16 de octubre 22

Computation of Maximum likelihood estimate

Maximiza $p(d|\theta) : (\theta_1, \ldots, \theta_M) = arg\ max_{\theta_1,\ldots,\theta_M}\ \ p(d|\theta) = arg\ max_{\theta_1,\ldots,\theta_M}\ \ \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$
Max. log-likelihood

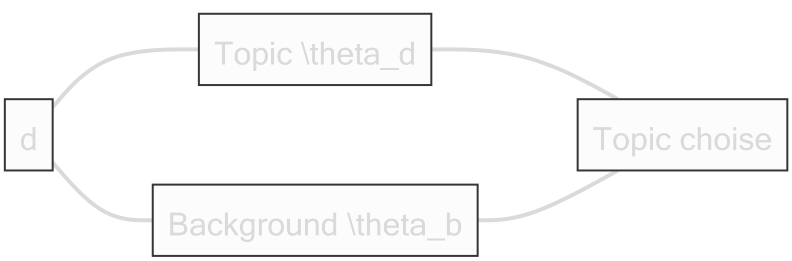
$$p(w_i|\theta) = \frac{(w_i, d)}{\sum_{i=1}^{M}(w_i, d)} = \frac{c(w_i, d)}{|d|}$$

What does the Topic look like?

d = text mining paper
can we get rid of these common words?

Generate d Using two word Distributions

d = text mining papel



where topic choise = $p(\theta_d) + p(\theta_B) = 1$ dado que $p(\theta_d) = 0.5$ y $p(\theta_B) = 0.5$

$$p(the) = p(\theta_d)\ \ p(the|\theta_d) + p(\theta_B)\ \ p(the|\theta_B) = 0.5 * 0.000001 + 0.5 * 0.03$$

$$p(tex) = p(\theta_d)\ \ p(tex|\theta_d) + p(\theta_B)\ \ p(the|\theta_B) = 0.5 * 0.04 + 0.5 * 0.04$$

Formally defines the following generative model:

$$w \to p(w) = p(\theta_d)\ \ p(w|\theta_d) + p(\theta_B)\ \ p(w|\theta_B)$$

what if p = 1

Likelihood function:

$$p(d|\Lambda) = \prod_{i=1}^{|d|} p(x_i|\Lambda) = \prod_{i=1}^{|d|} [p(\theta_d)\ p(x_i|\theta_d) + p(\theta_B)\ p(x_i|\theta_B)]$$

Ecuacion lineal:

$$0.5 * p(text|\theta_d) + 0.5 * 0.1 = 0.5 * p(the|\theta_d) + 0.5 * 0.9$$

quedando p(text|θ_d) = 0.9 >> the = p(the|θ_d) = 0.1

from θ_d $(Z=0)$ $p(\theta_d)$ $p(text|\theta_d)$
para calcular si la palabra esta en θ_d o en θ_B usamos una variable Z
θ_d	θ_B
$z=0$	$z=1$

p(text|θ_d)
p(text|θ_B)

$$p(z=0|w=text) = \frac{p(\theta_d)\ p(text|\theta_d)}{p(\theta_d)\ p(text|\theta_d) + p(\theta_B)p(text|\theta_B)}$$

The expectation-Maximization (EM) Algoithm

$$p^n = \frac{p(\theta_d)p^n(w|\theta_d)}{p(\theta_d)p^{(n)}(w|\theta_d) + p(\theta_B)p(w|\theta_B)} \quad \rightarrow E-step$$

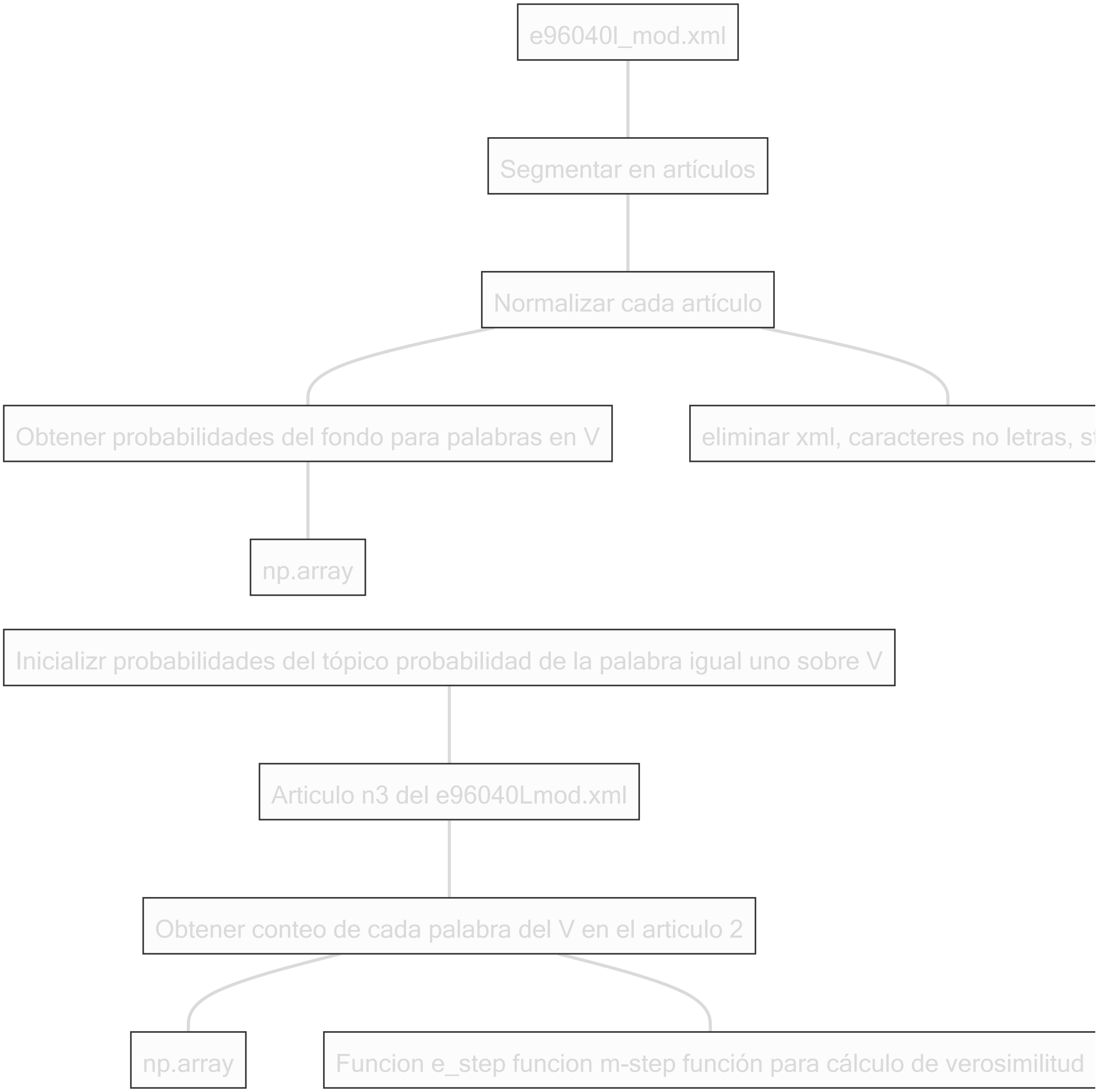
how likely w is from θ_d

$$p^{(n+1)}(w|\theta_d) = \frac{c(w,d)p^{(n)}(z=0|w)}{\sum_{w' \in V} c(w',d)p^{(n)}(z=0|w')} \quad \rightarrow M-step$$

Assume $p(\theta_d) = p(\theta_B) = 0.5$ and $p(w|\theta_B)$ is kwnown

	conteos de palabrea de en articulo 2	probabilidad de palabras de fondo	Probabilidad de palabras del tópico					
word	#	p(w| θ_B)	Iteracion 1		Iteracion 2		Iteracion 3	
			P(w| θ)	p(z=0| w)	p(w| θ)	p(z=0| w)	p(w| θ)	p(z=0| w)
The	4	0.5	**0.25**	0.33	**0.20**	0.29	**0.18**	0.26
Paper	2	0.3	**0.25**	0.45	**0.14**	0.32	**0.10**	0.25
Text	4	0.1	**0.25**	0.71	**0.44**	0.81	**0.50**	0.93
Mining	2	0.1	**0.25**	0.71	**0.22**	0.69	**0.22**	0.69
Long-likelihood				-16.96		-16.13		-16.02

17 de octubre del 2022



Funcion de verosimilitud:

$log(p(articulo_2|modelo)) = \sum_{i=1}^{|V|}(conteodew_i) + log(p(\theta_B) * p(W_2|\theta_B) + p(\theta_d) * p(w_i|\theta_d) \quad \rightarrow algoritmoEM(tr$

Ordenar probabilidades del fondo y del topico e imprimir las primeras 10

```
f = open('articule_lemmatize')
text = f.read()
f.close()

words = nltk.word_tokenize(text)

count = []
```

```
for voc in vocabulary count
```