GRAPHICAL MODEL : LYRIC-BASED GENRE CLASSIFICATION

DAVID VAN ERKELENS (10264019)
SHARON GIESKE (6167667)
ELISE KOSTER (5982448)

1. Topic distribution over genres using LDA

To model topic distributions over genres, the dataset is split into the individual genres. Then, for each genre topic distributions per document are averaged, to provide a final topic distribution over genres.

2. Plate Diagram

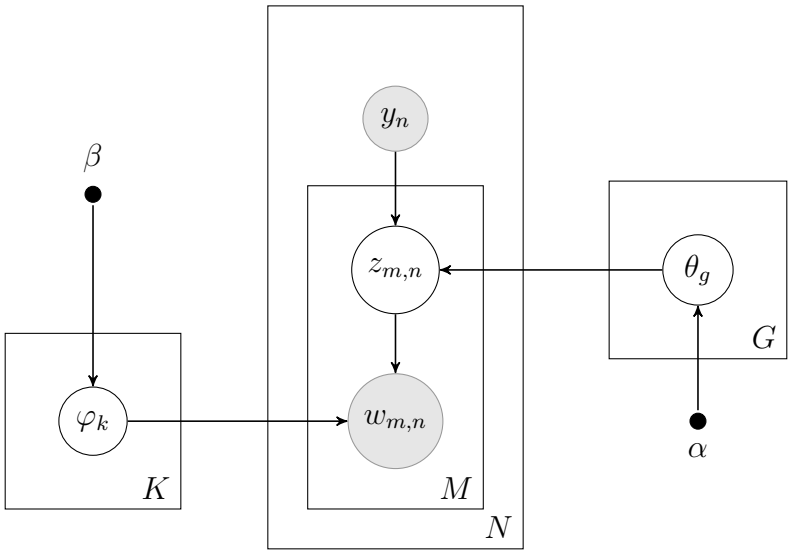

Figure 1. Graphical model

SYMBOL	DESCRIPTION
K	# of topics
G	# of genres
N	# of documents
M_i	# of words in document i
y_i	genre of document i
$w_{i,j}$	word in document i at position j
$z_{i,j}$	topic assignment for word $w_{i,j}$
θ_g	probabilities of topics given genre g
φ_k	probability of words given topic k
α	Dirichlet prior for genre distribution
β	Dirichlet prior for topic distribution

3. Derivation

$$P(w, z, \varphi, \theta|\alpha, \beta, y) = P(\varphi|\beta)P(\theta|\alpha)P(z|\theta, y)P(w|\varphi, z) \tag{1}$$

Separate probabilities:

$$P(\varphi|\beta) = \prod_{k=1}^{K} P(\varphi_k|\beta) \tag{2}$$

$$P(\theta|\alpha) = \prod_{g=1}^{G} P(\theta_g|\alpha) \tag{3}$$

$$P(z|\theta, y) = \prod_{i=1}^{N}\prod_{j=1}^{M_i} P(z_{i,j}|\theta_{y_i}) \tag{4}$$

$$P(w|\varphi, z) = \prod_{i=1}^{N}\prod_{j=1}^{M_i} P(w_{i,j}|\varphi_{z_{i,j}}) \tag{5}$$

Joined together:

$$P(w, z, \varphi, \theta|\alpha, \beta, y) = \prod_{k=1}^{K} P(\varphi_k|\beta)\prod_{g=1}^{G} P(\theta_g|\alpha)\prod_{i=1}^{N}\prod_{j=1}^{M_i} P(z_{i,j}|\theta_{y_i})P(w_{i,j}|\varphi_{z_{i,j}}) \tag{6}$$

Integrate over joint:

$$\int\int P(w, z, \varphi, \theta|\alpha, \beta, y)d\varphi d\theta = \int \prod_{k=1}^{K} P(\varphi_k|\beta)\prod_{i=1}^{N}\prod_{j=1}^{M_i} P(w_{i,j}|\varphi_{z_{i,j}})d\varphi \tag{7}$$

$$\times \int \prod_{g=1}^{G} P(\theta_g|\alpha)\prod_{i=1}^{N}\prod_{j=1}^{M_i} P(z_{i,j}|\theta_{y_i})d\theta \tag{8}$$

3.1. Derivation of $\int \prod_{k=1}^{K} P(\varphi_k|\beta)\prod_{i=1}^{N}\prod_{j=1}^{M_i} P(w_{i,j}|\varphi_{z_{i,j}})d\varphi$.

Use $C(k, w)$ as the number of times word w is assigned to topic k in any document.

$$\prod_{k=1}^{K} P(\varphi_k|\beta)\prod_{i=1}^{N}\prod_{j=1}^{M_i} P(w_{i,j}|\varphi_{z_{i,j}})d\varphi \tag{9}$$

$$= \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{w=1}^{V} \beta)}{\prod_{w=1}^{V} \Gamma(\beta)} \prod_{w=1}^{V} \varphi_k(w)^{\beta-1} \prod_{i=1}^{N} \prod_{j=1}^{M_i} \varphi_{z_{i,j}}(w_{i,j})) d\varphi \tag{10}$$

$$= \prod_{k=1}^{K} \int \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V} \prod_{w=1}^{V} \varphi_k(w)^{\beta-1} \prod_{w=1}^{V} \varphi_k(w)^{C(k,w)} d\varphi \tag{11}$$

$$= \prod_{k=1}^{K} \int \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V} \prod_{w=1}^{V} \varphi_k(w)^{\beta+C(k,w)-1} d\varphi \tag{12}$$

$$= \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V} \frac{\prod_{w=1}^{V} \Gamma(C(k,w)-1)}{\Gamma(\sum_{w=1}^{V} C(k,w)-1))} \int \frac{\Gamma(\sum_{w=1}^{V} C(k,w)-1))}{\prod_{w=1}^{V} \Gamma(C(k,w)-1} \prod_{w=1}^{V} \varphi_k(w)^{\beta+C(k,w)-1} d\varphi \tag{13}$$

$$= \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V} \frac{\prod_{w=1}^{V} \Gamma(C(k,w)-1)}{\Gamma(\sum_{w=1}^{V} C(k,w)-1))} \tag{14}$$

$$\propto \prod_{k=1}^{K} \frac{\prod_{w=1}^{V} \Gamma(C(k,w)-1)}{\Gamma(\sum_{w=1}^{V} C(k,w)-1))} \tag{15}$$

3.2. **Derivation of** $\prod_{g=1}^{G} P(\theta_g|\alpha) \prod_{i=1}^{N} \prod_{j=1}^{M_i} P(z_{i,j}|\theta_{y_i}) d\theta$**.**

Use $C(g,k)$ as the number of times topic k is assigned to genre g in any document.

$$\prod_{g=1}^{G} P(\theta_g|\alpha) \prod_{i=1}^{N} \prod_{j=1}^{M_i} P(z_{i,j}|\theta_{y_i}) d\theta \tag{16}$$

$$= \prod_{g=1}^{G} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha)}{\prod_{k=1}^{K} \Gamma(\alpha)} \prod_{k=1}^{K} \theta_g(k)^{\alpha-1} \prod_{i=1}^{N} \prod_{j=1}^{M_i} \theta_{y_i}(k_{i,j}))d\theta \tag{17}$$

$$= \prod_{g=1}^{G} \int \frac{\Gamma(K\alpha)}{(\Gamma(\alpha))^K} \prod_{k=1}^{K} \theta_g(k)^{\alpha-1} \prod_{k=1}^{K} \theta_g(k)^{C(g,k)} d\theta \tag{18}$$

$$= \prod_{g=1}^{G} \int \frac{\Gamma(K\alpha)}{(\Gamma(\alpha))^K} \prod_{k=1}^{K} \varphi_k(w)^{\alpha+C(g,k)-1} d\theta \tag{19}$$

$$= \prod_{g=1}^{G} \frac{\Gamma(K\alpha)}{(\Gamma(\alpha))^K} \frac{\prod_{k=1}^{K} \Gamma(C(g,k)-1)}{\Gamma(\sum_{k=1}^{K} C(g,k)-1))} \int \frac{\Gamma(\sum_{k=1}^{K} C(g,k)-1))}{\prod_{k=1}^{K} \Gamma(C(g,k)-1} \prod_{k=1}^{K} \varphi_k(w)^{\alpha+C(g,k)-1} d\theta \tag{20}$$

$$= \prod_{g=1}^{G} \frac{\Gamma(K\alpha)}{(\Gamma(\alpha))^K} \frac{\prod_{k=1}^{K} \Gamma(C(g,k)-1)}{\Gamma(\sum_{k=1}^{K} C(g,k)-1))} \tag{21}$$

$$\propto \prod_{g=1}^{G} \frac{\prod_{k=1}^{K} \Gamma(C(g,k)-1)}{\Gamma(\sum_{k=1}^{K} C(g,k)-1))} \tag{22}$$

3.3.

Use $\neg w_{i,j}$ as collection that not uses word from document i at position j. Use $\widetilde{C}$ as a count that does not include word from document i at position j.

$$P(z_{i,j} = k | Z_{\neg w_{i,j}}, \alpha, \beta, W, Y) \propto \frac{\beta + \widetilde{C}(k, w_{i,j})}{V\beta + \widetilde{C}(k)} \times \frac{\alpha + \widetilde{C}(y_i, k)}{K\alpha + \widetilde{C}(y_i)} \tag{23}$$