Rozdział 1

Baza teoretyczna

1.1. Podstawowe pojęcia

Tu będą definicje: sigma-ciało, funkcja mierzalna, miara, zmienna losowa, wektor losowy, rozkład prawdopodobieństwa, gęstość prawdopodobieństwa, zbieżność według rozkładu (a wtedy też pewnie trzeba zdefiniować miare i zbieżność według miary), rozkłady warunkowe, twierdzenie bayesa

Pytanie: czy to jest odpowiedni poziom szczegółowości?

1.2. Wybrane rozkłady prawdopodobieństwa

W niniejszym podrozdziale przedstawione są wykorzystywane w pracy rozkłady prawdopodobieństwa. Opisane są ich własności oraz interpretacje, które są istotne dla dobrego rozumienia przedmiotu pracy.

1.2.1. Rozkład kategoryczny oraz wielomianowy

W teorii matematycznej najczęściej rozpatruje się rozkłady prawdopodobieństwa zmiennych losowych, które przyjmują wartości w $\mathbb{R}^k$. W ten sposób modeluje się wiele zjawisk w rzeczywistości niezwiązanych z liczbami. Przykładem może być tutaj rzut monetą. W rozważaniach matematycznych naturalnym podejściem jest zamodelowanie eksperymentu w następujący sposób: mamy do czynienie ze zmienną losową, która przyjmuje wartość 0, gdy wypada reszta oraz wartość 1, w przypadku orła. To pozwala nam badać pewne własności świata przy użyciu aparatu matematycznego i wówczas możemy np. policzyć wartość oczekiwaną tej zmiennej, co dostarcza pewnego opisu rzeczywistości. W praktyce niekiedy zachodzi potrzeba modelowania zdarzeń losowych, dla których interpretacja liczbowa nie ma sensu lub po prostu nie jest potrzebna. Mając powyższe na uwadze, opiszemy poniżej pewne rozkłady prawdopodobieństwa.

Definicja 1.1 (Rozkład Bernoulliego i dwupunktowy).
Rozkład Bernoulliego opisuje prawdopodobieństwo p przyjęcia wartości 1 przez zmienną losową X, o wartościach w zbiorze $\{0,1\}$, tzn:

$$P(X=1) = p \tag{1.1}$$
$$P(X=0) = 1-p. \tag{1.2}$$

Przez rozkład dwupunktowy rozumiemy analog rozkładu Bernoulliego dla klas:

$$P(X \text{ jest klasy pierwszej}) = p \tag{1.3}$$
$$P(X \text{ jest klasy drugiej}) = 1-p. \tag{1.4}$$

Uwaga. W praktyce często powyższe nazwy są używane wymiennie.

Definicja 1.2 (Rozkład dyskretny i kategoryczny).
Rozkład dyskretny jest uogólnieniem rozkładu Bernoulliego na wiele wartości - jest to ciąg par (k, p_k), gdzie $p_k = P(X=k)$, $k = 1, ..., n$, $n \in \mathbb{N} \cup +\infty$, $\sum_{k=1}^{n} p_k = 1$. Rozkład oznaczamy przez Discr$(p_1, ..., p_n)$

Rozkład kategoryczny jest uogólnieniem rozkładu dwupunktowego na wiele klas - jest to ciąg par (K, p_K), gdzie $p_K = P(X=K)$, $K \in \{K_1, ..., K_n\}$, $n \in \mathbb{N} \cup +\infty$, $\sum_{K \in \{K_1,...,K_n\}} p_K = 1$. Rozkład oznaczamy przez Categ$(p_{K_1}, ..., p_{K_n})$.

Definicja 1.3 (Rozkład dwumianowy i wielomianowy).
Rozkład dwumianowy opisuje prawdopodobieństwo uzyskania k sukcesów w ciągu n niezależnych prób, w których każda ma prawdopodobieństwo sukcesu p. z Zmienna losowa X, pochodząca z rozkładu dwumianowego bin(p, n), ma rozkład postaci:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad n \in \mathbb{N}, \ k \in \{0, ..., n\}.$$

Rozkład wielomianowy opisuje prawdopodobieństwo uzyskania n_1 razy wartości 1, n_2 wartości 2, ..., n_m wartości m, dla $n = \sum_{i=1}^{m} n_i$ prób, w których każda ma prawdopodobieństwo otrzymania wartości k równe p_k, $k = 1, ..., m$. Wektor losowy $(X_1, ..., X_m)$, pochodzący z rozkładu wielomianowego Mult$((p_1, ..., p_m), n)$, ma rozkład postaci:

$$\begin{aligned} p(n_1, n_2, ...n_m) &= \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m} \\ &= \frac{\Gamma(n+1)}{\prod_{k=1}^{m} \Gamma(n_k+1)} p_1^{n_1} \cdots p_m^{n_m}, \quad n = \sum_{k=1}^{m} n_k \end{aligned} \tag{1.5}$$

Uwaga 1. Rozkład wielomianowy można stosować bezposrednio do danych kategorycznych, interpretując wartość k jako klasę K_k.

Uwaga 2. Rozkład wielomianowy dla $n = 1$ staje się rozkładem dyskretnym (lub kategorycznym). W literaturze podczas stosowania rozkładu kategorycznego, często oznacza się go jako

rozkład wielomianomy, bez zaznaczenia faktu $n = 1$, co może prowadzić do nieporozumień. W niniejszej pracy używane będzie oznaczenie Categ$(\cdot)$.

Uwaga 3. W pracy będziemy utożsamiać tzw. wektor indykatorowy (wektor o wartości 1 na jednej wpółrzędnej oraz 0 na pozostałych) z indeksem współrzędnej, na której przyjmowana jest wartość 1. Zatem definiując zbiór klas jako zbiór wektorów indykatorowych $\{K_1, ..., K_n\}$, gdzie K_i przyjmuje wartość 1 na i-tej współrzędnej, utożsamiać będziemy również rozkłady:

- rozkład kategoryczny Categ$(p_1, ..., p_n)$, gdzie p_k oznacza prawdopodobieństwo wylosowania klasy K_k,
- rozkład dyskretny $(p_1, ..., p_n)$, gdzie p_k oznacza prawdopodobieństwo wylosowania liczby k.

1.2.2. Rozkład Dirichleta

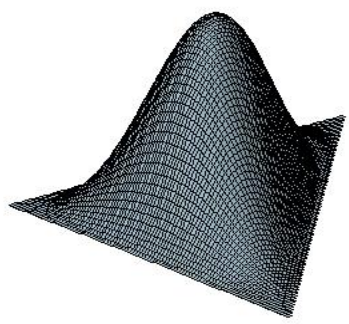
Kolejnym ważnym rozkładem, na którym opiera się model LDA, jest rozkład Dirichleta. Jest to wielowymiarowy rozkład ciągły, parametryzowany przez wektor dodatnich liczb rzeczywistych α (wymiar parametru jest jednocześnie wymiarem rozkładu). Jego gęstość wyraża się wzorem

$$f(x_1, ..., x_{K-1}; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

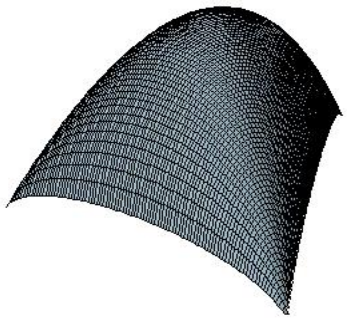
gdzie $x_K = 1 - x_1 - ... - x_{k-1}$, a nośnikiem gęstości jest wnętrze $(K-1)$-wymiarowego sympleksu, opisane nierównościami: $x_1, ..., x_K > 0$ oraz $x_1 + ... + x_{K-1} < 1$. Stałą normująca $B(\alpha) = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{K} \alpha_i)$ jest uogólnieniem funkcji Beta.

Z perspektywy modelu LDA ważna jest interpretacja tego rozkładu. K-wymiarowy rozkład Dirichleta opisuje gęstość prawdopodobieństwa na wnętrzu $(K-1)$-wymiarowego sympleksu. Innymi słowy wektor losowy $(X_1, ..., X_K)$ pochodzący z tego rozkładu jest wymiaru K, a wartościami, które przyjmuje są punkty należące do wnętrza tego sympleksu, które spełniają warunki: $x_1, ..., x_K > 0$ oraz $x_1 + ... + x_K = 1$. Te dwie własności sprawiają, że wartości wektora losowego pochodzącego z rozkładu Dirichleta, mogą zostać wykorzystane do zdefiniowania skończonego dyskretnego rozkładu prawdopodobieństwa, co jest wykorzystane w modelu LDA.

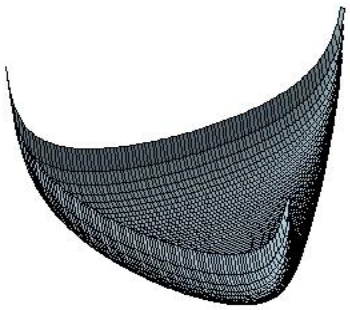
Opiszemy teraz wpływ parametru α na kształt gęstości. Rozpatrzmy najpierw sytację, gdy $\alpha_1 = ... = \alpha_K = \alpha$. Wówczas gęstość jest symetryczna i środek ciężkości masy prawdopodobieństwa pokrywa się ze środkiem geometrycznym sympleksu. Jeżeli $\alpha > 1$, to gęstość jest jedno-modalna (z modą leżącą na środku sympleksu), przy czym im większa wartość α, tym gęstość jest bardziej skupiona wokół mody. Gdy $\alpha = 1$, gęstość jest funkcją stałą, a zatem rozkład jest równomierny. Natomiast gdy $\alpha < 1$, to wówczas gęstość jest funkcją wklęsłą, przyjmując największe wartości w narożnikach sympleksu, przy czym im α jest mniejsza, tym bardziej masa prawdopodobieństwa skupia się w narożnikach. Przykłady dla rozkładu 3-wymiarowego zostały przedstawione na rysunku 1.1. W sytuacji, gdy współrzędne α mają różne wartości, gęstość wygląda podobnie, z tą różnicą, ze środek masy prawdopodobieństwa jest przesunięty i przez to gęstość nie jest symetryczna, a kierunek przesunięcia zależy od tego, na których współrzędnych stoją większe wartości. Wpływ wartości współrzędnych wektora α jest analogiczny - im większe wartości tym masa bardzie skupiona wokół mody, a im mniejsze, tym bardziej skupiona w narożnikach.



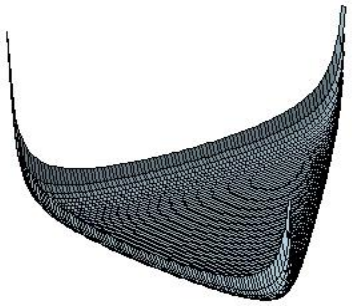
(a) $\alpha = 3$



(b) $\alpha = 1.5$



(c) $\alpha = 0.95$



(d) $\alpha = 0.5$

Rysunek 1.1: Symetryczny 3-wymiarowy rozkład Dirichleta

Wektor losowy $(X_1, ..., X_k) \sim \text{Dir}(\alpha)$ ma następujące własności:

1. $EX_i = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}$,
2. $E \ln X_i = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{k} \alpha_j), \quad i = 1, ..., k.$

Ponadto dla rozkładu Dirichleta zachodzi następujący lemat:

Lemat 1.1. *Niech $\alpha \in \mathbb{R}^k$ oraz $p = (p_1, ..., p_k) \sim Dir(\alpha)$. Niech $X_1, ..., X_n$ będzie próbą iid z rozkładu p oraz $n_i = \#\{j : X_j = i\}$, $i = 1, ..., k$. Wówczas rozkład warunkowy p pod warunkiem parametru α oraz liczności n_i jest rozkładem $Dir(\alpha_1 + n_1, ..., \alpha_k + n_k)$.*

Powyższy lemat jest bardzo użyteczny w praktyce i wykorzystywany w modelu LDA.

1.2.3. Rozklad wielomianowy Dirichleta

Załóżmy, że losujemy wektor $p = (p_1, ..., p_k)$ z rozkładu $\text{Dir}(\alpha)$, $\alpha \in \mathbb{R}^k$. Traktując wektor p jako dyskretny rozkład prawdopodobieństwa, losujemy z niego n-elementową próbkę $X_1, ..., X_n$. Rozkład wygenerowanego według tego schematu wektora $(X_1, ..., X_n)$, parametryzowany przez α, nazywamy rozkładem wielomianowym Dirichleta i oznaczamy $\text{DirMult}(\alpha)$.

Ma on postać:

$$P(X_1 = x_1, ..., X_n = x_n) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(n + \sum_{i=1}^k \alpha_i)} \prod_{i=1}^k \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)},$$

gdzie $n_i = \#\{j : x_j = i\}$, $\sum_{i=1}^k n_i = n$.

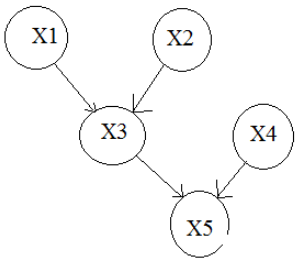
W opisanym schemacie generowania próby ważne jest również pytanie o rozkład liczności konkretnych wartości otrzymanych w próbie $X_1, ..., X_n$. Oznaczając przez N_i, $i = 1, ..., k$ zmienną losową określającą liczność zaobserwowanych wartości i - $N_i = \#\{j : X_j = i\}$, rozkład $P(N_1 = n_1, ..., N_k = n_k), \sum_{i=1}^k n_i = n$, również nazywa się rozkładem wielomianowym Dirichleta i ma on postać:

$$\begin{aligned} P(N_1 = n_1, ..., N_k = n_k) &= \frac{n!}{\prod_{i=1}^k n_k!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(n + \sum_{i=1}^k \alpha_i)} \prod_{i=1}^k \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)} \\ &= \frac{\Gamma(n+1)}{\prod_{i=1}^k \Gamma(n_k + 1)} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(n + \sum_{i=1}^k \alpha_i)} \prod_{i=1}^k \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)}. \end{aligned} \tag{1.6}$$

Uzasadnimy postać pierwszego rozkładu. Rozkład ten można łatwo otrzymać, całkując rozkład łączny $P((x_1, ..., x_n), p)$ względem p. Próbę $X_1, ..., X_n$ losujemy z rozkładu dyskretnego, stąd $P((X_1, ..., X_n) = (x_1, ..., x_n)|p) = p_{x_1} \cdots p_{x_n}$, a $p \sim \text{Dir}(\alpha)$. Zatem:

$$\begin{aligned} P(X_1 = 1_1, ..., X_n = x_n) &= \int P((x_1, ..., x_n), p)\ dp \\ &= \int P((x_1, ..., x_n)|p) P(p|\alpha)\ dp \\ &= \int p_{x_1} \cdots p_{x_n} \frac{1}{B(\alpha)} \prod_{i=1}^k p_i^{\alpha_i - 1}\ dp \\ &= \int p_1^{n_1} \cdots p_k^{n_k} \frac{1}{B(\alpha)} \prod_{i=1}^k p_i^{\alpha_i - 1}\ dp \\ &= \frac{1}{B(\alpha)} \int \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i}\ dp \\ &= \frac{B(\alpha_1 + n_1, ..., \alpha_k + n_k)}{B(\alpha)} \int \frac{1}{B(\alpha_1 + n_1, ..., \alpha_k + n_k)} \prod_{i=1}^k p_i^{\alpha_i + n_i - 1}\ dp \\ &= \frac{B(\alpha_1 + n_1, ..., \alpha_k + n_k)}{B(\alpha)} \\ &= \prod_{i=1}^k \Gamma(\alpha_i + n_i) / \Gamma(\sum_{i=1}^K (\alpha_i + n_i)) \times \Gamma(\sum_{i=1}^k (\alpha_i)) / \prod_{i=1}^k \Gamma(\alpha_i) \\ &= \frac{\Gamma(\sum_{i=1}^k (\alpha_i))}{\Gamma(\sum_{i=1}^k (\alpha_i + n_i))} \prod_{i=1}^k \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)} \end{aligned} \tag{1.7}$$

Drugi rozkład jest bezpośrednią konsekwencją pierwszego - jest to suma po wszystkich możliwych kombinacjach o ustalonych licznościach $n_1, ..., n_k$.



Rysunek 1.2

Zauważmy, że wartość funkcji prawdopodobieństwa $P(X_1 = x_1, ..., X_n = x_n)$ rozkładu wielomianowego Dirichleta nie zależy od kolejności wartości $x_1, ..., x_n$. Własność ta nazywa się wymienialnością:

Definicja 1.4. *Mówimy, że skończony ciąg zmiennych losowych $x_1, ..., x_n$, pochodzących z pewnego rozkładu p, jest wymienialny, jeśli dla dowolnej kombinacji $\sigma(x) = (\sigma(x_1), ..., \sigma(x_n))$, zachodzi tożsamość*

$$p(x_1, ..., x_n) = p(\sigma(x_1), ..., \sigma(x_n)).$$

Ciąg nieskończony nazwiemy wymienialnym, jeśli każdy jego podciąg skończony jest wymienialny.

1.3. Graficzne modele probabilistyczne

Graficzne modele probabilistyczne są reprezentacją zależności grupy zmiennych losowych przy pomocy grafu. Model definiują dwa elementy:

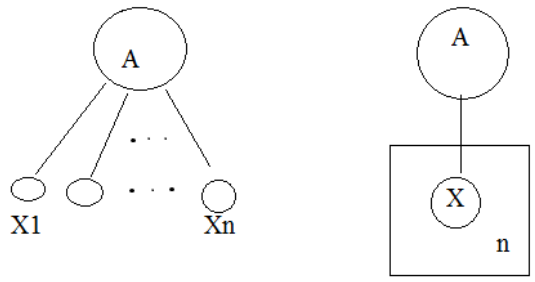
1. graf, którego wierzchołki reprezentują zmienne losowe, a krawędzie oznaczają istnienie bezpośredniej zależności miedzy dwoma zmiennymi,
2. lokalne rozkłady prawdopodobieństwa, definiujące warunkowe rozkłady zmiennych pod warunkiem ich sąsiadów.

Model graficzny, w którym graf jest nieskierowany, nazywa się siecią Markowa. W przypadku, gdy graf w modelu jest acykliczny i skierowany, to wówczas model nazywa się siecią Bayesowską, przy czym kierunki krawędzi wyznaczają jednostronną bespośrednią zalezność rozkładu zmiennej losowej od jej rodzica. Zaznaczmy, że w reprezentacji graficznej modelu można uwzględnić również nielosowe parametry rozkładów.

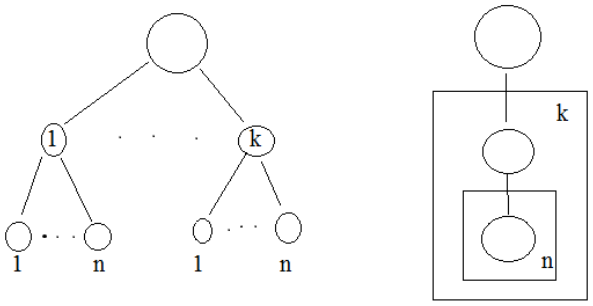
Dysponując warunkowymi rozkładami w siecie Bayesowskiej możemy w prosty sposób wyznaczyć rozkład łączny zmiennych, który wynika z faktoryzacji. Przykładowo dla sieci przedstawionej na rys. 1.2, w której określone są rozkłady $P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4), P(X_5|X_3, X_4)$, zachodzi:

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_5|X_3, X_4)P(X_4)P(X_3|X_1, X_2)P(X_1)P(X_2)$$

Przedstawienie graficznej graficzne bardzo dużej ilości zmiennych jest praktycznie niemożliwe. W tym celu przyjęła się pewna konwencja rysowania ("plate notation"), która pozwala reprezentować duże modele w sposób zwięzły (o ile istnieją pewne regularności w grafie modelu).



Rysunek 1.3



Rysunek 1.4

Polega ona na przedstawieniu jedynie powatarzającego się schematu i zaznaczeniu sposobu jego powielania. Rys. 1.3 przedstawia zasadę działania konwencji.

Na rysunku 1.4 przedsawiony jest bardziej złożony przykład:

1.4. Dywergencja Kullbacka-Leiblera

Definicja 1.5. *Niech* $\mathbf{P}, \mathbf{Q}$ *będą dyskretnymi rozkładami prawdopodobieństwa. Wówczas dywergencja Kullbacka-Leiblera zdefiniowana jest wzorem:*

$$D_{KL}(\mathbf{P}\|\mathbf{Q}) = \sum_i P_i \ln \frac{P_i}{Q_i}.$$

Dla rozkładów ciągłych o gęstościach $p(x), q(x), x \in X$*, dywergencja Kullbacka-Leiblera zdefiniowana jest wzorem:*

$$D_{KL}(p\|q) = \int_X p(x) \ln \frac{p(x)}{q(x)}\ dx.$$

Dywergencję Kullbacka-Leiblera można interpretować jako pewna miara podobieństwa (odmienności) między rozkładami prawdopodobieństwa. Zaznaczmy, że nie jest to metryka, ponieważ nie spełnia ona warunku symetryczności. Jednakże, następujący fakt pozwala stosować ją jako "odległość" miedzy rozkładami:

Twierdzenie 1.2. *Niech P będzie pewnym rozkładem prawdopodobieństwa, a $\{P_1, P_2, ...\}$ będzie ciągiem rozkładów. Wówczas*

$$D_{KL}(P_n \| P) \to 0 \implies P_n \xrightarrow{d} P, \quad \text{przy } n \to \infty.$$

1.5. Wybrane własności funkcji Gamma

Definicja 1.6.
Funkcja Gamma zdefiniowana jest wzorem:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt \quad dla\ Rez > 0.$$

Definicja 1.7.
Stałą Eulera-Mascheroniego γ definiujemy wzorem:

$$\gamma = \lim_{n\to\infty} \big(\sum_{k=1}^{n} \frac{1}{k} - \log n \big) = 1 + \sum_{k=1}^{\infty} (\frac{1}{k} + \log(1 - \frac{1}{k}))$$

Twierdzenie 1.3 (Wzór Weistrassa).
Zachodzi następująca tożsamość:

$$\frac{1}{\Gamma(z)} = z e^{\gamma z} \prod_{n=1}^{\infty} (1 + \frac{z}{n} \exp(-\frac{z}{n})$$

Definicja 1.8.
Funkcją Ψ nazywany pochodną loagrytmiczną funkcji Gamma:

$$\Psi(z) = \frac{\partial}{\partial z} \Gamma(z) \quad dla\ Rez > 0.$$

Lemat 1.4.
Zachodzą następujące tożsamości:

$$\log \Gamma(z) = -\log z - \gamma z + \sum_{n=1}^{\infty} (\frac{z}{n} - \log(1 + \frac{z}{n}))) \tag{1.8}$$

$$\Psi(z) = -\gamma + \sum_{n=0}^{\infty} (\frac{1}{n+1} - \frac{1}{n+z}) \tag{1.9}$$

$$\Psi'(z) = \sum_{n=0}^{\infty} \frac{1}{(n+z)^2} \tag{1.10}$$

Jest to bezpośrednia konsekwencja wzoru Weistrassa. Z powyższego lematu wynikają natychmiast następujące wnioski:

Wniosek 1.5. *Dla zmiennej rzeczywistej x:*

- *Funkcja $\Psi(x)$ jest rosnąca,*
- *Funkcja $\Psi'(x)$ jest malejąca.*

Rozdział 2

Model LDA

2.1. Latent Dirichlet Allocation

Model LDA jest przykładem probabilistycznego modelu graficznego, a konkretniej sieci Bayesowskiej, który został zaproponowany do modelowania danych tekstowych. Ideą LDA jest reprezentacja tekstu jako mieszanka tematów, definiowanych jako rozkłady prawdopodobieństwa na zbiorze słów. W modelu zakłada się, że każde słowo w dokumencie pochodzi z pewnego tematu i to pochodzenie staramy się odkryć, by ostatecznie opisać dokument jako procentowy rozkład zawartych w nim tematów. LDA można zatem traktować jako pewną metodę redukcji wymiaru, gdyż staramy się przedstawić zbiór tekstów, przy pomocy pewnej (mniejszej niż liczba tekstów) ilość tematów. Ponadto, z praktycznego punktu widzenia często ważna jest również interpretacja tematów.

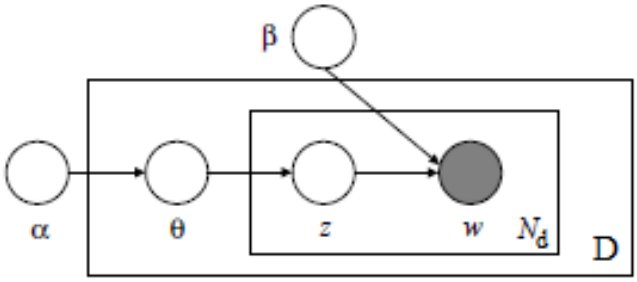
W niniejszej pracy opisujemy model LDA w zastosowaniu do danych tekstowych. Do formalnego opisu modelu, będzie potrzebne kilka pojeć, które teraz zdefiniujemy:

- Słowo w - podstawowa jednostka danych. Formalnie wektor V-wymiarowy, gdzie V - liczba unikalnych słów, przyjmujący wartość 1 na jednej współrzędnej oraz 0 na pozostałych. Fakt $w^i = 1$ można interpretować tak, że w jest i-tym słowem (przy założeniu ustalonej kolejości słów) ze zbioru wszystkich V słów.
- Dokument - skończony ciąg słów.
- Korpus - zbiór dokumentów.
- Temat - dyskretny rozkład prawdopodobieństwa wymiaru V, opisujący rozkład na zbiorze słów. W pracy będziemy wymiennie używać sformułowań "rozkład słów w temacie", "rozkład tematu", "temat".

W modelu LDA występują trzy parametry:

- K - liczba tematów.
- α - wektor sterujący rozkładami tematów w dokumentach.
- β - macierz wymiaru $K \times V$, której i-ty wiersz opisuje rozkład i-tego tematu.

Jedynym parametrem, którego wartość trzeba zadać jest K. Oznacza to, że musimy a priori założyć ile tematów w danych chcemy wykryć. Wartości α oraz β można ustalić z góry lub wyestymować ich wartości, w zależności od problemu, do którego wykorzystujemy model.



Rysunek 2.1

Graficzna reprezentacja modelu LDA jest przedstawiona na rysunku 2.1.

Model reprezentuje on następujący proces powstawania dokumentu tekstowego d:

1. Ustal liczbę słów w dokumencie N_d (nie zakładamy rozkładu, z którego ta wartość pochodzi, ponieważ dla modelu nie ma to żadnego znaczenia).
2. Wylosuj rozkład tematów w dokumencie $\theta_d \sim \text{Dir}(\alpha)$.
3. Dla każdego $i = 1, ..., N_d$:
   - Wylosuj temat $z_{di} \sim \text{Discr}(\theta_d)$, z którego zostanie wygenerowane słowo.
   - Wylosuj słowo $w_{di} \sim \text{Categ}(\beta_{z_{di}})$.

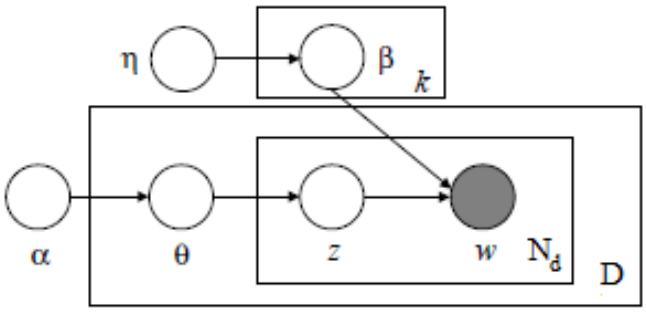
$\beta_{z_{di}}$ jest wierszem macierzy β odpowiadającym wylosowanemu tematowi z_{di}.

Rozkład łączny korpusu o zaobserwowanych słowach, gdzie $\mathbf{w}$ oznacza wektor wszystkich słów w korpusie oraz $\mathbf{w}_d, \mathbf{z_d}$ to odpowiednio słowa i przypisania w dokumencie d, ma następującą postać:

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) &= p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{z}|\theta)p(\theta|\alpha) \\ &= \prod_{d=1}^{M} p(\mathbf{w}_d|\mathbf{z}_d, \beta)p(\mathbf{z}_d|\theta_d)p(\theta_d|\alpha) \\ &= \prod_{d=1}^{M}\prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \beta)p(z_{dn}|\theta_d) \cdot p(\theta_d|\alpha), \\ &= \prod_{d=1}^{M}\prod_{n=1}^{N_d} p(w_{dn}|\beta_{z_{dn}})p(z_{dn}|\theta_d) \cdot p(\theta_d|\alpha), \end{aligned} \tag{2.1}$$

gdzie $w_{dn} \sim \text{Categ}(\beta_{z_{dn}})$, tzn. $p(w_{dn} = i|z_{dn}, \beta) = \beta_{z_{dn}i}$, $z_{dn} \sim \text{Discr}(\theta_d)$ oraz $\theta_d \sim \text{Dir}(\alpha)$ dla wszystkich $i = 1, ..., V$, $n = 1, ..., N_d$, $d = 1, ..., M$, przy czym N_d oznacza liczność słów w dokumencie d. W zapisie skorzystaliśmy z utożsamienia wektora indykatorowego z indeksem jego niezerowej współrzędnej. Rozkład ten faktoryzuje się w ten sposób, ponieważ odpowienie zmienne są warunkowo niezależne między sobą:

- słowa w_{dn} dokumencie d są warunkowo niezależne pod warunkiem ustalonych przypisań $\mathbf{z}_d$ oraz tematów β, $n = 1, ..., N_d$,
- przypisania $\mathbf{z}_d$ są warunkowo niezależne między dokumentami przy ustalonych θ_d, $d = 1, ..., M$,



Rysunek 2.2

- rozkłady tematów θ_d są warunkowo niezależne między dokumentami pod warunkiem α, $d = 1, ..., M$.

W procesie tym pojawiają się zmienne losowe θ_d, stanowiąca rozkład tematów w dokumencie d oraz z_{di}, oznaczająca temat, z jakiego pochodzi słowo w_{di}. Nazywane są one parametrami ukrytymi. To właśnie do nich odnosi się nazwa modelu, mówiąca o "ukrytej" alokacji Dirichleta, ponieważ nie obserwujemy ich realizacji bezpośrednio, a jedynie możemy wnioskować o nich na podstawie obserwowanych słów, które od nich zależą. Parametry θ_d są głównym obiektem zainteresowania i ich estymacja jest celem stosowania LDA.

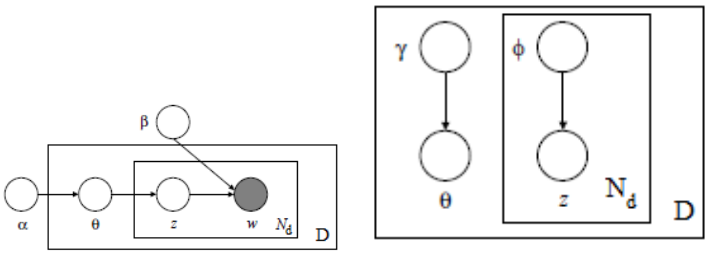
2.2. Wygładzane LDA

W sytuacji, gdy nasz zbiór danych tekstowych składa się z bardzo wielu dokumentów, zazwyczaj skutkuje to również dużym zbiorem słownictwa. To z kolei rodzi istotny problem, polegający na tym, że w danych, do których chcemy zastosować wcześniej zbudowany model, pojawią się słowa, które nie pojawiły się w danych treningowych. Oczywiście według modelu, prawdopodobieństwo pojawienia się tych słów jest zerowe (ponieważ w danych treningowych ich nie było), a zatem prawdopodobieństwo takiego dokumentu również jest zerowe. Jest to istony problem, ponieważ wówczas nie można dla tych danych prowadzić wnioskowania opartego na metodzie największej wiarogodności. Wygładzanie rozwiązuje ten problem, sprawiając, że każde słowo ma dodatnie prawdopodobieństwo pojawienia się w każdym temacie.

Wygładzanie w LDA dotyczy prawdopodobieństw pojawiania się słów. W klasycznej wersji zakładamy, że rozkłady słów w tematach są parametrem swobodnym, natomiast w wersji z wygładzaniem zakłada się, że wszystkie te rozkłady pochodzą niezależnie ze wspólnego rozkładu Dirichleta. Dokładnie mówiąc, każdy wiersz macierzy β pochodzi z rozkładu Dir(η). To zapewnia dodatniość wszystkich współrzędnych rozkładu słów w temacie pozwala przeprowadzić poprawne obliczenia numeryczny. Wygładzamy model LDA przedstawia rysunek 2.2.

Rozkład łączny w wygładzanej wersji LDA ma postać:

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}, \theta, \beta | \alpha, \eta) &= p(\mathbf{w}|\mathbf{z}, \beta) \cdot p(\beta|\eta) p(\mathbf{z}|\theta) \cdot p(\theta|\alpha) \\ &= \prod_{i=1}^{K} p(\beta_i|\eta) \cdot \prod_{d=1}^{M} \prod_{n=1}^{N_d} p(w_{dn}|\beta_{z_{dn}}) p(z_{dn}|\theta_d) \cdot p(\theta_d|\alpha). \end{aligned} \tag{2.2}$$



(a) Model LDA (b) Model przybliżony

Rysunek 2.3: Klaszcyne LDA. tu jest N_d a w obliczeniach N)

2.3. Wnioskowanie i estymacja

Głownym celem stosowania modelu LDA jest estymacja parametrów $\theta_d, d = 1, ..., M$ (estymacja parametrów ukrytych nazywana jest często "wnioskowaniem"). Idealnym rozwiązaniem (według statystki bayesowskiej) byłoby wyznaczenie ich rozkładu a posteriori względem obserwacji $\mathbf{w}$ i przyjęcie jako estymator ich wartości oczekiwanej w tym rozkładzie. Niestety nie da się analitycznie wyznaczyć tego rozkładu i dlatego rozwija się metody omijające ten problem. W rozdziale zaprezentowane zostaną dwie obecnie najpopularniejsze metody estymacji parametrów.

2.3.1. Wnioskowanie wariacyjne

Wnioskowanie wariacyjne opiera się na aproksymacji rozkładu zmiennych ukrytych. Upraszczając model LDA, poprzez pominięcie niektórych zależności, otrzymuje się model, w którym estymacja staje się wykonalna. Wówczas, wsród rodziny rozkładów tego modelu, poszukuje się rozkładu, który jest najlepszym (w pewnym sensie) przybliżeniem rozkładu prawdziwego. Ostatecznie, jako estymator rozkładów tematów w dokumentach, przyjmuje się ich wartość oczekiwaną rozkładzie przybliżonym.

Klasyczne LDA

W modelu LDA mamy $\theta \sim \text{Dir}(\alpha)$ oraz $z \sim \text{Multinomial}(\theta)$, co oznacza, że z jest bezpośrednio zalezne od θ. Upraszczamy model poprzez usunięcie połączenia między $\mathbf{z}$ a θ oraz wprowadzenie swobodnych parametrów wariacyjnych γ i ϕ - k-wymiarowego wektora oraz macierzy wymiaru $N \times k$. Otrzymany model jest przedstawiony na rysunku 2.3b, przy czym w graficznej reprezentacji pomijamy również obserwacje $\mathbf{w}$. W modelu uproszczonym zachowuje się rodziny rozkładów warunkowych zmiennych ukrytych. Zauważmy, że w modelu uproszczonym dokumenty są od siebie niezależne (pod warunkiem α, β), zatem wnioskowanie można przeprowadzać dla każdego dokumentu osobno. Rozkład łączny zmiennych ukrytych w dokumencie (zawierającym N słów) ma teraz postać

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n), \tag{2.3}$$

gdzie $\theta \sim \mathrm{Dir}(\gamma)$ oraz $z_n \sim \mathrm{Discr}(\phi_n)$. Mamy zatem rodzinę rozkładów q, parametryzowaną przez γ oraz ϕ, wśród której będziemy szukać najlepszego przybliżenia rozkładu prawdziwego $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$, a za kryterium podobieństwa rozkładów przyjmujemy dywergencję Kullbacka-Leiblera. Formalnie, naszym celem jest zatem znalezienie parametrów (γ^*, ϕ^*), będących rozwiązaniem problemu optymalizacyjnego:

$$(\gamma^*, \phi^*) = \arg\min_{\gamma,\phi} D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi) \;||\; p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)).$$

Zaznaczmy, że optymalne parametry są naturalnie zależne od zaobserwowanych słów $\mathbf{w}$, czego nie uwzględniamy w zapisie.

Minimalizacja dywergencji Kullbacka-Leiblera jest numerycznie trudna. Dlatego sprowadzimy to zadanie do problemu równoważnego, który jest łatwiejszy obliczeniowo. Zaczynamy od wykorzystania nierówności Jensena do ograniczenia z dołu logarytmu funkcji wiarogodności dla dokumentu w prawdziwym modelu:

$$\begin{aligned}
\log p(\mathbf{w}|\alpha, \beta) &= \log \int \sum_z p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) d\theta \\
&= \log \int \sum_z \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z})} q(\theta, \mathbf{z}) d\theta \\
&= \log E_q\Big[\frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z})}\Big] \\
&\geq E_q[\log \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z})}] \\
&= E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})].
\end{aligned} \tag{2.4}$$

Czyli

$$\log p(\mathbf{w}|\alpha, \beta) \geq E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})] \tag{2.5}$$

Oznaczmy przez $L(\gamma, \phi, \alpha, \beta)$ prawą stronę nierównosci 2.5. Zauważmy, że:

$$\begin{gathered}
\log p(\mathbf{w}|\alpha, \beta) - L(\gamma, \phi, \alpha, \beta) = \\
= \log p(\mathbf{w}|\alpha, \beta) - \big(E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})]\big) \\
= E_q[\log q(\theta, \mathbf{z})] - E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] + E_q[\log p(\mathbf{w}|\alpha, \beta)] \\
= E_q \log(\frac{q(\theta, \mathbf{z})}{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)} p(\mathbf{w}|\alpha, \beta)) = E_q \log(\frac{q(\theta, \mathbf{z})}{p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)}) \\
= D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi) \;||\; p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)),
\end{gathered}$$

gdzie D_{KL} to dywergencja Kullbacka-Leiblera. Mamy zatem

$$\log p(\mathbf{w}|\alpha, \beta) - L(\gamma, \phi, \alpha, \beta) = D(q(\theta, \mathbf{z}|\gamma, \phi) \;||\; p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

Oznacza to, że minimalizacja dywergencj Kullbacka-Leiblera jest równoważna maksymalizacji ograniczenia dolnego funkcji wiarogodności $L(\gamma, \phi, \alpha, \beta)$. Dzięki temu, zamiast minimalizować dywergencję K-L, wystarczy maksymalizować funkcję L, co numerycznie jest znacznie łatwiejsze.

Podkreślmy, że z niezależności warunkowej dokumentów, funkcja wiarogodności korpusu jest iloczym wiarogodności dokumentów. Oznacza to, że możemy rozpatrywać dokumenty osobno, ponieważ maksymalizacja wyprowadzonego ograniczenia wiarogodności poszczególnych dokumentów prowadzi do maksymalizacji ograniczenia dla korpusu.

Estymacja parametrów modelu (proces przebiega analogicznie dla wersji wersji z wygładzaniem) polega na iteracyjnym naprzemiennym powtarzaniu dwóch kroków, ustaloną ilość razy lub do uzyskania zadowalającej względem pewnej miary zbieżności:

1. Estymacja parametrów ukrytych (γ_d^*, ϕ_d^*), dla wszystkich dokumentów $d = 1, ...M$,
2. Estymacja parametrów swobodnych α oraz β.

Oczywiście kroki te są od siebie zależne, dlatego trzeba przyjąć pewne warunki początkowe. Naturalnym podejściem jest inicjalizacja parametrów swobodnych wartościami nieinformatywnymi, tzn:

$$\alpha_i = c, \quad i = 1, ..., K,$$
$$\beta_{ij} = 1/V \quad i = 1, ..., K, \ j = 1, ..., V,$$

gdzie $c > 0$ jest zadaną stałą (dobiera się ją na podstawie spodziewanej koncentracji rozkładu). Takie wartości początkowe nie wyróżniają na starcie żadnego tematu ani słów. Oba kroki wykonujemy mając dane wartości z kroku poprzedzającego, a zaczynamy od estymacji parametrów ukrytych.

Krok 1. Estymacja parametrów ukrytych

Przechodzimy do optymalizacji. Korzystając z faktoryzacji rozkładów (równania 2.1 i 2.3), rozpiszemy najpierw wartość $L(\gamma, \phi, \alpha, \beta)$:

$$\begin{aligned} L(\gamma, \phi, \alpha, \beta) &= E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})] \\ &= E_q \log p(\theta|\alpha) + E_q \log p(\mathbf{z}|\theta) + E_q \log p(\mathbf{w}|\mathbf{z}, \beta) \\ &\quad - E_q \log q(\theta) - E_q \log q(\mathbf{z}). \end{aligned} \tag{2.6}$$

Obliczmy kolejne wartości oczekiwane. Pierwsza:

$$\begin{aligned} &E_q \log p(\theta|\alpha) \\ &= E_q \log(\exp\Big\{ \sum_{i=1}^{k}(\alpha_i - 1)\log\theta_i + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i)\Big\}) \\ &= \sum_{i=1}^{k}(\alpha_i - 1)E_q[\log\theta_i] + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) \end{aligned}$$

Przypomnijmy, że w rozpatrywanym modelu $\theta \sim \mathrm{Dir}(\gamma)$, więc:

$$E_q[\log\theta_i] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j).$$

Zatem ostatecznie:

$$E_q \log p(\theta|\alpha) = \sum_{i=1}^{k}(\alpha_i - 1)\Big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\Big) + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i).$$

Analogicznie otrzymujemy czwartą wartość oczekiwaną.

Druga wartość oczekiwana (z_n^i oznacza i-tą współrzędną wektora z_n):

$$E_q \log p(\mathbf{z}|\theta) = E_q \log \Big(\prod_{n=1}^{N} p(z_n|\theta) \Big) = \sum_{n=1}^{N} E_q \log p(z_n|\theta)$$

$$\sum_{n=1}^{N} E_q[\log \prod_{i=1}^{k} \log \theta_i^{z_n^i}] = \sum_{n=1}^{N} \sum_{i=1}^{k} E_q[z_n^i \log \theta_i]$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{k} E_q[z_n^i] E_q[\log \theta_i] = \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \big),$$

Przypomnijmy, że w_n w indeksie dolnym interpretujemy jako indeks j, dla którego $w_n^j = 1$. Trzecia wartość oczekiwana:

$$E_q \log p(\mathbf{w}|\mathbf{z}, \beta) = E_q \log \Big(\prod_{n=1}^{N} p(w_n|z_n, \beta) \Big) = \sum_{n=1}^{N} E_q \log p(w_n|z_n, \beta)$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{k} \log p(w_n|z_n = i, \beta) \underbrace{P(z_n = i|\phi_n)}_{= \phi_{ni}} = \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \beta_{iw_n}$$

Piąta:

$$E_q \log q(\mathbf{z}) = \sum_{n=1}^{N} E_q \log q(z_n) = \sum_{n=1}^{N} \sum_{i=1}^{k} \log p(z_n = i|\phi) p(z_n = i|\phi) = \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}.$$

Ostatecznie:

$$L(\gamma, \phi, \alpha, \beta) = \log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1) \big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \big)$$

$$+ \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \big(\Psi(\gamma_i) - \Psi\big(\sum_{j=1}^{k} \gamma_j \big) \big)$$

$$+ \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \beta_{iw_n}$$

$$- \log \Gamma\big(\sum_{j=1}^{k} \gamma_j \big) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1) \big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \big)$$

$$- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}$$

W optymalizacji wykorzystamy metodę mnożników Lagrange'a. Naszym ograniczeniem są warunki $\sum_{i=1}^{k} \phi_{ni} = 1,\ \forall n = 1, ..., N$, gdyż ϕ_n są wektorami prawdopodobieństw. Definiujemy funkcję Lagrange'a (dla uproszczenie zapisu pomijamy argumenty):

$$Lag = L(\gamma, \phi, \alpha, \beta) + \sum_{n=1}^{N} \lambda_n (\sum_{j=1}^{k} \phi_{nj} - 1)$$

W celu skrócenia zapisu, oznaczmy przez $Lag_{[\xi]}$ składniki Lag zawierające ξ. Podczas różniczkowania funkcji Lag po kolejnych zmiennych, składniki nie zawierające danej zmiennej zerowałyby się, więc je pominiemy. Zaczynamy od maksymalizacji po ϕ_{ni}.

$$\begin{aligned} Lag_{[\phi]} = &\sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \big(\Psi(\gamma_i) - \Psi\big(\sum_{j=1}^{k} \gamma_j\big)\big) \\ &+ \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \beta_{iw_n} \\ &- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni} \\ &+ \sum_{n=1}^{N} \lambda_n (\sum_{j=1}^{k} \phi_{nj} - 1) \end{aligned} \tag{2.7}$$

$$Lag_{[\phi_{ni}]} = \phi_{ni} \big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\big) + \phi_{ni} \log \beta_{iw_n} - \phi_{ni} \log \phi_{ni} + \lambda_n (\sum_{j=1}^{k} \phi_{nj} - 1)$$

$$\frac{\partial Lag}{\partial \phi_{ni}} = \frac{\partial Lag_{[\phi_{ni}]}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \log \beta_{iw_n} - \log \phi_{ni} - 1 + \lambda_n$$

Przyrównujemy pochodną do zera.

$$\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \log \beta_{iw_n} - \log \phi_{ni} - 1 + \lambda_n = 0$$

Zatem

$$\begin{aligned} \phi_{ni} &= \exp\{\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \log \beta_{iw_n} - 1 + \lambda_n\} \\ &= \beta_{iw_n} \exp\{\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\} \exp\{-1 + \lambda_n\}. \end{aligned} \tag{2.8}$$

Możemy pominąć ostani czynniki, ponieważ nie zależy on od i. Dokładną wartość ϕ_{ni} otrzymamy poprzez normalizację otrzymanego wektora ϕ_n. Mamy zatem:

$$\phi^*_{ni} \propto \beta_{iw_n} \exp\big\{\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\big\}$$

Na koniec sprawdzamy czy to rzeczywiście jest maksimum:

$$\frac{\partial^2 Lag}{\partial \phi_{ni}^2} = -\frac{1}{\phi_{ni}}.$$

Dla $\phi_{ni} > 0$ druga pochodna przyjmuje wartość ujemną, zatem rzeczywiście jest to maksimum. Natomiast dla $\phi_{ni} = 0$ jest ona nieokreślona, a taka sytuacja ma miejsce wówczas, gdy $\beta_{iw_n} = 0$. Oznacza to, że gdy prawdopodobieństwo pojawienia się słowa w_n w temacie i jest zerowe, optymalną wartością ϕ_{ni}, czyli prawdopodobieństwo przypisania słowa w_n do tematu i również wynosi 0, a zatem jest to słuszny wynik.

Przechodzimy do maksymalizacji ze względu na γ_i.

$$\begin{aligned}
Lag_{[\gamma]} &= \sum_{i=1}^{k} \big[(\alpha_i - 1)\big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\big)\big] \\
&+ \sum_{n=1}^{N} \sum_{i=1}^{k} \big[\phi_{ni}\big(\Psi(\gamma_i) - \Psi\big(\sum_{j=1}^{k} \gamma_j\big)\big)\big] \\
&- \log \Gamma\big(\sum_{j=1}^{k} \gamma_j\big) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)\big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\big) \\
&= \sum_{i=1}^{k} \big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\big)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) \\
&- \log \Gamma\big(\sum_{j=1}^{k} \gamma_j\big) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) \\
&= \sum_{i=1}^{k} \Psi(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \sum_{i=1}^{k} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) \\
&- \log \Gamma\big(\sum_{j=1}^{k} \gamma_j\big) + \sum_{i=1}^{k} \log \Gamma(\gamma_i)
\end{aligned} \tag{2.9}$$

$$\begin{aligned}
Lag_{[\gamma_i]} &= \Psi(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \sum_{j=1}^{k} (\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j) \\
&- \log \Gamma\big(\sum_{j=1}^{k} \gamma_j\big) + \log \Gamma(\gamma_i)
\end{aligned} \tag{2.10}$$

Różniczkujemy

$$\frac{\partial Lag}{\partial \gamma_{i_0}} = \frac{\partial Lag_{[\gamma]}}{\partial \gamma_{i_0}} =$$

$$= \Psi'(\gamma_{i_0})(\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni_0} - \gamma_{i_0}) - \Psi(\gamma_{i_0}) - \Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{j=1}^{k} (\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j) + \Psi(\sum_{j=1}^{k} \gamma_j)$$

$$-\Psi(\sum_{j=1}^{k} \gamma_j) + \Psi(\gamma_{i_0})$$

$$= \Psi'(\gamma_{i_0})(\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni} - \gamma_{i_0}) - \Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{i=1}^{k} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)$$

$$= \big(\Psi'(\gamma_{i_0}) - \Psi'(\sum_{j=1}^{k} \gamma_j)\big)(\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni} - \gamma_{i_0}) - \Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{i \neq i_0} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)$$

Po przyrównaniu pochodnej do zera, nie da się wyznaczyć analitycznie wzoru na γ_{i_0}, a ponadto w równaniu zawarte są wszystkie pozostałe parametry γ_j dla $j \neq i_0$. Zauważmy jednak, że gdy $\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$ dla każdego $i = 1, ..., k$, to wówczas pochodna przyjmuje wartość 0. Oznacza to, że jeśli w momencie obliczania optymalnej wartości parametru γ_{i_0}, wszystkie pozostałe współrzędne wektora γ będą spełniały warunek $\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$, to pochodna przyjmuje wartość zero jedynie dla γ_{i_0} również równego $\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni_0}$. Kolejne współrzędne wyliczamy iteracyjnie, zatem ustalając odpowiednie wartości początkowe, optymalna wartość to

$$\gamma^*_{i_0} = \alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni_0}.$$

Uzasadnimy, że jest to maksimum:

$$\frac{\partial^2 Lag}{\partial \gamma_{i_0}^2} = \big(\Psi''(\gamma_{i_0}) - \Psi''(\sum_{j=1}^{k} \gamma_j)\big)(\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni} - \gamma_{i_0}) - \big(\Psi'(\gamma_{i_0}) - \Psi'(\sum_{j=1}^{k} \gamma_j)\big)$$

$$-\Psi''(\sum_{j=1}^{k} \gamma_j) \sum_{i \neq i_0} \underbrace{(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)}_{=0}$$

$$= \big(\Psi''(\gamma_{i_0}) - \Psi''(\sum_{j=1}^{k} \gamma_j)\big)(\alpha_{i_0} + \sum_{n=1}^{N} \phi_{ni} - \gamma_{i_0}) + \Psi'(\sum_{j=1}^{k} \gamma_j) - \Psi'(\gamma_{i_0})$$

Mamy zatem

$$\frac{\partial^2 Lag}{\partial \gamma_{i_0}^2}\Big|_{\gamma_{i_0} = \alpha_{i_0} + \sum\limits_{n=1}^{N} \phi_{ni_0}} = \Psi'(\sum_{j=1}^{k} \gamma_j) - \Psi'(\gamma_{i_0}) < 0,$$

ponieważ $\sum_{j=1}^{k} \gamma_j > \gamma_{i_0}$, a Ψ' jest malejąca (1.4). Oznacza to, że jest to maksimum.

Zauważmy, że nie ma potrzeby wyliczania parametrów λ_n. Mamy zatem jawne wzory do wyliczania parametrów, jednakże wystepuje tutaj zależność parametrów γ i ϕ. Z tego powodu, do wyznaczenia opytmalnych wartości (γ^*, ϕ^*), należy zastosować procedurę iteracyjną - wyliczamy naprzemiennie parametry z otrzymanych wzorów, do momentu uzyskania zbieżności. Trzeba również ustalić warunki początkowe. Dla rozkładów ϕ_n naturalnym wyborem jest inicjalizacja rozkładem równomiernym, zatem przyjmujemy $\phi_{ni} = 1/K$, dla wszystkich $n = 1, ..., N$ i $i = 1, ..., K$. Natomiast współrzędne parametru γ muszą spełniać warunki $\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$, dla każdego $i = 1, ..., K$, co daje wartości początkowe $\gamma_i = \alpha_i + \frac{N}{K}$.

Podsumowując, cały proces wnioskowania wariacyjnego można przedstawić jako algorytm:

TU BĘDZIE ALGORYTM W PSEUDOKODZIE WYLICZANIA W PETLACH TYCH PARAMETRÓW

Krok 2. Estymacja parametrów swobodnych

Nieznane parametry swobodne to parametr sterujący rozkładami tematów w dokumentach α oraz rozkłady słów w tematach β. Przypomnijmy, że wnioskowanie prowadzone było na poziomie dokumentu - optymalne wartości były estymowane w obrębie każdego dokumentu osobno. Parametry α oraz β są parametrami globalnymi - odnoszą się do całego korpusu i wszystkie dokumenty są od nich zależne. Do tego potrzebna jest globalna funkcja celu dla korpusu, a nie dokumentu, jak podczas wnioskowania o parametrach ukrytych. Zatem rozpatrujemy teraz dolne logarytmu wiarogodności korpusu, w którym pojawiające się rozkłady są teraz rozkładami dla całego korpusu.

$$L(\gamma, \phi, \alpha, \beta) = E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})]. \tag{2.11}$$

Z warunkowej niezależności dokumentów po warunkiem (α, β), powyższe rozkłady faktoryzują się, a stąd otrzymujemy

$$L(\gamma, \phi, \alpha, \beta) = \sum_{d=1}^{M} E_{q_d}[\log p_d(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)] - E_{q_d}[\log q_d(\theta_d, \mathbf{z}_d)]. \tag{2.12}$$

Oznacza to, że maksymalizowane ograniczenie L dla korpusu, jest sumą ograniczeń dla poszczególnych dokumentów. Zatem na mocy poprzednich obliczeń mamy:

$$\begin{aligned} L(\gamma, \phi, \alpha, \beta) = \sum_{d=1}^{M} \Big[& \log\Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log\Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big) \\ & + \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni}\big(\Psi(\gamma_{di}) - \Psi\big(\sum_{j=1}^{k} \gamma_{dj}\big)\big) \\ & + \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni} \log \beta_{iw_{dn}} \\ & - \log\Gamma\big(\sum_{j=1}^{k} \gamma_{dj}\big) + \sum_{i=1}^{k} \log\Gamma(\gamma_{di}) - \sum_{i=1}^{k} (\gamma_{di} - 1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big) \\ & - \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni} \log \phi_{dni} \Big]. \end{aligned}$$

Maksymalizujemy L przy warunku $\sum_{j=1}^{V} \beta_{ij} = 1, \ \forall i = 1, ..., k$. Wówczas:

$$L_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni} \log \beta_{iw_{dn}} + \sum_{i=1}^{k} \lambda_i\big(\sum_{j=1}^{V} \beta_{ij} - 1\big)$$

$$L_{[\beta_{ij}]} = \sum_{d=1}^{M} \sum_{\{n:w_{dn}=j\}} \phi_{dni} \log \beta_{ij} + \lambda_i(\beta_{ij} - 1) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} w_{dn}^{j} \phi_{dni} \log \beta_{ij} + \lambda_i(\beta_{ij} - 1)$$

$$\frac{\partial L}{\partial \beta_{ij}} = \frac{\partial L_{[\beta_{ij}]}}{\partial \beta_{ij}} = \frac{1}{\beta_{ij}} \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j} + \lambda_i = 0$$

Stąd

$$\beta_{ij} = -\lambda_i \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$$

Czyli

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$

Nie ma potrzeby wyliczania wartości λ_i, ponieważ dokładne wartości β_{ij} otrzymujemy normalizując wektory β_i.

Parametr β_{ij} może przyjąć wartość 0 tylko wówczas, gdy wszystkie wyrazy $\phi_{dni} w_{dn}^{j}$ są równe zero, czyli gdy $\phi_{dni} = 0$ dla n takich, że $w_{dn} = j$. Taka sytuacja oznacza, że w rozkładzie przybliżonym prawdopodobieństwa pojawienia się j-tego słowa w i-tym temacie jest zerowe. Zatem β_{ij} słucznie przyjmuje wrtość 0. W przeciwnym przypadku mamy:

$$\frac{\partial^2 L}{\partial \beta_{ij}^2} = -\frac{1}{\beta_{ij}^2} \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j} < 0,$$

a zatem otrzymana wartość jest argumentem maksymalizującym funkcję celu.

Przechodzimy do parametru α.

$$L_{[\alpha]} = \sum_{d=1}^{M} \Big(\log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\big(\Psi(\gamma_{dj}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big)\Big)$$

$$= M\Big(\log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \Gamma(\alpha_i)\Big) + \sum_{d=1}^{M} \sum_{i=1}^{k} \Big((\alpha_i - 1)\big(\Psi(\gamma_{dj}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big)\Big)$$

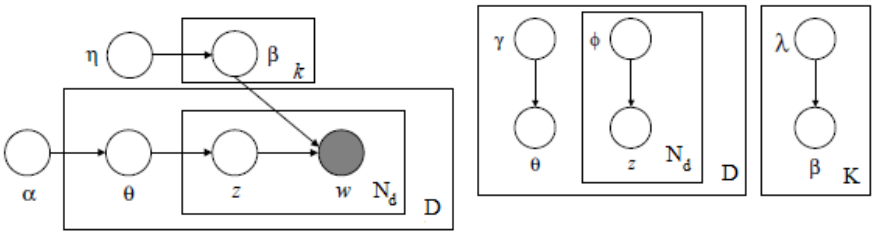
$$\frac{\partial L_{[\alpha]}}{\partial \alpha_i} = M\Big(\Psi\big(\sum_{j=1}^{k} \alpha_j\big) - \Psi(\alpha_i)\Big) + \sum_{d=1}^{M} \Big(\Psi(\gamma_{di} - \Psi\big(\sum_{j=1}^{k} \gamma_{dj}\big)\Big)$$

Dla układu $\frac{\partial L_{[\alpha]}}{\partial \alpha_i} = 0,\ i = 1, ..., K$ rozwiązanie analityczne nie istnieje. Dlatego optymalne wartości współrzędnych wektora α wyliczamy numerycznie, np. metodą Raphsona-Newtona. Punkty, w których pochodne zerują się są to maksima, ponieważ

$$\frac{\partial^2 L_{[\alpha]}}{\partial \alpha_i^2} = M\Big(\underbrace{\Psi'\big(\sum_{j=1}^{k} \alpha_j\big) - \Psi'(\alpha_i)}_{<\,0}\Big) + \sum_{d=1}^{M} \Big(\underbrace{\Psi'(\gamma_{di} - \Psi'\big(\sum_{j=1}^{k} \gamma_{dj}\big)}_{<\,0}\Big) < 0,$$

co wynika z tego, że Ψ' jest funkcją rosnącą, przy czym nierówność jest spełniona niezależnie od wartości parametru α.

Po powtórzeniu opisanych dwóch kroków odpowiedną ilość razy, pozostaje jedynie wyestymować udział tematów w dokumencie, czyli parametr ukryty θ. Jak zostało wspomniane



(a) Model LDA (b) Model przybliżony

Rysunek 2.4: Wygładzane LDA. (Widać że w paincie robione?)

na początku rozdziału, za jego estymator przyjmujemy wartość oczekiwaną w wyznaczonym rozkładzie:

$$\hat{\theta}_i = \frac{\gamma_i^*}{\sum_{j=1}^{K} \gamma_j^*}, \quad \forall\ i = 1, ..., K.$$

Wygładzane LDA.

Istotna część rachunków w modelu wygładzanym jest identyczna jak w wersji klasycznej, dlatego nie będziemy ich powtarzać, lecz będziemy się do nich odnosić. W przypadku wygładzanego LDA, uproszczenie modelu jest podobne. Usuwamy połączenia między $\mathbf{z}$ a θ, wprowadzamy swobodne parametry wariacyjne γ i θ oraz pomijamy obserwacje $\mathbf{w}$. Pojawia się jeden dodatkowy parametr wariacyjny λ - macierz wymiaru $K \times V$. Graficzna reprezentacja modelu jest przedstawiona na rysunku 2.4b. O ile w klasycznym przypadku wnioskowanie jest prowadzone dla każdego dokumentu niezależnie, tak w przypadku wersji z wygładzaniem wnioskowanie odbywa się globalnie, co jest oczywiście jest spowodowane faktem nieznanych rozkładów słów w tematach, od których zależą wszystki dokumenty.

Rozkład łączny parametrów ukrytych w całym korpusie w modelu uproszczonym ma postać:

$$q(\beta, \theta, \mathbf{z}|\lambda, \gamma, \phi) = \prod_{i=1}^{k} f(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\phi_d, \gamma_d),$$

gdzie q_d jest rozkładem zdefiniowanym w równości 2.3 i $\beta_i \sim \mathrm{Dir}(\lambda_i)$.

Krok 1. Estymacja parametrów ukrytych

Nierówność 2.4 przybiera tym razem postać postać:

$$\log p(\mathbf{w}|\alpha, \eta) \geq E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \eta)] - E_q[\log q(\beta, \theta, \mathbf{z})], \tag{2.13}$$

przy czym pojawiające się tutaj rozkłady są rozkładami dla całego korpusu. Korzystając z faktoryzacji rozkładów względem dokumentów, rozpisujemy prawą stronę:

$$L(\gamma, \phi, \lambda, \alpha, \eta) = \sum_{d=1}^{M} E_q \log p(\theta_d|\alpha) + \sum_{d=1}^{M} E_q \log p(\mathbf{z}_d|\theta_d) + \sum_{d=1}^{M} E_q \log p(\mathbf{w}_d|\mathbf{z}_d, \beta) + E_q \log p(\beta|\eta)$$

$$- E_q \log f(\beta) - \sum_{d=1}^{M} E_q \log q_d(\theta) - \sum_{d=1}^{M} E_q \log q_d(\mathbf{z}).$$

Pierwsze dwa składniki otrzymujemy wprost w wersji klasycznej:

$$\sum_{d=1}^{M} E_q \log p(\theta_d|\alpha) = \sum_{d=1}^{M} \big[\sum_{i=1}^{k} (\alpha_i - 1)\Big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\Big) + \log\Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} \log\Gamma(\alpha_i)\big],$$

$$\sum_{d=1}^{M} E_q \log p(\mathbf{z}_d|\theta_d) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni}\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big).$$

Trzeci jest prostą modyfikacją składnika odpowiadającego w w wersji klasycznej. Tym razem β jest zmienną losową, zatem zamiast $\log \beta_{iw_n}$ mamy $E_q \log \beta_{iw_n}$, a ponieważ w modelu uproszczonym $\beta_i \sim \mathrm{Dir}(\lambda_i)$, więc:

$$\sum_{d=1}^{M} E_q \log p(\mathbf{w}_d|\mathbf{z}_d, \beta) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni} E_q \log \beta_{iw_{dn}} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni}\Big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})\Big).$$

Wartości oczekiwane w czwartym i piątym składniku, są tym samym co wartość oczekwiana w pierwszym, ale z innymi parametrami. Zatem:

$$E_q \log p(\beta|\eta) = \sum_{i=1}^{k} E_q \log p(\beta_i|\eta) = \sum_{i=1}^{k} \sum_{j=1}^{V} (\eta_j - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})\Big) + \log\Gamma(\sum_{j=1}^{V} \eta_j) - \sum_{j=1}^{V} \log\Gamma(\eta_j),$$

$$E_q \log f(\beta) = \sum_{i=1}^{k} E_q \log f(\beta_i|\lambda_i) = \sum_{i=1}^{k} \sum_{j=1}^{V} (\lambda_{ij} - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})\Big) + \log\Gamma(\sum_{j=1}^{V} \lambda_{ij}) - \sum_{j=1}^{V} \log\Gamma(\lambda_{ij}).$$

Szósty i siódmy składnik otrzymujemy ponownie wprost z wersji klasycznej:

$$\sum_{d=1}^{M} E_q \log q_d(\theta) = \sum_{d=1}^{M} \big[\log\Gamma\big(\sum_{j=1}^{k} \gamma_{dj}\big) - \sum_{i=1}^{k} \log\Gamma(\gamma_{di}) + \sum_{i=1}^{k} (\gamma_{di} - 1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big)\big],$$

$$\sum_{d=1}^{M} E_q \log q_d(\mathbf{z}) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{dni} \log \phi_{dni}.$$

Ostatecznie:

$$
\begin{aligned}
L(\gamma,\phi,\lambda,\alpha,\eta) = &\sum_{d=1}^{M}\Big[\sum_{i=1}^{k}(\alpha_i-1)\Big(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\Big)+\log\Gamma(\sum_{i=1}^{k}\alpha_i)-\sum_{i=1}^{k}\log\Gamma(\alpha_i)\Big]\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\big(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}})-\Psi(\sum_{j=1}^{V}\lambda_{ij})\Big)\\
&+\sum_{i=1}^{k}\sum_{j=1}^{V}(\eta_j-1)\Big(\Psi(\lambda_{ij})-\Psi(\sum_{j=1}^{V}\lambda_{ij})\Big)+\log\Gamma(\sum_{j=1}^{V}\eta_j)-\sum_{j=1}^{V}\log\Gamma(\eta_j)\\
&-\sum_{i=1}^{k}\sum_{j=1}^{V}(\lambda_{ij}-1)\Big(\Psi(\lambda_{ij})-\Psi(\sum_{j=1}^{V}\lambda_{ij})\Big)-\log\Gamma(\sum_{j=1}^{V}\lambda_{ij})+\sum_{j=1}^{V}\log\Gamma(\lambda_{ij})\\
&-\sum_{d=1}^{M}\big[\log\Gamma\big(\sum_{j=1}^{k}\gamma_{dj}\big)-\sum_{i=1}^{k}\log\Gamma(\gamma_{di})+\sum_{i=1}^{k}(\gamma_{di}-1)\big(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\big]\\
&-\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\log\phi_{dni}
\end{aligned}
\tag{2.14}
$$

Definiujemy funkcję Lagrange'a:

$$
Lag = L(\gamma,\phi,\lambda,\alpha,\beta)+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\xi_{dn}(\sum_{i=1}^{k}\phi_{dni}-1)
$$

Maksymalizujemy po kolejnych parametrach:

$$
\begin{aligned}
Lag_{[\phi]} = &\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\big(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}})-\Psi(\sum_{j=1}^{V}\lambda_{ij})\Big)\\
&-\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\log\phi_{dni}\\
&+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\xi_{dn}(\sum_{i=1}^{k}\phi_{dni}-1)
\end{aligned}
\tag{2.15}
$$

$$
Lag_{[\phi_{dni}]} = \phi_{dni}\big(\Psi(\gamma_{di})-\Psi(\sum_{j=1}^{k}\gamma_{dj})\big)+\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}})-\Psi(\sum_{j=1}^{V}\lambda_{ij})\Big)-\phi_{dni}\log\phi_{dni}+\xi_{dn}(\sum_{j=1}^{k}\phi_{dnj}-1)
$$

Jest to zatem postać analogiczna do wersji klasycznej. Różnicą jest pojawienie się dodatkowych indeksów d oraz wartości $\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})$ w miejsce $\log \beta_{iw_{dn}}$. Zatem na mocy poprzednich obliczeń mamy:

$$\phi^*_{dni} \propto \big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})\big) \exp\big\{\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big\}.$$

Przechodzimy do parametru γ.

$$\begin{aligned} Lag_{[\gamma]} = &\sum_{d=1}^{M}\sum_{i=1}^{k}\big[(\alpha_i - 1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\big] \\ &+ \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\big[\phi_{dni}\big(\Psi(\gamma_{di}) - \Psi\big(\sum_{j=1}^{k}\gamma_{dj}\big)\big)\big] \\ &+ \sum_{d=1}^{M}[-\log\Gamma\big(\sum_{j=1}^{k}\gamma_{dj}\big) + \sum_{i=1}^{k}\log\Gamma(\gamma_{di}) - \sum_{i=1}^{k}(\gamma_{di}-1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big)] \end{aligned} \tag{2.16}$$

$$\begin{aligned} Lag_{[\gamma_{di}]} = &(\alpha_i - 1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big) \\ &+ \sum_{n=1}^{N_d}\phi_{dni}\big(\Psi(\gamma_{di}) - \Psi\big(\sum_{j=1}^{k}\gamma_{dj}\big)\big) \\ &- \log\Gamma\big(\sum_{j=1}^{k}\gamma_{dj}\big) + \log\Gamma(\gamma_{di}) - (\gamma_{di}-1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big) \\ &= \big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big)(\alpha_i + \sum_{n=1}^{N_d}\phi_{dni} - \gamma_{di}) - \log\Gamma\big(\sum_{j=1}^{k}\gamma_{dj}\big) + \log\Gamma(\gamma_{di}) \end{aligned} \tag{2.17}$$

Ponownie wartość ta ma analogiczą postać do wersji klasycznej, a jedyną różnicą jest pojawienie się dodatkowego indeksu d. Zatem maksimum jest przyjmowane dla

$$\gamma^*_{di} = \alpha_i + \sum_{n=1}^{N_d}\phi_{dni}.$$

Przechodzimy do parametru λ, który nie wystepował w wersji klasycznej.

$$\begin{aligned} L_{[\lambda]} = &\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\ &+ \sum_{i=1}^{k}\sum_{j=1}^{V}(\eta_j - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) + \log\Gamma(\sum_{i=1}^{V}\eta_i) - \sum_{i=1}^{V}\log\Gamma(\eta_i) \\ &- \sum_{i=1}^{k}\sum_{j=1}^{V}(\lambda_{ij} - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) - \log\Gamma(\sum_{j=1}^{V}\lambda_{ij}) + \sum_{j=1}^{V}\log\Gamma(\lambda_{ij}) \end{aligned} \tag{2.18}$$

$$
\begin{aligned}
L_{[\lambda_{ij}]} &= \sum_{d=1}^{M} \sum_{\{n:w_{dn}=1\}} \phi_{dni}\Big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&+ (\eta_j - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&- (\lambda_{ij} - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) - \log\Gamma(\sum_{j=1}^{V}\lambda_{ij}) + \log\Gamma(\lambda_{ij}) \\
&= \sum_{d=1}^{M}\sum_{n=1}^{N_d} w_{dn}^{j}\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&+ (\eta_j - \lambda_{ij})\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&- \log\Gamma(\sum_{j=1}^{V}\lambda_{ij}) + \log\Gamma(\lambda_{ij})
\end{aligned}
\tag{2.19}
$$

$$
\begin{aligned}
\frac{\partial L_{[\lambda_{ij}]}}{\partial \lambda_{ij}} &= \sum_{d=1}^{M}\sum_{n=1}^{N_d} w_{dn}^{j}\phi_{dni}\Big(\Psi'(\lambda_{iw_{dn}}) - \Psi'(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&- \Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) + (\eta_j - \lambda_{ij})\Big(\Psi'(\lambda_{iw_{dn}}) - \Psi'(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&- \Psi(\sum_{j=1}^{V}\lambda_{ij}) + \Psi(\lambda_{ij}) \\
&= \Big(\Psi'(\lambda_{iw_{dn}}) - \Psi'(\sum_{j=1}^{V}\lambda_{ij})\Big)(\eta_j - \lambda_{ij} + \sum_{d=1}^{M}\sum_{n=1}^{N_d} w_{dn}^{j}\phi_{dni})
\end{aligned}
\tag{2.20}
$$

Z równania $\frac{\partial L_{[\lambda_{ij}]}}{\partial \lambda_{ij}} = 0$ otrzymujemy natychmiast

$$\lambda_{ij}^{*} = \eta_j + \sum_{d=1}^{M}\sum_{n=1}^{N_d} w_{dn}^{j}\phi_{dni}$$

Uzasadniamy, że jest to maksimum:

$$
\begin{aligned}
&\frac{\partial^2 L_{[\lambda_{ij}]}}{\partial \lambda_{ij}^2}\Big|_{\lambda_{ij}=\eta_j+\sum_{d=1}^{M}\sum_{n-j}^{N_d} w_{dn}^{j}\phi_{dni}} \\
&= \Big[\Big(\Psi''(\lambda_{iw_{dn}}) - \Psi''(\sum_{j=1}^{V}\lambda_{ij})\Big)(\eta_j - \lambda_{ij} + \sum_{d=1}^{M}\sum_{n-j}^{N_d} w_{dn}^{j}\phi_{dni}) \\
&\quad - \Big(\Psi'(\lambda_{iw_{dn}}) - \Psi'(\sum_{j=1}^{V}\lambda_{ij})\Big)\Big]_{\lambda_{ij}=\eta_j+\sum_{d=1}^{M}\sum_{n-j}^{N_d} w_{dn}^{j}\phi_{dni}}
\end{aligned}
$$

$$= \Psi'(\sum_{j=1}^{V} \lambda_{ij}) - \Psi'(\lambda_{iw_{dn}}) < 0.$$

Zatem wartość optymalne wyliczamy ze wzorów:

$$\phi^*_{dni} \propto \big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V} \lambda_{ij})\big) \exp\big\{\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big\}$$

$$\gamma^*_{di} = \alpha_i + \sum_{n=1}^{N_d} \phi_{dni}.$$

$$\lambda^*_{ij} = \eta_j + \sum_{d=1}^{M}\sum_{n=1}^{N_d} w^j_{dn}\phi_{dni}$$

Narzucamy na ϕ oraz γ identyczne warunki początkowe jak w wersji klasycznej. Nie inicjujemy wartośc λ, ponieważ wyliczamy ją z danych wartości pozostałych parametrów. Iteracyjnie wyliczamy kolejne parametry do uzyskania zbieżności, według następującego algorytmu:

TU BĘDZIE ALGORYTM W PSEUDOKODZIE WYLICZANIA W PETLACH TYCH PARAMETRÓW

Krok 2. Estymacja parametrów swobodnych

W modelu z wygładzaniem parametrami swobodnymi są α oraz η. Przypomnijmy postać funkcji celu L:

$$\begin{aligned}
L(\gamma,\phi,\lambda,\alpha,\eta) = &\sum_{d=1}^{M}\Big[\sum_{i=1}^{k}(\alpha_i-1)\Big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\Big) + \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i)\Big] \\
&+ \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big) \\
&+ \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\Big(\Psi(\lambda_{iw_{dn}}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) \\
&+ \sum_{i=1}^{k}\sum_{j=1}^{V}(\eta_j-1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) + \log\Gamma(\sum_{j=1}^{V}\eta_j) - \sum_{j=1}^{V}\log\Gamma(\eta_j) \\
&- \sum_{i=1}^{k}\sum_{j=1}^{V}(\lambda_{ij}-1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) - \log\Gamma(\sum_{j=1}^{V}\lambda_{ij}) + \sum_{j=1}^{V}\log\Gamma(\lambda_{ij}) \\
&- \sum_{d=1}^{M}\Big[\log\Gamma\big(\sum_{j=1}^{k}\gamma_{dj}\big) - \sum_{i=1}^{k}\log\Gamma(\gamma_{di}) + \sum_{i=1}^{k}(\gamma_{di}-1)\big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\Big] \\
&- \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\phi_{dni}\log\phi_{dni}
\end{aligned} \tag{2.21}$$

Zaczynamy maksymalizację od parametru η.

$$L_{[\eta]} = \sum_{i=1}^{k}\sum_{j=1}^{V}(\eta_j - 1)\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) + \log\Gamma(\sum_{j=1}^{V}\eta_j) - \sum_{j=1}^{V}\log\Gamma(\eta_j)$$

$$\frac{\partial L_{[\eta]}}{\partial \eta_j} = \sum_{i=1}^{k}\Big(\Psi(\lambda_{ij}) - \Psi(\sum_{j=1}^{V}\lambda_{ij})\Big) + \Psi(\sum_{j=1}^{V}\eta_j) - \Psi(\eta_j)$$

W tym przypadku również nie istnieje analityczna postać rozwiązania problemu $\frac{\partial L_{[\eta]}}{\partial \eta_j} = 0,\ j = 1, ..., V$. Dlatego optymalną wartość parametru η wyznaczamy numerycznie np. metodą Newtona-Raphsona. Punkty zerowania się pochodnej są maksimami, ponieważ

$$\frac{\partial^2 L_{[\eta]}}{\partial \eta_j^2} = \Psi'(\sum_{j=1}^{V}\eta_j) - \Psi'(\eta_j) < 0.$$

Parametr α:

$$L_{[\alpha]} = \sum_{d=1}^{M}\Big(\log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{i=1}^{k}\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)\big(\Psi(\gamma_{dj}) - \Psi(\sum_{j=1}^{k}\gamma_{dj})\big)\Big).$$

Jest to ta sama wartość co poprzednio, zatem optymalną wartość otrzymujemy numerycznie.

W wersji z wygładzaniem nie są znane rozkłady słów w tematach, które możemy jednak łatwo wyestymować (na podstawie przyplizonego rozkładu). W uproszczonym modelu $\beta_i \sim \mathrm{Dir}(\lambda)$, $i = 1, ..., K$, zatem jako estymatory przyjmujemy ich wartości oczekiwane w tym rozkładzie. Ostatecznie:

$$\hat{\theta}_i = \frac{\gamma_i^*}{\sum_{j=1}^{K}\gamma_j^*}, \quad \forall\ i = 1, ..., K.$$

$$\hat{\beta}_{ij} = \frac{\lambda_j^*}{\sum_{j=1}^{V}\lambda_{ij}^*}, \quad \forall\ i = 1, ..., K,\ j = 1, ..., V.$$

2.3.2. Wnioskowanie przy użyciu próbkowania Gibbsa

Drugą spośród omawianym w niniejszej pracy metod wnioskowania jest metoda oparta na próbkowaniu Gibbsa. W przypadku zastosowania tej metody, cały proces wnioskowania i estymacji wygląda inaczej niż przy użyciu wnioskowania wariacyjnego. W poprzedniej metodzie wnioskowanie i estymacja opierały się na optymalizacji wspólnej funkcji celu. Przy próbkowaniu Gibbsa nie mamy żadnej funkcji celu, a parametry ukryte i swobodne otrzymuje się w tym sensie niezależnie. Ponadto inaczej niż wcześniej, nie wylicza się ich naprzemiane po jednym kroku, lecz estymacje powtarza się co zadaną liczbę powtórzeń wnioskowania, przy czym liczba ta jest dobierana indywidualnie. Powodem tego jest skrócenie czasu obliczeń, ponieważ wnioskowanie jest procesem losowych i zazwyczaj nie opłaca się przeprowadzać estymacji po każdym wnioskowaniu, gdyż w praktyce okazuje się, że zmiana wartości parametrów jest bardzo mała.

W niniejszej pracy nie będzie opisywana metoda estymacji parametrów swobodnych: α w wersji klasycznej oraz α i η w wersji z wygładzaniem, gdy wniokowanie przebiega metodą wykorzystującą próbkowanie Gibbsa. Mając dane wartości parametrów ukrytch, ich estymacja jest zagadnieniem estymacji parametru rozkładu Dirichleta. Jest to problem nietrywialny i istnieje wiele metod jego rozwiązania, opisanych m.in w [tu będą odnośniki do literatury].

Metoda oparta na próbkowaniu Gibbsa reprezentuje zupełnie inne podejście do problemu niż wnioskowanie wariacyjne, które opierało się na aproksymacji rozkładu parametrów ukrytych przy użyciu uproszczonego modelu, by ostatecznie jako estymator rozkładów tematów w dokumentach przyjąć wartość oczekiwaną tego parametru. Próbkowanie Gibbsa opiera się na iteracyjnym losowaniu wartości przypisań $z_{dn},\ n = 1, ..., N_d,\ d = 1, ..., M$, według wyestymowanego empirycznego rozkładu, do ustalonego momentu stopu, by ostatecznie jako estymator rozkładów tematów w dokumentach przyjąć ich wartość oczekiwaną w otrzymanym rozkładzie a posteriori przy ustalonych już przypisaniach. W losowaniu kolejnych przypisań wykorzystywana będzie informacja o pozostałych przypisaniach. Jest to istotny czynnik, ponieważ w naturalny sposób duża liczba dotychczasowych przypisań słów do tematu k w dokumencie d powinna zwiekszać pradopodobieństo, że kolejne słowa tego dokumentu również pochodzą z tematu k i odwrotnie. Oczekuje się, że udziały przypisań do poszczególnych tematów w dokumentach ustabilizują się, tzn. od pewnego momentu zaczną oscylować wokół pewnych wartości. Podkreślmy, że nigdy nie nastąpi stałe ustalenie się ich wartości, ponieważ każde przypisanie jest losowe. Działamy na poziomie całego korpusu, zatem $\mathbf{w}$ oraz $\mathbf{z}$ oznaczają tutaj odpowiednio układ wektorów słów i przypisań w korpusie.

Wersja klasyczna

Potrzebny rozkład prawdopodobieństa zdarzeń $z_{d_0 i} = k,\ k = 1, ..., K$ jest zależny od obserwacji $\mathbf{w}$ oraz od pozostałych przypisań w korpusie $\mathbf{z}_{-(d_0 i)}$. Zatem szukamy rozkładu

$$P(z_{d_0 i} = k_0 | \mathbf{z}_{-(d_0 i)}, \mathbf{w}, \alpha, \beta). \tag{2.22}$$

Korzystając z definicji rozkładu warunkowego

$$P(z_{d_0 n} = k_0 | \mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta) = \frac{P(z_{d_0 n} = k_0, \mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta)}{P(\mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta)},$$

otrzymujemy

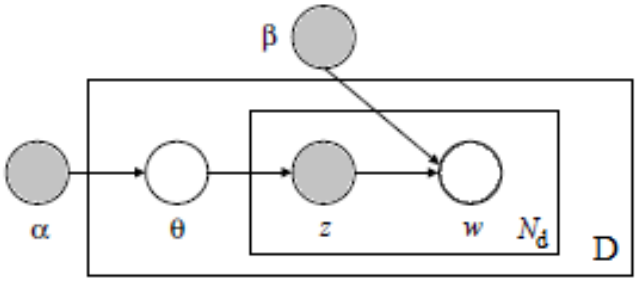
$$P(Z_{d_0 n} = k_0 | \mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta) \propto P(z_{d_0 n} = k_0, \mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta),$$

gdyż mianownik nie zależał od k_0. Wartość $P(z_{dn} = k_0, \mathbf{z}_{-(dn)}, \mathbf{w})$ otrzymamy z empirycznego rozkładu łącznego $P(\mathbf{w}, \mathbf{z} | \alpha, \beta)$, który zostanie wyznaczony przy użyciu faktoryzacji

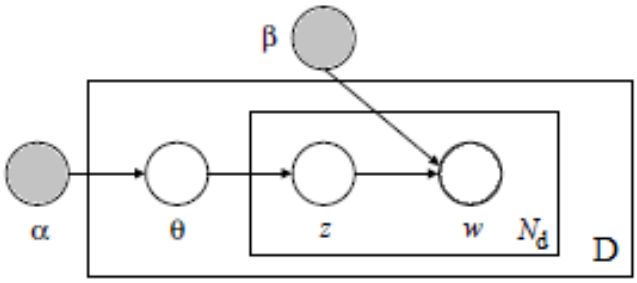
$$P(\mathbf{w}, \mathbf{z} | \alpha, \beta) = P(\mathbf{w} | \mathbf{z}, \alpha, \beta) P(\mathbf{z} | \alpha, \beta). \tag{2.23}$$

Dodajmy, że wyprowadzona w pracy postać rozkładu $P(Z_{d_0 n} = k_0 | \mathbf{z}_{-(d_0 n)}, \mathbf{w}, \alpha, \beta)$ będzie trywialna pod względem numerycznym. Z praktyczego punktu widzenia, jest to bardzo istotna zaleta tej metody wnioskowania.

Znajdziemy rozkład łączny obliczając osobno oba czynniki z równania 2.33. Zaczniemy od obliczenia $P(\mathbf{w} | \mathbf{z}, \alpha, \beta)$. Obecna sytuacja jest przedstawiona na ponższym rysunku (rys 2.5):



Rysunek 2.5



Rysunek 2.6

Mamy zatem trzy ustalone węzły parametry: α, β oraz przypisania $\mathbf{z}$. Zauważmy, że przy ustalonej wartości $\mathbf{z}$, wektor słów $\mathbf{w}$ nie zależy już od α oraz θ. Przejdźmy do analizy zależności względem tematów. Każde słowo w_{dn} pochodzi z już ustalonego tematu z_{dn} z rozkładu Categ$(\beta_{z_{dn}})$. Zatem słowa pochodzące z różnych tematów są od siebie niezależne, a każdy wektor słów $w_{i_1^k}, ..., w_{i_{r_k}^k}$, pochodzących ze wspólnego tematu k, ma rozkład wielomianowy. Oznaczając przez n_{kv}^d liczność przypisań słowa v w dokumencie d do tematu k, a przez $(\cdot)$ sumę po odpowiedniej współrzędnej, wpomniany rozkład wilomianowy jest postaci Mult$(n_{k(\cdot)}^{(\cdot)}, \beta_k)$, gdzie $n_{k(\cdot)}^{(\cdot)}$ jest licznością przypisań do tematu k w korpusie. Ze względu na niezależność słów pochodzących z różnych tematów (o różnych przypisaniach), całkowity rozkład $P(\mathbf{w}|\mathbf{z}, \beta)$ jest iloczynem rozkładów słów w obrębie każdego tematu:

$$P(\mathbf{w}|\mathbf{z},\beta) = \prod_{k=1}^{K}\left[\frac{\Gamma(n_{k(\cdot)}^{(\cdot)}+1)}{\prod\limits_{\mathbf{w}_n:\mathbf{z}_n=k}\Gamma(n_{k\mathbf{w}_n}^{(\cdot)}+1)}\prod_{\mathbf{w}_n:\mathbf{z}_n=k}\beta_{k\mathbf{w}_n}^{n_{k\mathbf{w}_n}^{(\cdot)}}\right] = \prod_{k=1}^{K}\left[\frac{\Gamma(n_{k(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{kv}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{kv}^{n_{kv}^{(\cdot)}}\right],$$

przy czym druga równość jest prawdziwa, ponieważ dla $v \notin \{\mathbf{w}_n : \mathbf{z}_n = k\}$ zachodzi $n_{kv}^{(\cdot)} = 0$, co implikuje $\Gamma(n_{kv}^{(\cdot)}+1) = \Gamma(1) = 1$ oraz $\beta_{kv}^{n_{kv}^{(\cdot)}} = \beta_{kv}^0 = 1$, więc wartości iloczynów nie zmieniają się.

Przechodzimy do rozkładu $P(\mathbf{z}|\alpha, \beta)$. Obecną sytuację ilustruje rys. 2.6

Zauważmy, że teraz $\mathbf{z}$ zależy tylko od α. Przypomnijmy, że θ_d pochodzi z rozkładu Dir(α), dla każdego $d = 1, ..., M$, a z_{di} z rozkładu Categ(θ_d), dla każdego $i = 1, ..., N_d$. Zatem wektor przypisań tematów w obrębie dokumentu pochodzi z rokładu wielomianowego Dirichleta. Ponadto, przypisania między dokumentami są warunkowo niezależne względem α, więc rozkład

całkowity jest iloczynem rozkładów w dokumentach, czyli:

$$P(\mathbf{z}|\alpha) = \prod_{d=1}^{D}\left[(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)}\right]. \tag{2.24}$$

Zatem szukany rozkład łączny jest postaci:

$$P(\mathbf{w},\mathbf{z}|\alpha,\beta) = \prod_{k=1}^{K}\left[\frac{\Gamma(n_{k,(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{k,v}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{kv}^{n_{k,v}^{(\cdot)}}\right]\times\prod_{d=1}^{D}\left[\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)}\right]. \tag{2.25}$$

Przedstawimy rozkład $P(z_{d_0n}=k_0,\mathbf{z}_{-(d_0n)},\mathbf{w},\alpha,\beta)$ jako iloczyn czynników, na które wpływa i nie wpływa fakt $z_{d_0n}=k_0$. To pozwoli nam znacząco uprościć obliczanie potrzebnego rozkładu. Zaczynamy od rozbicia $P(\mathbf{w}|\mathbf{z},\beta)$:

$$\begin{aligned}
P(\mathbf{w}|\mathbf{z},\beta) &= \prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{k,v}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{kv}^{n_{k,v}^{(\cdot)}}\\
&= \prod_{k\neq k_0}\frac{\Gamma(n_{k,(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{k,v}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{kv}^{n_{k,v}^{(\cdot)}}\times\frac{\Gamma(n_{k_0,(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{k_0,v}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{k_0v}^{n_{k_0,v}^{(\cdot)}}\\
&= \prod_{k\neq k_0}\frac{\Gamma(n_{k,(\cdot)}^{(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(n_{k,v}^{(\cdot)}+1)}\prod_{v=1}^{V}\beta_{kv}^{n_{k,v}^{(\cdot)}}\times\frac{1}{\prod_{v\neq w_n}\Gamma(n_{k_0,v}^{(\cdot)}+1)}\prod_{v\neq w_n}\beta_{k_0v}^{n_{k_0,v}^{(\cdot)}}\\
&\times\frac{\Gamma(n_{k_0,(\cdot)}^{(\cdot)}+1)}{\Gamma(n_{k_0,w_n}^{(\cdot)}+1)}\beta_{k_0w_n}^{n_{k_0,w_n}^{(\cdot)}}
\end{aligned} \tag{2.26}$$

Niech $\tilde{n}_{kv}^{d}$ oznacza liczność przypisania słowa v w dokumencie d do tematu k, wyliczoną bez uwzględniania przypisania $z_{d_0n}=k_0$. Innymi słowy są to liczności otrzymane tak, jakby z korpusu usunięto słowo w_{d_0n}. Wówczas mamy następujące zależności między $n_{d,v}^{k}$ a $\tilde{n}_{d,v}^{k}$:

$$n_{k,(\cdot)}^{(\cdot)} = \begin{cases} \tilde{n}_{k,(\cdot)}^{(\cdot)}+1, & \text{dla } k=k_0,\\ \tilde{n}_{k,(\cdot)}^{(\cdot)}, & \text{dla } k\neq k_0. \end{cases}$$

$$n_{k,v}^{(\cdot)} = \begin{cases} \tilde{n}_{k,v}^{(\cdot)}+1, & \text{dla } k=k_0 \text{ i } v=w_{d_0,n},\\ \tilde{n}_{k,v}^{(\cdot)}, & \text{wpp.} \end{cases}$$

$$n_{k,(\cdot)}^{d} = \begin{cases} \tilde{n}_{k,(\cdot)}^{d}+1, & \text{dla } d=d_0 \text{ i } k=k_0,\\ \tilde{n}_{k,(\cdot)}^{d}, & \text{wpp.} \end{cases}$$

oraz analogicznie dla liczności dokumentów

$$N_d = \begin{cases} \tilde{N}_d+1, & \text{dla } d=d_0\\ \tilde{N}_d, & \text{wpp.} \end{cases}$$

Poszczególne liczności uwzględniające $z_{d_0n} = k_0$ nie zmieniają się (gdy nie zależą od tego przypisania) lub mają wartość o jeden większą niż bez uwzględnienia tego faktu. Dzięki tym zależnościom możemy przedstawić rozpatrywany rozkład przy pomocy liczności niezależnych od z_{d_0n}. Wykorzystamy rozbicie z 2.26, w którym zmieni się tylko trzeci czynnik, a następnie skorzystamy z własności funkcji $\Gamma(x+1) = x\Gamma(x)$.

$$\begin{aligned}
P(\mathbf{w}|\mathbf{z},\beta) &= \prod_{k\neq k_0} \frac{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k,v}+1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}^{(\cdot)}_{k,v}} \times \frac{1}{\prod_{v\neq w_{d_0n}}\Gamma(\tilde{n}^{(\cdot)}_{k_0,v}+1)} \prod_{v\neq w_{d_0n}} \beta_{k_0v}^{\tilde{n}^{(\cdot)}_{k_0,v}} \\
&\times \frac{\Gamma(\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1+1)}{\Gamma(\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1+1)} \beta_{k_0w_nd_0n}^{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1} \\
&= \prod_{k\neq k_0} \frac{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k,v}+1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}^{(\cdot)}_{k,v}} \times \frac{1}{\prod_{v\neq w_{d_0n}}\Gamma(\tilde{n}^{(\cdot)}_{k_0,v}+1)} \prod_{v\neq w_{d_0n}} \beta_{k_0v}^{\tilde{n}^{(\cdot)}_{k_0,v}} \\
&\times \frac{\Gamma(\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1)}{\Gamma(\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1)} \beta_{k_0w_{d_0n}}^{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}} \times \frac{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1}{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1} \beta_{k_0w_{d_0n}}
\end{aligned} \tag{2.27}$$

Otrzymany iloczyn zwijamy do postaci podobnej do tej, od której rozpoczęliśmy:

$$\begin{aligned}
P(\mathbf{w}|\mathbf{z},\beta) &= \prod_{k\neq k_0} \frac{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k,v}+1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}^{(\cdot)}_{k,v}} \times \frac{1}{\prod_{v\neq w_{d_0n}}\Gamma(\tilde{n}^{(\cdot)}_{k_0,v}+1)} \prod_{v\neq w_{d_0n}} \beta_{k_0v}^{\tilde{n}^{(\cdot)}_{k_0,v}} \\
&\times \frac{\Gamma(\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1)}{\Gamma(\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1)} \beta_{k_0w_{d_0n}}^{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}} \times \frac{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1}{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1} \beta_{k_0w_{d_0n}} \\
&= \prod_{k\neq k_0} \frac{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k,v}+1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}^{(\cdot)}_{k,v}} \times \frac{\Gamma(\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k_0,v}+1)} \prod_{v=1}^{V} \beta_{k_0v}^{\tilde{n}^{(\cdot)}_{k_0,v}} \\
&\times \frac{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1}{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1} \beta_{k_0w_{d_0n}} \\
&= \prod_{k=1}^{K} \frac{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+1)}{\prod_{v=1}^{V}\Gamma(\tilde{n}^{(\cdot)}_{k,v}+1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}^{(\cdot)}_{k,v}} \times \frac{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+1}{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+1} \beta_{k_0w_{d_0n}}
\end{aligned} \tag{2.28}$$

Przechodzimy do drugiego czynnika - $P(\mathbf{z}|\alpha)$. Rozpiszemy jego wartość na czynniki w analogiczny sposób.

$$
\begin{aligned}
P(\mathbf{z}|\alpha) &= \prod_{d=1}^{D}(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)} \\
&= \prod_{d\neq d_0}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_{d_0}+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d_0}+\alpha_k)}{\Gamma(\alpha_k)} \\
&= \prod_{d\neq d_0}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(n_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \Gamma(\sum_{k=1}^{K}\alpha_k)\prod_{k\neq k_0}\frac{\Gamma(n_{k,(\cdot)}^{d_0}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \frac{1}{\Gamma(N_{d_0}+\sum_{k=1}^{K}\alpha_k)}\times\frac{\Gamma(n_{k_0,(\cdot)}^{d_0}+\alpha_{k_0})}{\Gamma(\alpha_{k_0})} \\
&= \prod_{d\neq d_0}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\tilde{N}_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \Gamma(\sum_{k=1}^{K}\alpha_k)\prod_{k\neq k_0}\frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d_0}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \frac{1}{\Gamma(\tilde{N}_{d_0}+1+\sum_{k=1}^{K}\alpha_k)}\times\frac{\Gamma(\tilde{n}_{k_0,(\cdot)}^{d_0}+1+\alpha_{k_0})}{\Gamma(\alpha_{k_0})} \\
&= \prod_{d\neq d_0}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\tilde{N}_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \Gamma(\sum_{k=1}^{K}\alpha_k)\prod_{k\neq k_0}\frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d_0}+\alpha_k)}{\Gamma(\alpha_k)} \\
&\times \frac{1}{\Gamma(\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k)}\times\frac{\Gamma(\tilde{n}_{k_0,(\cdot)}^{d_0}+\alpha_{k_0})}{\Gamma(\alpha_{k_0})}\times\frac{\tilde{n}_{k_0,(\cdot)}^{d_0}+\alpha_{k_0}}{\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k}
\end{aligned}
\tag{2.29}
$$

Podobnie jak w poprzednim przypadku, zwijając otrzymany iloczyn, mamy

$$
P(\mathbf{z}|\alpha) = \prod_{d=1}^{D}(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\tilde{N}_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d}+\alpha_k)}{\Gamma(\alpha_k)}\times\frac{\tilde{n}_{k_0,(\cdot)}^{d_0}+\alpha_{k_0}}{\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k} \tag{2.30}
$$

Łącząc 2.28 oraz 2.30 otrzymujemy:

$$\begin{aligned}
P(z_{d_0n} = k_0, \mathbf{z}_{-(d_0n)}, \mathbf{w}, \alpha, \beta) &= P(\mathbf{w}|\mathbf{z}, \beta) \cdot P(\mathbf{z}|\alpha) \\
&= \prod_{k=1}^{K} \frac{\Gamma(\tilde{n}_{k,(\cdot)}^{(\cdot)} + 1)}{\prod_{v=1}^{V} \Gamma(\tilde{n}_{k,v}^{(\cdot)} + 1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}_{k,v}^{(\cdot)}} \times \frac{\tilde{n}_{k_0,(\cdot)}^{(\cdot)} + 1}{\tilde{n}_{k_0,w_{d_0n}}^{(\cdot)} + 1} \beta_{k_0 w_{d_0n}} \\
&\times \prod_{d=1}^{D} (\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\tilde{N}_d + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d} + \alpha_k)}{\Gamma(\alpha_k)} \times \frac{\tilde{n}_{k_0,(\cdot)}^{d_0} + \alpha_{k_0}}{\tilde{N}_{d_0} + \sum_{k=1}^{K} \alpha_k} \\
&= \prod_{k=1}^{K} \frac{\Gamma(\tilde{n}_{k,(\cdot)}^{(\cdot)} + 1)}{\prod_{v=1}^{V} \Gamma(\tilde{n}_{k,v}^{(\cdot)} + 1)} \prod_{v=1}^{V} \beta_{kv}^{\tilde{n}_{k,v}^{(\cdot)}} \\
&\times \prod_{d=1}^{D} (\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\tilde{N}_d + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(\tilde{n}_{k,(\cdot)}^{d} + \alpha_k)}{\Gamma(\alpha_k)} \\
&\times \frac{\tilde{n}_{k_0,(\cdot)}^{(\cdot)} + 1}{\tilde{n}_{k_0,w_{d_0n}}^{(\cdot)} + 1} \beta_{k_0 w_{d_0n}} \cdot \frac{\tilde{n}_{k_0,(\cdot)}^{d_0} + \alpha_{k_0}}{\tilde{N}_{d_0} + \sum_{k=1}^{K} \alpha_k}.
\end{aligned} \tag{2.31}$$

Stąd

$$P(z_{d_0n} = k_0|\mathbf{z}_{-(d_0n)}, \mathbf{w}, \alpha, \beta) \ \propto \ \frac{\tilde{n}_{k_0,(\cdot)}^{(\cdot)} + 1}{\tilde{n}_{k_0,w_{d_0n}}^{(\cdot)} + 1} \beta_{k_0 w_{d_0n}} \cdot \frac{\tilde{n}_{k_0,(\cdot)}^{d_0} + \alpha_{k_0}}{\tilde{N}_{d_0} + \sum_{k=1}^{K} \alpha_k}, \tag{2.32}$$

gdyż pozostałe czynniki nie zależą od przypisania z_{d_0n}. Spójrzmy na interpretację otrzymanego wyniku. Przy założeniu $z_{d_0n} = k_0$ pierwszy ułamek jest odwrotnością empirycznego prawdopodobieństwa pojawienia się słowa w_{d_0n} w temacie k_0 - liczność przypisań słowa w_{d_0n} do tematu k_0 dzielona przez liczność pojawienie się tematu k_0. Zatem jeśli jest ono równe prawdopodobieństwu określonemu według modelu $\beta_{k_0 w_{d_0n}}$, to pierwszy czynnik wynosi 1. Natomiast gdy jest ono mniejsze (większe) to dzielenie $\beta_{k_0 w_{d_0n}}$ przez nie zwiększa (zmniejsza) prawdopodobieństwo przypinania słowa w_{d_0n} do tematu k_0, czyli koryguje odstępstwa wartości empirycznej od prawdziwej. Drugi czynnik jest przybliżeniem (wygładzeniem przez dodanie składników zależnych od α) empirycznego prawdopodobieństwa pojawienie się tematu k_0 w rozpatrywanym dokumencie d.

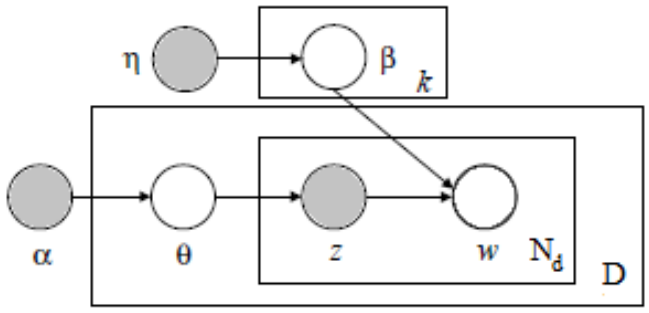
Pozostaje wyznaczyć estymatory parametrów θ_d, $d = 1, ..., M$. Jak zostało wspomniane na początku rozdziału, będzie to ich wartość oczekiwana w rozkładzie a posteriori. Mamy zatem dane przypisania $\mathbf{z}_d$, a zatem również liczności $n_{k,(\cdot)}^{d}$, $k = 1, ..., K$, $d = 1, ..., M$ oraz parametr α. Z 1.1 wiemy, że $\theta_d|\alpha, \{n_{k,(\cdot)}^{d}, k = 1, ..., K\} \sim \text{Dir}(n_{1,(\cdot)}^{d} + \alpha_1, ..., n_{K,(\cdot)}^{d} + \alpha_K)$, więc z własności rozkładu Dirichleta:

$$\hat{\theta}_{dk} = \frac{n_{k,(\cdot)}^{d} + \alpha_k}{\sum_{k=1}^{K} (n_{k,(\cdot)}^{d} + \alpha_k)} = \frac{n_{k,(\cdot)}^{d} + \alpha_k}{N_d + \sum_{k=1}^{K} \alpha_k}, \quad k = 1, ..., K, \ d = 1, ..., M.$$

Wersja wygładzana

Analogicznie jak poprzednio szukamy rozkładu łącznego, w którym parametr η zajmuje miejsce parametru β, przy użyciu faktoryzacji:

$$P(\mathbf{w}, \mathbf{z}|\alpha, \eta) = P(\mathbf{w}|\mathbf{z}, \alpha, \eta) P(\mathbf{z}|\alpha, \eta). \tag{2.33}$$



Rysunek 2.7

Podobnie zaczniemy od obliczenia $P(\mathbf{w}|\mathbf{z},\alpha,\eta)$. Obecna sytuacja jest przedstawiona na poniższym rysunku (rys 2.7):

Ustalone węzły to: α,η oraz $\mathbf{z}$. Tak samo jak poprzednio, przy ustalonej wartości $\mathbf{z}$, $\mathbf{w}$ nie zależy od α oraz θ, natomiast zmieni się zależność od tematów. Każde słowo w_{dn} pochodzi z już ustalonego tematu z_{dn} (z rozkładu Categ$(\beta_{z_{dn}})$), ale tym razem tematy nie są ustalone, lecz pochodzą z rozkładu Dir(η). W tej sytuacji, słowa pochodzące z różnych tematów są od siebie warunkowo niezależne pod warunkiem η, a każdy wektor słów $w_{i_1^k},...,w_{i_{r_k}^k}$, pochodzących ze wspólnego tematu k, ma rozkład wielomianowy Dirichleta z parametrem η. Zatem całkowity rozkład $P(\mathbf{w}|\mathbf{z},\eta)$ ma postać

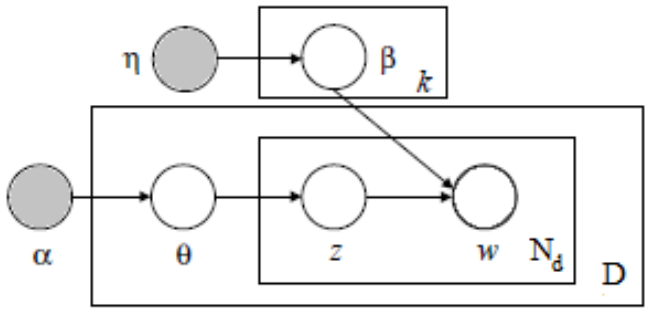
$$\begin{aligned} P(\mathbf{w}|\mathbf{z},\eta) &= \prod_{k=1}^{K} P\big((w_{i_1^k},...,w_{i_{r_k}^k}) : z_{i_1^k} = ... = z_{i_{r_k}^k} = k\big) \\ &= \prod_{k=1}^{K} \Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k,(\cdot)}^{(\cdot)} + \sum_{v=1}^{V}\eta_v)} \prod_{w_{dn}:z_{dn}=k} \frac{\Gamma(n_{kw_{dn}}^{(\cdot)} + \eta_{w_{dn}})}{\Gamma(\eta_{w_{dn}})}\Big) \\ &= \prod_{k=1}^{K} \Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k(\cdot)}^{(\cdot)} + \sum_{v=1}^{V}\eta_v)} \prod_{v=1}^{V} \frac{\Gamma(n_{kv}^{(\cdot)} + \eta_v)}{\Gamma(\eta_v)}\Big), \end{aligned} \tag{2.34}$$

przy czy ostatnia równość wynika z faktu, że $v \notin \{w_{dn} : z_{dn} = k\} \Rightarrow n_{kv}^{(\cdot)} = 0 \Rightarrow \frac{\Gamma(n_{kv}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)} = \frac{\Gamma(\eta_v)}{\Gamma(\eta_v)} = 1$, czyli iloczyn nie zmienia się. Analogicznie jak poprzednio rozpisujemy rozkład na

czynniki zależne i niezależne od $z_{d_0 n}$:

$$
\begin{aligned}
P(\mathbf{w}|\mathbf{z},\eta) &= \prod_{k=1}^{K}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(n_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&= \prod_{k\neq k_0}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(n_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&\times \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k_0,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(n_{k_0,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\\
&= \prod_{k\neq k_0}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(n_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&\times \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(n_{k_0,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v\neq w_{d_0 n}}\frac{\Gamma(n_{k_0,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\\
&\times \frac{\Gamma(n_{k_0,w_{d_0 n}}^{(\cdot)}+\eta_{w_{d_0 n}})}{\Gamma(\eta_{w_{d_0 n}})}\\
&= \prod_{k\neq k_0}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(\tilde{n}_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&\times \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}_{k_0,(\cdot)}^{(\cdot)}+1+\sum_{v=1}^{V}\eta_v)}\prod_{v\neq w_{d_0 n}}\frac{\Gamma(\tilde{n}_{k_0,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\\
&\times \frac{\Gamma(\tilde{n}_{k_0,w_{d_0 n}}^{(\cdot)}+1+\eta_{w_{d_0 n}})}{\Gamma(\eta_{w_{d_0 n}})}\\
&= \prod_{k\neq k_0}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(\tilde{n}_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&\times \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}_{k_0,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v\neq w_{d_0 n}}\frac{\Gamma(\tilde{n}_{k_0,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\\
&\times \frac{\Gamma(\tilde{n}_{k_0,w_{d_0 n}}^{(\cdot)}+\eta_{w_{d_0 n}})}{\Gamma(\eta_{w_{d_0 n}})}\times\frac{\tilde{n}_{k_0,w_{d_0 n}}^{(\cdot)}+\eta_{w_{d_0 n}}}{\tilde{n}_{k_0,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v}\\
&= \prod_{k=1}^{K}\Big(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}_{k,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(\tilde{n}_{k,v}^{(\cdot)}+\eta_v)}{\Gamma(\eta_v)}\Big)\\
&\times \frac{\tilde{n}_{k_0,w_{d_0 n}}^{(\cdot)}+\eta_{w_{d_0 n}}}{\tilde{n}_{k_0,(\cdot)}^{(\cdot)}+\sum_{v=1}^{V}\eta_v}
\end{aligned}
\tag{2.35}
$$

Przechodzimy do rozkładu $P(\mathbf{z}|\alpha,\eta)$. Obecną sytuację przedstawia poniższy rysunek (2.8):



Rysunek 2.8

Tak samo jak w wersji klasycznej, przypisania $\mathbf{z}$ zależą tylko od α i zalezność ta jest identyczna ja poprzednio. Zatem postać rozkładu nie zmienia się:

$$P(\mathbf{z}|\alpha) = \prod_{d=1}^{D}\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\tilde{N}_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(\tilde{n}^d_{k,(\cdot)}+\alpha_k)}{\Gamma(\alpha_k)}\times\frac{\tilde{n}^{d_0}_{k_0,(\cdot)}+\alpha_{k_0}}{\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k}\right., \tag{2.36}$$

Otrzymujemy zatem:

$$P(z_{d_0n}=k_0,\mathbf{z}_{-(d_0n)},\mathbf{w},\alpha,\beta) = \prod_{k=1}^{K}\left(\frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\Gamma(\tilde{n}^{(\cdot)}_{k,(\cdot)}+\sum_{v=1}^{V}\eta_v)}\prod_{v=1}^{V}\frac{\Gamma(\tilde{n}^{(\cdot)}_{k,v}+\eta_v)}{\Gamma(\eta_v)}\right)\times\frac{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+\eta_{w_{d_0n}}}{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+\sum_{v=1}^{V}\eta_v}$$
$$\prod_{d=1}^{D}\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\tilde{N}_d+\sum_{k=1}^{K}\alpha_k)}\prod_{k=1}^{K}\frac{\Gamma(\tilde{n}^d_{k,(\cdot)}+\alpha_k)}{\Gamma(\alpha_k)}\times\frac{\tilde{n}^{d_0}_{k_0,(\cdot)}+\alpha_{k_0}}{\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k}\right) \tag{2.37}$$

czyli

$$P(z_{d_0n}=k_0|\mathbf{z}_{-(d_0n)},\mathbf{w},\alpha,\beta) \;\propto\; \frac{\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}+\eta_{w_{d_0n}}}{\tilde{n}^{(\cdot)}_{k_0,(\cdot)}+\sum_{v=1}^{V}\eta_v}\times\frac{\tilde{n}^{d_0}_{k_0,(\cdot)}+\alpha_{k_0}}{\tilde{N}_{d_0}+\sum_{k=1}^{K}\alpha_k}$$

Również w tym przypadku, otrzymany wynik można łatwo zinterpretować. Zauważmy, że $\tilde{n}^{(\cdot)}_{k_0,w_{d_0n}}/\tilde{n}^{(\cdot)}_{k_0,(\cdot)}$ jest empirycznym (na podstawie danych $\mathbf{z}_{-(d_0n)}$) prawdopodobieństwem pojawieniem się słow w_{d_0n} w temacie k_0. Cały ułamek jet wygładzeniem tego prawdopodobieństwa, a więc pewnym jego przybliżeniem. Oznacza to, że im wieksze empiryczne prawdopodobieństwo wystąpienia słowa w_{d_0n} w temacie k_0, tym większe prawdopodobieństwo przypisania w_{d_0n} do tematu k_0. Łącząc oba czynniki (drugi czynnik jest przybliżeniem empirycznego prawdopodobieństa pojawienia się tematu k_0 w dokumencie d_0), mamy $P(z_{d_0n}=k_0,\mathbf{z}_{-(d_0n)},\mathbf{w},\alpha,\beta)\propto\tilde{p}(w_{d_0n}|z_{d_0n}=k)\cdot\tilde{(p)}(z_{d_0n}=k)=\tilde{p}(w_{d_0n},z_{d_0n}=k)$, gdzie $\tilde{p}(\cdot)$ oznacza przypliżenie rozkładu $p(\cdot)$.

Estymatory parametrów θ_d, $d=1,...,M$ są identyczna jak w wersji klasycznej. Jednakże w tym przypadku nie znamy tematów, które być może również chcielibyśmy poznać. Ich wartości możemy wyestymować w analogiczny sposób, ponieważ przy ustalonych przypisaniach $\mathbf{z}$ oraz parametrze η, rozkład a posteriori β_i jest rozkładem $\mathrm{Dir}(n^{(\cdot)}_{i1}+\eta_1,...,n^{(\cdot)}_{iV}+\eta_V)$. Zatem:

$$\hat{\theta}_{dk}=\frac{n^d_{k,(\cdot)}+\alpha_k}{N_d+\sum_{k=1}^{K}\alpha_k},\quad k=1,...,K,\;\; d=1,...,M,$$

$$\hat{\beta}_{ij} = \frac{n_{i,j}^{(\cdot)} + \eta_i}{n_{i,(\cdot)}^{(\cdot)} + \sum_{j=1}^{V} \eta_j}, \quad i = 1, ..., K, \ j = 1, ..., V.$$

2.4. *Perplexity*

TO DO

2.5. Predykcja przy użyci modelu LDA

TO DO

Rozdział 3

System rekomendacyjny

Stack Exchange jest grupą stron internetowych, służących wymianie wiedzy dziedzinowej, poprzez zadawanie pytań i udzielanie na nie odpowiedzi przez zarejestrowanych użytkowników. Ich najważniejszą cechą jest możliwość udzielania wielu odpowiedzi na jedno pytanie, których użyteczność jest oceniania przez użytkowników. Różne rodzaje aktywności są nagradzane zdobywaniem punktów. Ich suma stanowi "reputację" użytkownika, która jest główną motywację do angażowania się w życie danego portalu. Poszczególne strony skupiają społeczności profesjonalistów i amatorów związanych z danymi tematami, takimi jak matematyka, filozofia, lingwistyka, rodzicielstwo, poker i wiele innych. Najpopularniejszym portalem jest Stack Overflow, poświęcony programowaniu, który jest jednocześnie pierwowzorem i pierwszą stroną grupy Stack Exchange.

Do sierpnia 2015 r. na Stack Overflow w przybliżeniu zadano 10 milionów pytań oraz udzielono 16 milionów odpowiedzi, a liczba zarejestrowanych użytkowników wynosi 4,5 miliona. Ponadto liczby te stale rosną. Oczywistym jest, że żaden użytkownik nie jest w stanie na bieżąco analizować wszystkich pytań, aby znaleźć te, na które mógłby udzielić odpowiedzi. Dlatego wydaje się, że narzędzie wskazujące użytkownikom pytania, na które znają odpowiedź z dużym prawdopodobieństem, powinno przyspieszyć wymianę wiedzy oraz przyczynić się do wzrostu zadowolenia zarówno użytkowników zadających pytania, jak i odpowiadających. W szczególności, może przyczynić się to do uzyskania odpowiedzi na pytania, na które właściwe osoby nigdy nie trafiłyby w odpowiednim czasie (jest wiele pytać pozostających bez odpowiedzi przez bardzo długi czas). Konstrukcja i zbadania jakości działania takiego narzędzia - systemu rekomendacyjnego, jest zasadniczym celem niniejszej pracy. Proponowany w pracy system rekomendacyjny opiera się na modelowaniu tematyki dokumentów tekstowych. Idea jego działania jest następująca: system analizuje tematy wypowiedzi pojawiających się w portalu, by w przyszłości danemu użytkownikowi zarekomendować pytania o tematach podobnych do tematów jego wcześniejszych odpowiedzi. To podejście opiera się na naturalnym założeniu, że ilość i jakość odpowiedzi w danym temacie świadczy o wiedzy użytkownika w tym obszarze.

Pytanie: Napisać plan rozdziału - w pierwszym podrozdz to, drugim tamto... ?

3.1. Środowisko działania systemu

Środowiskiem działania systemu jest strona interentowa Stack Overflow. W niniejszym rozdziale opisane zostaną główne aspekty funkcjonowania portalu.

Stack Overflow to internetowy portal dla programistów formatu "question and answer", służący wymianie wiedzy eksperckiej poprzez zadawanie pytań i udzielanie odpowiedzi przez jego użytkowników, co stanowi zasadniczą zawartość strony. Ważną cechą jest to, że na pytanie można udzielić dowolnej ilości odpowiedzi. Naturalnym jest, że problemy mogą mieć wiele przyczyn oraz wiele rozwiązań. Ponadto, nawet jeśli rozwiązanie jest jedno (bo np. pytanie było tak precyzyjne), to każdy i tak przedstawiłby je w indywidualny sposób. Dlatego bardzo często trzeba dokonywać wyboru spośród zaproponowanych rozwiązań. Jedną z funkcjonalności portalu jest możliwości oceny jakości wypowiedzi przez użytkowników, co wspomaga wybór rozwiązania. Polega to na przyznawaniu głosów pozytywnych (*up-votes*), gdy ktoś chce poświadczyć użyteczność danej wypowiedzi oraz negatywnych (*down-votes*), gdy ktoś uważa, że dane rozwiązanie jest niepoprawne. Ich różnica (zwana *score*) dostarcza informacji o jakość poszczególnej wypowiedzi. Naturalnie, wiele odpowiedzi na jedno pytanie może mieć dodatnią wartość *score*. Ponadto głosy można również oddawać na pytania, co oznacza poparcie zasadności pytania lub jej braku (np. gdy w portalu pojawiło się już takie pytanie wcześniej). Istnieje jeszcze jedna funkcjonalność związana z oceną jakości odpowiedzi - pytający ma możliwość oznaczenia, którą odpowiedź wykorzystał do rozwiązania problemu. W praktyce ta informacja ma podobną wartość co oddanie pozytywnego głosu. Oprócz zadawania pytań i udzielania odpowiedzi na portalu można również dodawać komentarze do wypowiedzi. Zawierają one często sugestie poprawienia wypowiedzi (np. doprecyzowanie pytania) lub po prostu pewne dopowiedzenia.

Motywacją do aktywnego udzielania się na portalu jest tzw. reputacja użytkowników. Wiele z możliwych czynności jest nagradzana zdobywaniem punktów. Na ich podstawie określa się reputację, co naturalnie wiąże się z pewnym prestiżem. Ponadto od jej wartości zależy zakres dostępnych aktywności. Dopiero po przekroczeniu pewnego progu użytkownik otrzymuje prawo do głosowania czy dodawania komentarzy do wypowiedzi. Użytkownicy o bardzo wysokiej reputacji mają możliwość moderowania treści portalu m. in. poprzez edytowanie (celem poprawienia) wypowiedzi innych czy zgłaszanie do zablokowania niewłaściwych pytań.

Kolejnym istotnym elementem jest obowiązek dołączania do pytań tagów, które są słowami lub wyrażeniami charakteryzującymi przedmiot pytania. Mogą to być nazwy języków programowania, nazwy konkretnych programów/systemów, powszechnie używane określenie danego zagadnienia, itp. Tagi są jawnym podaniem obszarów jakich dotyczy pytanie.

3.2. Schemat działania

W niniejszym rozdziale przedstawiony zostanie dokładny opis konstrukcji proponowanego systemu rekomendacyjnego. Celem jego działania jest znalezienie dla danego pytania pojawiającego się na portalu użytkowników, którzy będą potrafili na nie odpowiedzieć.

Zaczniemy od przedstawienia ogólnej zasady działania, by następnie przejść do szczegółowego opisu poszczególnych aspektów. System składa się z dwóch elementów: bazy wiedzy użytkowników oraz modułu generującego rekomendacje, przy czym rekomendacje są generowane na podstawie posiadanej bazy. Pierwszym krokiem jest dopasowanie do danych, którymi dysponujemy wygładzanego modelu LDA. Otrzymujemy z niego wyestymowane rozkłady tematów dla każdej wypowiedzi. Jednocześnie na podstawie głosów użytkowników określamy jakości odpowiedzi. Łącząc te dwie informacje oceniamy wiedzę autora odpowiedzi w określonych tematach. Analizując wszystkie odpowiedzi danego użytkownika, otrzymujemy jego sumaryczną wiedzę w każdym temacie i w ten sposób tworzymy całą bazę. Generowanie rekomendacji rozpoczyna się od estymacji tematyki pytania, przy użyciu wcześniej dopasowanego modelu.

Następnie dokonuje się agregacji wiedzy użytkowników w tematach dotyczących pytania i na tej podstawie tworzony jest ranking. Rekomendacja polega na wybraniu użytkowników, którym pytanie miałoby zostać przedstawione. Zaznaczmy, że przedmiotem analiza zawartych w pracy będą generowane przez system rankingi.

Zwróćmy uwagę na pewien praktyczny aspekt działania systemu. Załóżmy, że mamy dane zgromadzone do pewnego momentu czasowego t_0, na podstawie, których będzie dokonywana rekomendacja dla pytań pojawiających się od tego momentu - dla $t > t_0$. W rzeczywistości środowisko działania systemu ewoluuje w czasie: pojawiają się nowi użytkownicy, udzielane są kolejne odpowiedzi, a ponadto intuicyjnie wzrasta liczba tematów poruszanych na portalu. Oczywistym jest, że wraz czasem jakość rekomendacji będzie spadać. Dlatego z praktycznego punktu widzenia, koniecznym byłaby aktualizacja "silnika" (zarówno bazy jak i modelu) systemu. Niniejsza praca nie porusza kwestii aktualizacji. Rozpatrywana jest sytuacja statyczna: na podstawie danych zgromadzonych do momentu t_0, generowane są rekomendacje dla pytań pojawiających się o okresie $(t_0, t_0+\delta)$ i na ich podstawie są prowadzone badania. W celu uniknięcia wpływu upływu czasu na jakość działania systemu, wartość δ jest stosunkowo mała.

Szczegółowy schemat pełnego procesu rekomendacji wygląda następująco:

Etap I. Zasilenie systemu danymi - stworzenie bazy wiedzy użytkowników.

1. Ocena jakość udzielonych odpowiedzi.
2. Przygotowanie danych - preprocessing.
3. Dopasowanie do danych modelu LDA w celu estymacji tematów zawartych w wypowiedziach.
4. Ocena wiedzy użytkowników w poszczególnych tematach.

Etap II. Generowanie rekomendacji dla nowych pytań.

5. Preprocessing pytania.
6. Predykcja tematyki pytania, przy użyciu modelu zbudowanego w etapie I.
7. Wygenerowanie rankingu użytkowników.
8. Zarekomendowanie pytania odpowiednim użytkownikom.

Omówimy teraz poszczególne kroki przedstawione w schemacie, za wyjątkiem punktów dotyczących obróbki danych (2. i 5.), którym poświęcony jest podrozdział 3.3.

Punkt 1. Zaczynamy od sposobu oceny jakości odpowiedzi. Przypomnijmy, że jedną z funkcjonalności portalu Stack Overflow jest udzielanie pozytywnych lub negatycznych głosów poszczególnym wypowiedziom (zarówno odpowiedziom jaki i pytaniom). Jest to naturalnie cenna informacja o tym, czy dana wypowiedź jest wartościowa dla użytkowników. Dodajmy, że głosy te są praktycznie jedyną informacją o wartości odpowiedzi jaką dysponujemy. Istnieje jeszcze funkcjonalość, pozwalająca na zaznaczenie przez autora pytania odpowiedzi, która pomogła mu rozwiązać problem, jednakże po pierwsze często pytający z tej mozliwości nie korzystają (być może dlatego, że żadna odpowiedź im nie pomogła), a po drugie wcale nie oznacza to, że dana odpowiedź zawiera rozwiązanie lepsze od innych. Dlatego jako miarę jakości odpowiedzi w systemie przyjmujemy różnicę pomiędzy liczbą głosów pozytywnych i negatywnych.

Punkt 3. Wykorzystywany model to wygładzana wersja LDA. Dokumentem tekstowym jest złączenie treści pytania i udzielonych na nie odpowiedzi (o ile istnieją), które będziemy nazywać konwersacją. Oznacza to, że tematyka jest określana na poziomie konwersacji i w ten sposób wszystkie znajdujące się w niej wypowiedzi mają wspólny rozkład tematów. Tytuły pytań traktowane są jak zwykły fragment ich treści. W treści konwersacji nie uwzględniamy komentarzy. Parametryzacja procesu dopasowania modelu zostanie opisane w późniejszej części pracy.

Punkt 4. Mając już określone tematy wypowiedzi oraz informacje o ich jakości, można na ich podstawie stworzyć bazę wiedzy użytkowników, którą będziemy oznaczać jako B. Formalnie będzie to macierz, w której wartość elementu B_{ij} określa wiedzę i-tego użytkownika w j-tym temacie. Elementy B_{ij} będą sumą (po wszystkich odpowiedziach użytkownika i) punktów wiedzy, przyznanych w temacie j. Przykładowo, jeśli użytkownik i udzielił trzech odpowiedzi, w których otrzymał odpowiednio $1, 2, 7$ punktów wiedzy w temacie j, to wiedza tego użytkownika w temacie j będzie wynosić $B_{ij} = 1 + 2 + 7 = 10$. Opiszemy teraz zasadę przyznawania punktów wiedzy. Niech k_{dj} oznacza punkty przyznane autorowi odpowiedzi d w temacie j. Mają one postać $k_{dj} = w_{dj} \cdot p_d$, gdzie w_{dj} jest wagą tematu j w odpowiedzi d, a p_d oznacza miarę jakości odpowiedzi d, która jest obliczana na podstawie głosów użytkowników. Sposób doboru wag oraz miary jakości jest elementem badań.

Punkt 6. Predykcja tematyki pytań jest dokonywana przy użyciu modelu otrzymanego w etapie pierwszym. Dokumentami są tym razem pojedyncze pytania.

Punkt 7. Ideą systemu jest generowanie rankingu na podstawie wiedzy zagregowanej pod kątem tematyki pytania. Mając wyestymowany rozkład tematów w pytaniu oraz dysponując wiedzą użytkowników, trzeba te dwie informacje połączyć, w celu określenia potencjalnej zdolność odpowiedzi na pytanie. Ranking dla pytania d będzie tworzony na podstawie wartości r_{di}, obliczonej dla każdego użytkownika i według formuły: $r_{di} = \sum_j w_{dj} \cdot B_{ij}$, gdzie w_{dj} są wagami poszczególnych tematów w pytaniu d (określonymi identycznie jak podczas obliczania bazy wiedzy). Zatem jest to po prostu ważona suma wiedzy w poszczególnych tematach. Oczywiście im wieksza wartość r_{di}, tym wyższa pozycja użytkownika i w rankingu. Dodajmy, że w przypadku jednakowych wartości r_{di} dla różnych użytkowników, ich pozycja w rankingu będzie również równa: przykładowo dla wartości $\{9, 7, 7, 3\}$ ranking będzie miał postać $\{1, 2, 2, 4\}$. Ponadto ranking obejmuje tylko użytkowników, dla których r_{di} jest dodatnie. W praktyce oznacza to, że z prawdopodobieństwem bliskim 1, nie będzie on obejmował wszystkich użytkowników znajdujących się w bazie. Jest to użyteczny filtr zawężający zbiór użytkowników uwzględnianych w rankingu. Można by było zwyczjnie włączyć użytkowników z $r_{di} = 0$ na odpowiedniej pozycji rankingu, natomiast zaburzyłoby to obraz działania systemu wynikający z zastosowania modelu. W skrajnym przypadku, gdyby dla danego pytania r_{di} wszystkich użytkowników wynosiłoby 0, to zajęliby oni ex aeqo pierwsze miejsce, a zatem właściwi użytkownicy też znaleźliby się na tym miejscu, co wcale nie wynikałoby z dobrego działania systemu.

Punkt 8. Zarekomendowanie pytań polega jedynie na wyborze użytkowników, do których pytanie zostałoby skierowane, a sposób ich wyboru zależy od efektu jaki chcemy uzyskać. Przykładowo, chcąc skrócić czas oczekiwania na odpowiedź zarekomendujemy pytania czołówce rankingu, natomiast chcąc dać szansę wypowiedzi "słabszym" użytkownikom rekomendacja uwzględni również dalsze pozycje rankingu. Punkt ten nie będzie przedmiotem rozważań zawartych w pracy.

3.3. Obróbka danych

Nieodłącznym elementem pracy z danymi, w szczególności tekstowymi, jest ich wstępna obróbka. Jednocześnie jest to kwestia całkowicie indywidualna - nie ma ogólnych, słusznych dla każdego problemu reguł przygotowania danych - metodę trzeba określić pod kątem danego przypadku. W przypadku środowiska działania systemu kwestia ta wydaje się nie mieć oczywistych rozwiązań. Naturalnie, można nie przykładać wagi do tego wagi i dokonać bardzo prostej obróbki danych lub nie robić jej wcale. Jednakże takie potraktowanie sprawy wiąże się z ryzykiem istotnego negatywnego wpływu na jakość działania systemu, a tym samym utrudnieniem oceny jakości systemu wynikającej z zastosowania modelowania tematycznego.

Jedną z kluczowych kwestii jest fakt, że posty składają się zazwyczaj z dwóch elementów: treści sformułowanej w języku naturalnym oraz kodu w pewnym języku programowania. Określenie relacji poszczególnych części w odniesieniu do całej treści jest jednym z trudniejszych problemów do rozwiązania. Przyjmijmy, że część wyrażoną w jęzuku naturalnym będziemy nazywać "naturalną", natomiast część, zawierającą kody programów będziemy nazywać po prostu "kodem". Zwązane z nimi są następujące aspekty (zaznaczmy, że rozważania dotyczą intuicji i nieformalnego podejścia do problemu, co jest integralną częścią praktycznej analizy danych):

- Zazwyczaj w części naturalnej zawarty jest opis problemu, zatem można się spodziewać, że ta część wnosi dużo informacji o tematyce pytania.
- W szczególność tytuł pytania jest jednozdaniowym streszczeniem problemu, często zawierającym istotne słowa odnoszące się do tematu.
- W części naturalnej pojawiają się też nazwy własne, które często samodzielnie mogłyby określać temat (w intuicyjnym rozumieniu).
- Model LDA, ze względu na proces generowania danych jaki zakłada, nie wydaje się być adekwatny dla tekstów będących kodami programów, które posiadają strukturę wynikającą z reguł danego języka programowania.
- Wydaje się, że "rozpoznanie" przez model języka programowania na podstawie kodu nie powinno być trudne, a informacja o języku sama w sobie powinna być cenna dla określenia tematu.
- Z drugiej strony duża część słów zawartych w kodzie są to stałe elementy kodów w danym języku, niezależne od tematyki zagadnienia.
- W kodzie mogą pojawiać się słowa charakterystyczna dla danego zagadnienia.

Pomimo tego, że tak naprawdę nie wiadomo ile wspólnego z rzeczywistym działaniem modelu mają oczekiwania oparte na intuicji, poddajemy tekst wstępnej obróbce uwzględniającej w pewien sposób powyższe aspekty. Opiszemy teraz modyfikacje przeprowadzone na obu częściach, a zaczniemy od opisu przygotowania języka naturalnego.

Szczególną kwestią związaną z tą częścią są nazwy własne. Wydaje się, że dla jakości rekomendacji, wykorzystywanie w jakiś sposób nazw własnych powinno przynosić korzyści. W szczególności, znaczenie może mieć poziom szczegółowości nazwy własnej. Przykładowo, jeśli użytkownik zna wersję pierwszą pewnego programu, to nie znaczy, że ma jakąkolwiek wiedzę na temat wersji drugiej, ale jednocześnie będzie wiedział coś na jej temat z większym prawdopodobieństwem, niż użytkownik niemający doświadczeń związanych z żadną wersją. Dlatego

działamy następująco: dokonujemy ekstrakcji (metoda zostanie opisana poniżej) nazw własnych zawierających element numeryczny, by następnie dodać do treści dokumentu słowo będące sklejenie tej nazwy w jeden wyraz. Przykładowo dla nazwy "NetBeans IDE 8.0.2" utworzone zostanie słowo "NetBeansIDE8.0.2". Taka operacja ma następujące zalety:

- Zwiększa się procentowy udział nazwy własnej w zbiorze słów stanowiących treść dokumentu, a tym samym zwiększa się wpływ nazwy na tematykę dokumentu.
- Wzrasta podobieństwo (mierzone liczbą wspólnych słów) między dokumentami zawierającymi np. tę samą wersję programu, co powinno się przełożyć na wzrost podobieństwa tematycznego.

Nazwy własne wykrywamy przy pomocy wyrażeń regularnych (biblioteka ICU), przy czym interesują nas tylko te, które zawierają element numeryczny. Nie jest to narzędzie doskonałe i nie wyłapiemy przy jego pomocy wszystkich nazw, natomiast pozwala ona w prosty sposób uzyskać zadowalający poziom ekstrakcji. Nazwy własne, które uwzględnimy definiujemy jako wyrażenia pasujące do jednej z dwóch następujących formuł (przez "słowo" rozumiemy ciąg liter):

- ciąg od jednego do czterech słów rozdzielonych spacjami, zawierających przynajmniej jedną wielką literę oraz ewentualnie znak "+", po którym następuje ciąg liczb ewentualnie rozdzielonych kropkami: ([a-zA-Z+]*[A-Z][a-zA-Z+]*\\s){1,4}\\d+(\\.\\d+)*,
- słowo, zawierające przynajmniej jedną wielką literę poprzedzone kropką lub słowem i kropką (bez spacji), po którym następuje ciąg liczb ewentualnie rozdzielonych spacjami lub kropkami: [a-zA-Z]*\\.[a-z]*[A-Z]+[a-z]*(\\s{0,1}\\d+\\.*)+ .

Nie uwzględniamy tutaj słów rozpoczynających zdania, ponieważ ona zawsze zaczynają się wielką literą. Początki zdania identyfikujemy faktem, że przed spacją poprzedzającą słowo zaczynające się wielką literą, znajduje się kropka. Nie jest to podejście bezbłędne, ale wystrczająco skuteczne i proste.

Mamy zatem wydobyte (i sklejone w jedno słowo) nazwe własne. Zanim dołączymy je do dokumentu, poddajemy treść części naturalnej następującym modyfikacjom:

- Usuwamy wszystkie znaki inne niż litery lub cyfry, za wyjątkiem jednej sytuacji - nie usuwamy znaków "+" oraz "#" następujących po ciągu liter. Głównym uzasadnieniem są nazwy bardzo popularnych języków programowania - "C++" oraz "C#", które po usunięciu tych znaków zamieniłyby się w literę "C", która również jest nazwą języka programowania. Z perpektywy analizy tematycznej mogłoby to utrudnić wykrycie faktu, że są to różne języki. Ponadto istnieją również inne nazwy własne, kończące się znakami "+" lub "#". Usunięcie przebiega w taki sposób, że jeśli np. dwa słowa było rozdzielone myślinikiem, to po usunięciu myślnika słową pozostają oddzielone.
- Na całych danych przeprowadzamy stemming - sprowadzamy słowa do rdzenia. Zakładamy, że odmiana nie ma znaczenia z punktu widzenia tematyki, co prowadzi do znacznej redukcji zbioru różnych słów.

Na koniec dodajemy sklejone nazwy własne, przy czym jesli nazwa występuje n razy, to jej postać sklejona również pojawi się n razy. Następnie zamieniamy wszystkie litery na małe (również w nazwach). Nie chcemy rozróżniać słów ze względu na wielkości liter, ponieważ nie powinna ona mieć znaczenia dla tematytki wypowiedzi. Na tym kończy się przygotownie części naturalnej.

Przechodzimy do kodu. Pierwszą ważną kwestią jest tutaj sama definicja "słów", które zostaną uwzględnione w analizie. Określamy je jako co najmniej dwuelementowe spójne ciągi liter. Nie uwzględniamy pojedynczych liter, ponieważ praktycznie zawsze są to nic nie wnoszące nazwy zmiennych. Ponadto nie uwzględniamy również żadnych cyfr, ponieważ trudno wyobrazić sobie sytuację, w której cyfry w kodzie mówiłyby cokolwiek o temacie problemu. Nie uwzględniamy "otoczenia" tak zdefiniowanych słów - jeżeli dwa takie ciągi są rozdzielone jedynie jakimś znakiem, np. kropką czy myślnikiem, bedą to dla nas dwa odrębne słowa. Drugim elementem modyfikacji jest redukcja powtarzających się słów. Uwzględniając punkt rozważań mówiący o tym, że część słów w kodzie jest stałym elementem danego języka, niezależącym istotnie od tematyki, decydujemy się na sprowadzenie kodu do zbioru słów unikalnych. Oznacza to, że niezaleznie od liczności wystąpień słowa w oryginalnym kodzie, po modyfikacji każde słowo będzie występowało raz. Przykładowo kod "jeden dwa dwa trzy trzy trzy" zostanie przetworzony do "jeden dwa trzy". W ten sposób chcemy zmniejszyć odsetek słów stale pojawiających się w danym języku niezależnie od tematu, zwiększając jednocześnie udział słów bardziej charakterystycznych dla danego zagadnienia. Po trzecie, w kodzie również zamieniamy wszystkie litery na małe. Nie chcemy rozróżniać słów ze względu na wielkość liter, ponieważ nie powinna ona zależeć od tematyki.

Pytanie: dać przykłady wyłapanych nazw własnych?

Po przeprowadzeniu opisanych modyfikacji w korpusie występuje blisko 5 mln. słów (dokładnie jest to 4 969 278). Przypomnijmy, że dokumentów (konwersacji) jest niecałe 60 tysięcy. Dokonujemy zatem redukcji zbioru słów, pozbywając się dwóch typów słów:

- słowa uznane za nieznaczące z perspektywy tematu treści (tzw. stopwords) - jest ich 708, z czego 216 stanowią słowa wybrane spośród 300 najczęściej występujących słów w danych;

- słowa bardzo rzadkie, definiowane jako słowa, które pojawiły się w nie więcej niż trzech konwersacjach - takich słów jest 4 926 623, co stanowi 99% liczby unikalnych słów.

Po redukcji w danych pozostaje 41 947 unikalnych słów.

3.4. Wyniki/Opis/Omówienie? badań

3.4.1. Dane

Danymi tekstowymi, na których będziemy pracować są odpowiedzi oraz pytania. Komentarze, które również są treścią związaną z wypowiedziami nie są włączone do zbioru danych, ponieważ ich treść często nie odnosi się bezpośrednio do problemu. W zależności od momentu stosowania modelu, dokumenty w korpusie będą różnie zdefiniowane - dopasowanie modelu do danych odbywa się na korpusie złożonym z konwersacji (pytania połączone z odpowiedziami), natomiast predykcje z wykorzystaniem modelu będą dokonywane dla dokumentów, które będą stanowiły pojedyncze pytania. Przypomnijmy, że elementem treści pytań są ich tytuły. Ponadto w systemie wykorzystywane są również dane inne niż tekstowe, czyli głosy oddane na wypowiedzi.

Zbiór treningowy

Zbiór treningowy - dane na podstawie, których budowany jest model oraz tworzona baza wiedzy o użytkownikach, składa się z wszystkich danych zgromadzonych od początku istnienia portalu (wrzesień 2008 r.) do końca 2008 roku. W okresie tym zadanych zostało 59606 pytań (zatem tyle jest konwersacji) oraz udzielono 129890 odpowiedzi o dodatniej jakości, która jest oceniana na podstawie głosów oddanych do końca 2008 r. Autorami pozytywnych odpowiedzi jest 13530 różnych użytkowników. Dokumentami tekstowych, do których dopasowywany jest model są konwersacje. Łączna liczba słów w korpusie wynosi 9 570 345, co daje średnio ok 160 słów w konwersacji.

Zbiór testowy

Załóżmy, że baza wiedzy systemu powstała na podstawie danych zebranych do momentu t_0. W celu zminimalizowania wpływu upływu czasu na jakość działania systemu, zbiór testowy zostanie utworzony z pierwszych pytań, które pojawiły się po czasie t_0 i spełniają pewne warunki. Jak zaraz uzasadnimy, dobór zbioru testowego nie jest kwestią oczywistą i w celu usunięcia wpływu czynników niezależnych od systemu, należy odpowiednio dobierać pytania do zbioru testowego. Ocenę jakosci działania systemu, mogą utrudnić następujące aspekty:

- Istnieje dużo pytań, na które nigdy nie została udzielona odpowiedź. Uwzględnienie ich w zbiorze testowym nie ma sensu, ponieważ nie da się ocenić jakości rekomendacji dla nich wygenerowanej.
- Również nie ma sensu uwzględnianie pytań, na które odpowiedzieli tylko użytkownicy, którzy nie udzielili żadnej odpowiedzi do czasu t_0, ponieważ niezależnie od doboru parametrów systemu, potencjalna zdolność odpowiedzi na jakiekolwiek pytanie będzie zerowa. Dla takich pytań również nie da się ocenić jakości rekomendacji.
- Ponadto zgodnie z konwencją przyjętą podczas gromadzenie wiedzy, polegającą na uwzględnianiu tylko odpowiedzi dodatniej jakości, tu również ta zasadna powinna być zastosowana z tych samych powodów. Mimo iż sam fakt udzielenia odpowiedzi mówi coś o wiedzy użytkownika, formalnie nie będziemy uwzględniać odpowiedzi o zerowej jakości.

Powyższe kwestie można łatwo rozwiązać. Z wyżej wymienionych powodów, zbiór testowy będzie składał się z pytań, które otrzymały przynajmniej jedną odpowiedź o dodatniej jakości, autorstwa użytkownika, który udzielił przynajmniej jednej odpowiedzi (nie muszą być pozytywnej jakości) przed momentem t_0. Nie nakładamy ograniczenia na czas udzielenia odpowiedzi, ponieważ każda z nich jest cenną informacją. Narzucenie takiego warunku pozwala sądzić, że zostały zneutralizowane najważniejsze czynniki zoburzające bezpośrednią ocenę jakości działania systemu, co pozwoli obiektywniej porównywać wyniki. Zaznaczmy, że najistotniejsze jest tutaj porównywanie jakości w zależności od parametrów, a nie globalna ocena systemu, więc konieczne jest, aby ocena była obiektywna względem parametrów, a mniej ważną kwestią jest konkretny dobór miar. Liczność zbioru wynosi 1500 pytań, przy czym są to pierwsze pytania od momentu t_0 spełniające wspomniany warunek. Zostały one zadane w ciągu pierwszych sześciu (jest to wspominana wcześniej wartość δ) dni roku 2009. Elementem zbioru testowego są również wszystkie (do połowy roku 2014) odpowiedzi do tych pytań oraz oddane głosy. Rozkład ilości odpowiedzi do pytań ze zbioru tekstowego przedstawia rysunek **??**.

TUTAJ TEN RYSUNEK

Dokumentami tekstowymi, dla których estymowania jest tematyka przy użyciu wcześniej zbudowanego modelu są poszczególne pytania.

3.4.2. nie wiem jak to nazwać, jakieś sugestie?

tu analiza o której wspominałem - popatrzymy jak wyglądają tematy zawierajce dane słowo w zalezności od liczby tematów w modelu

3.4.3. Analiza wyników

Rekomendajca podlegająca ocenie jest to wygenerowany przez system ranking, który nie obejmuje zwykle wszystkich użytkowników znajdujących się w bazie. Potrzebna jest zatem miara jakości rankingu. Istnieją w literaturze takie miary, które powstały z myślą o ocenie systemów wyszukiwania informacji (*information retrieval*), które najczęściej polegają na wygenerowaniu rankingu obiektów (np. stron internetowych) uporządkowanych względem szacowanej adekwatności dla danego zapytania. Są dwa główne typy takich miar: pierwszy opiera się na założeniu, że każdy obiekt jest adekwatny lub nie i wówczas dysponujemy etykietami z tą informacją, natomiast drugi typ wykorzystuje etykiety liczbowe, oznaczające stopień adekwatności każdego obiektu dla danego zapytania. Byłoby wygodnie, gdybyśmy o każdym użytkowniku wiedzieli, na które pytania zna on odpowiedź i na które nie zna. Niestety dane, którymi dysponujemy nie odpowiadają żadnemu z wspomnianych dwóch przypadków, ponieważ tak naprawdę zawierają one tylko częściową informację. Z danych wiemy kto udzielił poztywnych odpowiedzi na dane pytanie, co daje nam informację potwierdzającą, że użytkownik posiada wiedzę w tej tematyce. Natomiast fakt, że użytkownik nie udzielił odpowiedzi na dane pytanie, nie oznacza braku wiedzy w danym temacie. Mając powyższe na uwadze, zdefiniujemy teraz miary jakość rekomendacji, które będziemy wykorzystywać. Przypomnijmy, że w zbiorze testowym do każdego pytania mamy przynajmniej jedną odpowiedź.

Załóżmy, że system rekomenduje dane pytanie k najlepszym użytkownikom z wygenerowanego rankingu. Naturalnie, dobrze działający system powinien przydzielać wysokie pozycje użytkownikiem, którzy w rzeczywistości odpowiedzieli na dane pytanie. Zatem z praktycznego punktu widzenia cenną informacją będzie ile procent użytkowników, którzy odpowiedzieli na dane pytanie, znalazło się wśród k pierwszych pozycji rankingu. Bedzie to pierwsza wykorzystywana w pracy miara. Niech d będzie indeksem pytania, r_d rankingiem wygenerowanym dla pytania d, $r_d(u)$ pozycją użytkownika u w rankingu r_d, a $ans(d)$ zbiorem użytkowników, którzy odpowiedzieli na pytanie d. Wówczas pierwszą ocenę jakości rankingu P_k definiujemy jako:

$$P_k(r_d) = \frac{|\{u:\ r_d(u) <= k\} \cap \{u : u \in ans(d)\}|}{|\{u : u \in ans(d)\}|}.$$

Podczas analizy wyników będziemy uwzględniać $k = 10, 50, 100$. Dla ustalonego k można tę miarę nazywać precyzją, ponieważ niesie ona onformację o tym, jaką część użytkowników, którzy na pewno znają odpowiedź na pytanie, udało się zawrzeć w rekomendacji. (UZASADNIĆ statystycznie, że traktując znajomość odp zero-jedynkowo, niezależnie od liczby odpowiadajcych, jeżeli każdy odpowiadający ma takie samo pstow zajęcia danej pozycji (czyli maja takie same rozklady wszysc) to w.o. tego procentu nie zalezy od liczby odpowiadających?) UWAGI? PRZYKŁAD? (do wszystkich miar?)

Miara P_k jest oceną pojedynczego rankingu. Ocena jakości działania systemu na podstawie całego zbioru testowego, będzie opierać się na rozkładzie jej wartości na przestrzeni wszystkich pytań - analizowane będą kwantyle rzędów $10, 25, 50$, a ponadto zostanie uwzględniona również ocena graficzna.

Przechodzimy do drugiej miary. Zauważmy najpierw, że praktycznego punktu widzenia ważne jest, aby pytanie trafiło do przynajmniej jednego użytkownika, który zna na nie odpowiedź, ponieważ pierwsze rozwiązanie jest często wystarczające dla pytającego. Tym razem załóżmy, że liczba użytkowników, którym rekomendowane są poszczególne pytania nie jest stała i z góry jej nie znamy, ale ponownie użytkownicy są wybierani z czołówki rankingu. W tym przypadku słusznym wydaje się ocenianie rankingu na podstawie najlepszej z pozycji zajętych przez użytkowników, którzy odpowiedzieli na dane pytanie. Wówczas najlepszy będzie ten ranking, w którym odpowiadający użtytkownik z najlepszą pozycją będzie wyżej niż w innych rankingach. Zatem formalnie definiujemy te ocenę rankingu jako:

$$\mathcal{B}(r_d) = \min\{r_d(u) : u \in ans(d)\},$$

czyli jest to najlepsza z pozycji w rankingu r_d, zajętych przez użytkowników, którzy odpowiedzieli na pytanie d. Tym razem mniejsza wartość oznacza lepszy ranking.

Przykład. Załóżmy, że na pewne pytanie ze zbioru testowego odpowiedziało trzech użytkowników. W rankingu pierwszym zajęli oni pozycje $\{5, 21, 63\}$, natomiast w drugim $\{7, 15, 41\}$. Wówczas wartość $\mathcal{B}$ wynosi 5 dla pierwszego z nich oraz 7 dla drugiego. Jeżeli nie wiemy ilu pierwszym użytkownikom zostanie to pytanie zarekomendowane, to uznanie pierwszego rangingu za lepszy jest uzasadnione, ponieważ lepsza pozycja oznacza większe prawdopodobieństwo, że dany użytkownik zostanie (słusznie) uwzględniony w rekomendacji.

Miara $\mathcal{B}$ również jest oceną pojedynczego rankingu, a ocena systemu na podstawie całego zbioru testowego będzie przebiegać analogicznie jak dla miary P_k.

Zauważmy, że konstrukcja opisanych miar nie uwzględnia wpominanej kwestii braku pełnej informacji. Skutkuje to brakiem możliwości obiektywnej oceny działania systemu, ponieważ nie da się stwierdzić jak powinny rozkładać się ich wartości w przypadku, gdyby system działał doskonale. Przykładowo jeśli wartość miar $\mathcal{B}$ dla danego pytania wynosi 5, to nie można obiektywnie ocenić czy system zadziałał dobrze czy źle, bowiem nie wiemy czy użytkownicy zajmujący wyższe pozycje w tym rankingu znają odpowiedź na pytanie (to by znaczyło, że system zadziałał poprawnie) czy też nie. Zatem miary te można jedynie wykorzystywać do porównywania systemów, przyjmując, że lepszy będzie ten, który na zbiorze testowym daje w pewnym sensie lepsze wartości tych miar.

Trzecia miara uwzględni fakt posiadanie niepełnej informacji, a ponadto przy pewnych założeniach będziemy wiedzieli jak w przybliżeniu powinny się zachowywać jej wartości. Nie będziemy w stanie formalnie zweryfikować hipotezy o poprawnym działaniu systemu, ale jeżeli będzie działał dobrze to będzie dało się to zauważyć. Miara ta reprezentuje zupełnie inne podejście - ocena systemu nie będzie opierała się na ocenach poszczególnych rankingów, lecz na ocenie rekomendacji z perspektywy poszczególnych użytkowników. Załóżmy, że każdy użytkownik u potrafi odpowiedzieć na pytanie d z prawdopodobieństwem $p_u(d)$. Załóżmy rówież, że jeżeli u odpowiedział na pytanie d to przyjmujemy, że $p_u(d) = 1$, natomiast gdy nie odpowiedział to wartości tej nie znamy, ale przyjmujemy, że $p_u(d) < 1$. Przy takich założeniach zdefiniujemy trzecią miarę, ale zaczniemy przedstawienia konsekwencji przyjęcia takich założeń na przykładzie. Niech X jest użytkownikiem, który odpowiedział na 10 pytań: $d_1, ..., d_{10}$, a Y odpowiedział na pytania $d_1, ..., d_9$. Zauważmy, że wartość oczekiwana liczby pytań, na które użytkownik Y zna odpowiedź (spośród rozważanych dziesięci) wynosi

$9 + p_Y(d_{10})$, ponieważ już wiemy, że na 9 pierwszych zna odpowiedź. Co ważne, wartość ta jest mniejsza niż 10, co wynika z założenia, że jeżeli ktoś nie udzielił odpowiedzi na pytanie, to zna na nie odpowiedź z prawdopodobieństwem mniejszym od 1. Użytkownik X ma wartość oczekiwana liczby pytań (spośród $d_1, ..., d_{10}$), na które zna odpowiedź wynosi 10, ponieważ potwierdził to udzielając dziesięciu odpowiedzi.

Generowane przez system rankingi są tworzone na podstawie wiedzy w poszczególnych tematach. Naturalnym wydaje się być założenie, że prawdopodobieństwo znajomości odpowiedzi na pytania zależy monotonicznie od wiedzy w tym obszarze. Zatem jeśli system poprawnie ocenia wiedzę użytkowników, to generowane rankingi powinny odzwierciedlać również porządek prawdopodobieństwa znajomosci odpowiedzi przez poszczególnych użytkowników. Idąc dalej, jeśli ranking dla danego pytania odzwierciedla prawdopodobieństwo znajomości odpowiedzi na nie, to na ustalonym zbiorze pytań, średnia pozycja w rankingach dla nich wygenerowanych powinna odzwierciedlać w pewnym przybliżeniu oczekiwaną wartość liczby pytań, na które użytkownicy znają odpowiedź. Oczywiście jest to duże uproszczenie, ponieważ nie są znane dokładne wartości prawdopodobieństw, lecz jedynie ich pozycje w rankingu. Ważne jest jednak to, że dokładność tego przybliżenia nie jest bardzo istotna, ponieważ kluczowy jest tutaj następujący fakt: dla dowolnego użytkownika innego niż X, który nie odpowiedział na wszystkie pytania, na które odpowiedział X, wartość oczekiwana liczby pytań, na które zna on odpowiedź, przy poczynionych założeniach jest mniejsza niż 10. To oznacza, że średnia pozycja użytkownika X w rankingach dla pytań $d_1, ..., d_{10}$, czyli tych na które odpowiedział, powinna być najlepsza spośród średnich pozycji poszczególnych użytkowników. Innymi słowy, jeśli średnia X będzie najlepsza, można to uznać za dowód na poprawne działanie systemu (oczywiście w odniesieniu do pytań $d_1, ..., d_{10}$), a jeśli jest w czołówce najlepszych średnich, to ocena jest dyskusyjna. Natomiast gdy średnia ta będzie gorsza od średnich wielu innych użytkowników, to można uznać, że system działa źle. Zatem definiujemy trzecią miarę, tym razem oceniającą rankingi z perspektywy użytkowników, jako:

$$\mathcal{M}(u) = \frac{\sum\limits_{d\ :\ u \in ans(d)} r_d(u)}{|\{d : u \in ans(d)\}|},$$

czyli jest to średnia pozycja użytkownika u w rankingach dla pytań, na które u odpowiedział. Przy tak ogólnych założeniach nie jesteś w stanie ocenić istotności wyników, ale miara ta daje szansę na spostrzeżenie obiektywnie dobrego działania systemu. Żeby taka miara miała sens, należy uwzględnić w zbiorze testowym tylko użytkowników, którzy odpowiedzieli na większą liczbe pytań. W analizach rozważymy uzwględnianie dwie werscje - użytkowników z przynajmniej dziesięcioma lub pięcioma odpowiedziami. Całościowa ocena systemu będzie opierała się ponownie na analizie rozkładu wartości miary dla poszczególnych użytkowników na zbiorze testowym.

Cele badania

Przypomnijmy najpierw, że wiedza użytkowników w poszczególnych tematach jest sumą punktów wiedzy k_{dj}, obliczanych według formuły $k_{dj} = w_{dj} \cdot p_d$, gdzie w_{dj} jest wagą tematu j w odpowiedzi d, a p_d oznacza miarę jakości odpowiedzi d. Przedmiotem badań jest wpływ na jakość działania systemu następujących czynników:

- Zakładana liczba tematów w modelu.
- Definicja wag tematów w_{dj}, określanych na podstawie ich wyestymowanych rozkładów.

- Definicja miar jakości odpowiedzi p_d, wykorzystywanych w gromadzenu wiedzy.

trzy akapity z opisem propozycji

ZROBIC DRUGĄ WERSJĘ REKOMENDACJI PO TAGACH ZEBY BYLO DZIELENIE PUNKTOW PRZEZ LICZBE TAGOW? (ZEBY W TEN SPOSÓB UWZGLEDNIC UDZIAŁY TAGÓW TAK JAK TEMATÓW W WAZONEJ WERSJI) Tu opis miar (chyba będą trzy różne miary)

Następnie opis punktu odniesienia - rekomendacja po tagach

parametryzacja dopasowania modelu

analiza wyników -tabele/wykresy

Pytanie: jaki układ opisu miar i analizy wyników - czy tak jak powyżej (najpierw opisacwszystkie miary), czy parami - opis miary,analiza wyników przy jej uzyciu, opis miary, analiza...