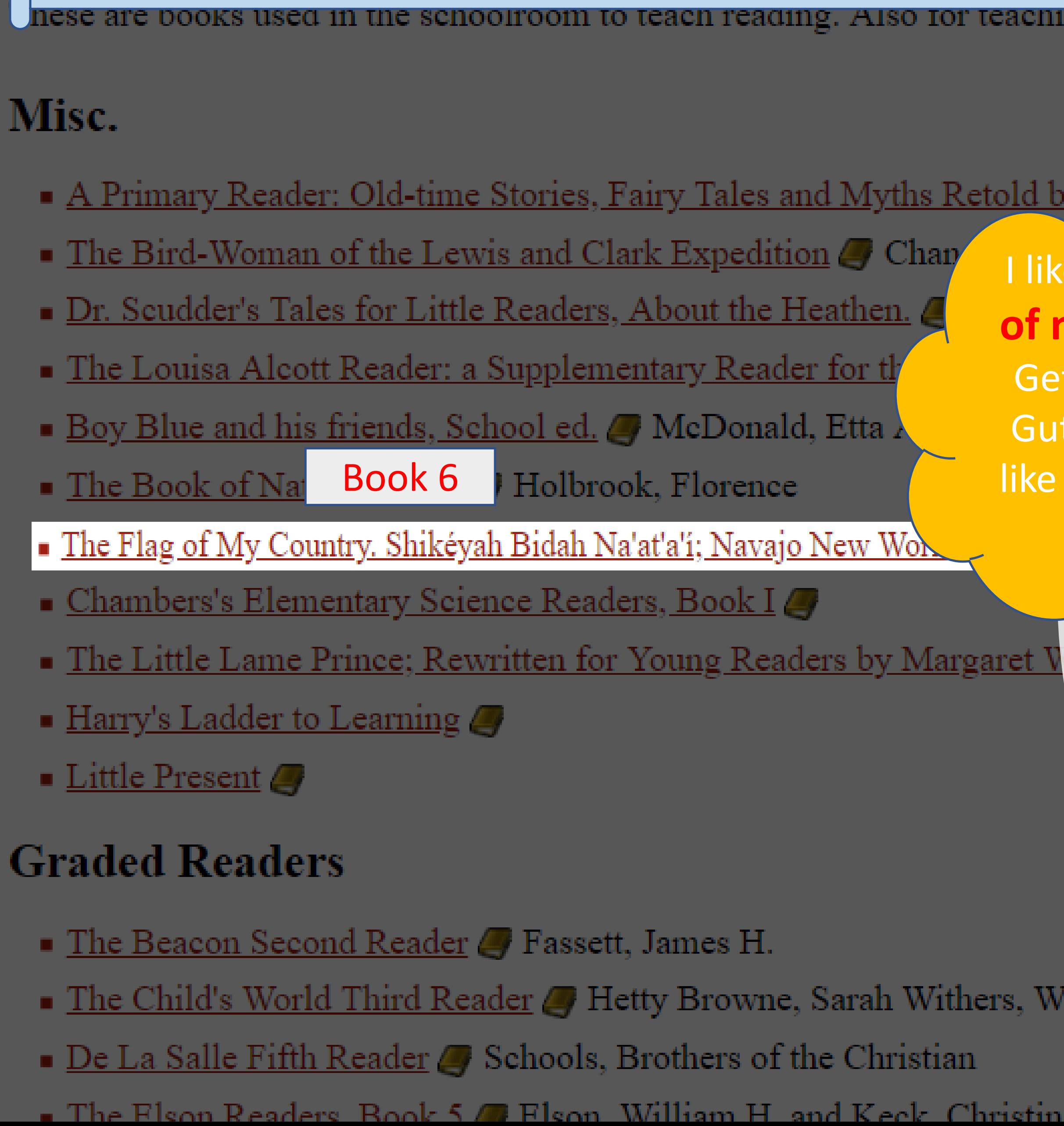
Topic Modeling – Gutenberg Books

Kevin Okiah

4/15/2019

Problem Statement – Find a book Similar to “The Flag of My Country”?



hese are books used in the schoolroom to teach reading. Also for teaching older children abou ... der as yet.

Misc.

- A Primary Reader: Old-time Stories, Fairy Tales and Myths Retold by Childr
- The Bird-Woman of the Lewis and Clark Expedition Chan
- Dr. Scudder's Tales for Little Readers, About the Heathen.
- The Louisa Alcott Reader: a Supplementary Reader for th
- Boy Blue and his friends, School ed. McDonald, Etta A
- The Book of Na Holbrook, Florence
- The Flag of My Country. Shikéyah Bidah Na'at'a'í; Navajo New Wor
- Chambers's Elementary Science Readers, Book I
- The Little Lame Prince; Rewritten for Young Readers by Margaret Wa
- Harry's Ladder to Learning
- Little Present

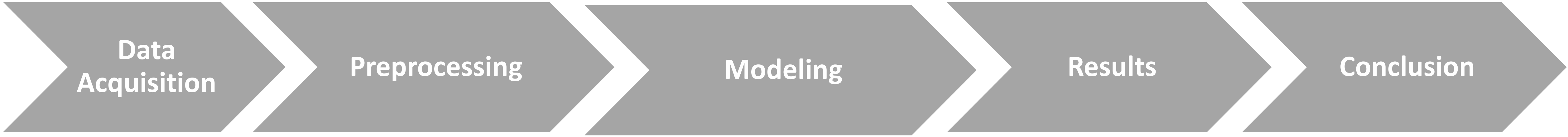
Book 6

Graded Readers

- The Beacon Second Reader Fassett, James H.
- The Child's World Third Reader Hetty Browne, Sarah Withers, W.K. Tate
- De La Salle Fifth Reader Schools, Brothers of the Christian
- The Elson Readers, Book 5 Elson, William H. and Keck, Christine M.

I liked “The Flag of my Country”, Get me another Gutenberg book like that and I will read?

NLP Pipeline Approach



Data Acquisition

Preprocessing

Modeling

Results

Conclusion

104 Books Scrapped from Gutenberg Website

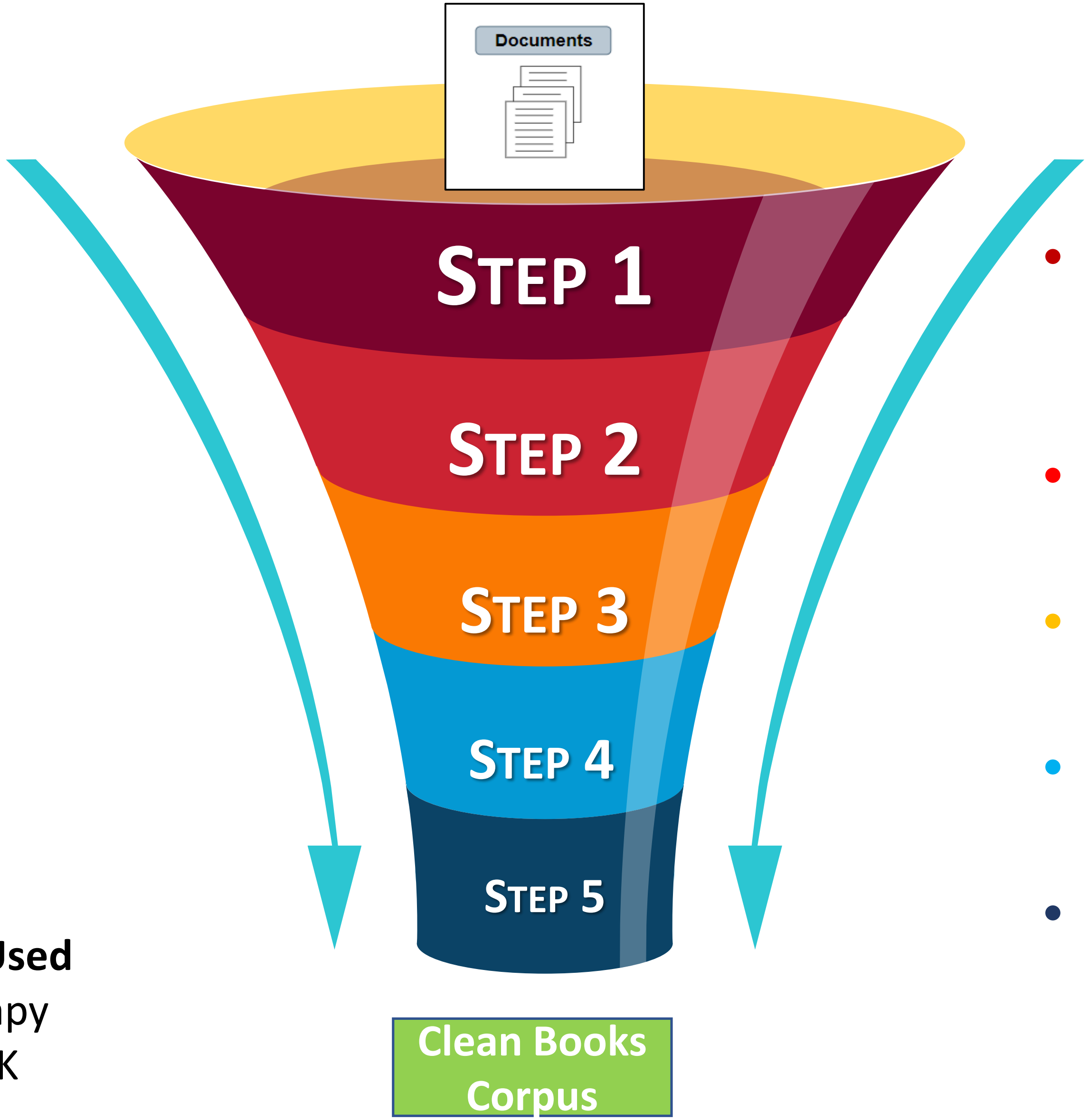
Proceed with Caution

Web Scrapping is Dangerous

Tools Used
- Python
- Beautiful Soup
- LXML
- Selenium
- Urllib2

https://towardsdatascience.com/web-scraping-using-selenium-python-8a60f4cf40ab

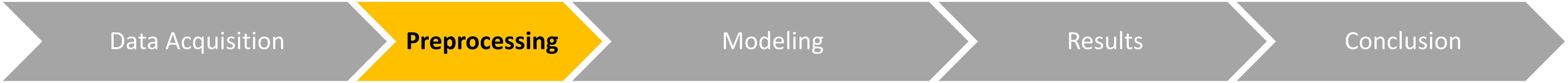
Text Preprocessing Steps



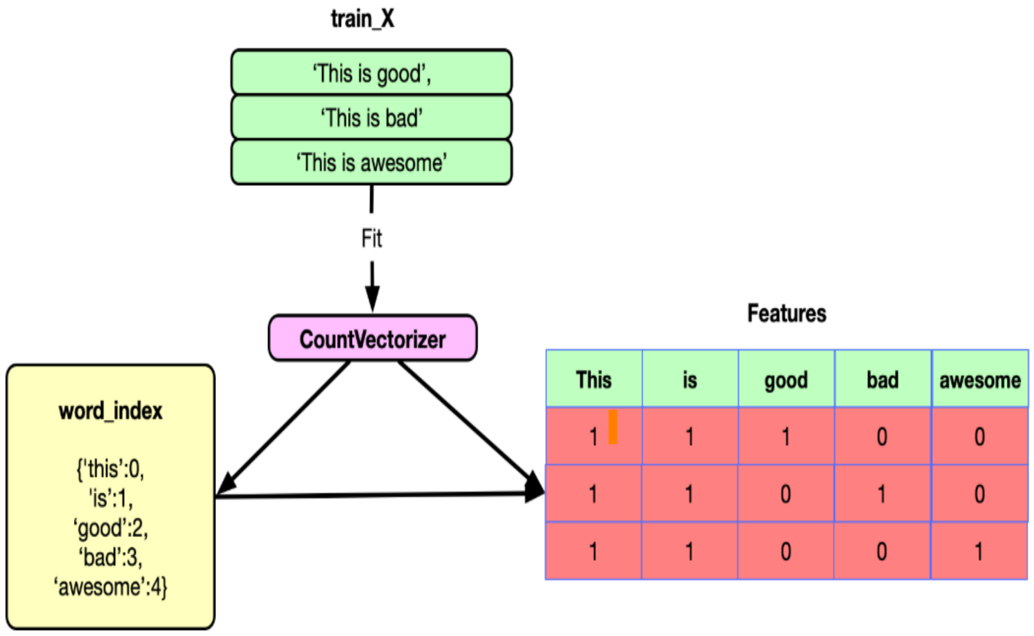
- **STEP 1:**
  - Break text into sentences and remove Gutenberg Header and Footer
- **STEP 2:**
  - Expand Contractions
- **STEP 3:**
  - Convert to Lower and Tokenize
- **STEP 4:**
  - Remove Punctuations and stop words
- **STEP 5:**
  - Lemmatization to base form

Tools Used

- Scrapy
- NLTK



Tokens converted to vectors using Count Vectorizer



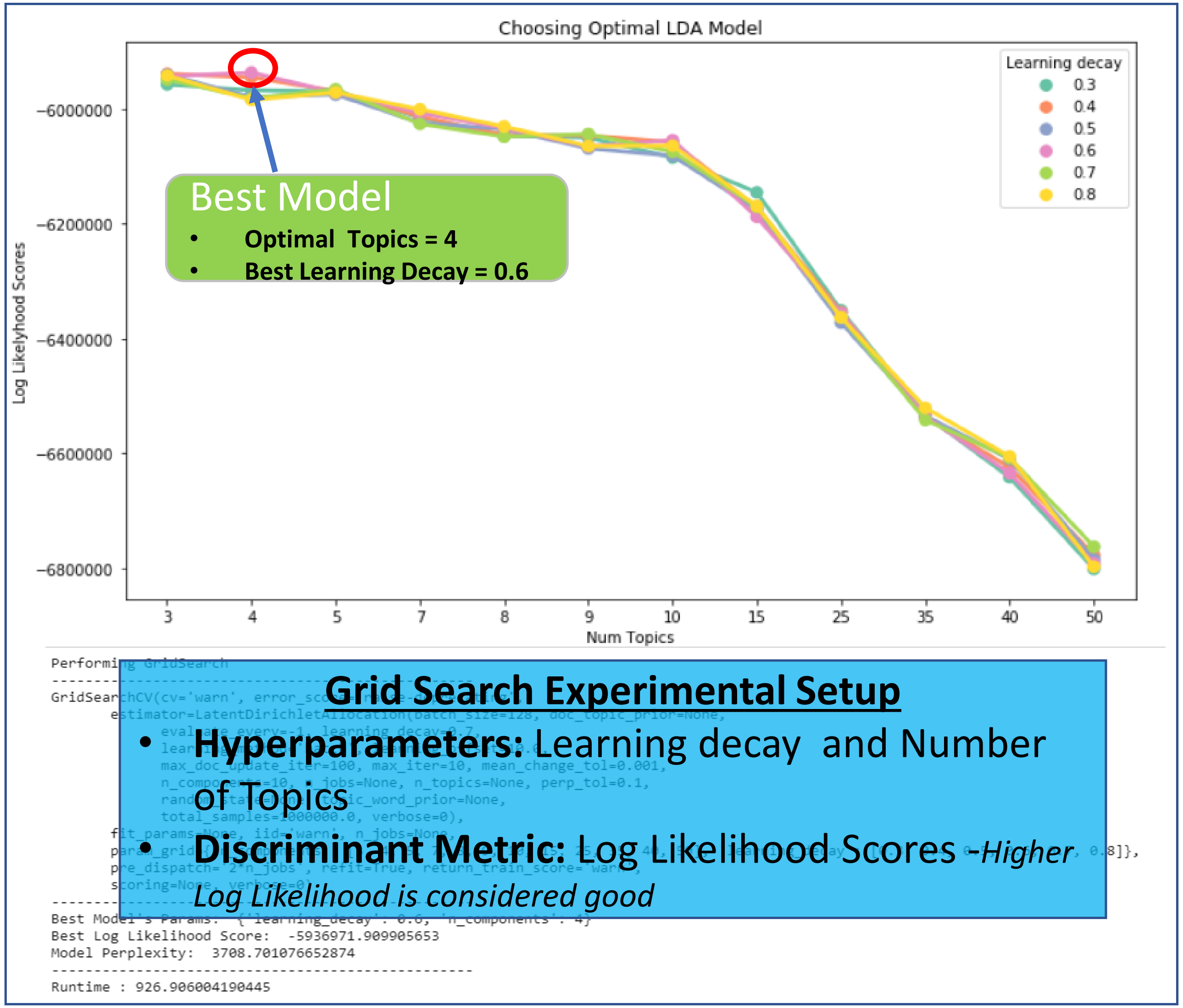
This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

https://mlwhiz.com/blog/2019/02/08/deeplearning_nlp_conventional_methods/

```
# Counter Vectorizer Object
cv = CountVectorizer(max_df=0.95, min_df=2)
```

Filter unique Words and High frequency words

latent Dirichlet allocation (LDA) GridSearch for optimal number of Topics



Best Model

- Optimal Topics = 4
- Best Learning Decay = 0.6

```
Performing GridSearch
-----------------------------------------------
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=LatentDirichletAllocation(batch_size=128, doc_topic_prior=None,
             evaluate_every=-1, learning_decay=0.7,
             learning_method='batch', learning_offset=10.0,
             max_doc_update_iter=100, max_iter=10, mean_change_tol=0.001,
             n_components=10, n_jobs=None, n_topics=None, perp_tol=0.1,
             random_state=None, topic_word_prior=None,
             total_samples=1000000.0, verbose=0),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'n_components': [3, 4, 5, 7, 8, 9, 10, 15, 25, 35, 40, 50], 'learning_decay': [0.3, 0.4, 0.5, 0.6, 0.7, 0.8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=0)
-----------------------------------------------
Best Model's Params:  {'learning_decay': 0.6, 'n_components': 4}
Best Log Likelihood Score:  -5936971.909905653
Model Perplexity:  3708.701076652874
-----------------------------------------------
Runtime : 926.906004190445
```

Grid Search Experimental Setup

- **Hyperparameters:** Learning decay and Number of Topics
- **Discriminant Metric:** Log Likelihood Scores -*Higher Log Likelihood is considered good*

Data Acquisition

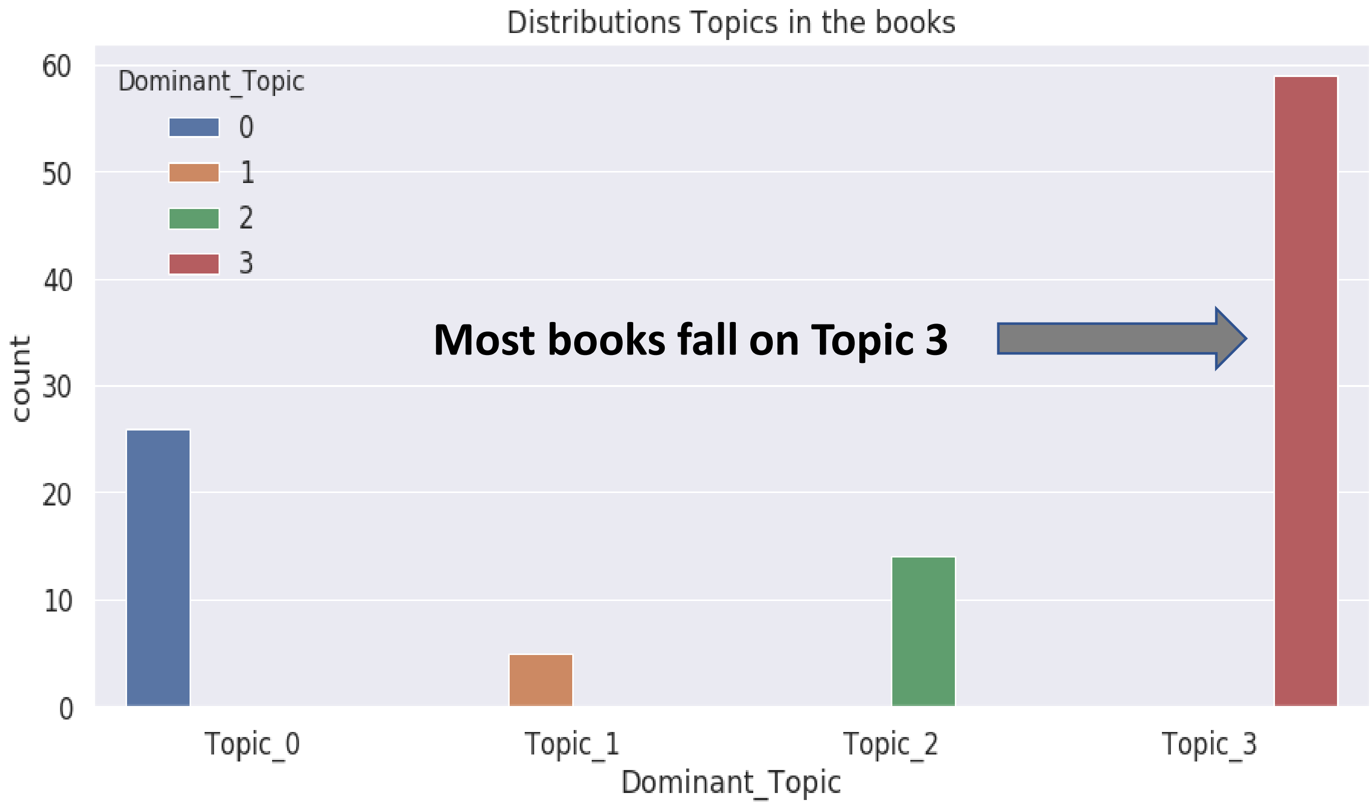
Preprocessing

Modeling

Results

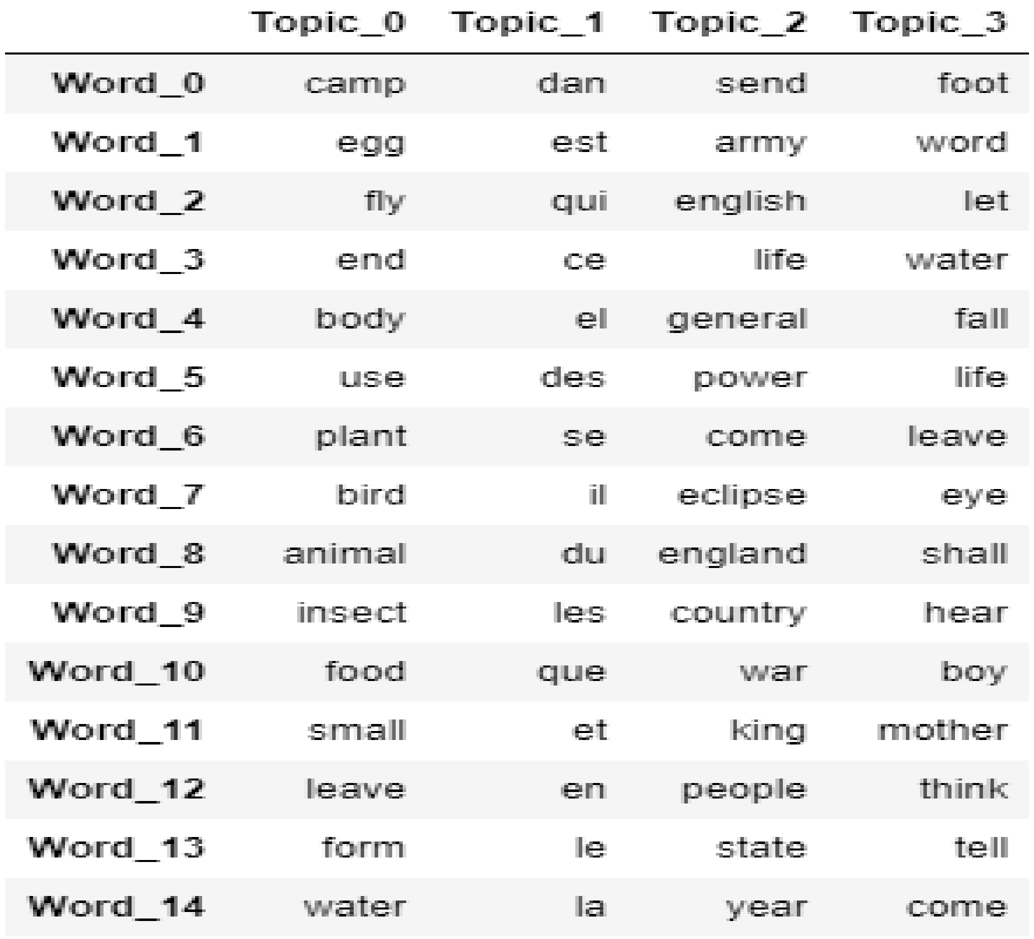
Conclusion

Distribution of books in the 4 topics



Most books fall on Topic 3

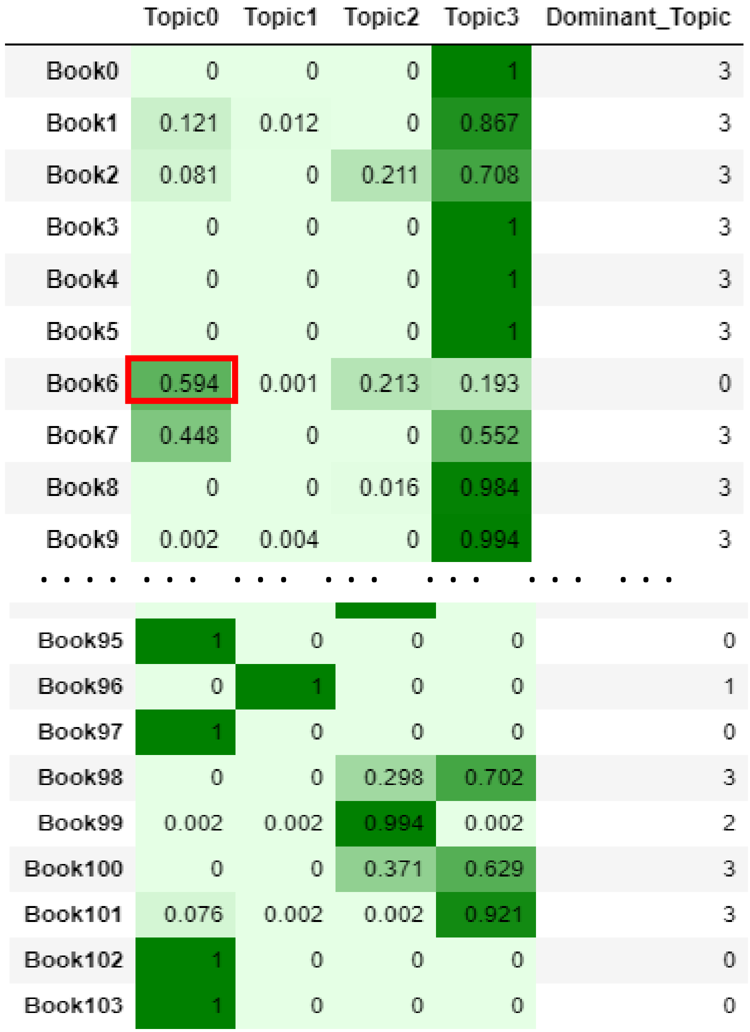
Top 15 words by Topic



	Topic_0	Topic_1	Topic_2	Topic_3
Word_0	camp	dan	send	foot
Word_1	egg	est	army	word
Word_2	fly	qui	english	let
Word_3	end	ce	life	water
Word_4	body	el	general	fall
Word_5	use	des	power	life
Word_6	plant	se	come	leave
Word_7	bird	il	eclipse	eye
Word_8	animal	du	england	shall
Word_9	insect	les	country	hear
Word_10	food	que	war	boy
Word_11	small	et	king	mother
Word_12	leave	en	people	think
Word_13	form	le	state	tell
Word_14	water	la	year	come

Science	Non English	History & Geography	General

Books list heatmap of Topic probabilities



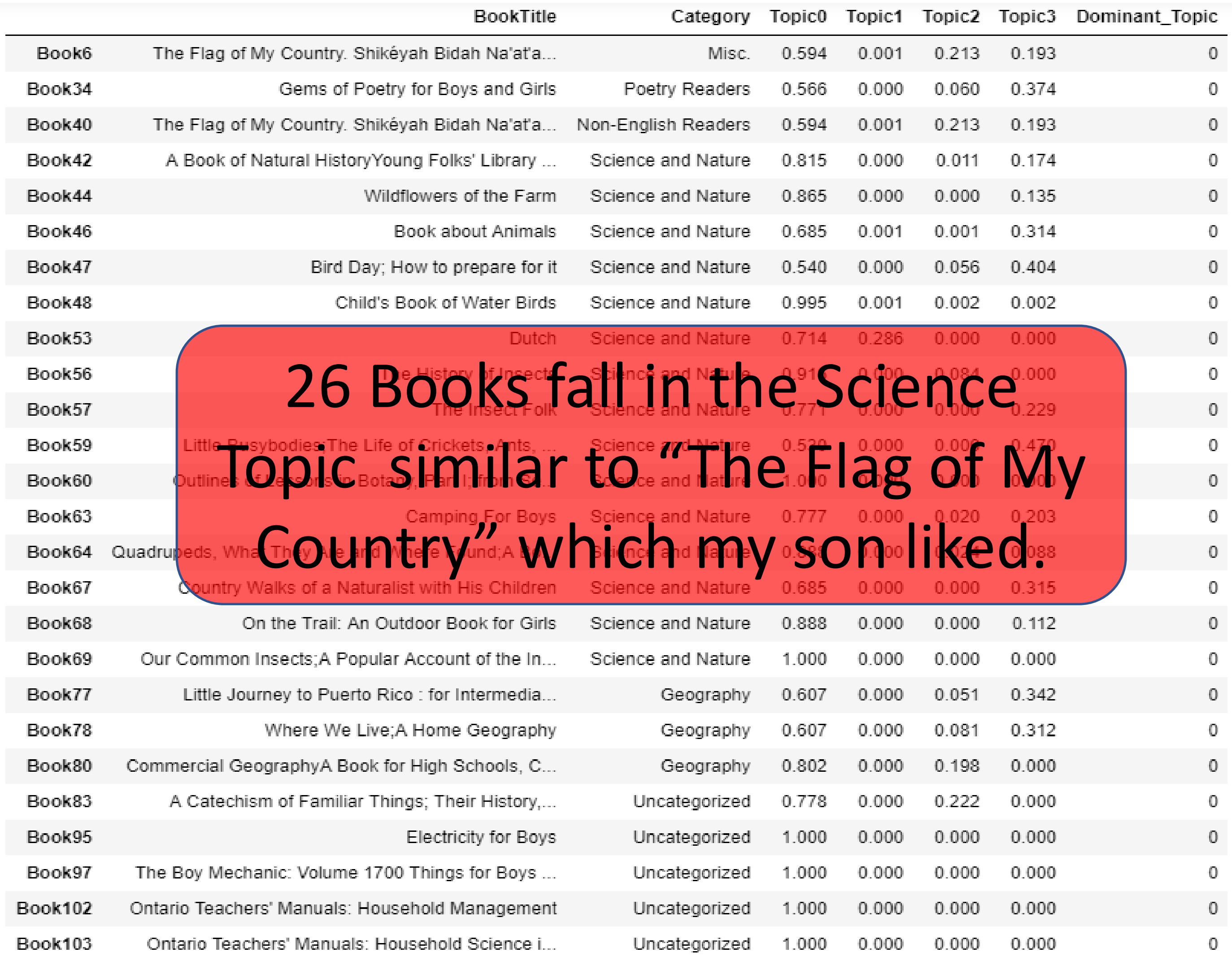
	Topic0	Topic1	Topic2	Topic3	Dominant_Topic
Book0	0	0	0	1	3
Book1	0.121	0.012	0	0.867	3
Book2	0.081	0	0.211	0.708	3
Book3	0	0	0	1	3
Book4	0	0	0	1	3
Book5	0	0	0	1	3
Book6	0.594	0.001	0.213	0.193	0
Book7	0.448	0	0	0.552	3
Book8	0	0	0.016	0.984	3
Book9	0.002	0.004	0	0.994	3

. . . .

	Topic0	Topic1	Topic2	Topic3	Dominant_Topic
Book95	1	0	0	0	0
Book96	0	1	0	0	1
Book97	1	0	0	0	0
Book98	0	0	0.298	0.702	3
Book99	0.002	0.002	0.994	0.002	2
Book100	0	0	0.371	0.629	3
Book101	0.076	0.002	0.002	0.921	3
Book102	1	0	0	0	0
Book103	1	0	0	0	0

Science

Other books similar to Book6 (Topic0 - Science)



	BookTitle	Category	Topic0	Topic1	Topic2	Topic3	Dominant_Topic
Book6	The Flag of My Country. Shikéyah Bidah Na'at'a...	Misc.	0.594	0.001	0.213	0.193	0
Book34	Gems of Poetry for Boys and Girls	Poetry Readers	0.566	0.000	0.060	0.374	0
Book40	The Flag of My Country. Shikéyah Bidah Na'at'a...	Non-English Readers	0.594	0.001	0.213	0.193	0
Book42	A Book of Natural HistoryYoung Folks' Library ...	Science and Nature	0.815	0.000	0.011	0.174	0
Book44	Wildflowers of the Farm	Science and Nature	0.865	0.000	0.000	0.135	0
Book46	Book about Animals	Science and Nature	0.685	0.001	0.001	0.314	0
Book47	Bird Day; How to prepare for it	Science and Nature	0.540	0.000	0.056	0.404	0
Book48	Child's Book of Water Birds	Science and Nature	0.995	0.001	0.002	0.002	0
Book53	Dutch	Science and Nature	0.714	0.286	0.000	0.000	0
Book56	The History of Insects	Science and Nature	0.916	0.000	0.084	0.000	0
Book57	The Insect Folk	Science and Nature	0.771	0.000	0.000	0.229	0
Book59	Little Busybodies;The Life of Crickets, Ants, ...	Science and Nature	0.530	0.000	0.000	0.470	0
Book60	Outlines of Lessons in Botany, Part I; from Se...	Science and Nature	1.000	0.000	0.000	0.000	0
Book63	Camping For Boys	Science and Nature	0.777	0.000	0.020	0.203	0
Book64	Quadrupeds, What They Are and Where Found;A Bo...	Science and Nature	0.888	0.000	0.024	0.088	0
Book67	Country Walks of a Naturalist with His Children	Science and Nature	0.685	0.000	0.000	0.315	0
Book68	On the Trail: An Outdoor Book for Girls	Science and Nature	0.888	0.000	0.000	0.112	0
Book69	Our Common Insects;A Popular Account of the In...	Science and Nature	1.000	0.000	0.000	0.000	0
Book77	Little Journey to Puerto Rico : for Intermedia...	Geography	0.607	0.000	0.051	0.342	0
Book78	Where We Live;A Home Geography	Geography	0.607	0.000	0.081	0.312	0
Book80	Commercial GeographyA Book for High Schools, C...	Geography	0.802	0.000	0.198	0.000	0
Book83	A Catechism of Familiar Things; Their History,...	Uncategorized	0.778	0.000	0.222	0.000	0
Book95	Electricity for Boys	Uncategorized	1.000	0.000	0.000	0.000	0
Book97	The Boy Mechanic: Volume 1700 Things for Boys ...	Uncategorized	1.000	0.000	0.000	0.000	0
Book102	Ontario Teachers' Manuals: Household Management	Uncategorized	1.000	0.000	0.000	0.000	0
Book103	Ontario Teachers' Manuals: Household Science i...	Uncategorized	1.000	0.000	0.000	0.000	0

26 Books fall in the Science Topic similar to "The Flag of My Country" which my son liked.

Data Acquisition · Preprocessing · Modeling · Results · **Conclusion**

End

References

https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/