 2
Words: 0.013*"@giorgiojackson" + 0.007*"@izqautonoma" + 0.007*"días" +
0.006*"concerta" + 0.005*"conflicto" + 0.004*"stgo" + 0.004*"facultad" +
0.004*"buenos" + 0.004*"@camila_vallejo" + 0.004*"huelga"
Topic: 3
Words: 0.016*"abrazo" + 0.015*"gracias" + 0.009*"aguante" + 0.008*"muchas" +
0.008*"saludos" + 0.006*"estudiantil" + 0.006*"@jschaulsohn" + 0.006*"allá" +
0.006*"compa" + 0.005*"siempre"
Topic: 4
Words: 0.008*"mismo" + 0.007*"@jen_abate" + 0.006*"dice" + 0.005*"@cbellolio"
+ 0.005*"mayoría" + 0.005*"@jleytong" + 0.005*"sólo" + 0.004*"terrible" +
0.004*"@jacques_roux_v" + 0.004*"concertación"
```

Figure: Sin eliminar los @

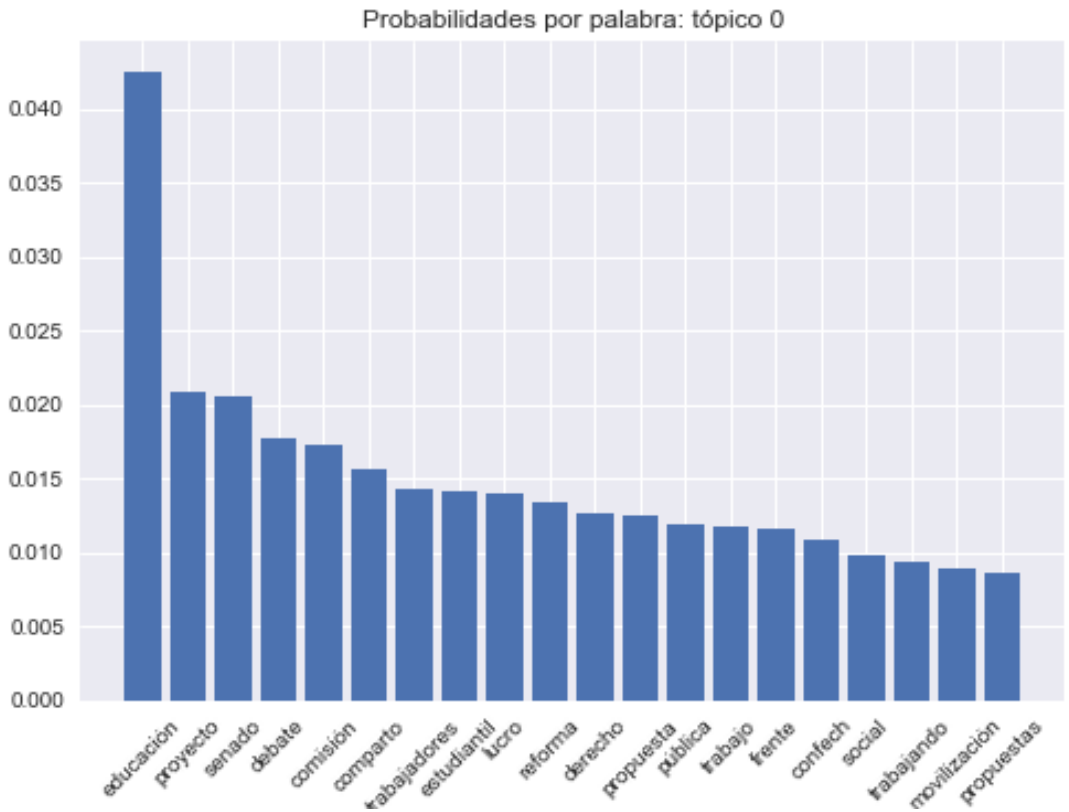
Resultados

```
Topic: 0
Words: 0.012*"chile" + 0.011*"educación" + 0.010*"política" + 0.007*"acuerdo"
+ 0.007*"ahora" + 0.007*"gobierno" + 0.006*"derecha" + 0.006*"puede" +
0.006*"proyecto" + 0.006*"senado"
Topic: 1
Words: 0.013*"buena" + 0.011*"recomiendo" + 0.009*"universidad" +
0.009*"columna" + 0.009*"comparto" + 0.008*"estudiantil" + 0.008*"entrevista"
+ 0.008*"gran" + 0.007*"lucro" + 0.007*"cabros"
Topic: 2
Words: 0.012*"izquierda" + 0.010*"fuerza" + 0.009*"movimiento" +
0.009*"trabajadores" + 0.009*"reunión" + 0.008*"mucha" + 0.007*"ahora" +
0.007*"magallanes" + 0.007*"confech" + 0.007*"nacional"
Topic: 3
Words: 0.021*"fech" + 0.007*"estimad@s" + 0.006*"tolerancia0" + 0.005*"stgo"
+ 0.005*"puerta" + 0.005*"superior" + 0.005*"financiamiento" +
0.004*"educacional" + 0.004*"onda" + 0.004*"pasar"
Topic: 4
Words: 0.019*"vamos" + 0.018*"abrazo" + 0.017*"gracias" + 0.014*"arenas" +
0.014*"punta" + 0.011*"vivo" + 0.010*"aguante" + 0.010*"muchas" +
0.008*"radio" + 0.008*"pega"
```

Figure: Eliminando los @

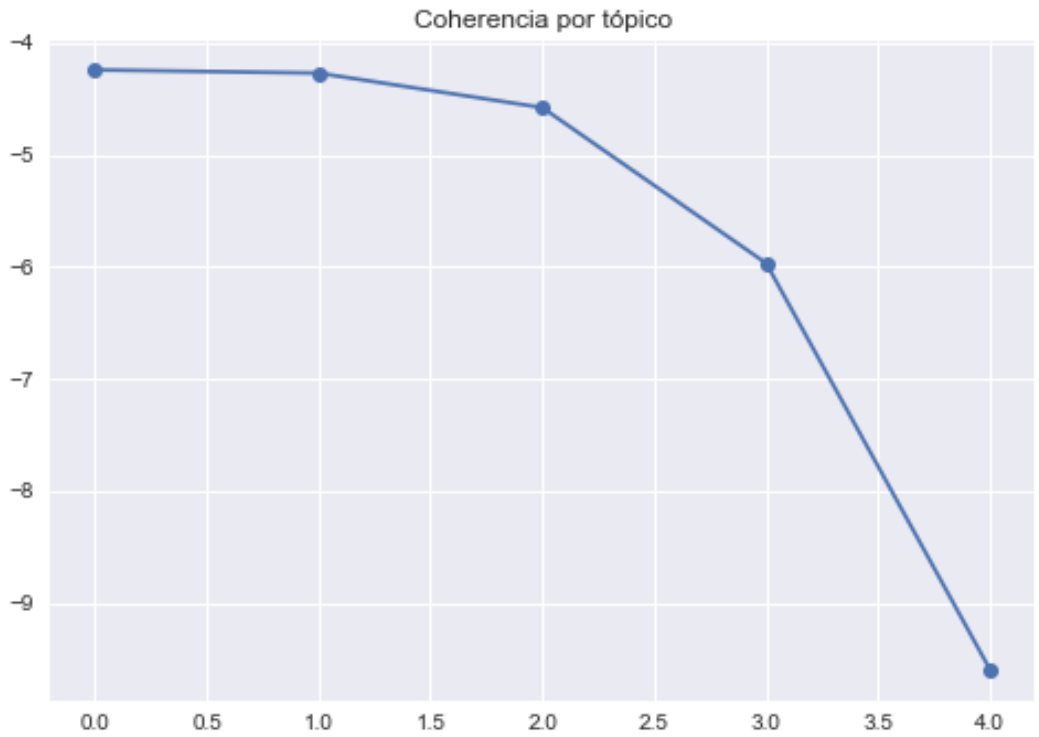
Resultados

Se muestra la distribución para el primer tópico.



Resultados

Se muestra la coherencia de cada tópico.



Ideas y lo que falta

- Calibrar los hiperparámetros en base a alguna métrica (número de tópicos, α y β).
- Hacer un limpieza más fina a los tweets (eliminar palabras de muy baja frecuencia, trabajar el problema de palabras mal escritas).
- Usar bigrams o similiares.
- Quedarse sólo con adjetivos y sustantivos (quizá sólo aplicar esto con un corpus en inglés).
- Probar lo mismo con tweets de otros contextos.

Bibliografía

- **Latent Dirichlet Allocation**, Blei, Ng, Jordan, Journal of Machine Learning Research, 2003.
- `radimrehurek.com/gensim/models/ldamodel.html`
- **Exploring the Space of Topic Coherence Measures**, Michael Röder, Andreas Both, Alexander Hinneburg, 2015.

Fin

Muchas Gracias!