Week 3 Coding: Clustering and Topic Modeling

Joseph Denby

Computational Content Analysis

January 19, 2018

Corpus

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear. Psychological Science, 26(9), 14891496. http://doi.org/10.1177/0956797615594361

Alexander and the Terrible, Horrible, No Good, Very Bad Day by Judith Viorst
Angelina Ice Skates by Katharine Holabird
Are You My Mother? by P. D. Eastman
Arnie the Doughnut by Laurie Keller
Arthur Writes a Story by Marc Brown
A Bad Case of Stripes by David Shannon
Bark, George by Jules Feiffer
Bear Wants More by Karma Wilson
The Berenstain Bears and the Green-Eyed Monster by Stan Berenstain and Jan Berenstain
The Berenstain Bears Forget Their Manners by Stan Berenstain and Jan Berenstain
Blueberries for Sal by Robert McCloskey
Bread and Jam for Frances by Russell Hoban
Brown Bear, Brown Bear, What Do You See? by Bill Martin, Jr.
Bunny Party by Rosemary Wells
Caps for Sale by Esphyr Slobodkina
The Carrot Seed by Ruth Krauss
The Cat in the Hat by Dr. Seuss
Charlie and the New Baby by Ree Drummond
Chicka Chicka 1-2-3 by Bill Martin, Jr., Michael Sampson, and Lois Ehlert
How Do Dinosaurs Say Good Night? by Jane Yolen and Mark Teague
How to Train a Train by Jason Carter Eaton
If You Give a Moose a Muffin by Laura Joffe Numeroff
If You Give a Mouse a Cookie by Laura Joffe Numeroff
I'm a Big Sister by Joanna Cole
The Keeping Quilt by Patricia Polacco
Knuffle Bunny by Mo Willems
Ladybug Girl at the Beach by David Soman and Jacky Davis
Lilly's Purple Plastic Purse by Kevin Henkes
Little Blue Truck Leads the Way by Alice Schertle
The Little Engine That Could by Watty Piper
The Little House by Virginia Lee Burton
Llama Llama Home With Mama by Anna Dewdney
Llama Llama Red Pajama by Anna Dewdney
The Lorax by Dr. Seuss
Love You Forever by Sheila McGraw
Madeline by Ludwig Bemelmans
Maisy Goes Camping by Lucy Cousins
Maisy Goes to the Library by Lucy Cousins
Make Way for Ducklings by Robert McCloskey
Mike Mulligan and His Steam Shovel by Virginia Lee Burton
Miss Rumphius by Barbara Cooney

Preliminaries

Term Frequency-Inverse Document Frequency (tf-idf)

- Word frequency scaled by the inverse of document frequency
- Used to assess informativeness of word frequency in a document

Preliminaries

Term Frequency-Inverse Document Frequency (tf-idf)

- Word frequency scaled by the inverse of document frequency
- Used to assess informativeness of word frequency in a document

```
[('george', 0.8173804012841295),
 ('mother', 0.0799363521813551),
 ('said', 0.05453320198192945),
 ('bark', 0.300642324953546),
 ('went', 0.09932712262216196),
 ('meow', 0.07340212851243166),
 ('no', 0.0275645668960507 43),
 ('cats', 0.02335372575097 5128),
 ('go', 0.05280143106100 4864),
 ('dogs', 0.06478954728663815),
 ('arf', 0.12025692998141839),
 ('now', 0.028488687043473303),
 ('quack', 0.14680425702486333),
 ('ducks', 0.02583037412878 3657),
 ('oink', 0.07749112238635097),
 ('pigs', 0.02583037412878 3657),
 ('moo', 0.05517516823509214),
 ('took', 0.01332272536355 9183),
 ('to', 0.02592901090972215),
 ('the', 0.08722467419426641)]
```

PCA

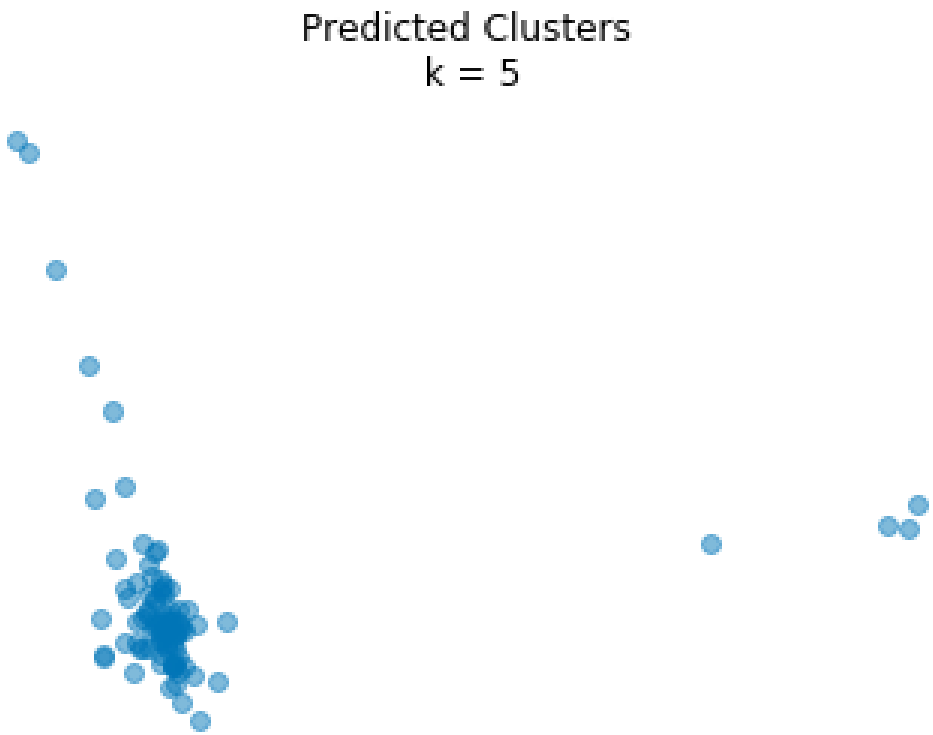
Principle Components Analysis

- Reduce variance in n_{docs} by n_{words} matrix to 2 dimensions
- Much easier to visualize

PCA

Principle Components Analysis

- Reduce variance in n_{docs} by n_{words} matrix to 2 dimensions
- Much easier to visualize



Flat Clustering

- Algorithmically group documents according to their tf-idf vectors
- Clusters are documents that use similar words similarly

Flat Clustering

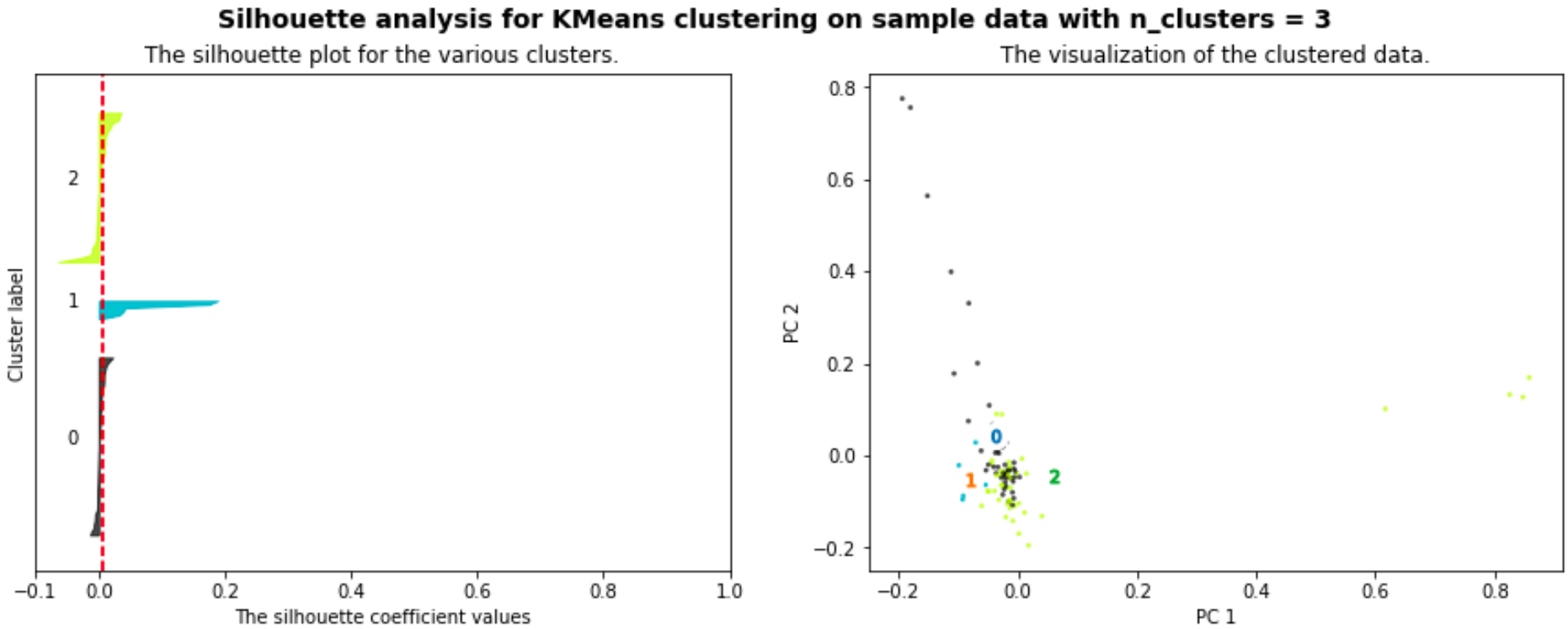
```
Top terms per cluster:
Cluster 0:
 ll
 ask
 train
 big
 little
 like
 baby
 bus
 eat
 just


Cluster 1:
 said
 george
 mother
 little
 good
 dog
 night
 went
 came
 house


Cluster 2:
 mama
 llama
 train
 papa
 fence
 sister
 said
 bike
 brother
```

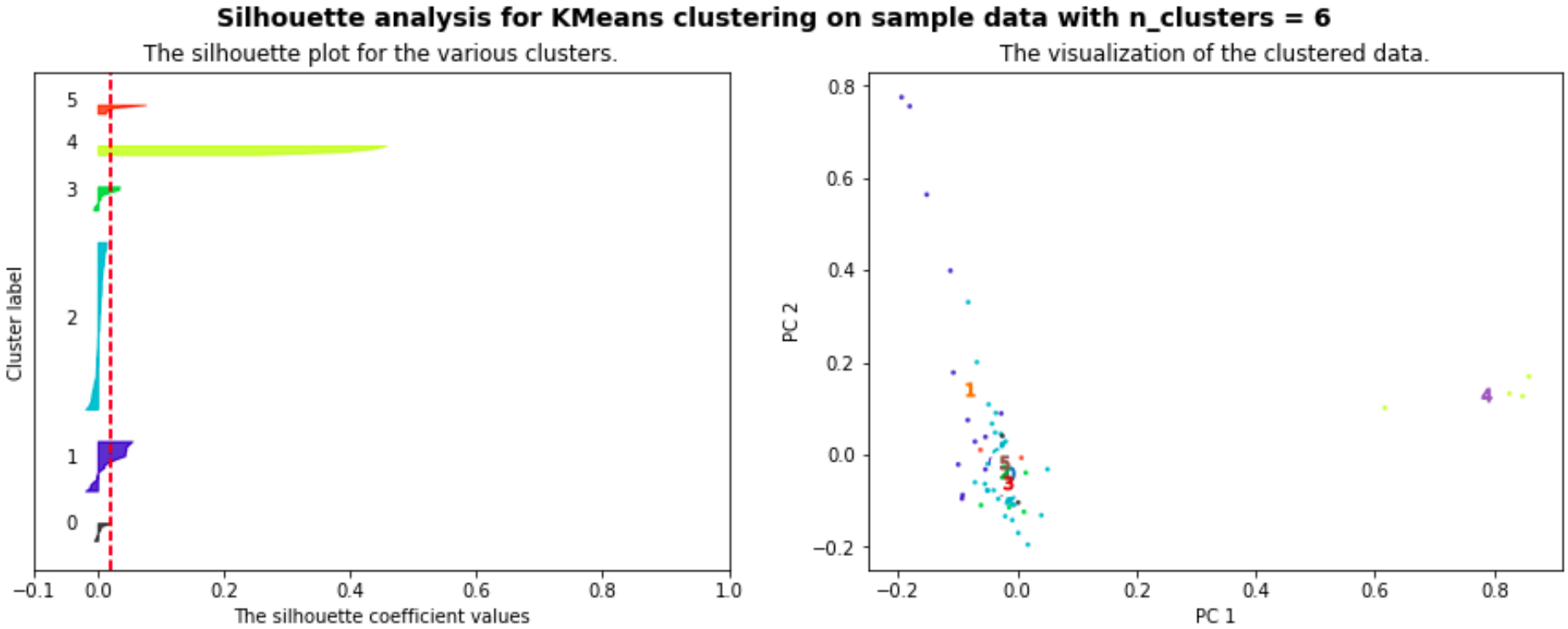
Silhouette

Allows us to determine the optimal number of clusters




```
For n_clusters = 3, The average silhouette_score is : 0.006
```

Silhouette



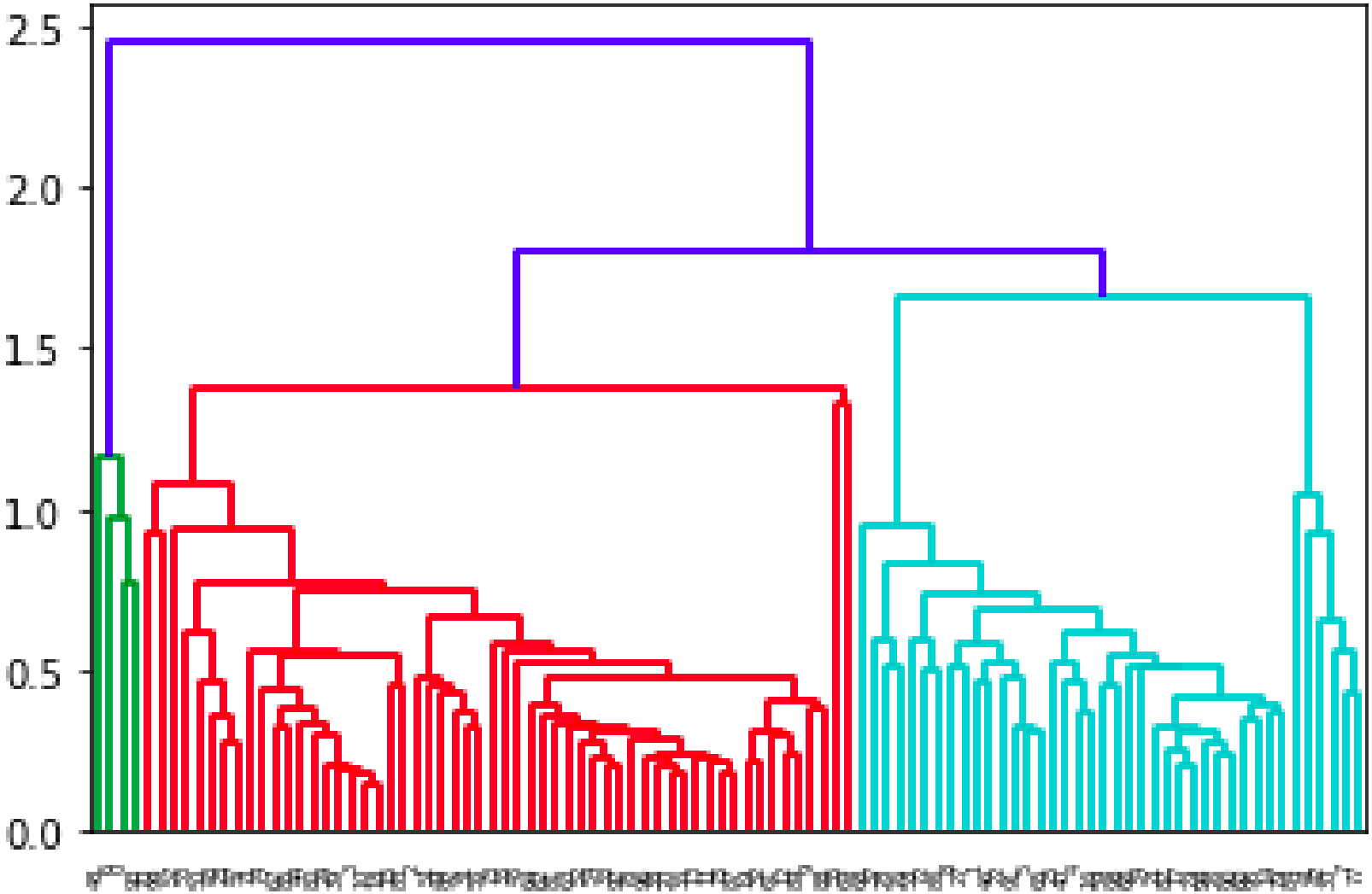

```
For n_clusters = 6, The average silhouette_score is : 0.021
```

No obvious clustering to the books

Hard to determine given outliers

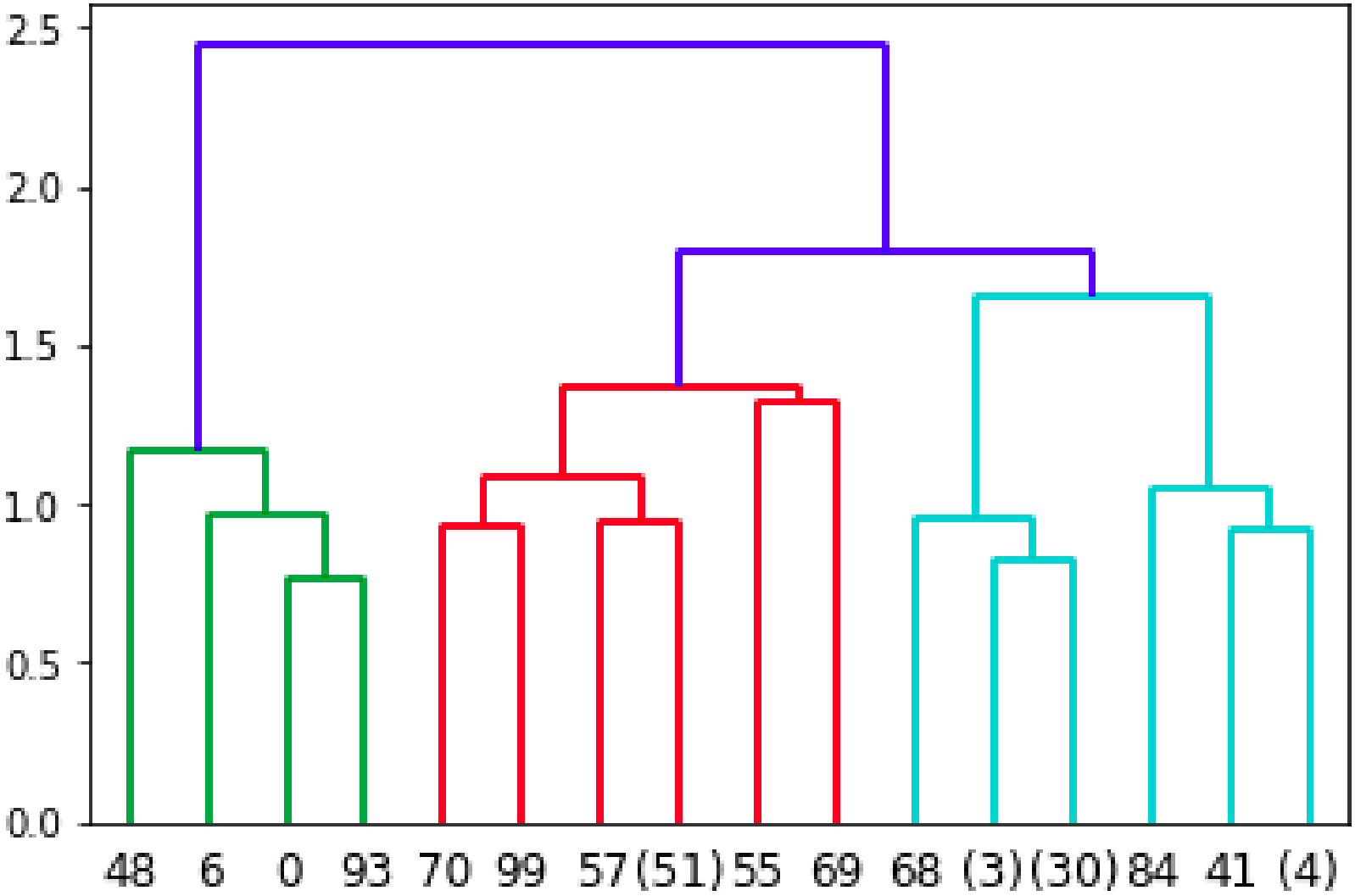
Hierarchical Clustering

A different approach that creates clusters at various points of resolution



Hierarchical Clustering

A different approach that creates clusters at various points of resolution



Topic Modeling

Assign words to topics based on frequency and co-occurence

	title	topics	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7
0	Bark, George	[(4, 0.7254582), (7, 0.26434156)]	0.000000	0.000000	0.000000	0.0	0.725458	0.000000	0.0	0.264342
10	Caps for Sale	[(0, 0.994875)]	0.994875	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
20	Stellaluna	[(5, 0.9977239)]	0.000000	0.000000	0.000000	0.0	0.000000	0.997724	0.0	0.000000
30	Don't Let the Pigeon Drive the Bus	[(1, 0.9742126)]	0.000000	0.974213	0.000000	0.0	0.000000	0.000000	0.0	0.000000
40	Little Blue Truck Leads the Way	[(1, 0.55645955), (2, 0.0349048), (4, 0.401254...	0.000000	0.556460	0.034905	0.0	0.401255	0.000000	0.0	0.000000

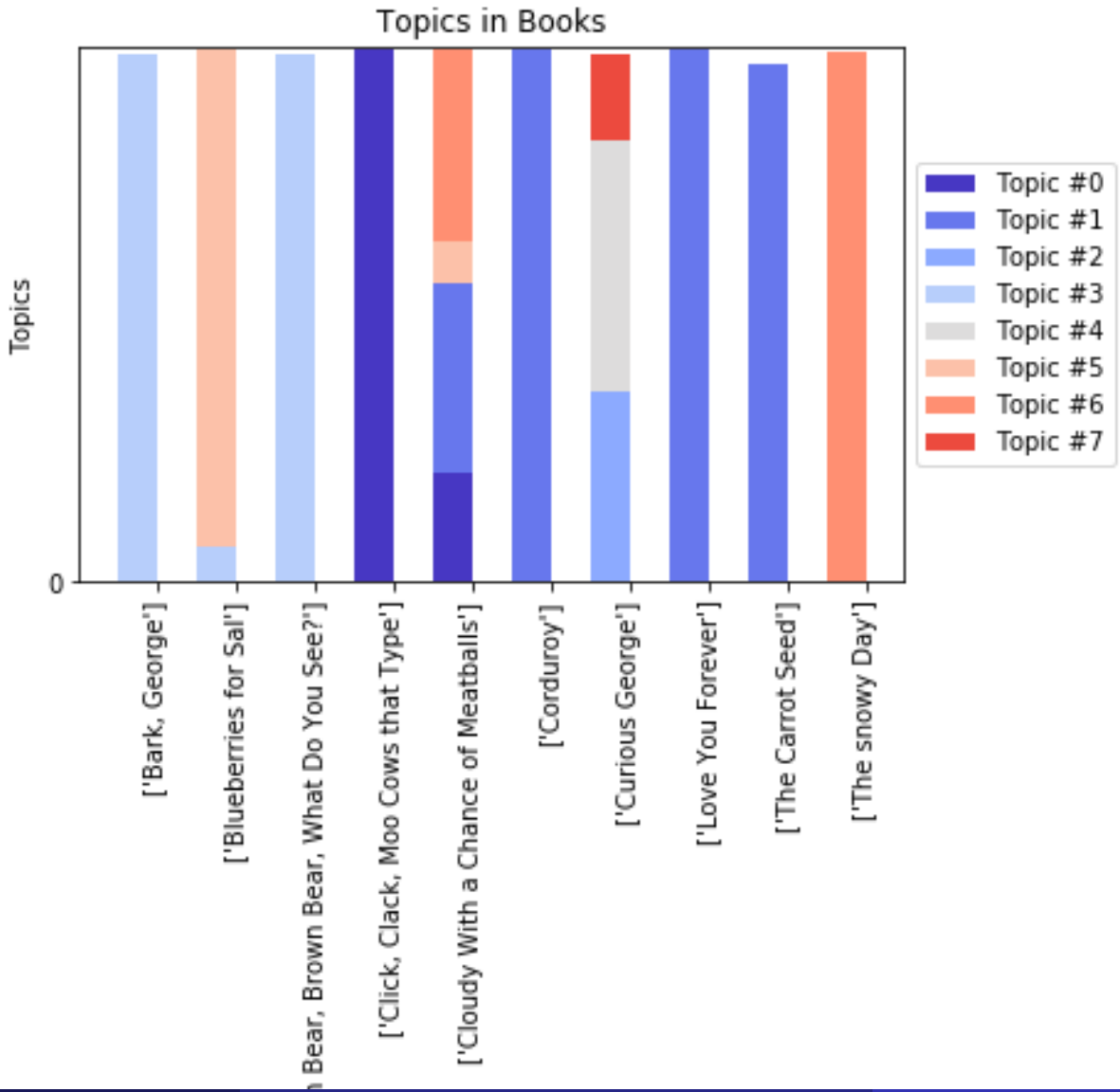
Topic Modeling

Each topic is a probability distribution of all words in the corpus.

	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7
0	said	said	train	llama	said	said	said	said
1	look	dragon	like	like	big	look	like	look
2	mother	look	big	mama	went	day	mother	good
3	cap	good	said	said	look	mother	day	llama
4	thing	like	look	long	day	want	chrysanthemum	time
5	like	love	love	day	like	stellaluna	want	ask
6	want	train	time	mother	time	big	look	duck
7	cat	dog	think	come	tree	like	tree	like
8	dog	bear	say	good	mother	thing	boy	night
9	ask	come	thing	place	sister	friend	ladybug	mother

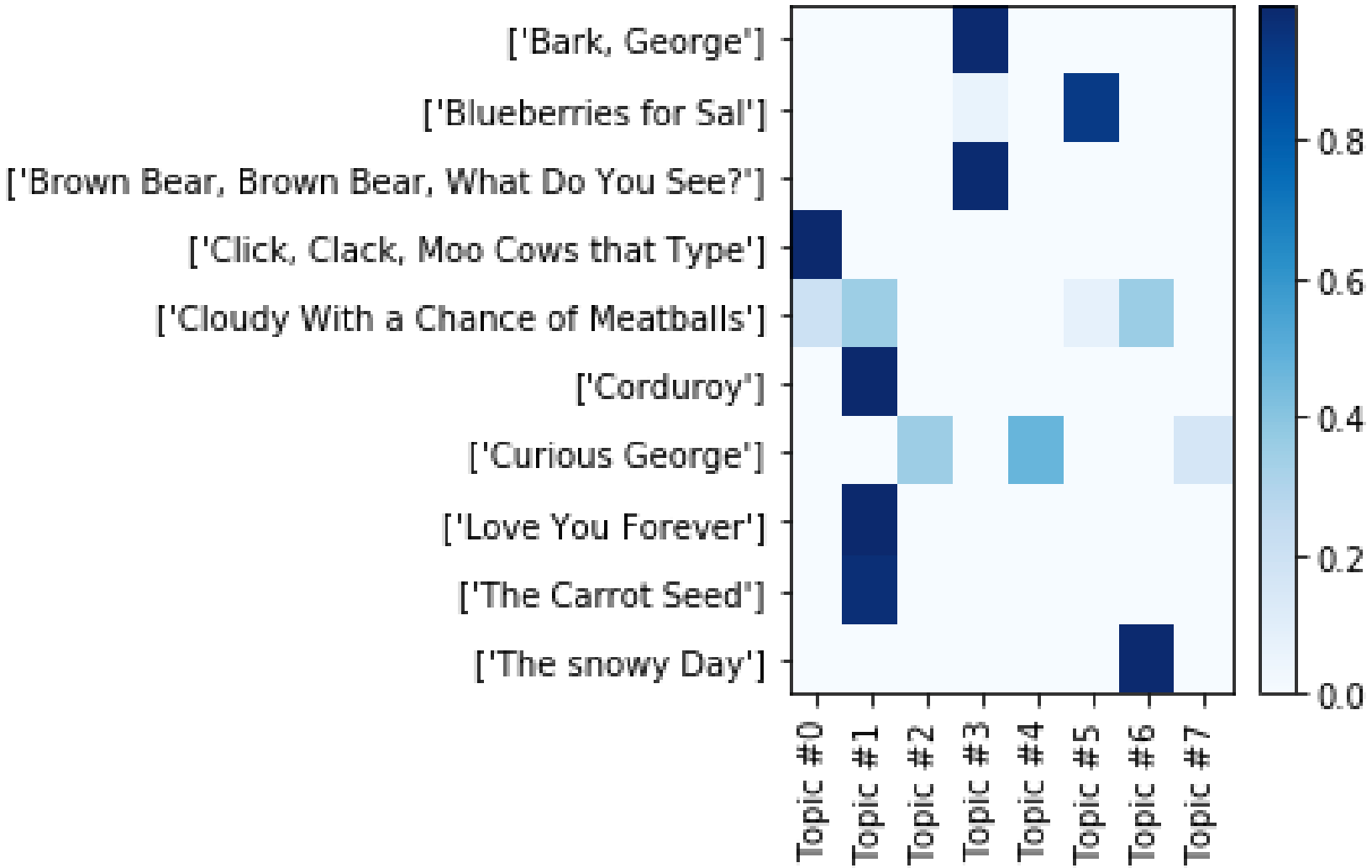
Topic Modeling

Each document is a probability distribution of all topics.



Topic Modeling

Each document is a probability distribution of all topics.

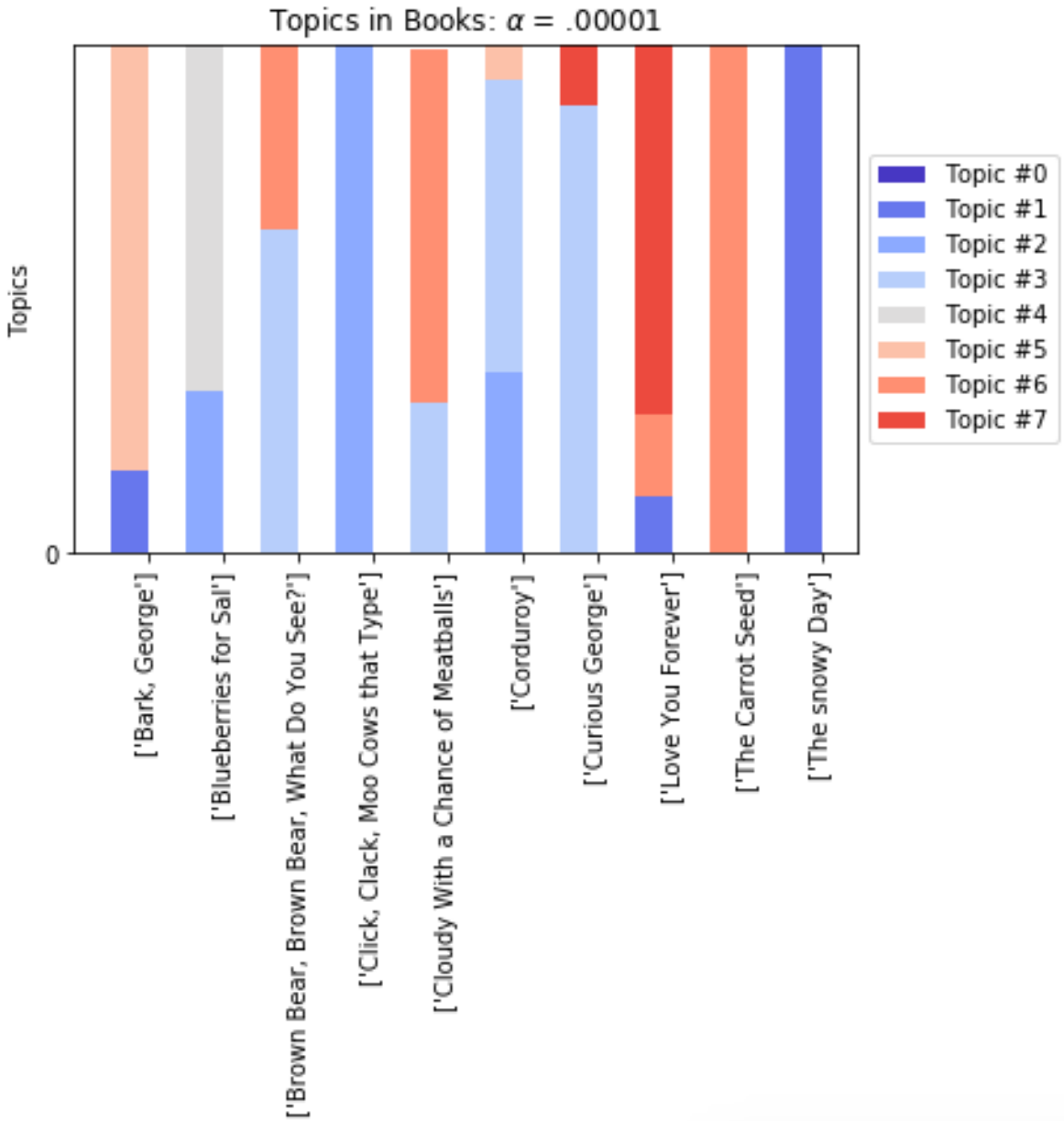


Topic Modeling

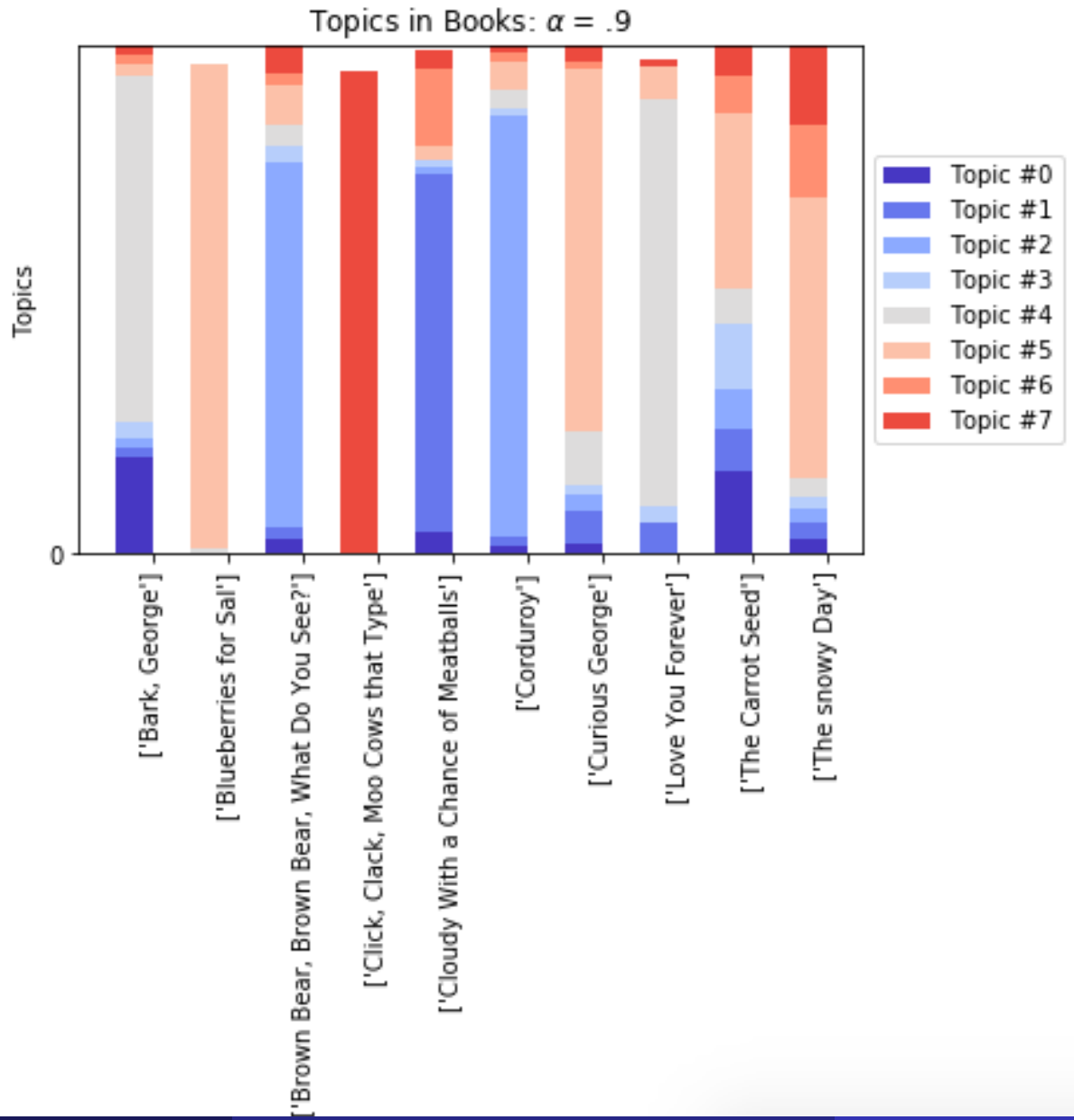
Can adjust the topic modeling algorithm parameters as well

- α - sparsity of document-topic loadings
- η - sparsity of topic-word loadings

Topic Modeling



Topic Modeling



Topic Modeling