
# BEST SELLING BOOKS

*What topics are people reading?*

# OBJECTIVE: CATEGORIZE POPULAR BOOKS SINCE 2008

**BOOKS**

## The New York Times Best Sellers

Authoritatively ranked lists of books sold in the United States, sorted by format and genre.

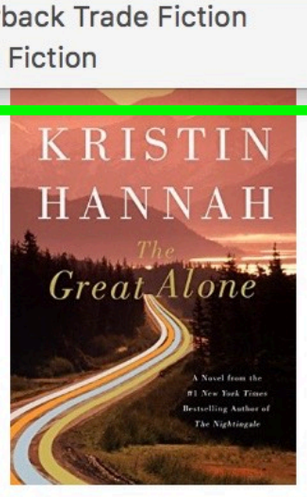
✓ FICTION

Combined Print & E-Book Fiction
Hardcover Fiction
Paperback Trade Fiction
Audio Fiction

E'S | MONTHLY LISTS

Fiction >

1



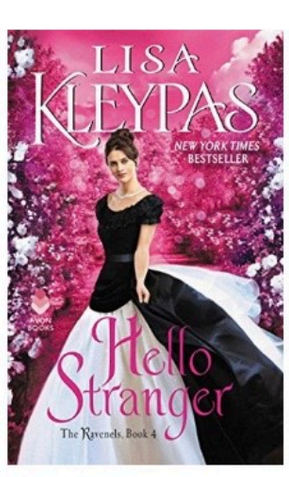

4 WEEKS ON THE LIST

**THE GREAT ALONE**

by Kristin Hannah

A former prisoner of war returns from Vietnam and moves his family to Alaska, where they face tough conditions.

2



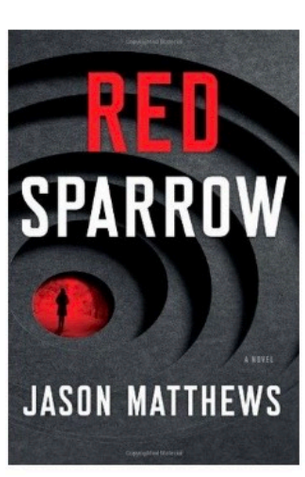

NEW THIS WEEK

**HELLO STRANGER**

by Lisa Kleypas

Book 4 of the Ravenels series.

3




2 WEEKS ON THE LIST

**RED SPARROW**

by Jason Matthews

A Russian intelligence officer trained in the art of seduction becomes entangled with a young C.I.A. officer.

# BUSINESS CASE & HYPOTHESIS

**Potential Uses:**

- ➤ Identify demand for authors/publishers
- ➤ Content-based recommendation for readers

**Hypothesis:**

- ➤ We can find what people are most interested in by reducing the number of categories
  - 20+ for NYT articles
  - 30+ for Amazon books

# METHODOLOGY

New York Times & Google Books APIs

⬇

MongoDB, Pandas merge

⬇

Clean, Lemmatize, Pre-process (31% net terms)

⬇

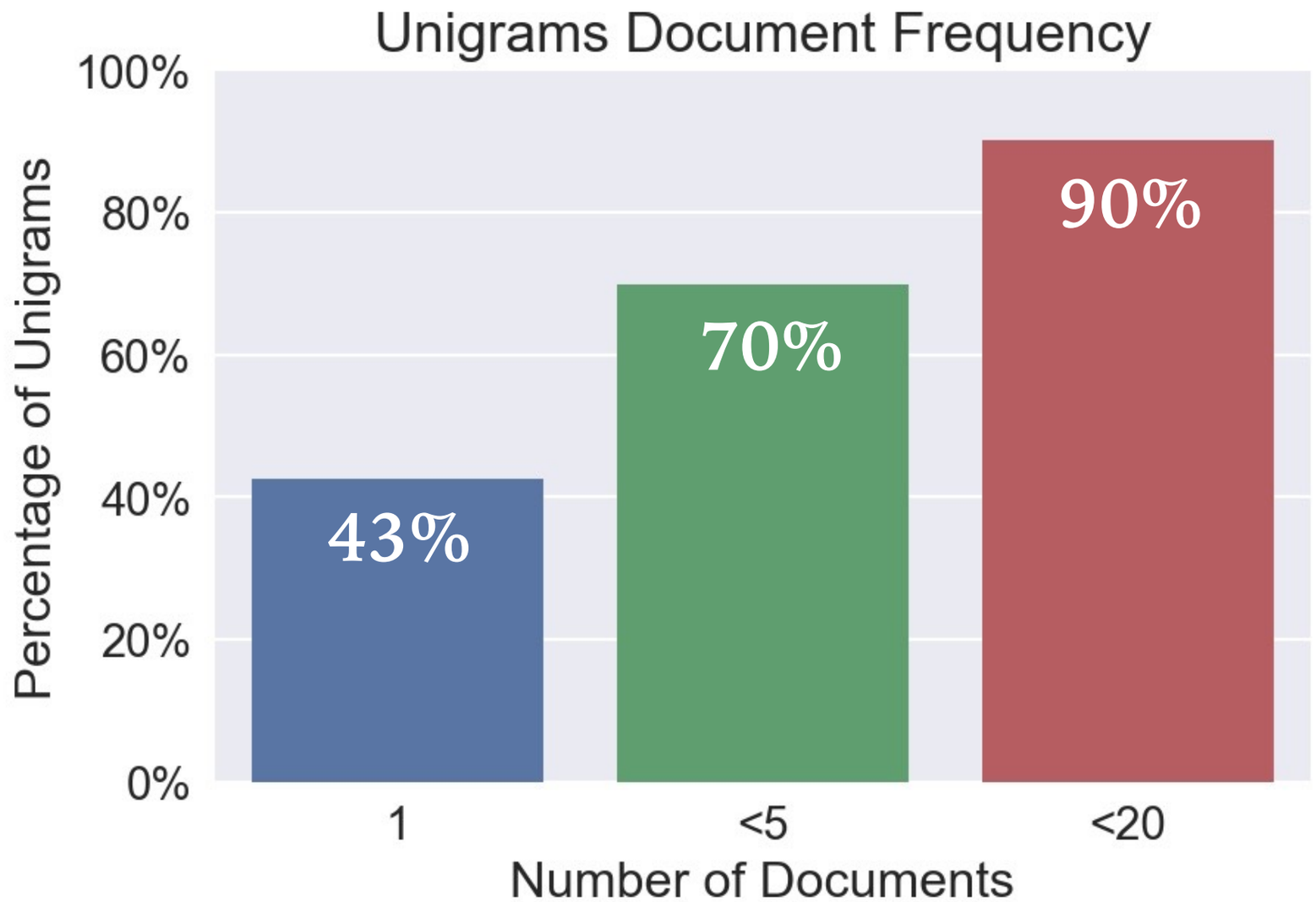
LSA dimension reduction

⬇

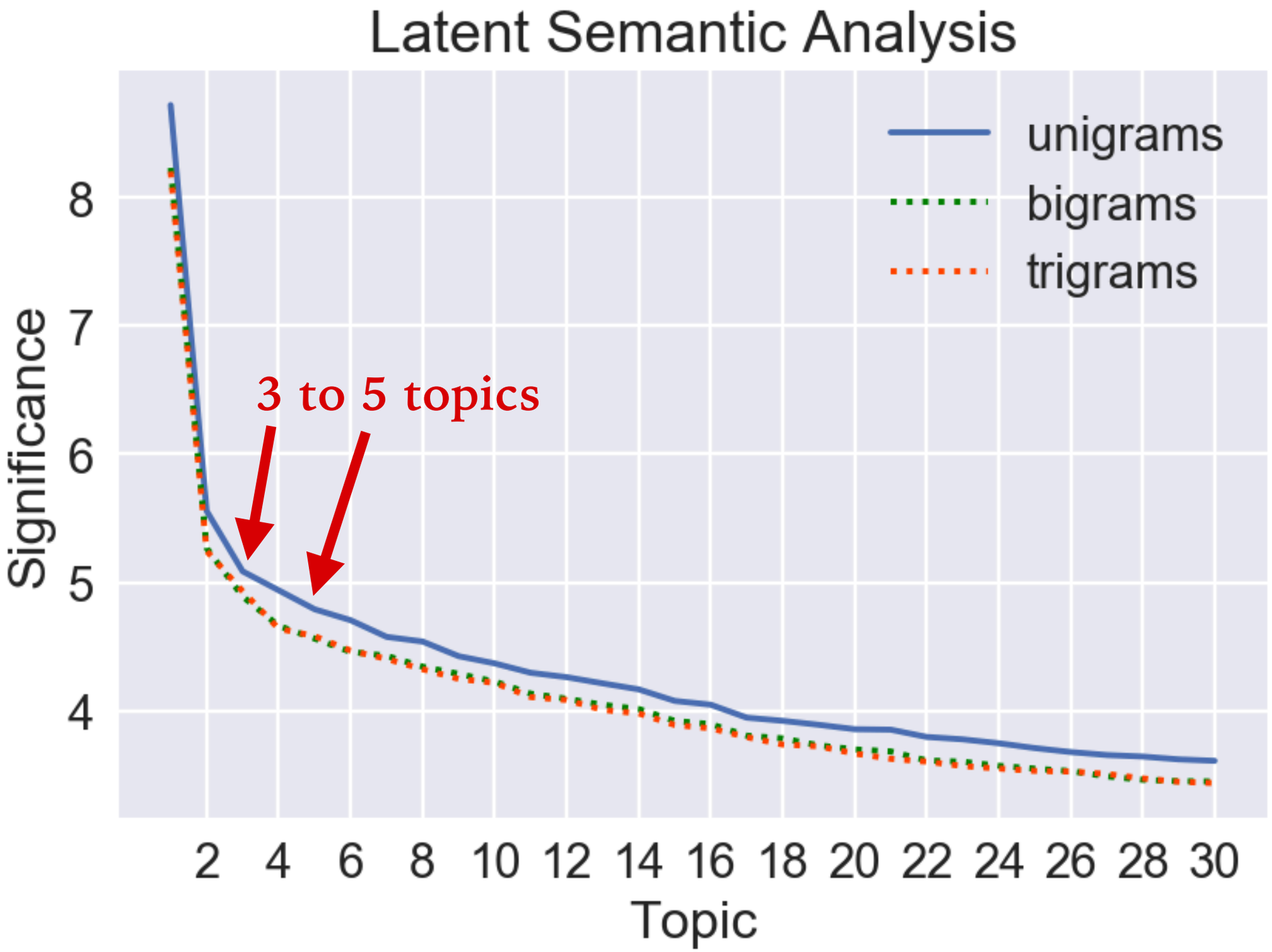
LDA topic labels

# 11,000 BOOKS: MOSTLY LONG TAIL N-GRAMS

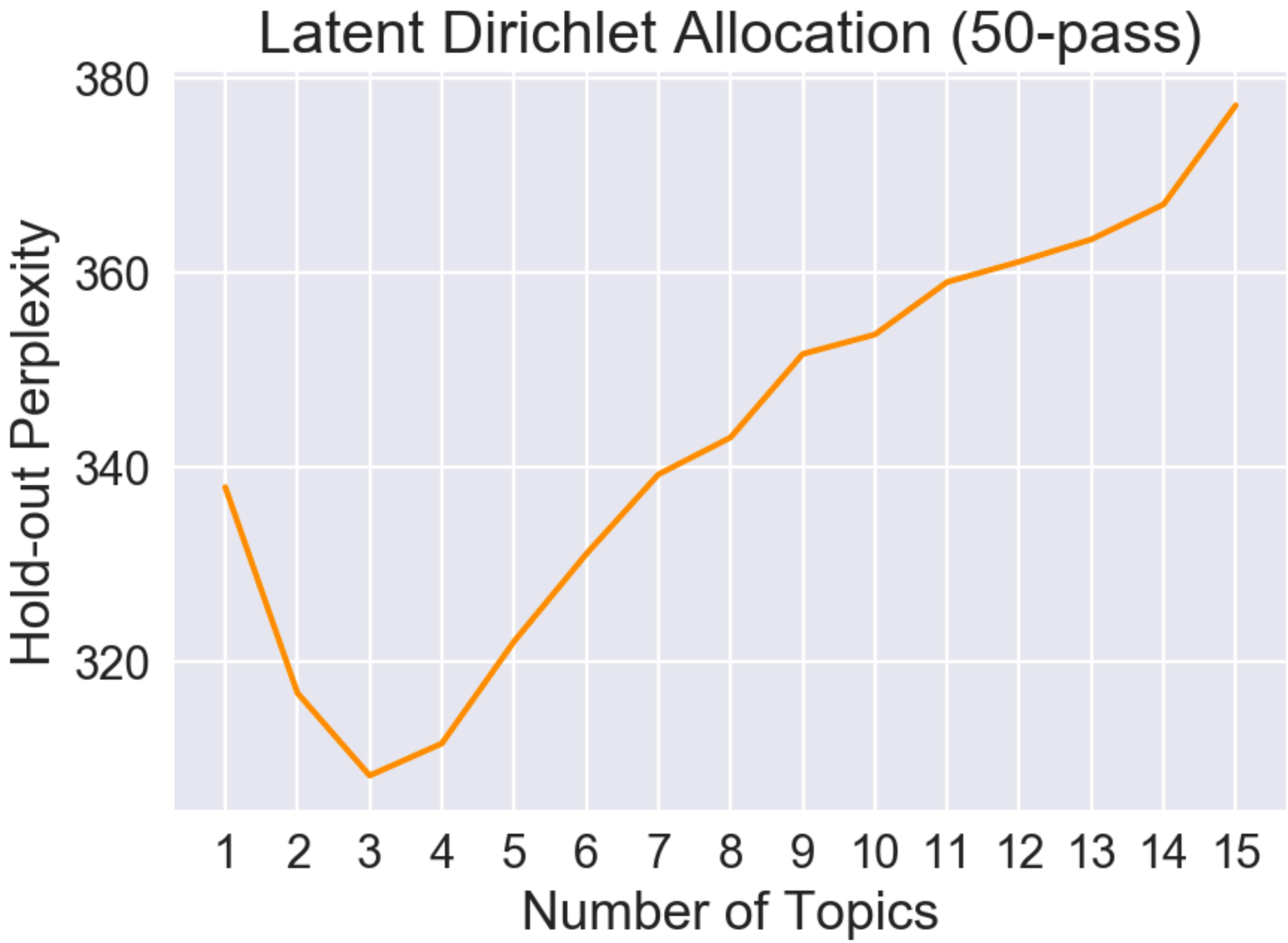
- Only 3 words occur in more than 10% of documents
- 20 documents represent 0.1% of corpus



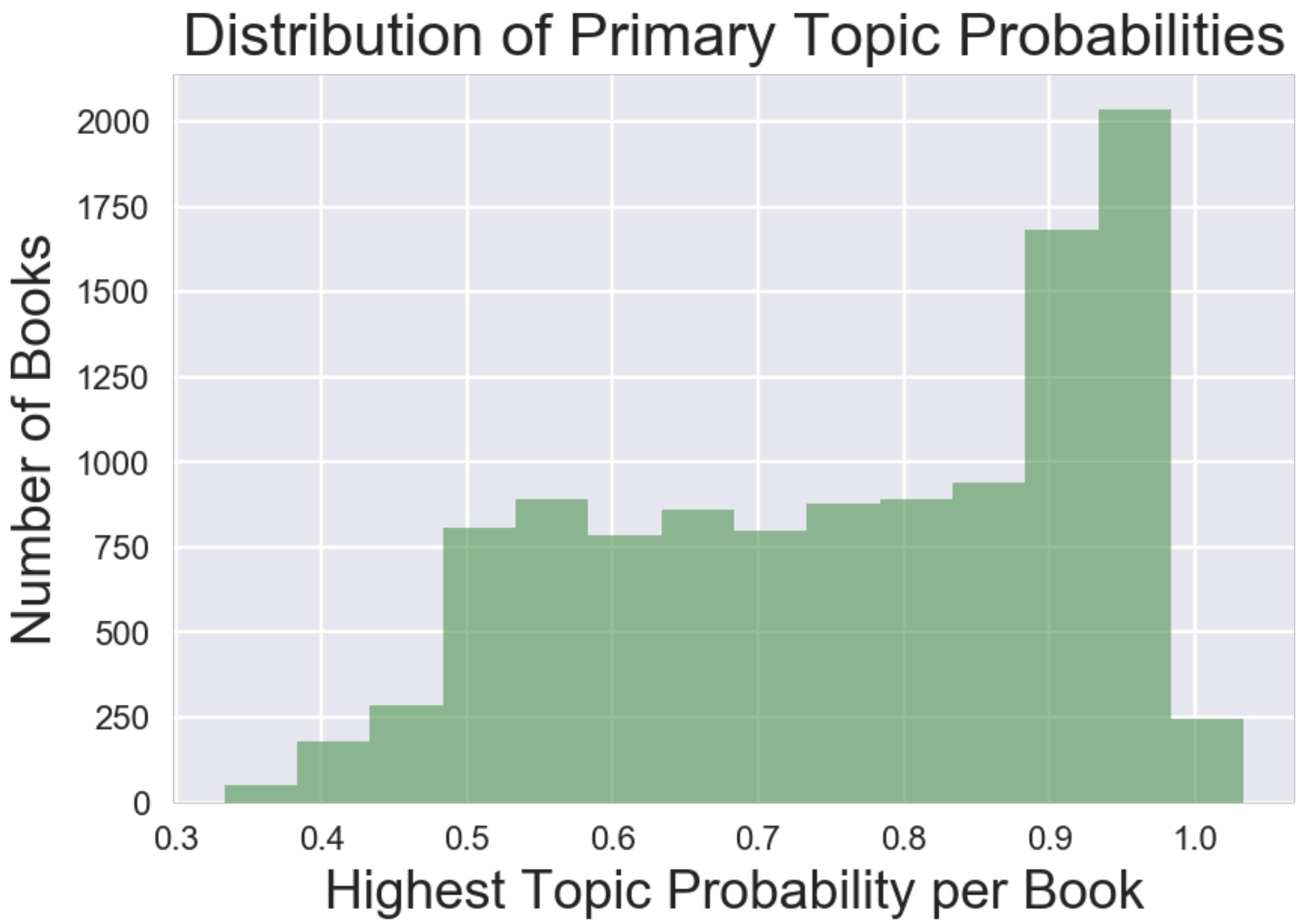
# OPTIMAL DIMENSIONALITY



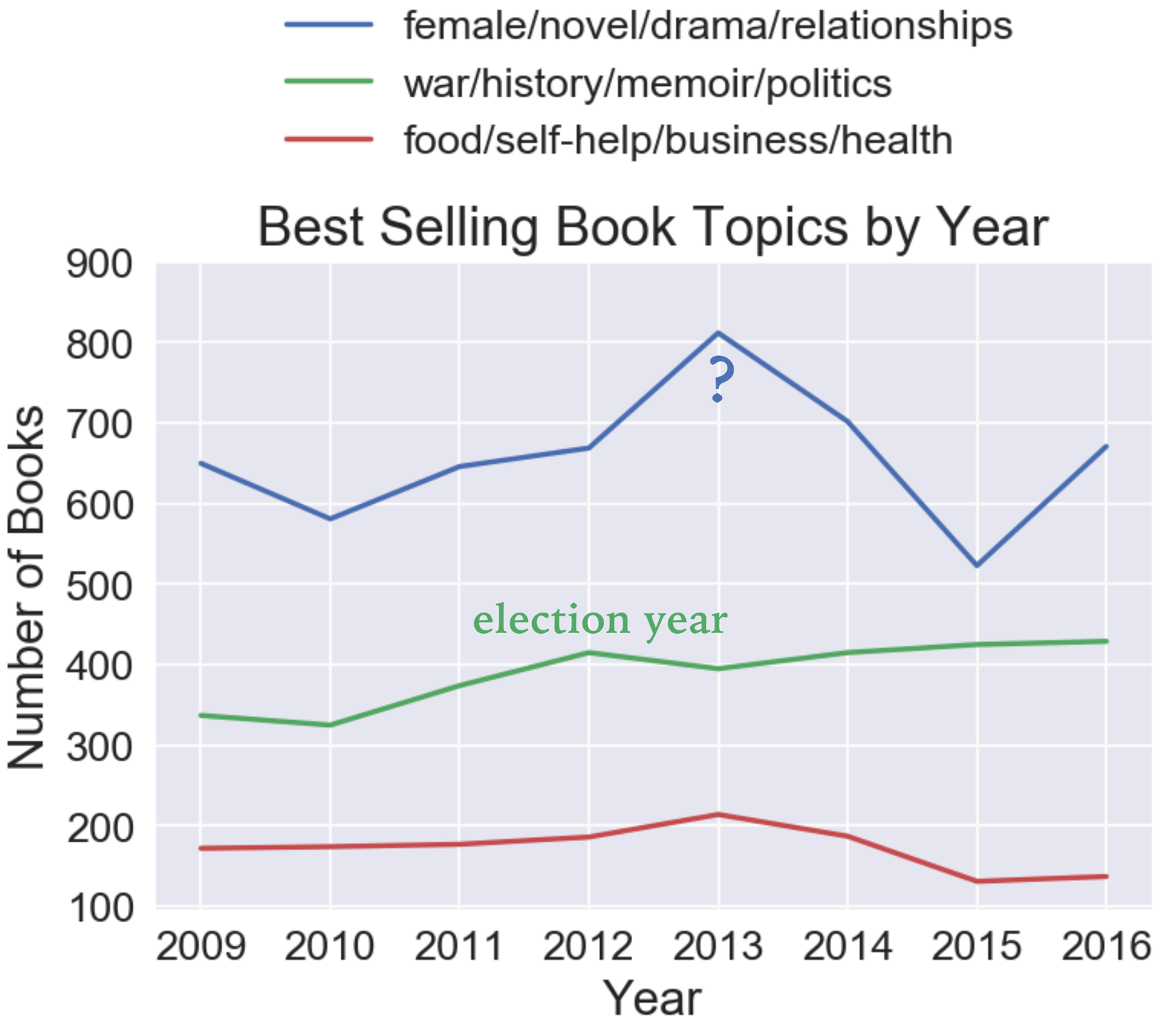
# OPTIMAL DIMENSIONALITY



# LDA TOPIC PROBABILITIES



# LDA TOPIC TRENDS




# TOPIC QUALITY & ISSUES

- Topics make sense, but limited use
- Term overlap, need more granularity
- Long tail n-grams, signal vs. noise
- Small dataset with mysterious ranking criteria
- Algorithm selection


Thank you