ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN TIN HỌC

TIỂU LUẬN
SEMINAR KHOA HỌC DỮ LIỆU

MÔ HÌNH CHỦ ĐỀ TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GV hướng dẫn: Ngô Minh Mẫn
SV thực hiện: Nguyễn Hoàng Nguyên – 19110132

THÁNG 1 NĂM 2023, THÀNH PHỐ HỒ CHÍ MINH

Tóm tắt nội dung

Ở bài báo cáo lần này em tập trung vào khái niệm Mô hình chủ đề, liệt kê một số vấn đề mà ta có thể áp dụng để giải quyết trong lĩnh vực Xử lý ngôn ngữ tự nhiên. Em cũng có đề cập đến Mô hình Chủ đề truyền thống là Phân bố Dirichlet tiềm ẩn, triển khai ý tưởng từ bài báo gốc công bố bởi Blei D. và các cộng sự vào năm 2003, đồng thời nêu một số phương pháp để giải quyết một số vấn đề gặp phải với mô hình ban đầu. Cuối cùng em có nhắc đến hai phương pháp để đánh giá chất lượng một chủ đề có được khi sử dụng mô hình và thảo luận thêm về sự phát triển của mô hình chủ đề trong những năm gần đây.

1 Giới Thiệu

Mô hình chủ đề là một trong những mô hình được sử dụng rộng rãi nhất cho việc học không giám sát các biểu diễn của văn bản, với rất nhiều những phiên bản khác nhau đáp ứng nhiều vấn đề như: Sự tương quan giữa các chủ đề trong văn bản khoa học (Mô hình chủ đề tương quan, Blei và Lafferty năm 2007), phát hiện xu hướng trong tập văn bản (Bolelli và các cộng sự, năm 2009).

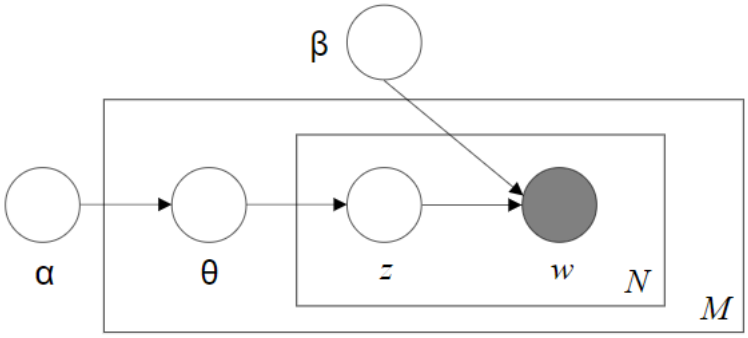
Phân bổ Dirichlet tiềm ẩn là một trong những mô hình chủ đề đầu tiên vẫn còn được sử dụng nhiều cho đến hiện nay và là ý tưởng cho các mô hình sau này với sự tinh chỉnh một số giả định trong mô hình để có được những mô hình tốt hơn theo từng tác vụ khác nhau. Vì lý do này nên em chọn chủ đề cho bài tiểu luận của mình. Khái niệm về Mô hình chủ đề được đề cập ở chương 2 Chương 2; mô hình Phân bổ Dirichlet tiềm ẩn (LDA) truyền thống (Blei và các cộng sự, nằm 2003) được trình bày ở Chương 3. Chương 4, 5, 6 nói về các phương pháp suy diễn cho phân phối hậu nghiệm của mô hình LDA để có thể sử dụng mô hình này. Chương 7 em đề cập đến 2 phương pháp dùng để đánh giá kết quả của một mô hình chủ đề và Chương 9 thảo luận thêm về hướng pháp triển gần đây của Mô hình Chủ đề.

2 Mô Hình Chủ Đề

Mô hình Chủ đề là một dạng mô hình toán học dùng để khám phá các chủ đề chính, ẩn trong các bộ dữ liệu văn bản lớn mà con người không thể tự đọc hết được. Và sau khi đã tìm ra được các chủ đề chính thì các Mô hình chủ đề có thể cấu trúc lại bộ dữ liệu theo các chủ đề đó.

Các chủ đề được đưa ra từ Mô hình chủ đề có dạng là một phân phối xác suất trên một tập từ điển cố định. Vì vậy, kết quả cuối cùng cần con người tự đặt tên cho các chủ đề được đưa ra; một vấn đề nữa là ta không có con số chính xác cho số lượng chủ đề để Mô hình cho ra kết quả phân loại tốt nhất. Số lượng chủ đề hầu hết phụ thuộc vào yêu cầu của người nghiên cứu.

Một ví dụ mà ta có thể sử dụng Mô hình Chủ đề để có được kết quả mong muốn là: Từ các bài báo điện tử được đăng trên https://vietnamnet.vn/ trong quý 4 năm 2022, ta muốn tìm ra những vấn đề đã xảy ra ở nước ta hiện nay qua 20 chủ đề tìm được qua các bài báo. Khi áp dụng Mô hình Chủ đề ta sẽ được các chủ đề ứng với từng văn bản, kết hợp với lượt xem của từng bài báo ta có thể biết được những vấn đề mà Xã hội nước ta qua tâm trong 3 tháng vừa qua.



Hình 1: *Mô hình LDA được biểu diễn dưới dạng đồ thị, được trích từ bài báo gốc của Blei và các cộng sự năm 2003.*

3 Phân Bổ Dirichlet Tiềm Ẩn (Latent Dirichlet Allocation)

Phân bổ Dirichlet tiềm ẩn (LDA) là một trong những Mô hình Chủ đề đầu tiên, cơ sở cho rất nhiều Mô hình chủ đề sau này. Ý tưởng khởi nguồn của LDA là: *Một văn bản là một hỗn hợp của các chủ đề*, với các chủ đề được định nghĩa là *phân phối các từ trên một tập từ điển cố định*. Blei và các cộng sự vào năm 2003 đã đưa ra mô hình LDA - một mô hình thống kê cho các tập văn bản có thể đáp ứng được ý tưởng trên.

Ý tưởng chính của mô hình LDA được mô tả qua quá trình tạo ra văn bản sau:

- Cho một tập văn bản D, ta có phân phối đồng thời sẽ bằng tích phân phối của các văn bản, với phân phối của từng văn bản $\boldsymbol{\theta}_d$ trên tập các chủ đề được lấy từ phân phối Dirichlet với tham số $\boldsymbol{\alpha}$.
- Với mỗi từ n, ta chọn 1 chủ đề $z_{d,n}$ từ phân phối chủ đề $\boldsymbol{\theta}_d$, và từ chủ đề này, ta chọn một từ $w_{d,n}$ từ phân phối $\boldsymbol{\beta}_{z_{d,n}}$

Tác giả cũng mô tả thuật toán trên dưới dạng đồ thị (Xem **Hình 1**). Từ thuật toán trên ta có xác suất đồng thời cho tham số chủ đề $\boldsymbol{\theta}$, tập N chủ đề $\mathbf{z}$ và tập N từ $\mathbf{w}$ được tính như sau:

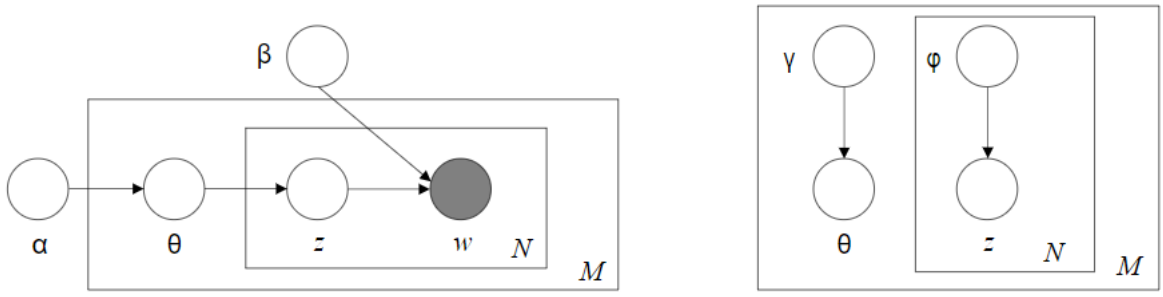
$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{k} p(\boldsymbol{\theta_d}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_{d,n}|\boldsymbol{\theta}) p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}), \tag{1}$$

với

- $p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_i - 1}$ (Phân phối Dirichlet).
- $p(z_{d,n}|\theta_d) = \theta_{d,z_{d,n}}$ (Lấy từ phân phối đa thức).
- $p(w_{d,n}|\boldsymbol{\beta}, z_{d,n}) = \beta_{z_{d,n},w_{d,n}}$ (Lấy từ phân phối đa thức).

Từ công thức trên ta có hàm hợp lý lề cho một văn bản $\mathbf{w}$ là

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta}. \tag{2}$$



Hình 2: *Bên trái: Mô hình LDA gốc, Bên phải: Mô hình LDA sau khi đã bỏ đi nốt w. Ảnh được lấy từ bài báo gốc của Blei và các cộng sự năm 2003.*

Và hàm phân phối đồng thời cho một tập văn D là:

$$p(D|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{d=1}^{M} \int \left(\prod_{n=1}^{N} \sum_{z_{d,n}=1}^{k} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|z_{d,n},\boldsymbol{\beta}) \right) p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) d\boldsymbol{\theta}_d. \tag{3}$$

Sau khi kết thúc quá trình tạo sinh, ta cần xây dựng một phương pháp suy diễn cho mô hình LDA để có thể sử dụng được mô hình. Vấn đề bây giờ là ta cần tính được phân phối hậu nghiệm của các biến $\boldsymbol{\theta}, \mathbf{z}$ là các biến ẩn được tìm thấy trong một văn bản, theo công thức sau:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}.$$

Tuy nhiên, phân phối này ta không thể có được trực tiếp thông qua tính toán vì phân phối $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ đã được tác giả Dickey chỉ ra là không thể tính được vào năm 1983. Vì thế, ta cần áp dụng một số phương pháp suy diễn để xấp xỉ phân phối hậu nghiệm trên, và phương pháp được đưa ra ở bài báo gốc là Suy diễn biến phân.

4 Suy Diễn Biến Phân Cho Mô Hình LDA

Ý tưởng cơ bản của phương pháp suy diễn cơ sở lồi là tận dụng bất đẳng thức Jensen để tìm được một chặn dưới cho hàm hợp lý logarit. Xét một họ các chặn dưới được đánh số theo các tham số biến phân, ta áp dụng các thủ tục tối ưu hóa để tìm ra các giá trị của tham số biến phân khiến chặn dưới này là chặt nhất.

Một cách đơn giản để tìm một họ các chặn dưới có thể tính toán được đó là xét một phiên bản đơn giản hơn được bỏ đi một vài cạnh hoặc nốt từ mô hình đồ thị ban đầu. Phiên bản điều chỉnh cho mô hình LDA được mô tả ở **hình 2**. Vấn đề dẫn đến việc không thể tính toán trực tiếp phân phối hậu nghiệm là vì kết nối giữa θ và β, thông qua liên kết θ, z, w. Với việc bỏ đi các cạnh và nốt đó, thêm vào 2 tham số biến phân tự do ta có được một họ phân phối trên các biến ẩn. Họ phân phối này đặc trưng bởi phân phối sau:

$$q(\boldsymbol{\theta}, \mathbf{z}) = \prod_{d=1}^{k} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{d,n}|\varphi_{d,n}), \tag{4}$$

với 2 tham số biến phân tự do $\boldsymbol{\gamma}$ là tham số Dirichlet và $\boldsymbol{\varphi}_n$ là các tham số đa thức. Sau các phép biến đổi, nhóm tác giả đưa bài toán tìm chặn dưới chặt nhất về bài toán cực

(1) initialize $\varphi_{ni}^0 := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) **repeat**
(4) **for** $n = 1$ **to** N
(5) **for** $i = 1$ **to** k
(6) $\varphi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
(7) normalize φ_n^{t+1} to sum to 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \varphi_n^{t+1}$
(9) **until** convergence

Hình 3: *Thuật toán suy diễn cho LDA. Ảnh được lấy từ bài báo gốc của Blei và các cộng sự năm 2003.*

tiểu hóa phân kỳ Kullback - Leibler sau:

$$(\boldsymbol{\gamma}*, \boldsymbol{\varphi}*) = \arg\min_{(\boldsymbol{\gamma},\boldsymbol{\varphi})} \mathrm{D}\left(q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\varphi}) \parallel p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}\boldsymbol{\beta})\right). \tag{5}$$

Với việc tính đạo hàm của Phân kỳ KL và gán bằng 0, ta thu được công thức tính cho 2 tham số trên:

$$\varphi_{ni} \propto \beta_{i,w_n} \exp\left\{\Psi(\gamma_i) - \Psi\left(\sum_i \gamma_i\right)\right\}, \tag{6}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni}, \tag{7}$$

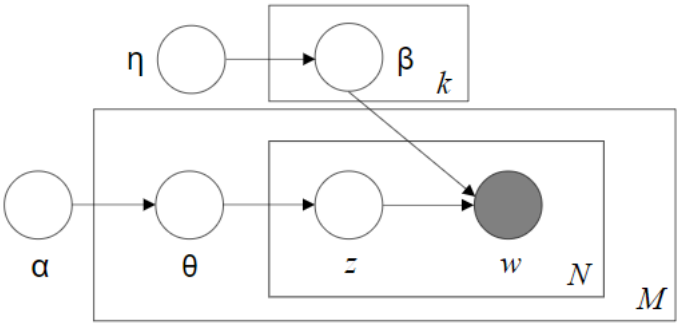
với hàm Ψ là đạo hàm bậc 1 của hàm $\log\Gamma$ được tính thông qua xấp xỉ Taylor.

Quá trình suy diễn biến phân được mô tả ở **Hình 3**. Từ thuật toán ta có thể thấy rõ mỗi lần lập lại quá trình suy diễn biến phân cho LDA ta cần $\mathrm{O}((N+1)k)$ toán tử. Tiếp theo, ta tiến hành quá trình xấp xỉ tham số trong mô hình. Với một tập văn $D = \{\mathbf{w}_1, ..., \mathbf{w}_M\}$, ta cần tìm các tham số $\boldsymbol{\alpha}, \boldsymbol{\beta}$ tối đa hóa hàm hợp lý lề của bộ dữ liệu:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{M} \log p(\mathbf{w}_d|\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{8}$$

Như đã đề cập ở trên thì hàm $p(\mathbf{w}_d|\boldsymbol{\alpha}, \boldsymbol{\beta})$ không thể tính toán được. Tuy nhiên, từ phương pháp suy diễn trên ta đã có một chặn dưới có thể tính được cho hàm hợp lí logarit, và ta có thể tiến hành xấp xỉ 2 tham số $\boldsymbol{\alpha}, \boldsymbol{\beta}$ để cực đại hóa hàm này. Vì thế ta có thể tìm được các xấp xỉ thực nghiệm cho các tham số trong mô hình LDA thông qua phương pháp cực đại hóa kỳ vọng biến phân. Với việc trước tiên cực đại hóa chặn dưới theo các tham số biến phân $\boldsymbol{\gamma}, \boldsymbol{\varphi}$, sau đó cố định hai tham số biến phân này, tiếp tục cực đại hóa chặn dưới theo các tham số của mô hình.

Các bước thực hiện phương pháp cực đại hóa kỳ vọng biến phân như sau:



Hình 4: *Biểu diễn đồ thị của mô hình LDA đã áp dụng phương pháp làm mịn. Ảnh được lấy từ bài báo gốc do Blei và các cộng sự công bố vào năm 2003.*

(1) Bước (E): Với mỗi văn bản, ta cực đại hóa các tham số biến phân $\{\boldsymbol{\gamma}_d^*, \boldsymbol{\varphi}_d^*\}$.

(2) Bước (M): Tối ưu hóa chặn dưới theo các tham số trong mô hình $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Việc này tương ứng với việc tìm các ướng lượng hợp lí cực đại với kỳ vọng thống kê đủ cho mỗi văn bản theo phân phối hậu nghiệm xấp xỉ được tính ở bước (E).

Cuối cùng ta có công thức để cập nhật biến $\boldsymbol{\beta}$ là:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi_{d,n,i}^* w_{d,n}^j. \tag{9}$$

5 Phương Pháp Làm Mịn Mô Hình LDA

Khi ta làm việc với tập văn gồm rất nhiều văn bản, một vấn đề có thể thường gặp phải là kích thước của tập từ điển sẽ rất lớn. Điều này dẫn đến việc đối với những văn bản mới có khả năng cao chứa những từ chưa từng xuất hiện trước đó trong quá trình huấn luyện. Ước lượng hợp lý cực đại của một phân phối đa thức gán xác suất xuất hiện bằng 0 cho các từ chưa từng xuất hiện, và cho cả một văn bản mới chứa nó. Hướng tiếp cận điển hình cho vấn đề này là làm mịn các tham số trong phân phối đa thức, gán xác suất dương cho tất cả các phần tử thuộc từ điển mặc dù nó đã được quan sát hay chưa trong tập huấn luyện. Phương pháp làm mịn Laplace cũng thường được sử dụng, áp dụng phương pháp này khiến trung bình của phân phối hậu nghiệm có được từ một tiên nghiệm Dirichlet đều trên các tham số đa thức.

Tuy nhiên, trong mô hình là hỗn hợp các mô hình nhỏ thì phương pháp làm mịn Laplace đơn giản không còn đáp ứng được với vai trò là phương pháp cực đại hậu nghiệm. Trong thực tế, việc thay tế tiên nghiệm Dirchlet trên các tham số đa thức ta thu được một hậu nghiệm không thể tính được với lý do tương tự như từ mô hình LDA. Giải pháp cho vấn đề này được tác giả đưa ra là áp dụng phương pháp suy diễn biến phân trên một phiên bản mở rộng của mô hình LDA bao gồm phương pháp làm mịn Dirichlet trên các tham số đa thức.

Với mô hình LDA,ta có được mô hình mở rộng dạng đồ thị ở **Hình 4**. Ta xem $\boldsymbol{\beta}$ bây giờ là một ma trận ngẫu nhiên kích thước $k \times V$ với giả định rằng mỗi hàng là độc lập được lấy từ một phân phối Dirichlet có thể hoán đổi được. Xét một phương pháp biến phân cho suy diễn Bayes, sử dụng các phân phối có thể tách được trên các biến ngẫu

nhiên $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}$:

$$q(\boldsymbol{\beta}_{1:k}, \mathbf{z}_{1:M}, \boldsymbol{\theta}_{1:M}|\boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\gamma}) = \prod_{i=1}^{k} \mathrm{Dir}(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, \mathbf{z}_d|\boldsymbol{\varphi}_d, \boldsymbol{\gamma}_d), \tag{10}$$

với $q_d(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\varphi}, \boldsymbol{\gamma})$ là phân phối biến phân được định nghĩa co mô hình LDA. Phương trình cập nhật cho 2 tham số $\boldsymbol{\gamma}, \boldsymbol{\varphi}$ tương tự như phương trình (6) và (7), và thêm một phương trình cập nhật cho tham số biến phân mới $\boldsymbol{\lambda}$:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \varphi^*_{d,n,i} w^j_{d,n}. \tag{11}$$

6 Phương pháp lấy mẫu Gibb cho mô hình LDA

Lẫy mẫu Gibb là một phương pháp đã được chứng minh có tính hiệu quả cao và đơn giản trong việc suy diễn xấp xỉ phân phối của một hậu nghiệm không thể tích được trực tiếp. Ý tưởng cơ bản của phương pháp này là khi ta gặp một phân phối khó tính hay không có thông tin về dạng đúng của phân phối đó, tuy nhiên ta lại có một biểu diễn của phân phối điều kiện thì ta có thể sử dụng Lấy mẫu Gibb để xấp xỉ cho phân phối ban đầu.

Đối với mô hình LDA, ta cần tìm được xác suất $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, và đã được đề cập ở trên về việc không thể tính được phân phối này. Nên điều ta cần quan bây giờ là tham số $\mathbf{z}$, hay cụ thể là xác suất một chủ đề được cho từ thứ i khi tất cả các từ còn lại đều đã được gán chủ đề. Mô tả vấn đề dưới dạng công thức sau:

$$p(z_i|\mathbf{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}),$$

với $\mathbf{z}_{-i}$ nghĩa là các biến chủ đề ngoại trừ z_i. Ta có khai triển công thức trên thành dạng tích của một hàm hợp lý với một tiên nghiệm như sau:

$$p(z_i|\mathbf{z}_{-i}, \mathbf{w}) \propto p(w_i|z_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}) p(z_i|\mathbf{z}_{-i}). \tag{12}$$

Các tham số $\boldsymbol{\theta}, \boldsymbol{\beta}$ không được đề cập tới trong công thức trên vì ta có thể thu được xác suất điều kiện cho chủ đề z_i mà chỉ dựa trên $\mathbf{z}_{-i}$ và $\mathbf{w}$ bằng việc lấy tích phân trên từng cụm bên phải phương trình. Với cụm thứ nhất, ta có

$$p(w_i|z_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int p(w_i|z_i, \beta_{z_i}) p(\beta_{z_i}|\mathbf{z}_{-i}\mathbf{w}_{-i}) \, d\beta_{z_i}, \tag{13}$$

và phân phối cuối cùng bên trong dấu tích phân có thể được viết lại theo quy luật Bayes

$$p(\beta_{z_i}|\mathbf{z}_{-i}\mathbf{w}_{-i}) \propto p(\mathbf{w}_{-i}|\beta_{z_i}, \mathbf{z}_{-i}) p(\beta_{z_i}). \tag{14}$$

Vì phân phối $p(\beta_{z_i})$ là Diriclet, và phân phối $p(\mathbf{w}_{-i}|\beta_{z_i}, \mathbf{z}_{-i})$ là phân phối loại, vậy nên theo tính chất phân phối tiên nghiệm liên hợp thì phân phối hậu nghiệm cũng sẽ là một phân phối Dirichlet có dạng $\mathrm{Dirichlet}(\eta + n^{(w)}_{-i,z_i})$ với η là tham số Diriclet tạo nên β, hay $\beta \sim \mathrm{Dirichlet}(\eta)$, và $n^{(w)}_{-i,z_i}$ là số lần từ w được gán cho chủ đề z_i không tính từ hiện tại. Ý nghĩa của biến $\mathbf{z}_{-i}$ trong mô hình là để phân vùng các từ thành các tập từ được gán vào

các chủ đề khác nhau, và chỉ có những từ được gán vào chủ đề z_i mới có thể ảnh hưởng đến hậu nghiệm của β_{z_i}.

Cụm đầu tiên trong tích phân ở phương trình (13) đơn giản là β_{i,z_i}, từ đây ta có thể hoàn thành bài toán tích phân và thu được

$$p(w_i|z_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{\eta + n^{(w)}_{-i,z_i}}{M\eta + n^{(\cdot)}_{-i,z_i}}, \tag{15}$$

với $n^{(\cdot)}_{-i,z_i}$ là tổng số từ được gán cho chủ đề z_i, không bao gồm từ hiện tại. Và với $p(z_j|\mathbf{z}_{-i})$, bằng kỹ thuật tương tự ta cũng thu được:

$$p(z_i|\mathbf{z}_{-i}) = \int p(z_i|\theta_d)p(\theta_d|\mathbf{z}_{-i})d\theta_d = \frac{n^d_{-i,z_i} + \alpha}{n^d_{-i} + K\alpha}, \tag{16}$$

với n^d_{-i,z_i} là số từ trong văn bản d được gán cho chủ đề z_i ngoại trừ từ hiện tại và n^d_{-i} là số lượng từ có trong văn bản d trừ đi từ hiện tại. Gộp 2 phương trình (15), (16) ta được

$$p(z_i|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{\eta + n^{(w)}_{-i,z_i}}{M\eta + n^{(\cdot)}_{-i,z_i}} \frac{n^d_{-i,z_i} + \alpha}{n^d_{-i} + K\alpha}. \tag{17}$$

Từ công thức trên ta có thể dễ dàng áp dụng Lấy mẫu Gibb. z_i được khởi tạo ngẫu nhiên trong khoảng 1 đến K, là trạng thái khởi đầu của chuỗi Markove, sau đó chuỗi Markov bắt đầu chạy theo một số lần lặp, mỗi lần lặp sẽ thêm vào một trạng thái mới bằng một chủ đề z_i được tính từ phương trình (17).

7 Các Phương Pháp Đánh Giá Mô Hình Chủ Đề

7.1 Đánh giá thông qua độ hỗn loạn (perplexity)

Các tập văn dùng để huấn luyện là chưa được gắn nhãn, vì thế muc tiêu của chúng ta là đạt được kết quả từ hàm hợp lí cao nhất có thể trên một tập kiểm thử. Để đạt được điều này ta sử dụng *độ hỗn loạn* trên một tập kiểm để đánh giá mô hình. Độ hỗn loạn được trên một tập dữ liệu được mô tả là trung bình âm của hàm hợp lí logarit trên tập dữ liệu được lỹ thừa theo hệ số e. Và theo công thức trên thì ta có được độ hỗn loạn sẽ tỷ lệ nghịch với hàm hợp lí, vậy nên mô hình có độ hỗn loạn thấp sẽ được xem là tốt hơn theo tiêu chuẩn này. Công thức chi tiết của độ hỗn loạn được tính như sau:

$$\text{perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

7.2 Đánh giá thông qua độ đồng nhất của chủ đề

Mô hình được Lau và các cộng sự giới thiệu vào năm 2014, với giả định rằng mỗi chủ đề sẽ có duy nhất một từ ngoại lai hay từ có tính đại diện thấp nhất cho chủ đề đó, ta tính toán đặc trưng thể hiện mối liên kết giữa các từ trong N từ có xác suất cao nhất trong 1 chủ đề; sau khi tính toán các đặc trưng này ta đưa vào mô hình Máy véc tơ hỗ trợ hồi

quy phân hạng để tìm ra các từ là từ ngoại lai. Ta có một số độ đo mối liên kết giữa các từ như: *PMI, CP1, CP2* (Lau và các cộng sự, năm 2010), và *NPMI* (Bouma, năm 2009). Công thức của các độ đo cụ thể như sau:

$$\begin{aligned}
\mathbf{PMI}(w_i) &= \sum_{j\neq i} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}, \\
\mathbf{CP1}(w_i) &= \sum_{j\neq i} \frac{P(w_i, w_j}{P(w_j)}, \\
\mathbf{CP2}(w_i) &= \sum_{j\neq i} \frac{P(w_i, w_j}{P(w_i)}, \\
\mathbf{NPMI}(w_i) &= \sum_{j\neq i} \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.
\end{aligned}$$

8 Kết Luận

Trong bài tiểu luận này em đã giới thiệu về Mô hình chủ đề và mô hình phân bổ Dirichlet tiềm ẩn với 2 phương pháp suy diễn cho phân phối hậu nghiệm của mô hình, và cách ta có thể đánh giá kết quả cho ra từ mô hình.

9 Thảo Luận Thêm

Khi áp dụng các phương pháp suy diễn cho các mô hình chủ đề chung, hầu hết ta sẽ cân nhắc giữa hai phương pháp là suy diễn biến phân và lẫy mẫu Gibb. Khi nói đến lấy mẫy Gibb, lợi thế của phương pháp này là cài đặt đơn giản và thời gian chạy nhanh, tuy nhiên khi hội tụ ta sẽ không thể biết được những gì đã xảy ra trước đó. Với phương pháp suy diễn biến phân, điểm cộng là ta có thể áp dụng rộng rãi, và ta có đủ tài nguyên để tính toán thì phương pháp này có thể nhanh hơn lấy mẫu Gibb nhờ khả năng có thể tính toán song song. Tuy nhiên vào năm 2014, D.P. Kingma và Welling M. đã giới thiệu mô hình Bộ tự mã hóa biến phân, là một mô hình mạng nơ ron, với hàm mục tiêu là phân kỳ Kullback - Leibler dùng để đo khoảng cách giữa phân phối biến phân và phân phối hậu nghiệm từ đó điều chỉnh các tham số trong mạng nơ ron để cho ra phân phối xấp xỉ tốt nhất.

Một điều nữa ta cần đề cập về các mô hình chủ đề truyền thống là cách mà mô hình truyền thống hâu như chỉ quan tâm đến các giá trị thống kê liên quan đến sự suất hiện của từ đó tuy nhiên lại bỏ qua mỗi liên kết về mặt ngữ nghĩa giữa các từ trong cùng một đoạn văn bản. Với từ "ngủ" trong câu "Tôi đi ngủ lúc 9 giờ" và "Tôi ngủ dậy lúc 5 giờ" mang nghĩa trái ngược nhau hoàn toàn, và nếu ta xem xét các từ này theo nghĩa đơn lẻ từ mô hình BoW thì gần như ta sẽ mất đi thông tin về mặt ngữ nghĩa giữa các từ. Vào năm 2017, mô hình Transfomrer được Vaswani và các cộng sự giới thiệu vào năm 2017 đã mở ra hướng giải quyết cho vấn đề nêu trên, và đến năm 2019, Delvin và các cộng sự tại Google đã cho ra mô hình BERT, với cơ chế *tự chú ý* và lượng dữ liệu khổng lồ từ Google đã có thể học được các thông tin về mặt ngữ nghĩa và hiểu được khi ta nhắc đến từ "ngủ" ở câu đầu tiên thì điều ta cần chú ý là từ "đi" hay ở câu thứ 2 là từ "dậy". Từ

đó mở ra một hướng mới cho các mô hình chủ đề, được gọi là Mô hình chủ đề được bối cảnh hóa.

10 Tài Liệu Tham Khảo

1. Dickey, J.M. 1983. Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. *Journal of the American Statistical Association*, 78, trang 628-637.

2. Blei D. và các cộng sự. 2003. Latent Dirichlet Allocation. *Journals of Machine Learning Research 3 (January 2003)*, trang 993 - 1022.

3. David M. Blei và John D. Lafferty. 2007. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35.

4. Bolelli, S. Ertekin, và C.L. Giles. 2009. Topicand trend detection in text collections using Latent Dirichlet Allocation. In *Proceedings of ECIR* 2009, trang 776–780, Toulouse, France.

5. Bouma. 2009.Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, trag 31–40, Potsdam, Germany.

6. J.H. Lau và các cộng sự. 2010. Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Posters Volume, trag 605-613, Beijing, China.

7. J.H. Lau và các cộng sự. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, trang 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

8. Diederik P Kingma và Max Welling. 2014. Auto-encoding Variational Bayes. *The International Conference on Learning Representation (ICLR), Banff.*

9. Vaswani A. và các cộng sự. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, trang 6000 - 6010.

10. Delvin và các cộng sự. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *arXiv preprint arXiv:1810.04805.*