# PLSI

`Image annotation/tagging`

`Probabilistic Latent Sematic Indexing (PLSI)`

## Some Problems

- **one concept can be represented by several different words**
- **two documents might not contain similar terms but refer to a single concept**
- **query can contain words not present in a document bu still be very relevant to that document**

## PLSI model elements

1. A set of documents $\{d_1, \dots, d_N\}$
2. A set of concepts, classes or topics$\{z_1, \dots, z_K\}$
3. A set of words $w_1, \dots, w_M$

### Some Assumptions

1. Each concept is a distribution over words
2. Each document is a mixture of corpus-wide topics
3. Each word is drawn from one of these topics
4. We only observe the word within the documents and the other structures are hidden variables.

## Model Pipeline

- Select a document with probability $P(d)$
- Pick a latent class z with probability $P(z|d;\theta)$
- Generate a wotd w with probability $P(w|z;\pi)$

$$P(d,w) = P(d)P(w|d)$$
$$P_{LSA}(w|d) = \sum_{z \in Z} P(w|z;\theta)P(z|d;\pi) ==>$$
$$P_{LSA}(d,w) = \sum_{z \in Z} P(d|z)P(z)P(w|z)$$

PS. 上式中的三部分分别是1.pLSA document probabilities 2.concept probabilities 3.pLSA term probabilities

# 核心公式推导如下：

我们在已知document和word的情况下，求解topic

$$P(topic|w,d) = \frac{P(w|topic)P(topic|d)}{P(w|d)}$$
$$Due\ to \quad P(topic|d) \propto P(d|topic)P(topic)$$
$$Hence \quad P(topic|w,d) \propto P(w|topic)P(d|topic)P(topic)$$

$topic \in Topic$

$$P(topic|w,d) = \sum_{topic \in Topic} P(w|topic)P(d|topic)P(topic)$$

求上式的MLE，然后取log得到

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(topic|w_j,d_i)^{n(w_j,d_i)}$$
$$logL = \sum_{i=1}^{N}\sum_{j=1}^{M} n(w_j,d_i)logP(topic|w_j,d_i)$$

EM to find optimal solution