

Predicción de calificación en reseñas de Goodreads

Competencia de Kaggle

- Facultad de Ciencias Exactas y Naturales
 - Ignacio J. López, Patricio Sanchez y Kenneth Syddall
-

goodreads



Summer Reading



Deciding what to read next?





You're in the right place. Tell us what titles or genres you've enjoyed in the past, and we'll give you surprisingly insightful recommendations.

What are your friends reading?


Chances are your friends are discussing their favorite (and least favorite) books on Goodreads.

What will you discover?

Because  Meagan  liked...



She discovered:





Historical Fiction, Book Club


News & Int


45 New Books

Discover & read more

 Continue with Facebook

 Continue with Amazon





By creati
Goodread

Alre

Readers' Most Anticipated Summer Books

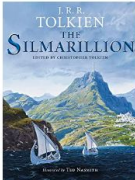
Discover these eagerly awaited reads! >

goodreads


Home My Books Browse ▼ Community ▼

Search books

Bilbo Bolson's Reviews > The Silmarillion




Want to Read ▼




★★★★★

review

Sep 07, 2011

 Share on Facebook


Warning, this book is for Tolkien junkies only. It is not for casual readers of Tolkien... not--the Hobbit was kinda fun, wasn't Bilbo cute--sort of readers. In fact I believe it might be prerequisite that in order to enjoy The Silmarillion, one must have read The Lord of the Rings a minimum of three times. I am one such dedicated dweeb so I love it.


40 likes · Like ·  flag


Sign into Goodreads to see if any of your friends have read *The Silmarillion*.

Sign In »

READING PROGRESS

 September 7, 2011 - Started Reading

 September 7, 2011 - Shelved

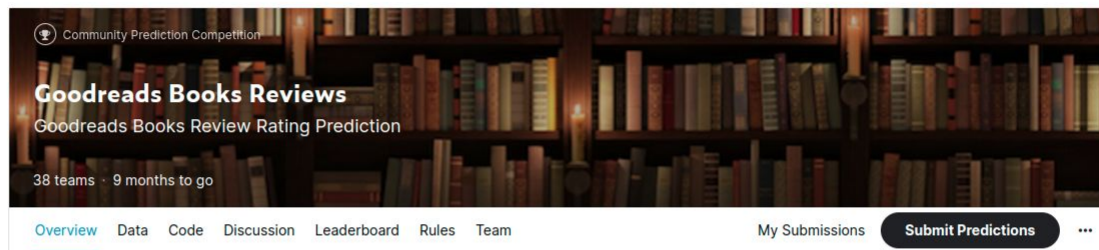
 September 13, 2011 - Finished Reading

COMMENTS Showing 1-12 of 12 (12 new)

Post a comment »

Dataset

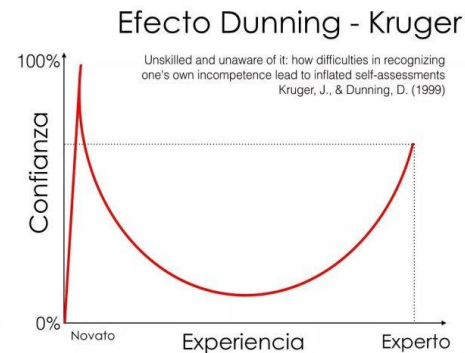
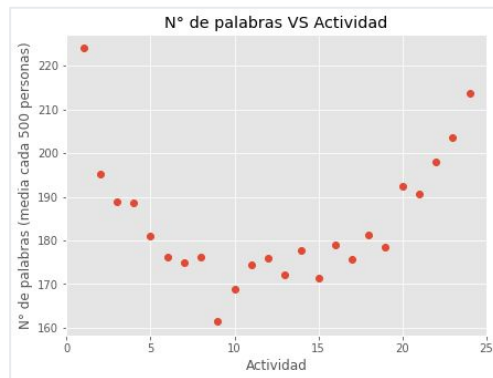
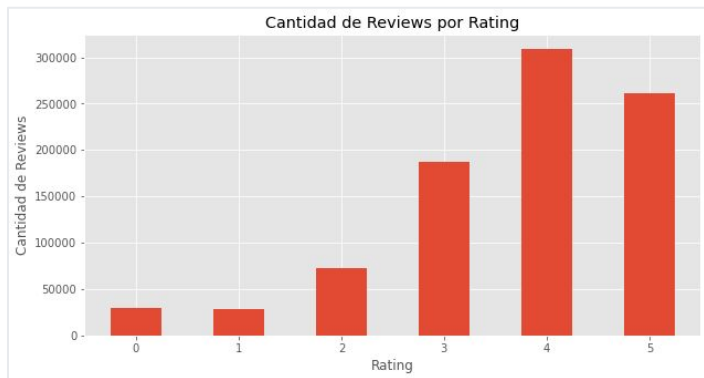
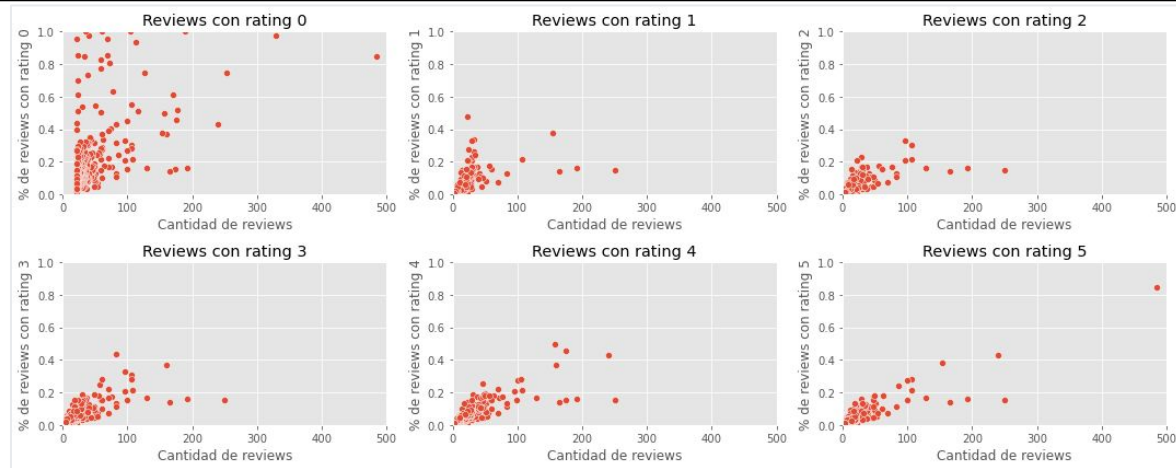
- 1.3 M de reviews totales
- 900k de reviews en el train set
- 478k de reviews en el test set



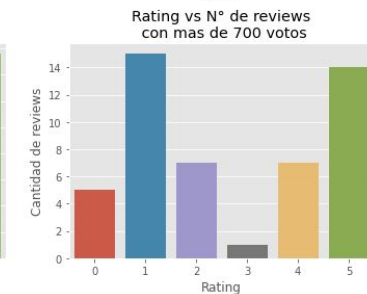
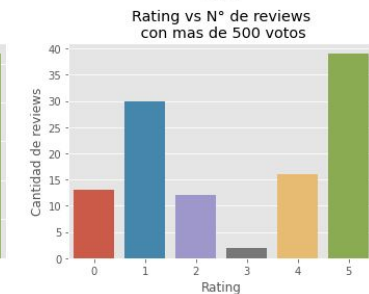
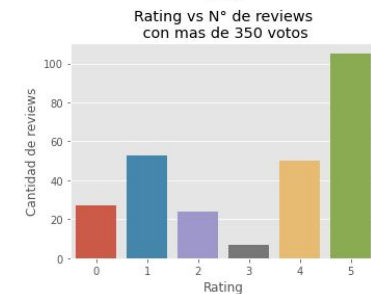
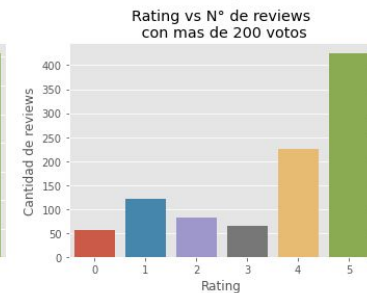
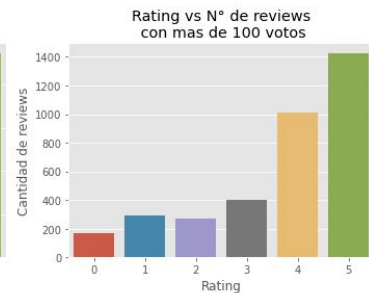
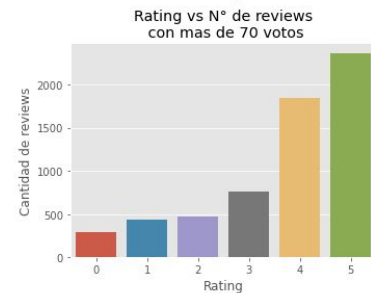
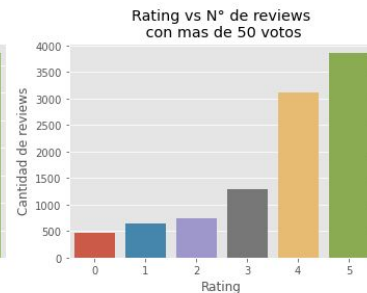
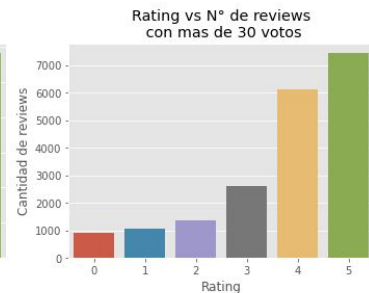
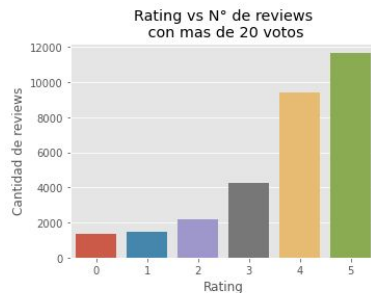
kaggle

	user_id	book_id	review_id	rating	review_text	date_added	date_updated	read_at	started_at	n_votes	n_comments
0	8842281e1d1347389f2ab93d60773d4d	18245960	dfdbb7b0eb5a7e4c26d59a937e2e5feb	5	This is a special book. It started slow for ab...	Sun Jul 30 07:44:10 -0700 2017	Wed Aug 30 00:00:26 -0700 2017	Sat Aug 26 12:05:52 -0700 2017	Tue Aug 15 13:23:18 -0700 2017	28	1
1	8842281e1d1347389f2ab93d60773d4d	16981	a5d2c3628987712d0e05c4f90798eb67	3	Recommended by Don Katz. Avail for free in Dec...	Mon Dec 05 10:46:44 -0800 2016	Wed Mar 22 11:37:04 -0700 2017	NaN	NaN	1	0
2	8842281e1d1347389f2ab93d60773d4d	28684704	2ede853b14dc4583f96cf5d120af636f	3	A fun, fast paced science fiction thriller. I ...	Tue Nov 15 11:29:22 -0800 2016	Mon Mar 20 23:40:27 -0700 2017	Sat Mar 18 23:22:42 -0700 2017	Fri Mar 17 23:45:40 -0700 2017	22	0
3	8842281e1d1347389f2ab93d60773d4d	27161156	ced5675e55cd9d38a524743f5c40996e	0	Recommended reading to understand what is goin...	Wed Nov 09 17:37:04 -0800 2016	Wed Nov 09 17:38:20 -0800 2016	NaN	NaN	5	1
4	8842281e1d1347389f2ab93d60773d4d	25884323	332732725863131279a8e345b63ac33e	4	I really enjoyed this book, and there is a lot...	Mon Apr 25 09:31:23 -0700 2016	Mon Apr 25 09:31:23 -0700 2016	Sun Jun 26 00:00:00 -0700 2016	Sat May 28 00:00:00 -0700 2016	9	1

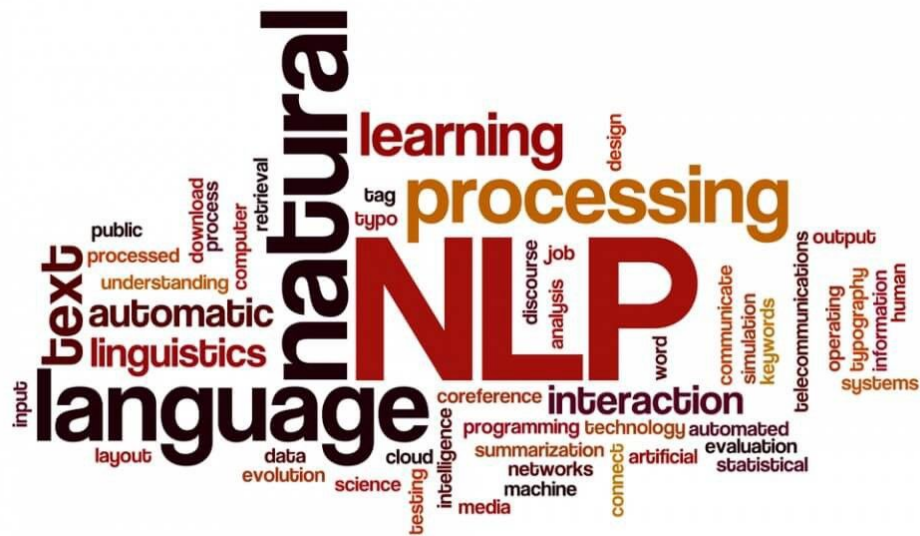
Análisis Preliminar



Análisis Preliminar



Modelos de NLP

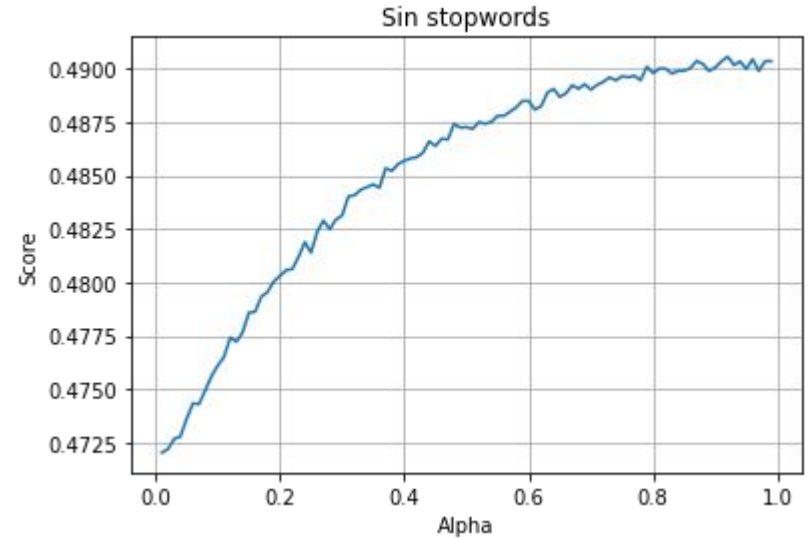
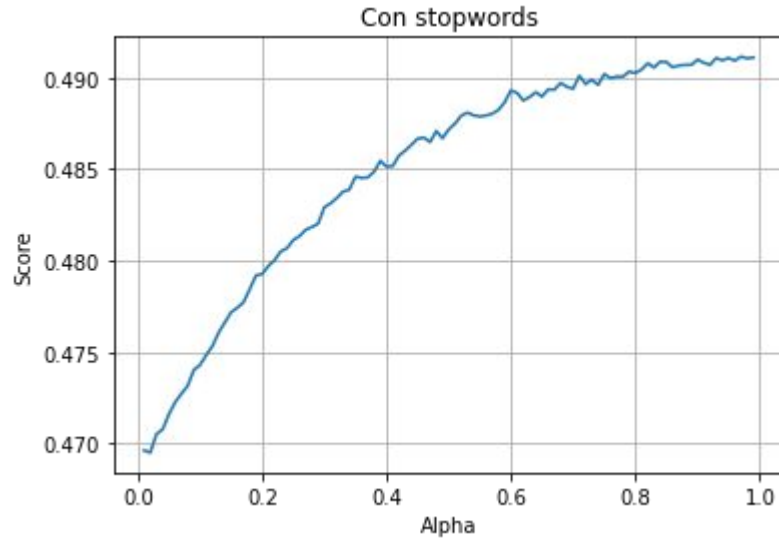


Naive Bayes

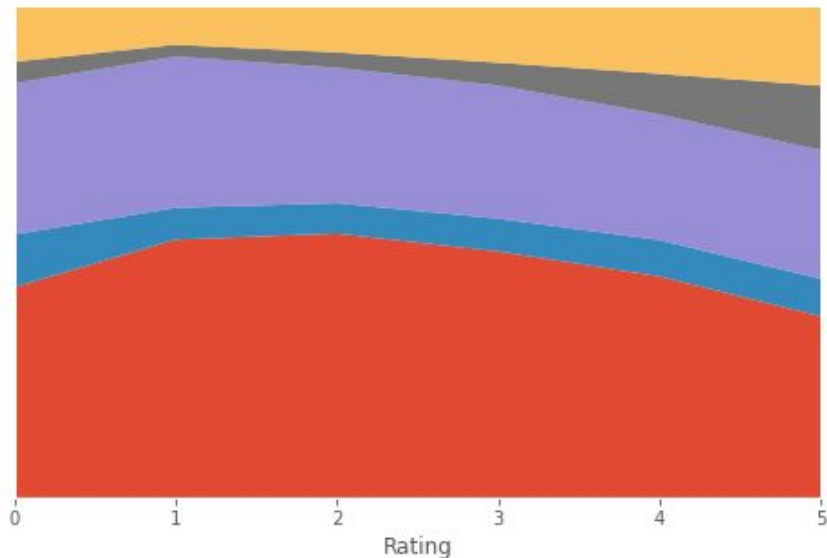
F1-Score: 0.49

Suavizado Laplaciano. $\alpha = 0,9$. $\frac{x_i}{N} \rightarrow \frac{x_i + \alpha}{N + \alpha K}$

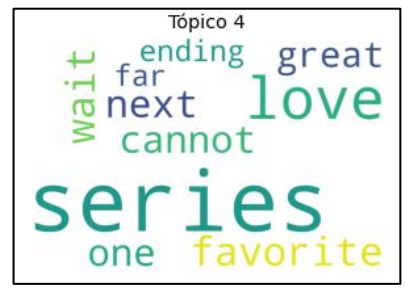
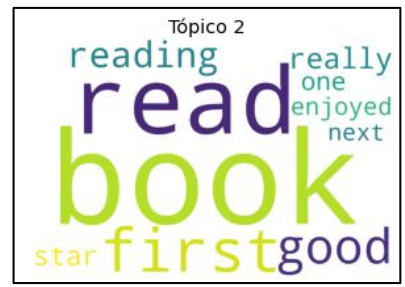
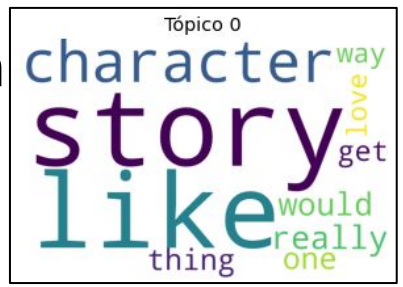
Validación cruzada.



Non-Negative Matrix Factorization



F1-Score: 0.36



Transformer RoBERTa











RoBERTa es un modelo pre-entrenado y reconfigurado de BERT.
Los cambios respecto a BERT son:

- Se entrenó al modelo por más tiempo, con muchos más datos.
- Eliminar la capacidad de predecir la siguiente palabra.
- Se entrenó con secuencias de texto más largas.
- Se entrenó enmascarando partes del texto de forma dinámica.

El modelo que utilizamos se le hizo un fine-tuning para clasificar texto como positivo, negativo o neutro.

F1-Score: 0.34

Resultados - Kaggle

#	Team	Members	Score	Entries	Last
1	martinerrazquin		0.63486	3	1mo
13	Zeyad Abdelreheem		0.53668	1	1mo
14	Maxine Attobrah		0.50856	8	2mo
15	Thành Minh		0.50687	3	1mo
16	Lopez_Sanchez_Syddall	  	0.48429	3	3m
 Your Best Entry! Your most recent submission scored 0.48429, which is an improvement of your previous score of 0.46218. Great job!					
17	cacc14		0.48398	3	6d
18	Jun Chen		0.47365	1	2mo

Top 43%

16/38

Conclusiones

- Naive-Bayes es el modelo de mayor performance (F1-Score=0.49).
 - El modelo aparentemente más sofisticado (RoBERTa) y reciente no obtuvo una performance competitiva (F1-Score=0.38).
-

Cómo seguimos

- Análisis preliminar más profundo.
 - LSA(Limitación en la RAM).
 - Transformer con toda la data (GPU Limitado).
 - Buscar hiperparametros.
-