Text Mining w R

WhyR? 2017 - warsztat

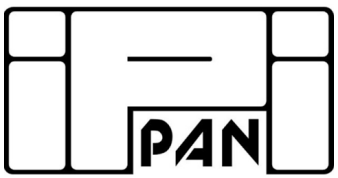
Norbert Ryciak

Politechnika Warszawska, Sages

4 października 2017

Prowadzący

- Wydział Matematyki i Nauk Informacyjnych PW
  - doktorant (zastosowanie deep learningu w analizie tekstu)



- Instytut Podstaw Informatyki PAN
  - projekt CLARIN (zadanie analizy sentymentu)

sages

- Sages
  - opiekun bootcampu Data Science, trener

Latent Dirichlet Allocation

Model LDA - wprowadzenie

Motywacja: przedstawienie tekstu jako mieszanki tematów.

Model LDA - wprowadzenie

Motywacja: przedstawienie tekstu jako mieszanki tematów.
Temat - rozkład prawdopodobieństwa na zbiorze słów.

Model LDA - wprowadzenie

Motywacja: przedstawienie tekstu jako mieszanki tematów.
Temat - rozkład prawdopodobieństwa na zbiorze słów.
Przykład:

- ~~Mam~~ gorączkę ~~i~~ katar.
- Graliśmy ~~w~~ siatkówkę.
- Sport ~~to~~ zdrowie.

Model LDA - wprowadzenie

Motywacja: przedstawienie tekstu jako mieszanki tematów.
Temat - rozkład prawdopodobieństwa na zbiorze słów.
Przykład:

- ~~Mam~~ gorączkę ~~i~~ katar.
- Graliśmy ~~w~~ siatkówkę.
- Sport ~~to~~ zdrowie.

Załóżmy, że mamy dwa tematy: „zdrowie” i „sport”. Wówczas:

Model LDA - wprowadzenie

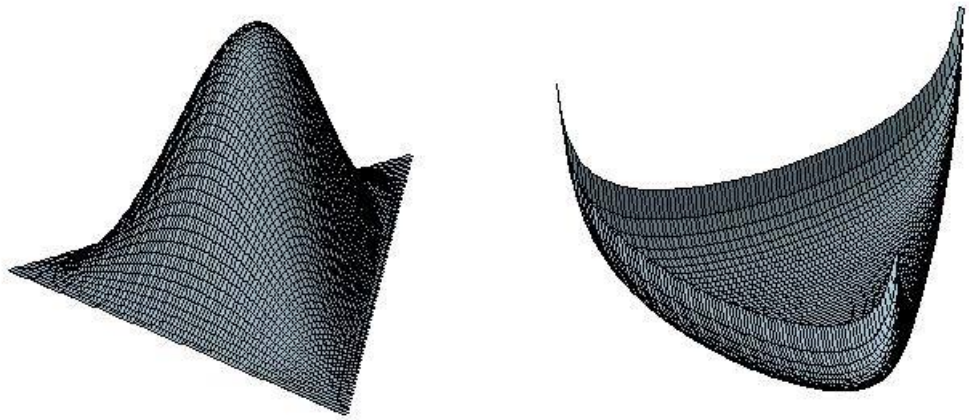
Motywacja: przedstawienie tekstu jako mieszanki tematów.
Temat - rozkład prawdopodobieństwa na zbiorze słów.
Przykład:

- ~~Mam~~ gorączkę ~~i~~ katar.
- Graliśmy ~~w~~ siatkówkę.
- Sport ~~to~~ zdrowie.

Załóżmy, że mamy dwa tematy: „zdrowie” i „sport”. Wówczas:

- Pierwsze zdanie = 100% zdrowie
- Drugie zdanie = 100% sport
- Trzecie zdanie = 50% sport + 50% zdrowie

Rozkład Dirichleta



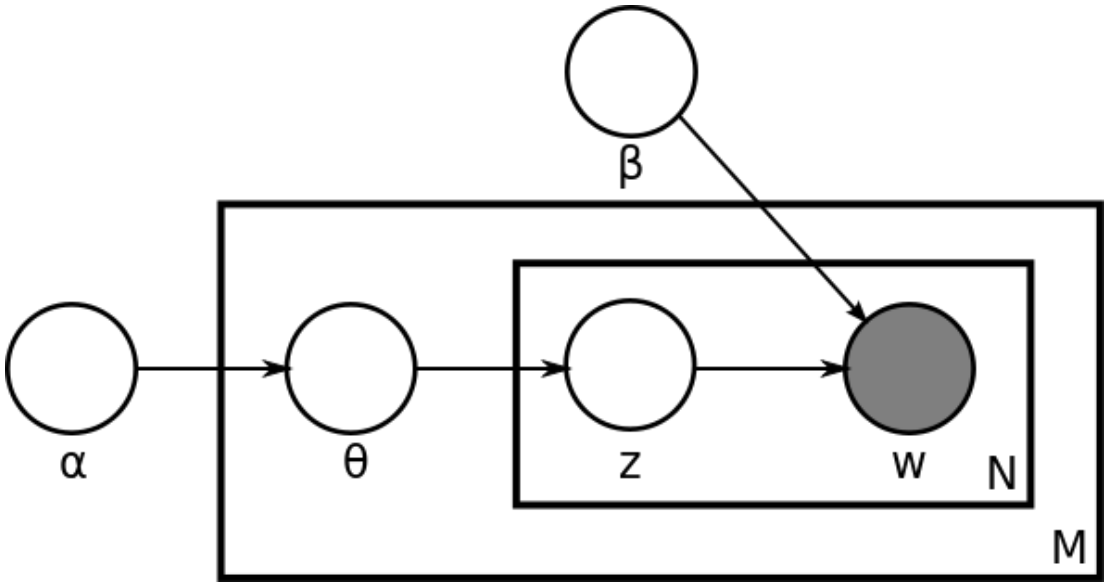
(a) $\alpha = 3$ (b) $\alpha = 0.95$

Rysunek: Gęstość trójwymiarowego rozkładu Dirichleta Dir(α).

Wektor losowy $(x_1, ..., x_K)$ z K-wymiarowego rozkładu Dirichleta to punkt na $(K-1)$-wymiarowym sympleksie, czyli $x_1 + ... + x_K = 1$, $x_i \geqslant 0$.

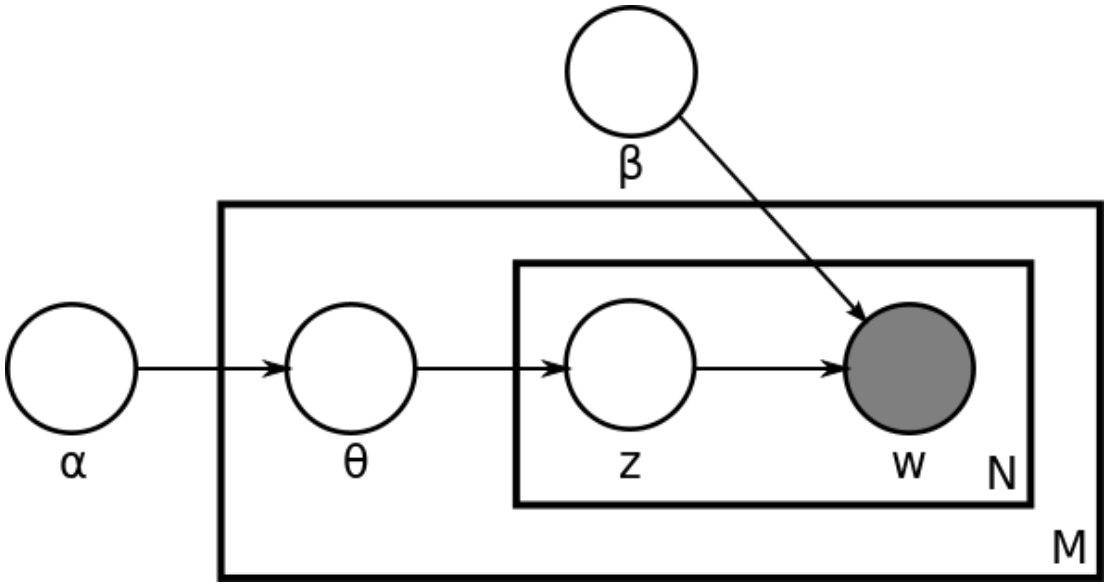
Model LDA - sformułowanie

LDA - Latent Dirichlet Allocation.



Model LDA - sformułowanie

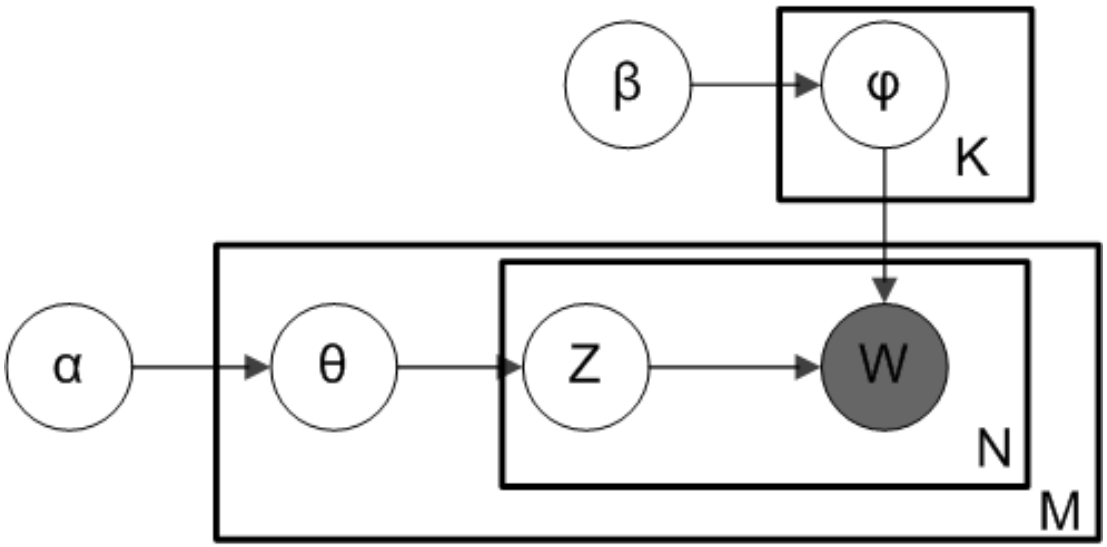
LDA - Latent Dirichlet Allocation.



Proces generowania dokumentu d:

1. Ustal liczbę słów w dokumencie N_d.
2. Losuj rozkład tematów w dokumencie $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$.
3. Dla $i = \{1, ..., N_d\}$:
   1. Losuj temat z_{di} z rozkładu dyskretnego $\boldsymbol{\theta}_d$.
   2. Losuj słowo w_{di} z rozkładu dyskretnego $\boldsymbol{\beta}_{z_{di}}$

Wygładzany model LDA



$\varphi_k \sim \text{Dir}(\boldsymbol{\beta})$ dla $k = 1, ..., K$,
$\boldsymbol{\beta} = (\beta_1, ..., \beta_V)$, $\beta_i \in \mathbb{R}$.

Estymacja w modelu LDA - próbkowanie Gibbsa

Estymacja w modelu LDA - próbkowanie Gibbsa

1. Inicjalizacja: każde słowo przypisz do losowego tematu.

Estymacja w modelu LDA - próbkowanie Gibbsa

1. Inicjalizacja: każde słowo przypisz do losowego tematu.
2. Dla każdego słowa w_{dj}, $j = 1, ..., N_d$, $d = 1, ..., M$, powtarzaj:
   - Oblicz $P(z_{dj} = k | \mathbf{w}, \mathbf{z}_{-dj}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ dla $k = 1, ..., K$.
   - Wylosuj z_{dj} z powyższego rozkładu.

Estymacja w modelu LDA - próbkowanie Gibbsa

1. Inicjalizacja: każde słowo przypisz do losowego tematu.
2. Dla każdego słowa w_{dj}, $j = 1, ..., N_d$, $d = 1, ..., M$, powtarzaj:
   - Oblicz $P(z_{dj} = k | \mathbf{w}, \mathbf{z}_{-dj}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ dla $k = 1, ..., K$.
   - Wylosuj z_{dj} z powyższego rozkładu.

$$P(z_{dn} = k | \mathbf{z}_{-(dn)}, \mathbf{w}, \boldsymbol{\alpha}, \beta) \propto \frac{n^{(\cdot)}_{k,w_{dn}} + \beta}{n^{(\cdot)}_{k,(\cdot)} + \beta V} \cdot \frac{n^{d}_{k,(\cdot)} + \alpha_k}{N_d + \sum_{k=1}^{K} \alpha_k},$$

$n^{d}_{k,w}$ - liczność przypisań słowa w do tematu k w dokumencie d.

Estymacja w modelu LDA - próbkowanie Gibbsa

1. Inicjalizacja: każde słowo przypisz do losowego tematu.
2. Dla każdego słowa w_{dj}, $j = 1, ..., N_d$, $d = 1, ..., M$, powtarzaj:
   - Oblicz $P(z_{dj} = k | \mathbf{w}, \mathbf{z}_{-dj}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ dla $k = 1, ..., K$.
   - Wylosuj z_{dj} z powyższego rozkładu.

$$P(z_{dn} = k | \mathbf{z}_{-(dn)}, \mathbf{w}, \boldsymbol{\alpha}, \beta) \propto \frac{n^{(\cdot)}_{k,w_{dn}} + \beta}{n^{(\cdot)}_{k,(\cdot)} + \beta V} \cdot \frac{n^{d}_{k,(\cdot)} + \alpha_k}{N_d + \sum_{k=1}^{K} \alpha_k},$$

$n^{d}_{k,w}$ - liczność przypisań słowa w do tematu k w dokumencie d.

$$\hat{\theta}_{di} = \frac{n^{d}_{i,(\cdot)} + \alpha_i}{N_d + \sum_{i=1}^{k} \alpha_i}, \quad i = 1, ..., k, \ d = 1, ..., M,$$

$$\hat{\varphi}_{ij} = \frac{n^{(\cdot)}_{i,j} + \beta}{n^{(\cdot)}_{i,(\cdot)} + \beta V}, \quad i = 1, ..., K, \ j = 1, ..., V.$$

Predykcja tematów w nowych dokumentach

Próbkowanie Gibbsa:

1. Inicjalizacja tematami najbardziej prawdopodobnymi dla odpowiednich słów.
2. Analogiczne losowanie sekwencyjne: losujemy na podstawie

$$P(z_{dn} = k|\mathbf{z}_{-(dn)}, \mathbf{w}, \boldsymbol{\alpha}, \beta) \propto \frac{n^d_{k,(\cdot)} + \alpha_k}{N_d + \sum_{k=1}^K \alpha_k}.$$

- $\frac{n^{(\cdot)}_{k,w_{dn}} + \beta}{n^{(\cdot)}_{k,(\cdot)} + \beta V}$ jest ustalone.