

The top banner features the Fudan University logo on the left, which includes the university's name in English ('FUDAN UNIVERSITY') and Chinese ('復旦大學') around a central emblem. To the right of the logo are several blue gears of varying sizes. Some gears contain white icons: a building, a graduation cap, a medical cross, a person silhouette, and an atomic symbol. The background of the banner is light blue with a dark blue curved shape on the right side.

Big Data Analytics & Applications

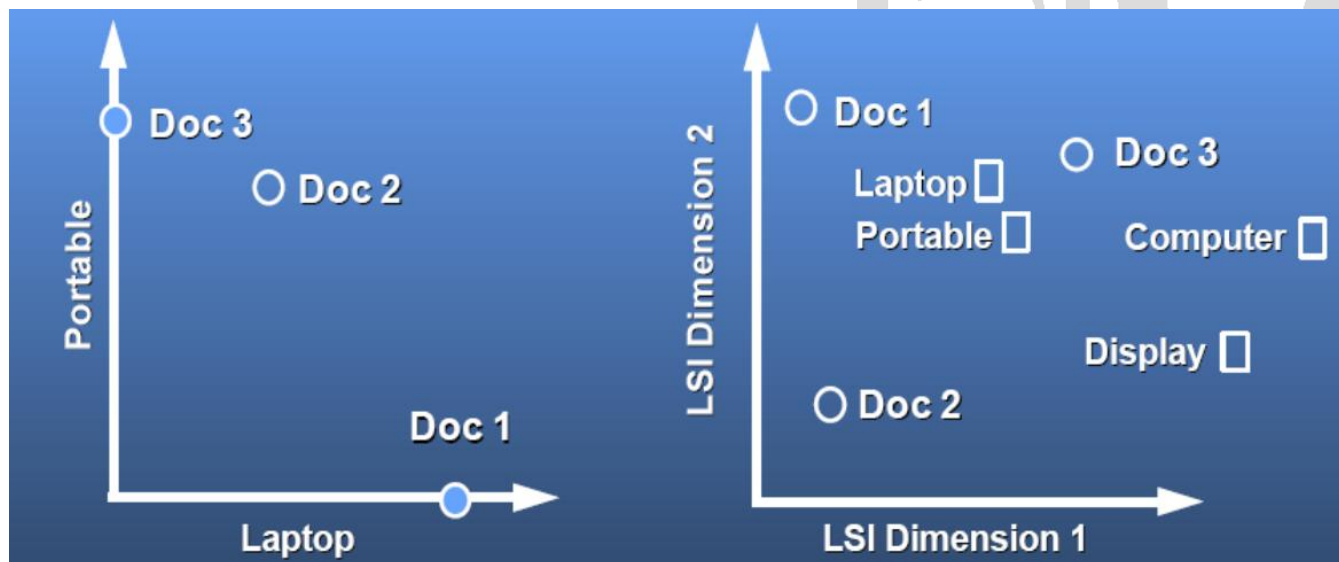
Bin Li

School of Computer Science

Fudan University

Latent Semantic Analysis

- After applying SVD to the word-document co-occurrence matrix and obtain the factorization $A = USV^T$
 - ▣ U : similar words have large inner products
 - ▣ V : similar documents have large inner products
 - ▣ Related word and document have large inner products



sis

sition (SVD) to find

nce matrix

representing words

representing documents

rix

approximation of A

x		Vt				
	f_4		d_1	d_2	d_3	d_4
3	0	f_1	0.37	0.38	0.65	0.53
0	0	f_2	-0.55	-0.63	0.37	0.38
5	0	f_3	-0.69	0.59	0.27	-0.21
0	1.5	f_4	0.26	-0.29	0.59	-0.69

- ❑ A : $m \times n$ word-document co-occurrence matrix
- ❑ U : $m \times k$ orthogonal matrices for representing words
- ❑ V : $n \times k$ orthogonal matrices for representing documents
- ❑ S : $k \times k$ diagonal singular value matrix
- ❑ Select $k' \ll n, k' \ll m$ for a low-rank approximation of A

A = U x S x Vt

	d1	d2	d3	d4
a	6	7	1	0
b	8	6	0	1
c	6	9	8	5
d	0	1	8	8
e	2	0	9	7
f	2	0	7	7

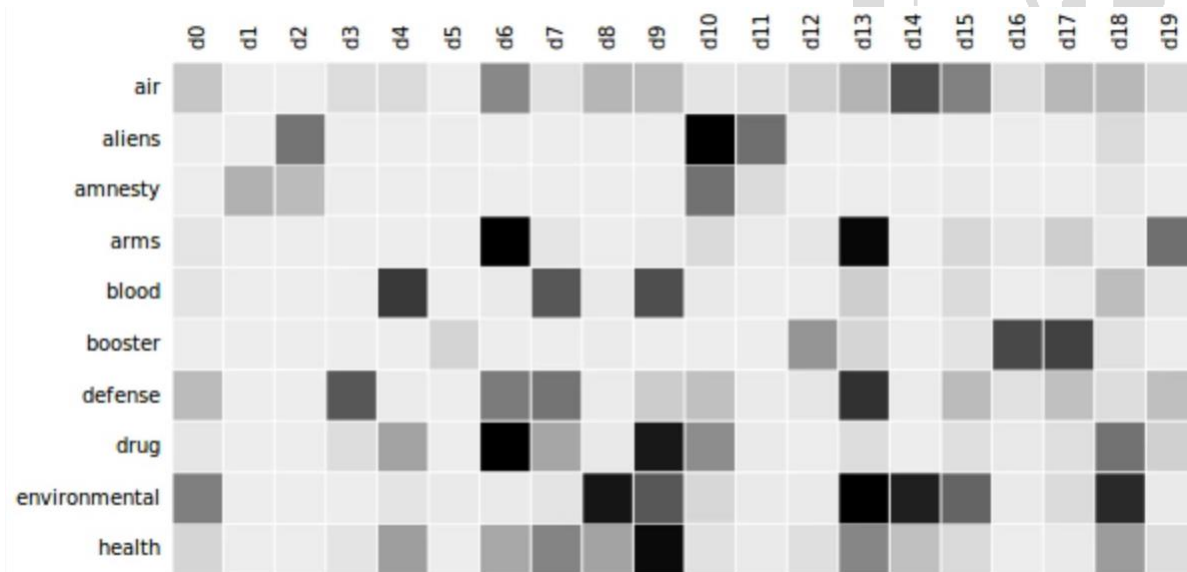
	f1	f2	f3	f4
a	0.24	-0.51	0.08	0.06
b	0.25	-0.54	-0.64	-0.23
c	0.58	-0.28	0.57	0.13
d	0.42	0.37	0.16	-0.68
e	0.44	0.34	-0.24	0.66
f	0.39	0.29	-0.40	-0.09

	f1	f2	f3	f4
f1	23.1	0	0	0
f2	0	14.3	0	0
f3	0	0	3.5	0
f4	0	0	0	1.5

	d1	d2	d3	d4
f1	0.37	0.38	0.65	0.53
f2	-0.55	-0.63	0.37	0.38
f3	-0.69	0.59	0.27	-0.21
f4	0.26	-0.29	0.59	-0.69

Probabilistic LSA

- Probabilistic LSA (PLSA) is a statistical technique for the analysis of co-occurrence matrix.
- Compared to standard LSA stemming from a **low-rank decomposition** (SVD), PLSA is based on a **mixture decomposition** derived from a latent class model



PLSA Model

- Observations in the form of co-occurrences (w, d) of words and documents
- PLSA models the probability of (w, d) as a mixture of conditionally independent multinomial distributions

$$p(w, d) = \sum_z p(z)p(d|z)p(w|z) = p(d) \sum_z p(w|z)p(z|d)$$

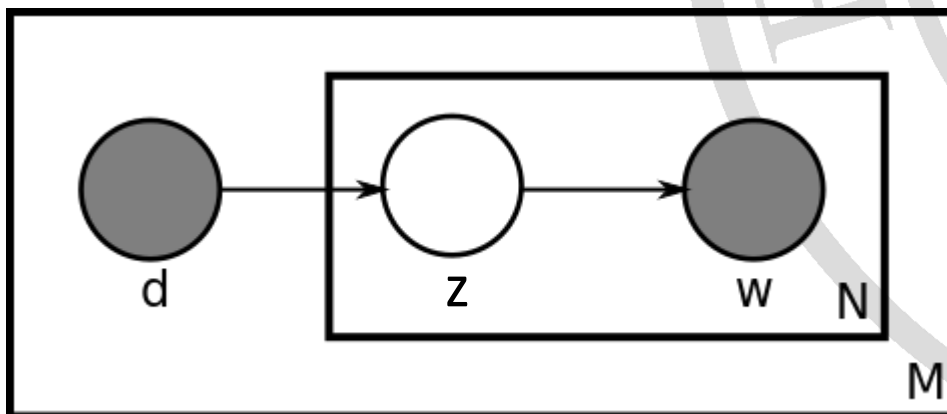
- An advantage of PLSA is that the latent variable z can be interpreted as a topic
 - z : topic (latent class)
 - $p(z|d)$: each document has a distribution over K latent topics
 - $p(w|z)$: each topic has a distribution over the vocabulary

PLSA Model

- PLSA is a generative model of the documents in the collection it is estimated on

$$p(w, d) = p(d) \sum_z p(w|z)p(z|d)$$

- For each document d , a topic z is generated conditionally to d according to $p(z|d)$
- A word is then generated from topic z according to $P(w|z)$



EM Algorithm for Latent Variable Models

- Given a joint distribution $p(X, Z|\theta)$ over observed variables X and latent variables Z , governed by parameters θ , the goal is to maximize the likelihood function $p(X|\theta)$ w.r.t. θ .
- The general EM algorithm:
 - ▣ Initialize the parameters θ^{old} ;
 - ▣ E-Step: Evaluate $p(Z|X, \theta^{\text{old}})$;
 - ▣ M-Step: Evaluate θ^{new} given by

$$\theta^{\text{new}} = \operatorname{argmax}_{\theta} \sum_Z p(Z|X, \theta^{\text{old}}) p(X, Z|\theta)$$

- ▣ Check the convergence of the parameter values; if not convergence condition not satisfied set $\theta^{\text{old}} = \theta^{\text{new}}$ and go to E-step.

Learning for PLSA

- The parameters $p(z|d)$ and $p(w|z)$ of PLSA can be learned by using the EM algorithm
- EM algorithm for PLSA:
 - E-Step: Evaluate $p(z_k | d_i, w_j; \theta^{\text{old}})$;

$$p(z_k | d_i, w_j; \theta^{\text{old}}) = \frac{p(w_j | z_k) p(z_k | d_i)}{\sum_{l=1}^K p(w_j | z_l) p(z_l | d_i)}$$

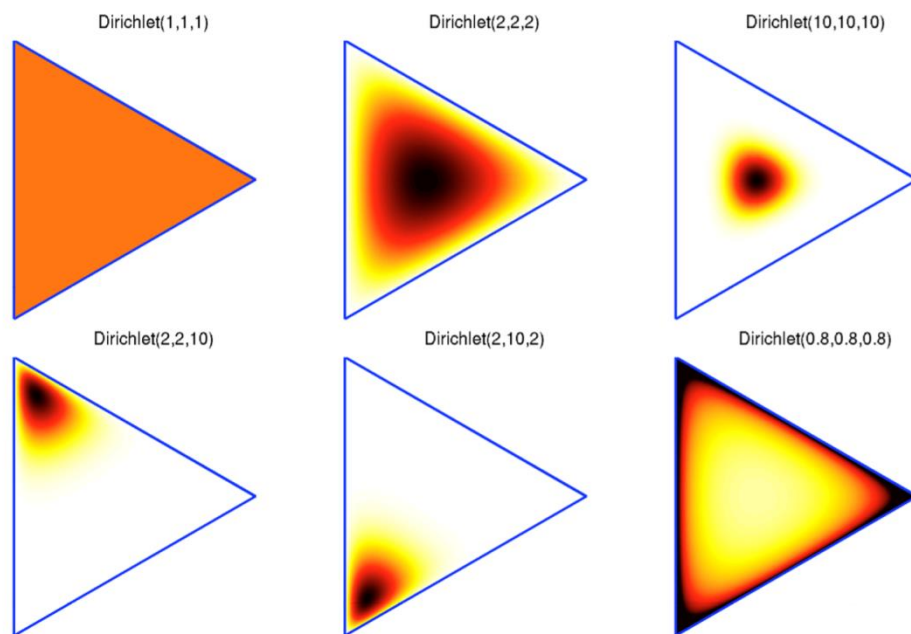
- M-Step: Evaluate θ^{new} given by

$$p(w_j | z_k) = \frac{\sum_{i=1}^M p(d_i, w_j) p(z_k | d_i, w_j)}{\sum_{n=1}^N \sum_{m=1}^M p(d_m, w_n) p(z_k | d_m, w_n)} = \frac{\sum_{i=1}^M \#(d_i, w_j) p(z_k | d_i, w_j)}{\sum_{n=1}^N \sum_{m=1}^M \#(d_m, w_n) p(z_k | d_m, w_n)}$$

$$p(z_k | d_i) = \sum_{j=1}^N p(w_j | d_i) p(z_k | d_i, w_j) = \frac{\sum_{j=1}^N \#(d_i, w_j) p(z_k | d_i, w_j)}{\#(d_i)}$$

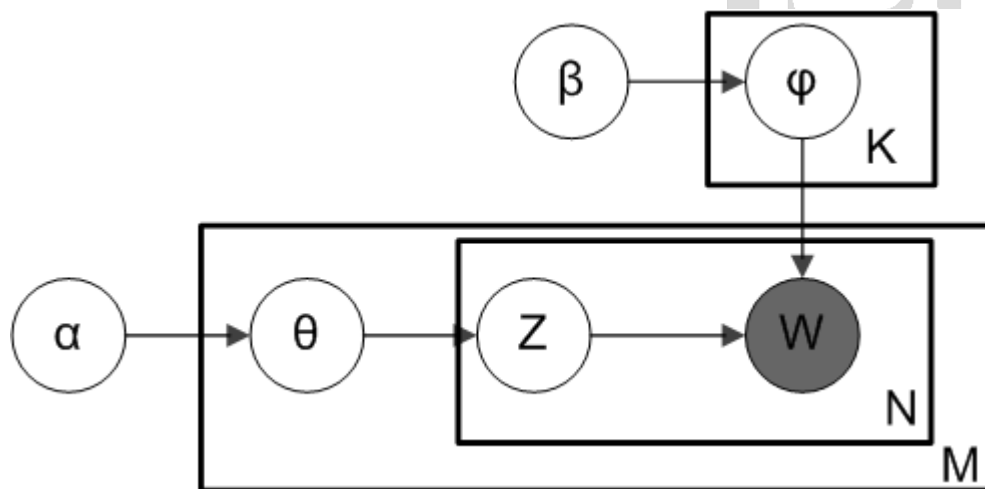
Latent Dirichlet Allocation

- PLSA is not a generative model of new documents
- LDA is identical to PLSA except that in LDA
 - Document-topic distribution \leftarrow sparse Dirichlet prior
 - Topic-word distribution \leftarrow sparse Dirichlet prior



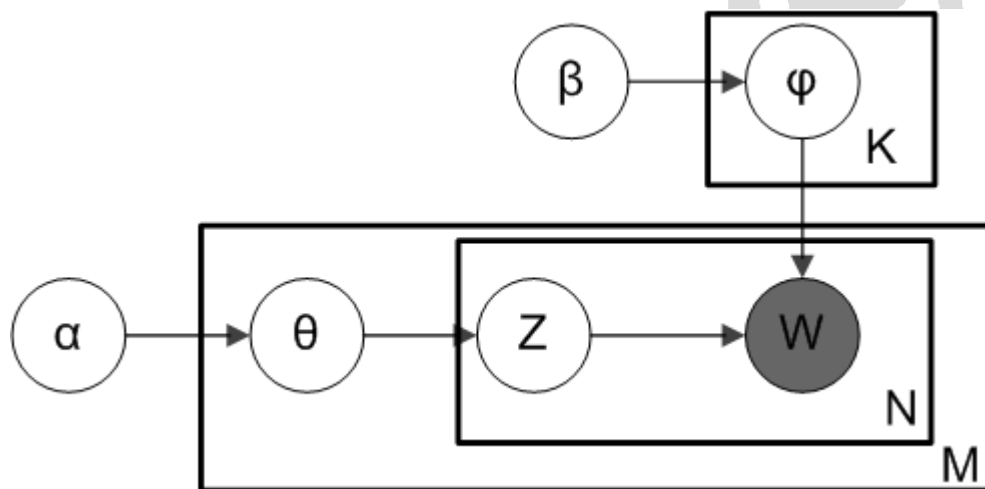
Latent Dirichlet Allocation

- θ and φ are matrices through decomposing the document-word co-occurrence matrix
 - ▣ θ : $N \times K$ matrix for distribution for a document over topics
 - ▣ φ : $K \times M$ matrix for distribution for a topic over words
 - ▣ α and β are fixed hyper-parameters
 - ▣ z is latent variable



Generative Process of LDA

- Sample $\theta_i \sim \text{Dir}(\alpha)$ for each document (typically $\alpha < 1$)
- Sample $\varphi_k \sim \text{Dir}(\beta)$ for each topic (typically $\beta < 1$)
- For each of the word positions (i, j)
 - ▣ Sample a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$
 - ▣ Sample a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$



Properties of Dirichlet

- Dirichlet distribution is the **conjugate prior** of the multinomial distribution

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_1 + \dots + \alpha_K)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k-1}$$

$$Multi(\mathbf{m}_1, \dots, \mathbf{m}_K | \mu, N) = \frac{n!}{m_1! \dots m_K!} \prod_{k=1}^K \mu_k^{m_k}$$

$$\begin{aligned} p(\mu|D, \alpha) &= Dir(\mu|\alpha + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_1 + \dots + \alpha_K + N)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_K + m_K)} \prod_{k=1}^K \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

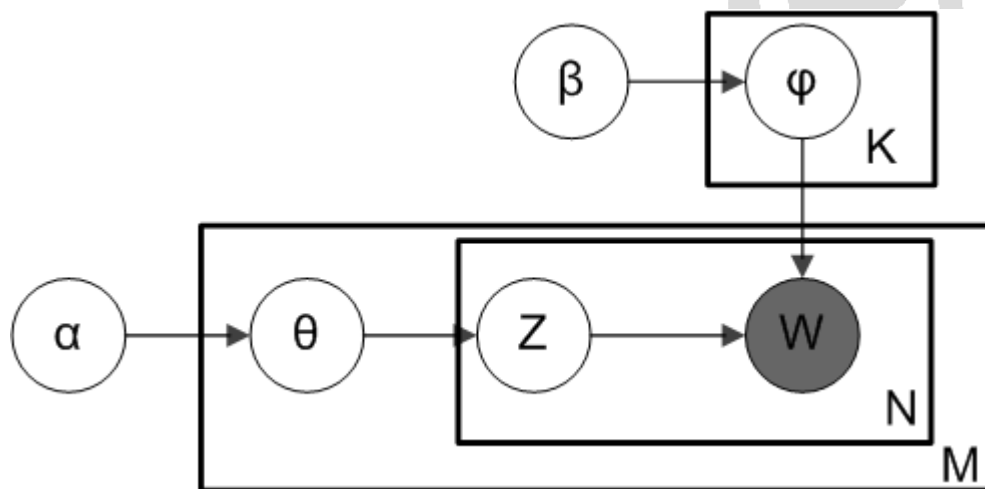
- The expectation of Dirichlet is $E(\mu_k) = \frac{\alpha_k}{\alpha_1 + \dots + \alpha_K}$

Gibbs Sampling

- Suppose we want to obtain k samples of (x_1, \dots, x_n) from a joint distribution $p(x_1, \dots, x_n)$, we can sample x_i in order in each iteration $p(x_i^{(t+1)} | x_1^{(t+1)}, \dots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \dots, x_n^{(t)})$
- Gibbs sampling
 - Sample x_1 conditioned on x_2, x_3
 - Sample x_2 conditioned on x_1, x_3
 - Sample x_3 conditioned on x_1, x_2
- Collapsed Gibbs sampling
 - Sample x_1 conditioned on x_3
 - Sample x_3 conditioned on x_1
 - x_2 is **collapsed** out during the sampling process

Collapsed Gibbs Sampling for LDA

- In the generative process of LDA, latent variable $z_{d,j}$ is used to choose a topic for the j th word of the d th document
 - ▣ If we know $\{z_{d,j}\}$ it is easy to estimate $\{\theta_d\}$ and $\{\varphi_k\}$
 - ▣ If we can integrate out $\{\theta_d\}$ and $\{\varphi_k\}$ the problem can be simplified to estimate $p(z_{d,j} = k | \{z_{-(d,j)}\}, \{w_{d,j}\})$



Collapsed Gibbs Sampling for LDA

- By integrating out $\{\theta_d\}$ and $\{\varphi_k\}$ the Gibbs sampling procedure boils down to estimate

$$p(z_i = k | \{z_{-i}\}, \{w_i\}) = \frac{p(z_i = k, \{z_{-i}\}, \{w_i\})}{p(\{z_{-i}\}, \{w_i\})}$$

$$\begin{aligned} p(z_i = k | \{z_{-i}\}, \{w_i\}) &\propto p(z_i = k, \{z_{-i}\}, \{w_i\}) \\ &\propto p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) p(z_i = k | \{z_{-i}\}, \{w_{-i}\}) \\ &= p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) p(z_i = k | \{z_{-i}\}) \end{aligned}$$

- The first term is the **likelihood**
- The second term is like a **prior**

Collapsed Gibbs Sampling for LDA

- Look at the first term in $p(z_i = k | \{z_{-i}\}, \{w_i\}) \propto p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) p(z_i = k | \{z_{-i}\})$

$$\begin{aligned} p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) &= \int p(w_i | z_i = k, \varphi_k) p(\varphi_k | \{z_{-i}\}, \{w_{-i}\}) d\varphi_k \\ &= \int \varphi_{k, w_i} p(\varphi_k | \{z_{-i}\}, \{w_{-i}\}) d\varphi_k \end{aligned}$$

where

$$p(\varphi_k | \{z_{-i}\}, \{w_{-i}\}) \propto p(\{w_{-i}\} | \varphi_k, \{z_{-i}\}) p(\varphi_k) \sim \text{Dir}(\#_{-i,k}^{(w_i)} + \beta)$$

By using the property of expectation of Dirichlet distribution

$$p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) = \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k} + W\beta}$$

Collapsed Sampling for LDA

- Look at the second term in $p(z_i = k | \{z_{-i}\}, \{w_i\}) \propto p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) p(z_i = k | \{z_{-i}\})$

$$\begin{aligned} p(z_i = k | \{z_{-i}\}) &= \int p(z_i = k | \theta_d) p(\theta_d | \{z_{-i}\}) d\theta_d \\ &= \int \theta_{d,z_i} p(\theta_d | \{z_{-i}\}) d\theta_d \end{aligned}$$

where

$$p(\theta_d | \{z_{-i}\}) \propto p(\{z_{-i}\} | \theta_d) p(\theta_d) \sim \text{Dir}(\#_{-i,k}^{(d)} + \alpha)$$

By using the property of expectation of Dirichlet distribution

$$p(z_i = k | \{z_{-i}\}) = \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

Collapsed Gibbs Sampling for LDA

- Recall that the Gibbs sampling for LDA

$$p(z_i = k | \{z_{-i}\}, \{w_i\}) \propto p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) \times p(z_i = k | \{z_{-i}\})$$
$$\propto \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k}^{(*)} + W\beta} \times \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

- Now the problem has been simplified to count
 - $\#_{-i,k}^{(w_i)}$: number of w_i appearing in topic k
 - $\#_{-i,k}^{(*)}$: number of words in topic k
 - $\#_{-i,k}^{(d)}$: number of words of document d in topic k
 - $\#_{-i,*}^{(d)}$: number of words of document d (constant)

PLSA vs LDA

■ PLSA

$$p(w_j|z_k) = \frac{\sum_{i=1}^M \#(d_i, w_j) p(z_k|d_i, w_j)}{\sum_{n=1}^N \sum_{m=1}^M \#(d_m, w_n) p(z_k|d_m, w_n)}$$

$$p(z_k|d_i) = \frac{\sum_{j=1}^N \#(d_i, w_j) p(z_k|d_i, w_j)}{\#(d_i)}$$

■ LDA

$$p(z_i = k | \{z_{-i}\}, \{w_i\}) \propto p(w_i | z_i = k, \{z_{-i}\}, \{w_{-i}\}) \times p(z_i = k | \{z_{-i}\})$$

$$\propto \frac{\#_{-i,k}^{(w_i)} + \beta}{\#_{-i,k}^{(*)} + W\beta} \times \frac{\#_{-i,k}^{(d)} + \alpha}{\#_{-i,*}^{(d)} + K\alpha}$$

Visualization?

The visualization displays 11 word clouds, each representing a different topic. The words are color-coded and sized according to their frequency in the topic.

- Topic2:** riding, met, like, dog, msg, eat, with, fair, time, ing, sunpat, fix, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631,

- How to interpret the following topic visualization?

$$\varphi_{k,w_i} = \frac{\#_{*,k}^{(w_i)} + \beta}{\#_{*,k}^{(*)} + W\beta}$$



Project: Topic Modeling

■ Dataset:

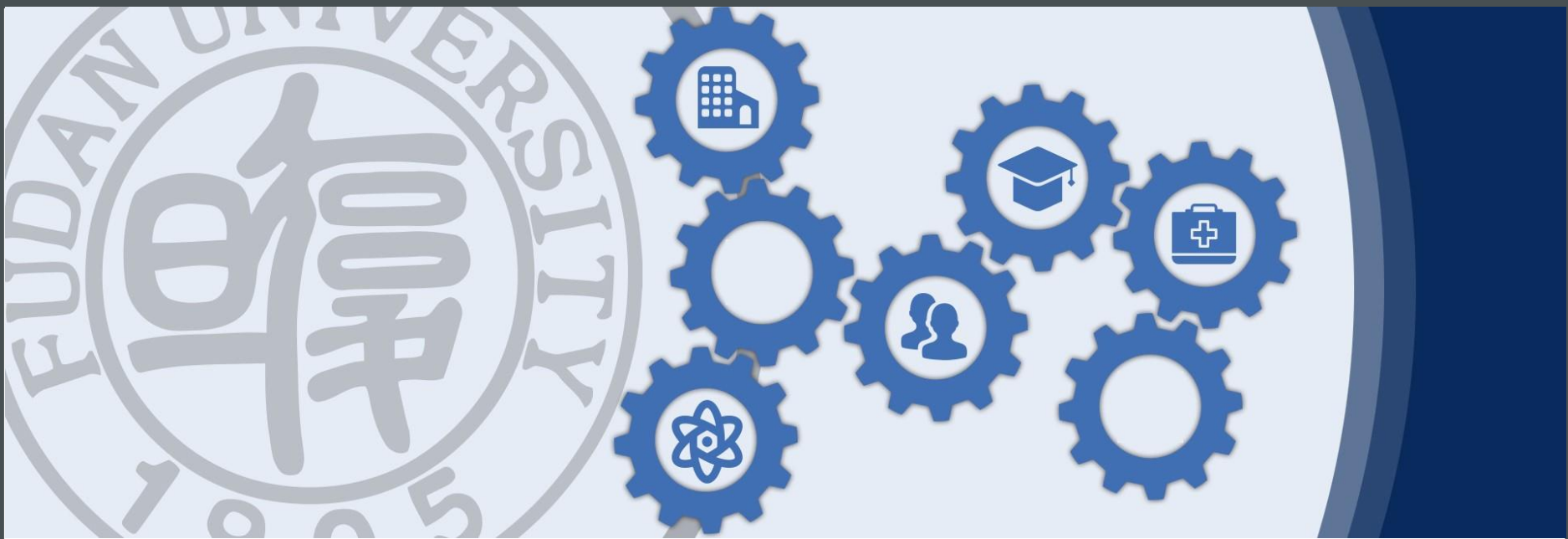
- ❑ Public available topic modeling datasets (e.g., <https://www.kaggle.com/jaykrishna/topic-modeling-enron-email-dataset>)
- ❑ Or document data collected by yourself

■ Method:

- ❑ Use **Probabilistic LSA** or **LDA with Collapsed Gibbs Sampling** for topic modeling

■ Experiments:

- ❑ Obtain the topic modeling results and visualize the topics using word clouds
- ❑ And discuss the observations from the visualization



Thanks

Email: libin@fudan.edu.cn