BAYESIAN STATISTICS

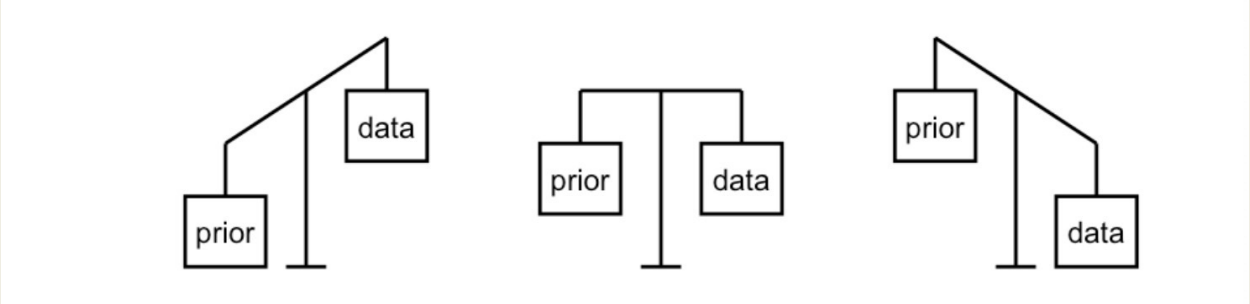
Mentor: Nicholas Irons
Mentee: Qianqian Yu

Contents :

- 1. Basic idea about Bayesian
- 2. Three components of the Bayesian model: prior, likelihood, and posterior
- 3. Latent Dirichlet Allocation (LDA) Model

Basic idea about Bayesian

- Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data



Bayesian Model

- U.S. Census figures provide prior information about the region in which people might live: the Midwest (M), Northeast (N), South (S), or West (W)
- For drinks, we have common regional terms including “pop,” “soda,” and “coke.”
- What is the distribution for different region use the term “pop”?

“U.S. Census figures provide prior information about the region in which they might live: the Midwest (M), Northeast (N), South (S), or West (W)”

- This is prior
- The information we know at first

TABLE 2.5: Prior model of U.S. region.

region	M	N	S	W	Total
probability	0.21	0.17	0.38	0.24	1

For drinks, we have common regional terms including “pop,” “soda,” and “coke.”

- This is data
- Eg. A boy who likely to use “pop”, maybe he is from south
- Then we introduce

  likelihood function: It indicates how likely a particular population is to produce an observed sample.

```r
# Load the data
data(pop_vs_soda)

# Summarize pop use by region
pop_vs_soda %>%
  tabyl(pop, region) %>%
  adorn_percentages("col")
```

```
   pop midwest northeast   south   west
 FALSE  0.3553    0.7266 0.92078 0.7057
  TRUE  0.6447    0.2734 0.07922 0.2943
```

By R

- A represent the use of “Pop”
- We can calculate likelihood Value:

$$L(M|A) = 0.6447, \quad L(N|A) = 0.2734, \quad L(S|A) = 0.0792, \quad L(W|A) = 0.2943$$

posterior probability

- **A posterior probability**, in Bayesian statistics, is the revised or updated probability of an event occurring after taking into consideration new information.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)L(B|A)}{P(A)}$$

BAYESIAN RULE

$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}}.$$

POSTERIOR CALCULATION

How we interpret posterior?

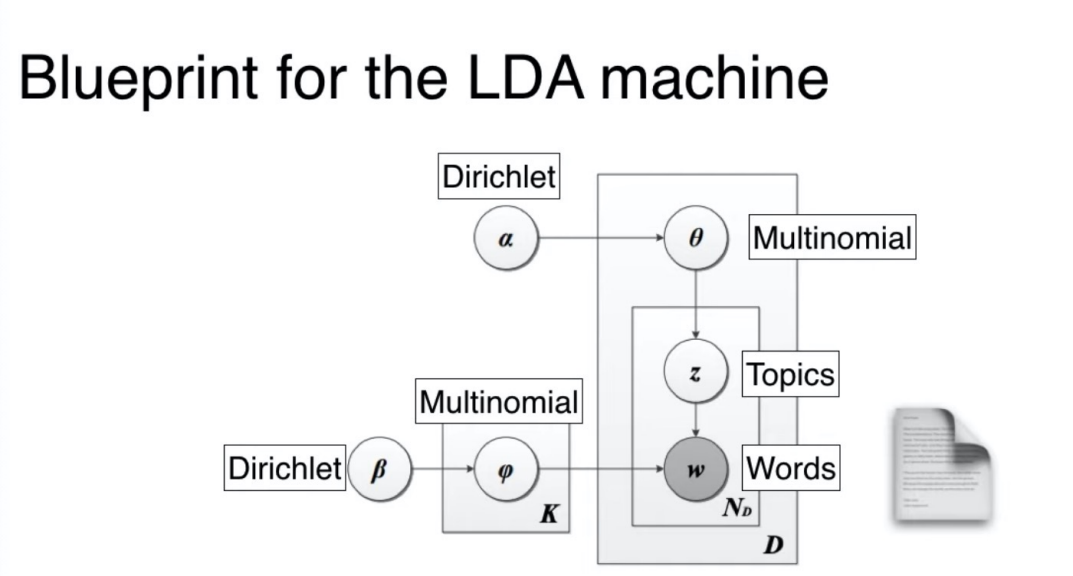
- A represent the use of “Pop”
- If we know one person use the term “Pop”, the person most likely from the Midwest.

$$P(S|A) = \frac{0.38 \cdot 0.0792}{0.2826} \approx 0.1065.$$

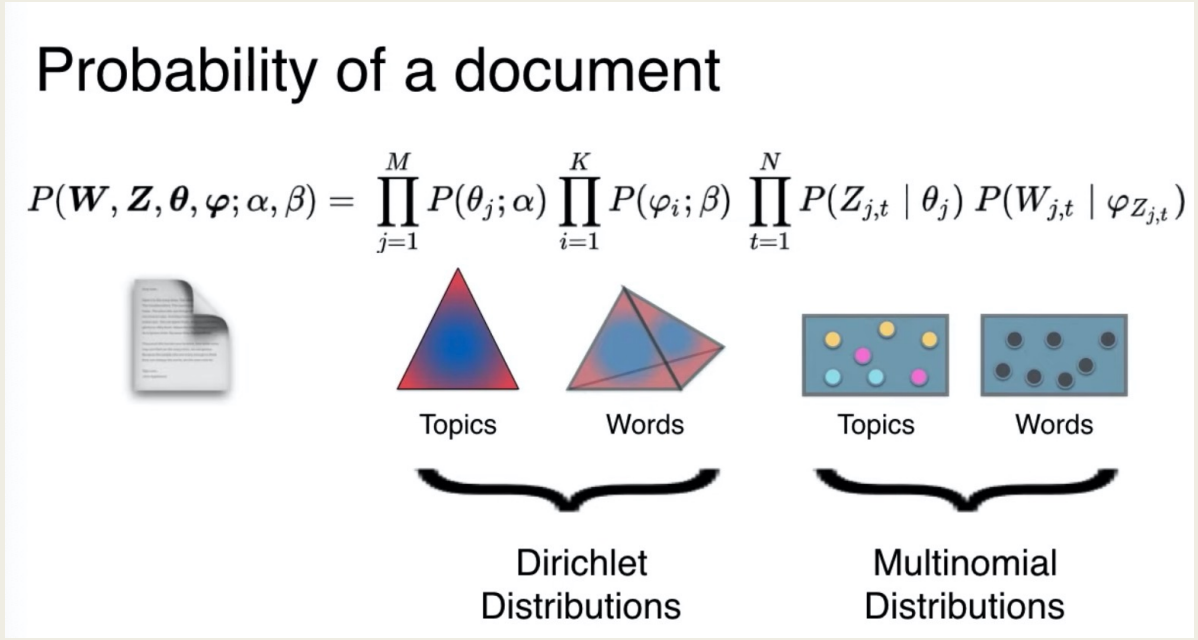
region	M	N	S	W	Total
prior probability	0.21	0.17	0.38	0.24	1
posterior probability	0.4791	0.1645	0.1065	0.2499	1

Latent Dirichlet Allocation (LDA) Model

- LDA is a Bayesian model used in natural language processing (NLP) to extract topics from a corpus of text documents



LDA



W --- words in document
Z --- Latent topic
θ — topic distribution
φ — topic-word distribution
α — per-document topic distribution(Dirichlet)
β — per topic word distribution(Dirichlet)

Party

- $\alpha = 1$

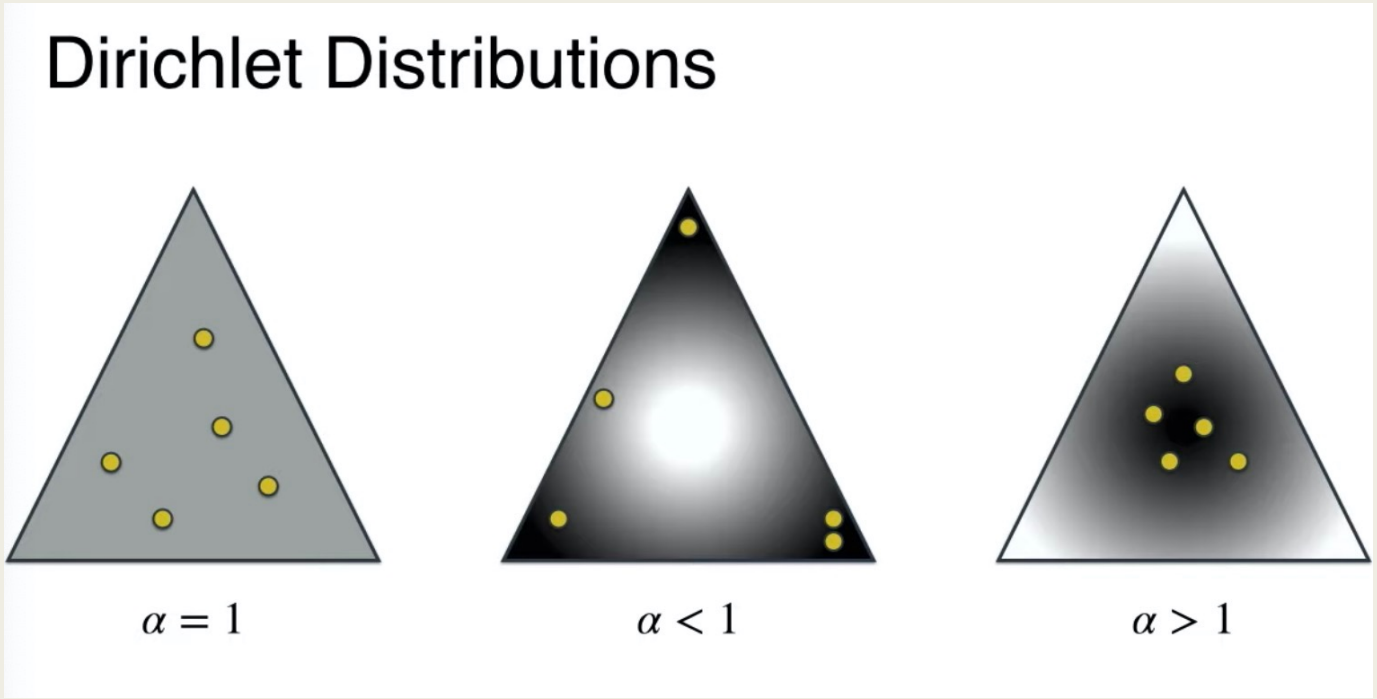
  no preference

- $\alpha < 1$

  movie, games, chips in the corner

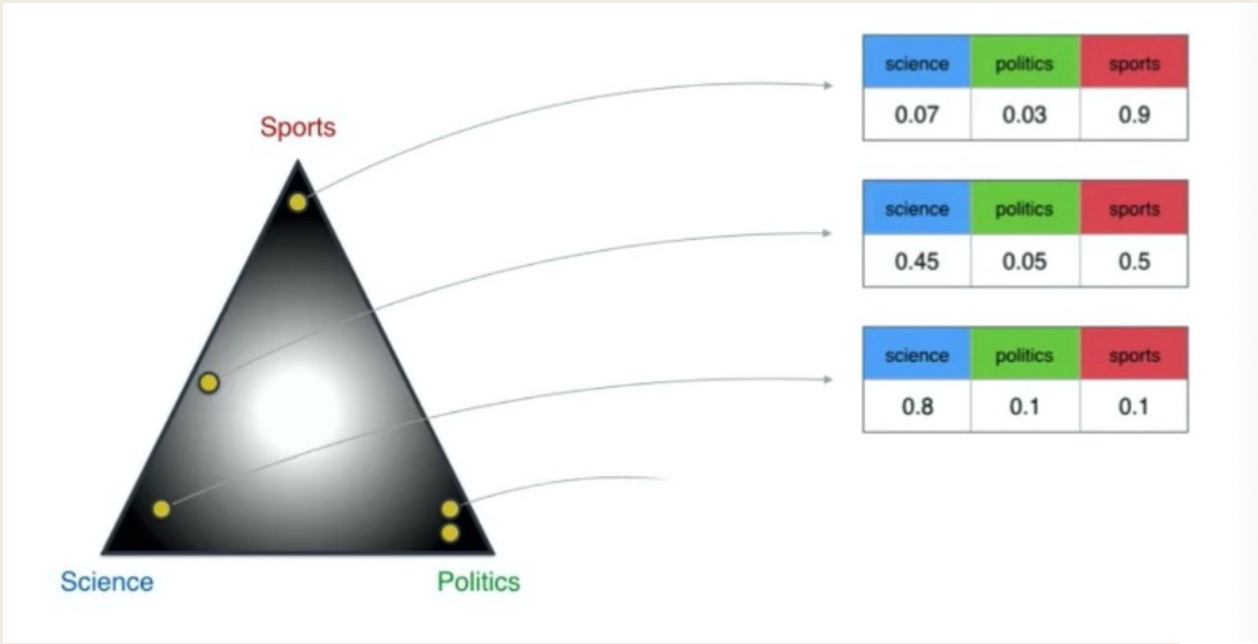
- $\alpha > 1$

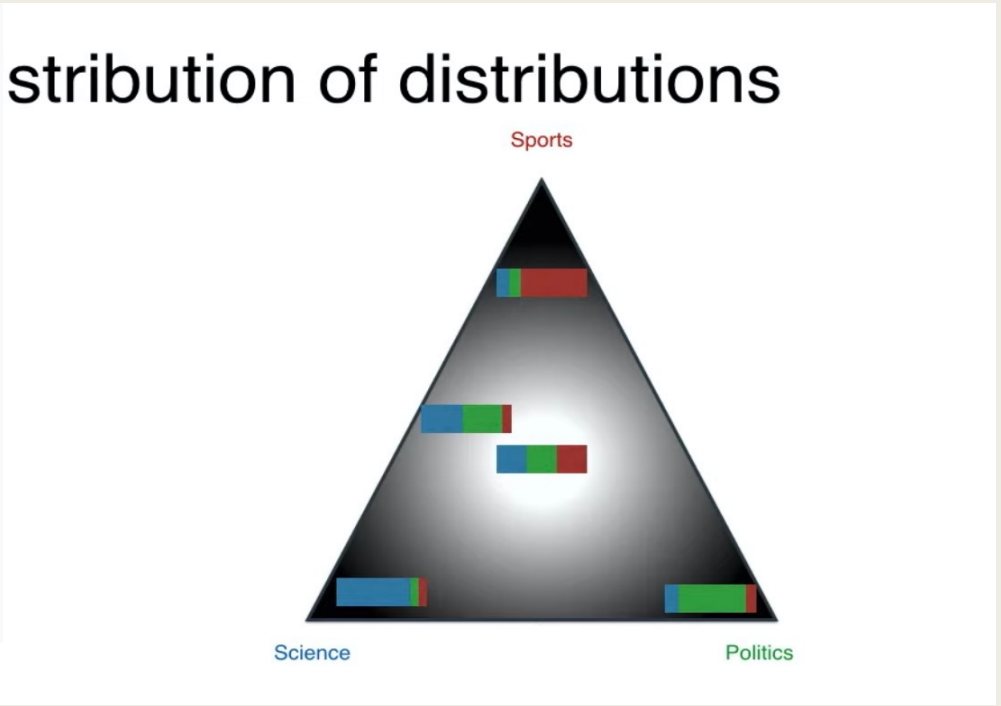
  tiger, wolf, in the corner



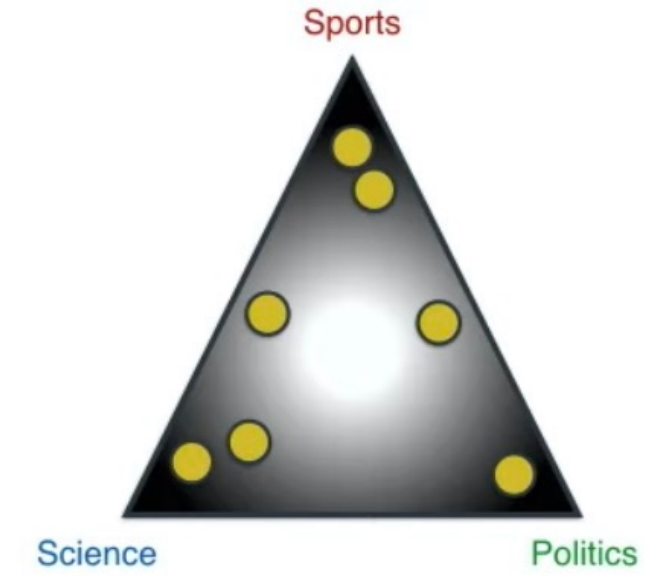
For document

- Point represents each document which contains different ratio of topics

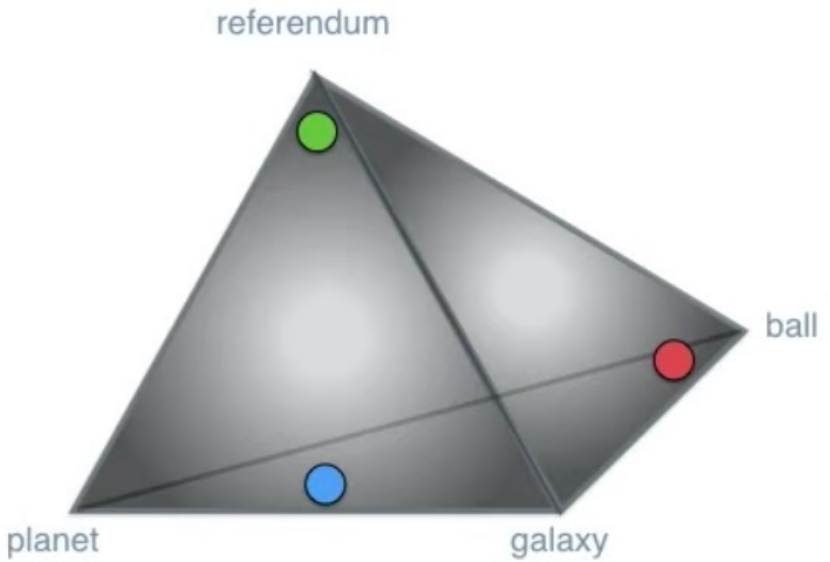




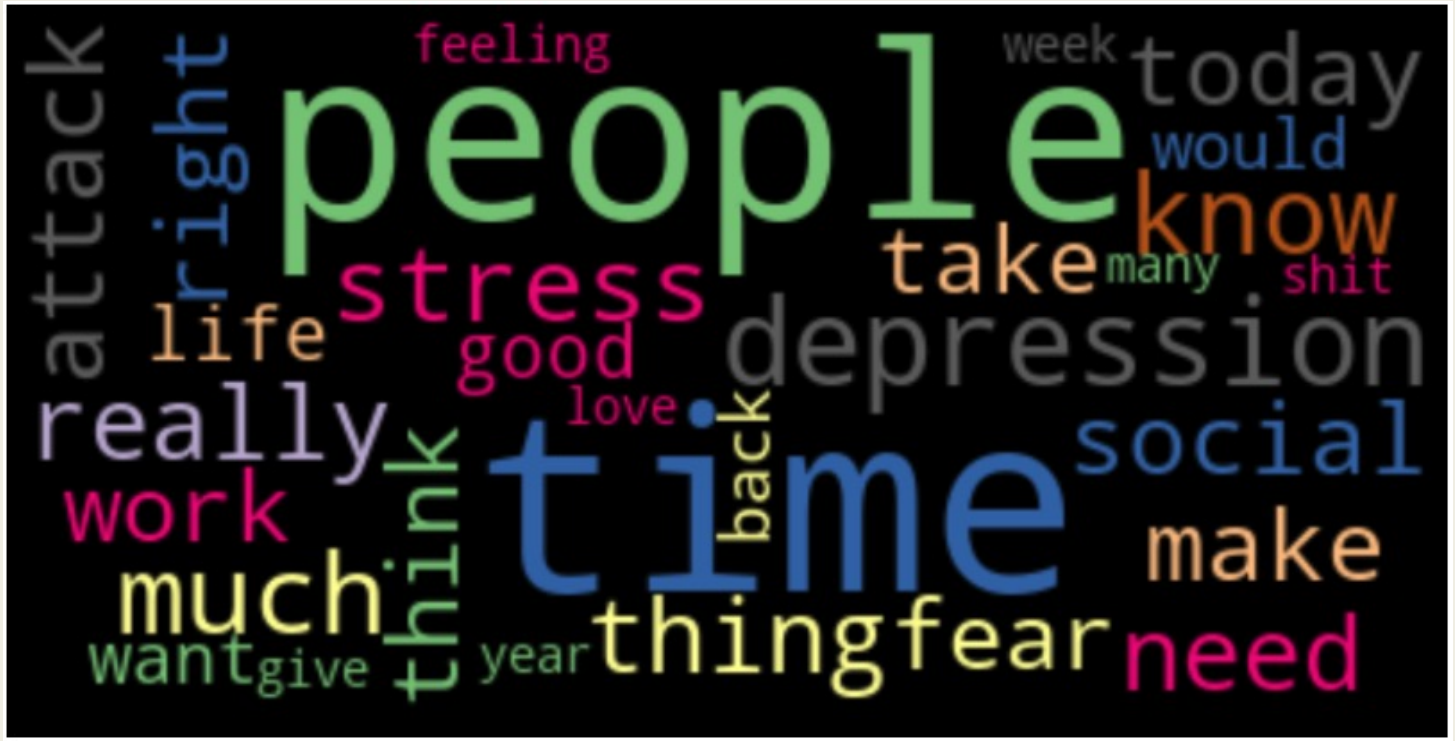
Two Dirichlet distributions



Documents-Topics



Topics-Words



Implication of LDA model

- Analysis unstructured text data
- How about tweet?
- Find the most frequent topic

Thanks