NATURAL LANGUAGE PROCESSING

LECTURE 2: Topic Modeling

goorm

KAIST AI
Graduate School of AI

DAVIAN

Bag-of-Words Encoding of Text Documents

- Bag-of-words vector
  - Document 1 = "John likes movies. Mary likes too."
  - Document 2 = "John also likes football."

Vocabulary	Doc 1	Doc 2
John	1	1
likes	2	1
movies	1	0
also	0	1
football	0	1
Mary	1	0
too	1	0

...

What is a Topic Modeling?

- A topic is a probability distribution over keywords.
  - A different keyword has a different probability.
- Alternatively, a topic is a weighted combination of keywords.
  - A different keyword has a different importance score (or simply a different weight).

- What is a topic modeling?
  - Topic modeling is a technique that extracts a set of topics out of document corpus.
  - Additionally, topic modeling represents a document as a probability distribution over topics.
  - More generally, topic modeling represents a document as a weighted combination of topics.

LDA on Reuters Data

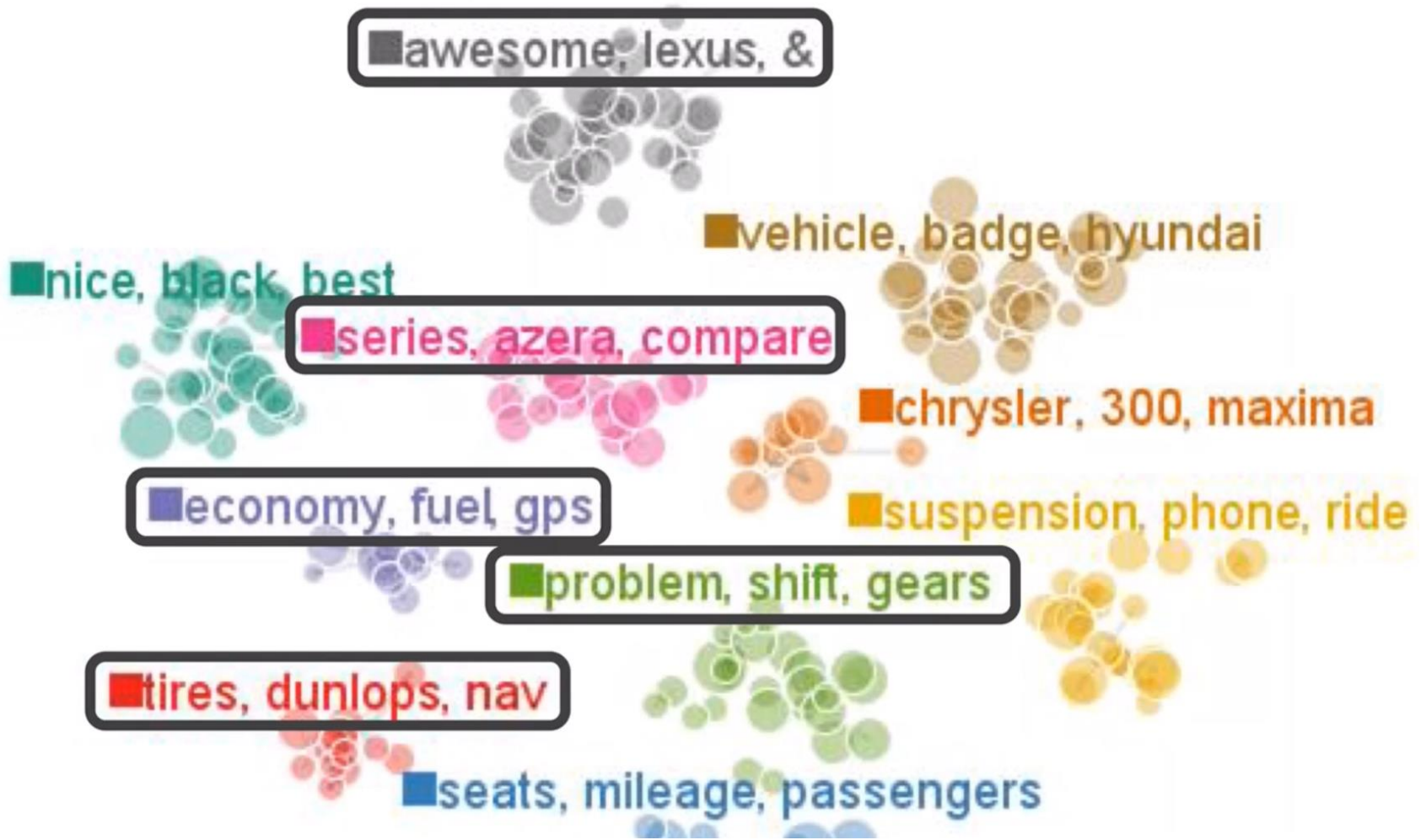
Recall:

- 100-topic LDA on 16,000 documents
- Some standard stopwords are removed.
- Top keywords for some $p(w|z)$.

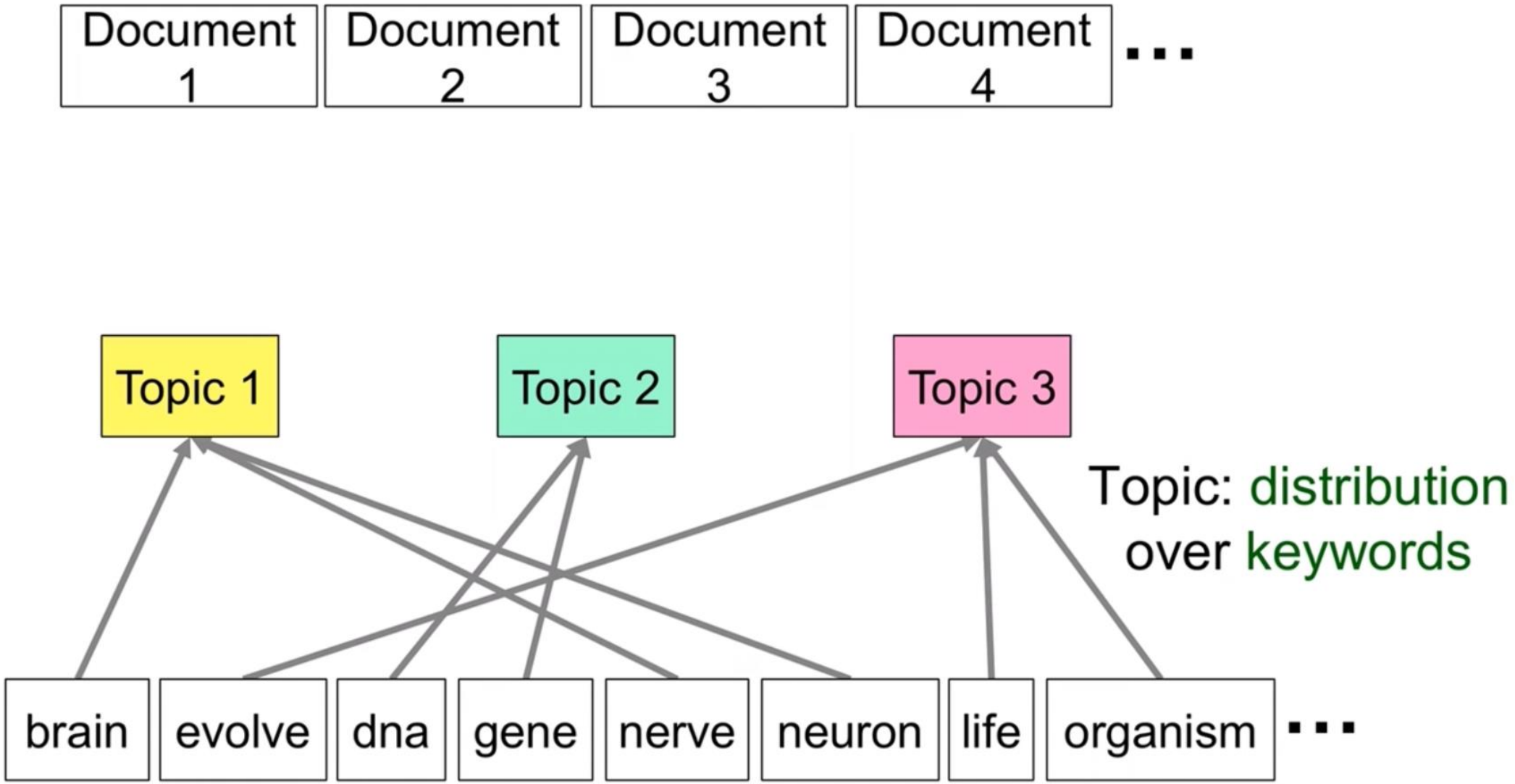
"Arts"	"Budgets"	"Children"	"Education"
new	million	children	school
film	tax	women	students
show	program	people	schools
music	budget	child	education
movie	billion	years	teachers
play	federal	families	high
musical	year	work	public

Visualization Example: Car Reviews

- Topic summaries are NOT perfect.



Topic Modeling: Overview



Topic Modeling: Overview