大數據在觀光行銷上的應用

林億雄[1]

2020 年 05 月 06 日 星期三

[1]任職單位: 台灣首府大學

故事的起頭

- 奧倫.埃齊奧尼 - Farecast 票價: 一位專研資訊科技的教授搭飛機參加弟弟的婚禮, 他很早上網訂票, 以爲賺到了, 沒想到鄰座比他晚買, 卻買得更便宜。他利用大數據分析, 確保消費者買到便宜機票。微軟以1.1億美元併購了。
- 消費者物價指數調查: 麻省理工學院開發出新軟體, 每天抓取全美逾50萬項產品價格, 透過大數據分析, 馬上發現通貨緊縮現象, 比官方早了2個月。這套分析工具, 比政府花大筆經費僱員查訪價格, 更有效率且精準。
- 細胞簡訊: 災防警告細胞廣播訊息系統, 在短時間內大量傳送疫情訊息、地震速報、土石流警戒、公路封閉等防災警示訊息到手機, 即時通知民眾的系統。

想像力創造大數據創新價值

- 想像力的重要：亨利.福特 如果使用大資料分析技術得到的結果會是消費者想要跑的更快的馬，你必須有想像力告訴消費者你需要的是一輛車子。
- 用數據做決策：Google 台灣 簡立峰總經理指出10年前由該公司開發的搜尋趨勢發現消費者搜尋PC 相關關鍵字的次數已經逐漸下降，然台灣還把 PC 當作電子業生產主軸。
- 異質資料結合：氣喘用的藥物吸入器與手機定位資料結合，找出氣喘發作的原因。
- 非接觸科技：防疫期間，非接觸科技的重視。EX: 區塊鏈與虛擬貨幣

用比特幣BTC 買 HTC 手機

商品名稱	價格	數量	小計
 EXODUS 1	BTC 0.21	1	BTC 0.21

訂單日期: 2019/1/6 上午12:22:04
訂單號碼: HTCEXDB-010329942
運費: BTC 0.00
訂單總額: BTC 0.21

收貨資訊:

虛擬貨幣付款資訊:
轉帳金額:
0.21000000 BTC
地址:
1Lixq99rTQ1Mu6GpHusS33Z5YyFRiBpD2o

大數據分析 - 行銷創新者

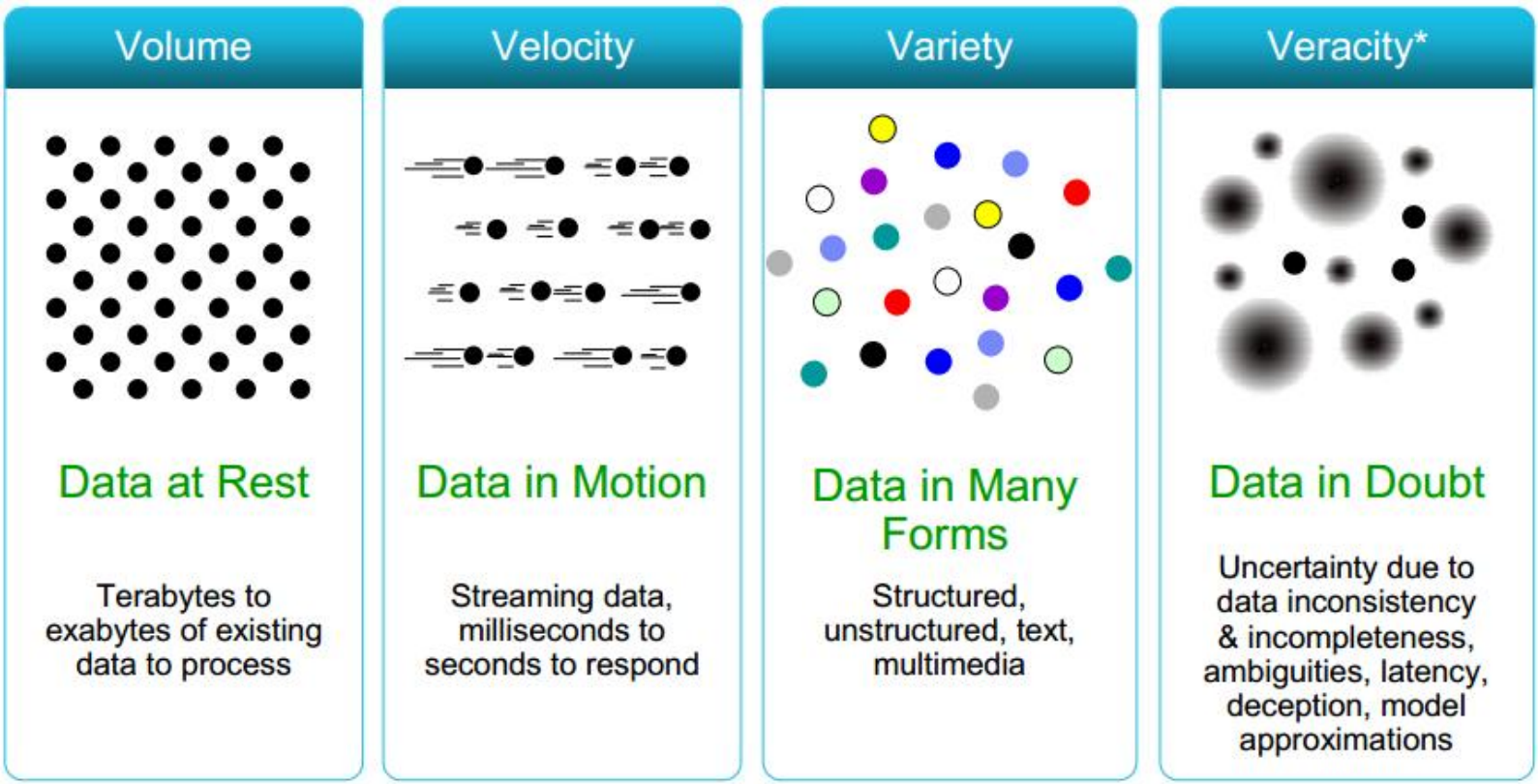
- 大數據分析 - 精準行銷: 精準行銷要了解用戶, 需透過第三方(例如：FB、Twitter、IG 等 APP) 收集使用者數據, 同時結合企業自身的數據, 才能了解用戶的樣貌, 推送的訊息才能更爲精準。
- 大數據分析 - 發現未知商機, 開發新產品或服務：賈伯斯認爲除非你給消費者看你的新產品, 否則消費者不知道自己要什麼?

Google

Google 台灣

整合全球範圍內的資訊
使人人皆可存取並從中受益
– Google 公司官方使命

Web 2.0 : The 4 Vs of Big Data



Big Data : Information Overload

Data sizes :

- IDC said that in 2011, the amount of information created and replicated surpassed 1.8ZB and estimated 35.2ZB in 2020.
- Social interactions, mobile devices, facilities, equipment, RD, simulations, and physical infrastructure all contribute to the flow. In aggregate, this is what is called Big Data.
- 1ZB =1,024EB; 1EB =1,024PB; 1PB =1,024TB; 1TB =1,024GB; 1ZB 澤位元 = 1 trillion GB 1兆吉位元
- ZB 有多大, 就像全世界海灘上的沙子數目一樣多

Big Data : 一分鐘的網路數據

- Google 上產生超過 2,000,000 則搜尋內容
- 超過 200,000,000 封電子郵件
- Facebook 上產生超過 680,000 則內容
- 超過27,000美元的網路購物內容
- APP Store 裡的 APP 被下載達 47,000 多次
- Flickr 用戶分享了 3,125 張照片
- 誕生了 217 名行動網路新用戶

Web 2.0 下無法逃避的趨勢

- Web2.0 已經影響改變：救災、教育、零售、製造、服務、金融、醫療、政治等傳統領域思維的運作。
- Web2.0 代表新世代與新秩序的崛起及誕生: 結合眾人的智慧及能力, 超越個人智慧而決定的新世代。
- 雲端應用與行動裝置的效益關鍵 ：說不清楚 講不明白的解決方案: 宏碁 ab App。
- 公民決策參與：開放政府, 全民表達意見, 網路世代善用新媒體如: PPT、部落格、臉書、LINE 對於現況發聲。
- 零售大賣場變數據中心: 線上購物日益興盛, 導致實體零售商店門可羅雀, 紛紛關門歇業, 空出來的店面轉爲數據中心, 專門處理線上網路購物資料。
- 防災救急的新思維： 過去做不到的災變預防轉爲有可能。

July, 29, 2008, LA Earthquake. First Twitter 5 Sec.


Los Angeles Times

33 Die, Many Hurt in 6.6 Quake

L.A. Area Freeways Buckle, Buildings Topple

Sylmar Jolted by Ghosts of Horror Past

2011 Japan Earthquake as seen through Twitter

2011 Japan Earthquake as seen through Twitter: 日本人 Rio Akasaka 利用地圖製作呈現在311日本大地震前後, 內含地震這個關鍵詞的Twitter 訊息如何傳遍世界。

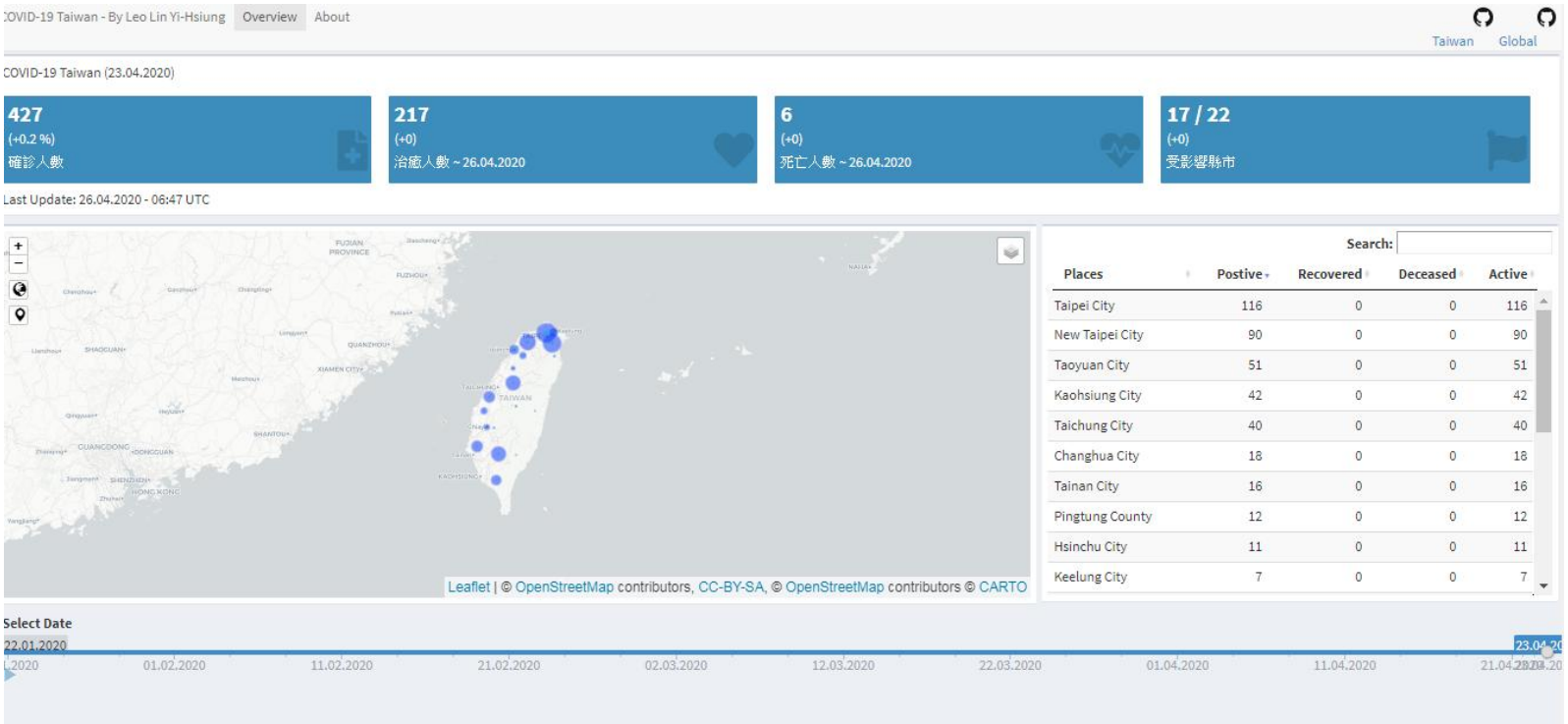
Google Flu Trends

Google追蹤全球流感趨勢:Google 流感趨勢能夠根據彙總的搜尋資料，提供近乎即時的全球流感疫情趨勢預測。

Johns Hopkins COVID-19 Outbreak

COVID-19 Outbreak: COVID-19 能夠根據各國彙總的資料, 提供近乎即時的新冠肺炎疫情傳播情形。

台灣首府大學 COVID-19 Taiwan



COVID-19 Taiwan: 台灣首府大學響應政府科技防疫根據彙總資料，提供台灣新冠肺炎疫情即時情形。

新型態的工作：資料科學家

- 新一波的資訊革命: 已由數位革命，進入資料革命。
- 典型資料科學家的五種特質：駭客, 科學家, 量化分析師, 可靠顧問, 及商業專家。
- 多重資料型態結合應用：大數據與小數據, 內部與外部資料來源, 結構與未結構化資料, 傳統資料分析與機器學習建模。

新型態工作 - 資料科學家

- 新型態工作 - 資料科學家: 你需要耙梳線上數位資料技術、數學統計專才、商業經營分析等多元能力的工作。
- 新型態工作 - 資料科學家: 你還需要有會說故事的能力。

新型態工作帶來的產業革命

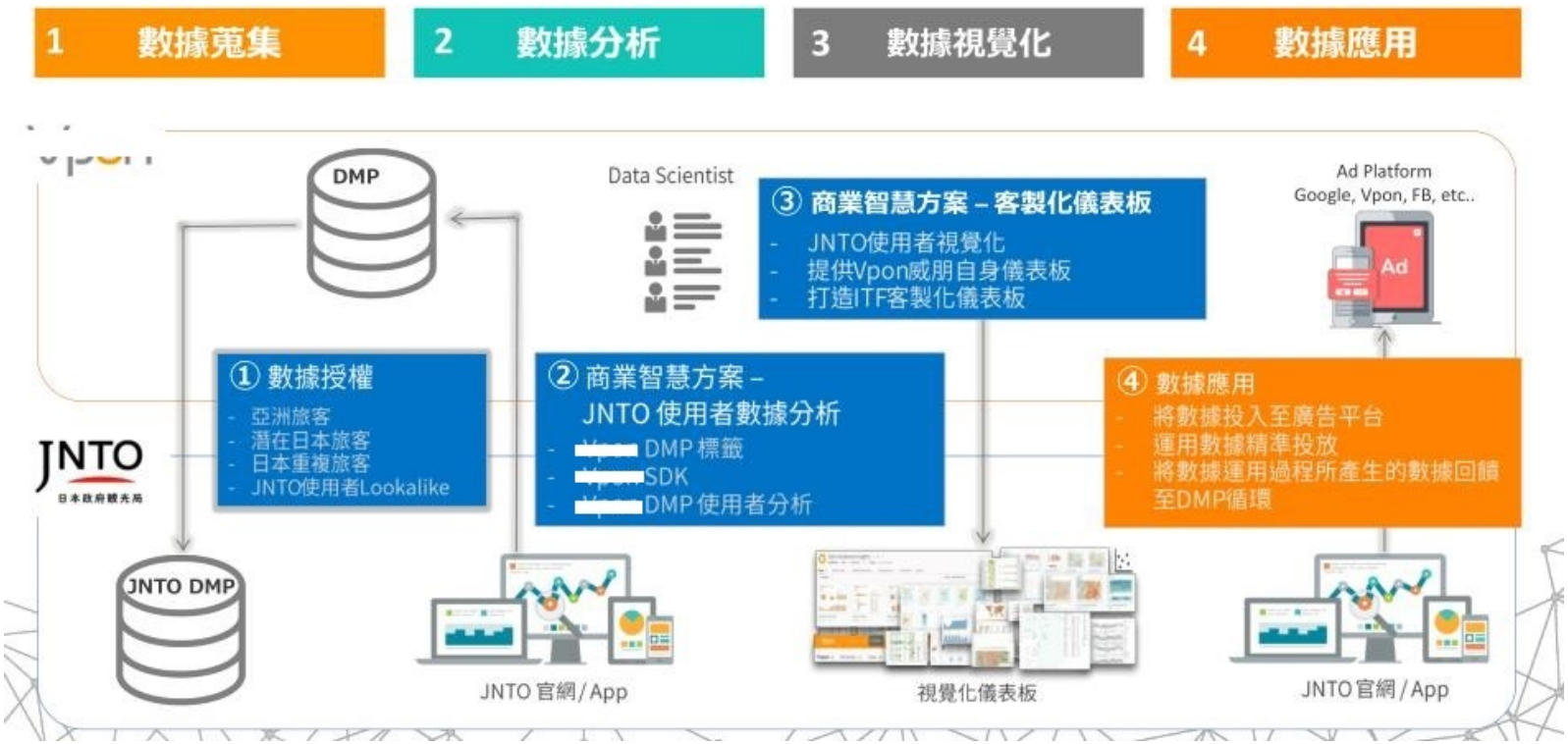
- 新型態工作 - 舊時代的拉扯: 航海時代雷達的新發明與天氣預報, 使得漁業進入新時代, 老船長經驗地位變得可以被取代。
- 新型態工作 - 舊時代的法令: 新技術的衝擊使得現行法令受到衝擊與挑戰, 例如 : 區塊鏈下虛擬貨幣的合法性。
- 新型態工作 - 舊時代的制度: 大數據分析之精準行銷部門需列入技術部門, 還是業務部門 ?

世界歷史上三次工業革命與第四次產業革命

- 第一次工業革命約從1760 - 1840年, 由鐵路建設和蒸汽機發明觸發的革命, 引領人類進入機械生產的時代。
- 第二次工業革命始於19世紀末, 延續至20世紀初, 隨著電力和生產線的出現, 規模化生產應運而生。
- 第三次工業革命始於20世紀中葉, 通常稱爲電腦革命、數位革命, 催生這場革命的是半導體技術、大型電腦 (60年代)、個人電腦 (70-80年代) 和網路 (90年代) 的發展。
- 第四次產業革命始於這個世紀, 是在數位革命的基礎上發展起來的, 其特點是: 網路變得無所不在, 移動性大幅提高; 感測器體積變得更小、性能更強大、成本也更低; 與此同時, 人工智慧和機器學習也開始嶄露鋒芒。

案例：觀光旅遊業數據解決方案 - 1

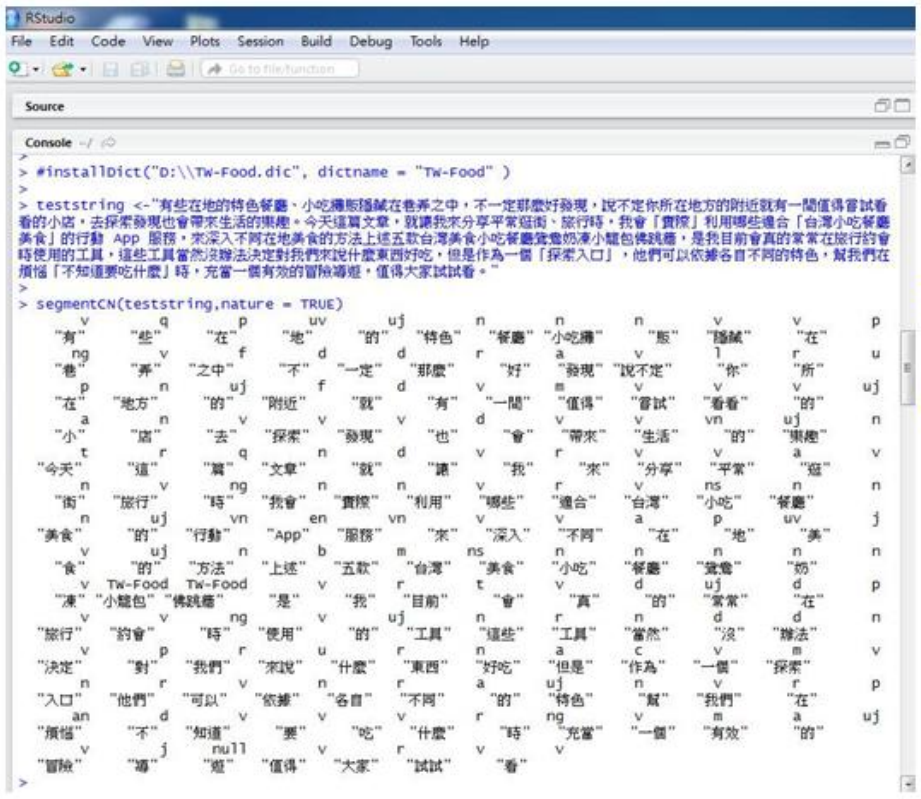
案例：觀光旅遊業數據解決方案 - 2



大數據分析 - 文字探勘

- (1): 耙梳網路線上數位資料
- (2): 文本清理與中文斷詞初探
- (3): 簡易主題模式-潛藏狄利克里分配

常見分析方法 - 中文斷詞與文字雲圖



目標網站 : TripAdvisor

- TripAdvisor 是全球最大的旅遊網站 [2], 旨在協助旅客規劃行程並享受最優質的旅遊體驗。
- TripAdvisor 所屬網站組成了世界上最大的旅遊社群,2014 年每月有超過 3 億 150 萬名非重複的訪客[3], 還有超過 2 億則評論和意見, 範圍囊括 440 萬個住宿服務、餐廳和景點物業。
- TripAdvisor 網站遍佈世界各地 45 個國家, 並擁有 TripAdvisor for Business, 專爲旅遊業提供與 TripAdvisor 每月數千萬名訪客接觸的商機。

[2] Source: comScore Media Metrix for TripAdvisor Sites, worldwide, August 2014
[3] Source: Google Analytics, average monthly unique users, Q3 2014

文本清理與中文斷詞

- 文本清理: 以詞性判斷篩選文本資料
- 中文斷詞: 分詞動作與詞庫管理

Introduction to Topic Models

Topic Models

- Topic models : a statistical method can be applied to large databases that can yield insight into human cognition. (Griffith and Steyvers, 2004)
- Topic models : a subfield of machine learning applied to computational linguistics, to bioinformatics, to political science and social network data. (Blei, Ng and Jordan, 2003)
- Topic models : based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. (Blei et al, 2003; Griffith and Steyvers, 2002;2003;2004; Hofmann, 1999;2001)

Textual Data : Computer.txt

Computer

From Wikipedia, the free encyclopedia

A computer is a machine that is able to take information (input), do some work on or make changes to the information, to make new information (output). Computers have existed for much of human history. Examples of early computers are the astrolabe and the abacus. Modern computers are very different from early computers. They are now very powerful machines that are able to do billions of calculations every second. Most people have used a personal computer in their home or at work. Computers are useful for many different jobs where automatic functions are useful. Some examples are controlling traffic lights, vehicle computers, security systems, Washing machines and Digital Televisions. A person (called a user) can control a computer by telling it to do things. Some ways of controlling a computer are with a keyboard, mouse, buttons, touch screen. Some very new computers can also be controlled with voice commands or hand gestures. Computers can be designed to do anything with information. Computers are used to control factories, which in the past were controlled by humans. They are also in homes, where they are used for things such as listening to music, reading the news, and writing.

Topic Models : Documents exhibit multiple topics

Computer

From Wikipedia, the free encyclopedia

A computer[1] is a machines[1] that is able to take information[1] (input), do some work on or make changes[2] to the information[1], to make new information[1] (output). Computers have existed[2] for much of human history[3]. Examples of early computers are the astrolabe and the abacus. Modern computer[1] are very different from early computers[1]. They are now very powerful machines[1] that are able to do billions of calculations[2] every second. Most people have used a personal computer[1] in their home or at work[2]. Computers[1] are useful for many different jobs where automatic[2] functions are useful. Some examples are controlling traffic lights[3], vehicle computers, security[2] systems, Washing[3] machines[1] and Digital Televisions[1]. A person (called a user) can control a computer[1] by telling it to do things. Some ways of controlling a computer[1] are with a keyboard, mouse, buttons, touch screen. Some very new computers[1] can also be controlled[2] with voice[1] commands or hand gestures. Computers can be designed[2] to do anything with information[1]. Computers[1] are used to control factories, which in the past were controlled[2] by humans. They are also in homes, where they are used for things such as listening to music[1], reading[3] the news, and writing[3].

Topic Models : Documents exhibit multiple topics

Computer From Wikipedia, the free encyclopedia

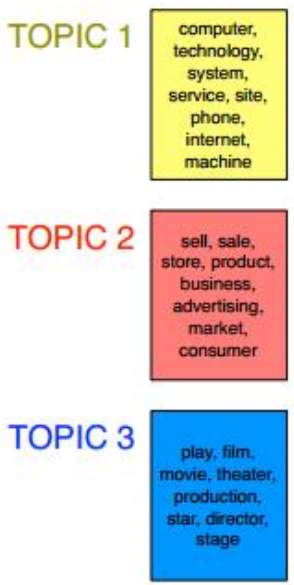
A computer[1] is a machines[1] that is able to take information[1] (input), do some work on or make changes[2] to the information[1], to make new information[1] (output). computers[1] have existed[2] for much of human history[3]. Examples of early computers are the astrolabe and the abacus. Modern computer[1] are very different from early computers[1]. They are now very powerful machines[1] that are able to do billions of calculations[2] every second. Most people have used a personal computer[1] in their home or at work[2]. Computers[1] are useful for many different jobs where automatic[2] functions are useful. Some examples are controlling traffic lights[3], vehicle computers, security[2] systems, Washing[3] machines[1] and Digital Televisions[1]. A person (called a user) can control a computer[1] by telling it to do things. Some ways of controlling a computer[1] are with a keyboard, mouse, buttons, touch screen. Some very new computers[1] can also be controlled[2] with voice[1] commands or hand gestures. Computers can be designed[2] to do anything with information[1]. Computers[1] are used to control factories, which in the past were controlled[2] by humans. They are also in homes, where they are used for things such as listening to music[1], reading[3] the news, and writing[3].

Topic 1: computer[1] (0.40) machines[1](0.25) information[1](0.15) Televisions[1](0.05)

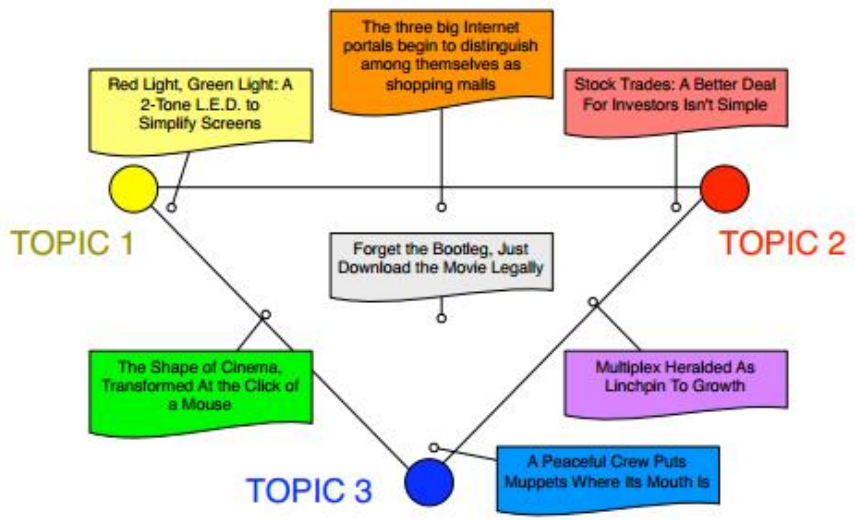
Topic 2: controlled[2](0.35) security[2](0.25) automatic[2](0.15) designed[2](0.07)

Topic 3: Washing[3](0.55) lights[3](0.25) reading[3](0.15) history[3](0.02)

Topic Models : Documents exhibit multiple topics

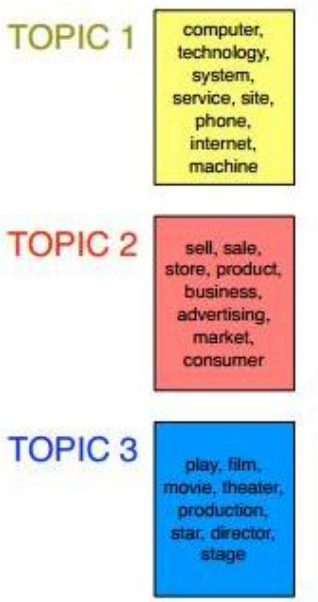


(a) Topics

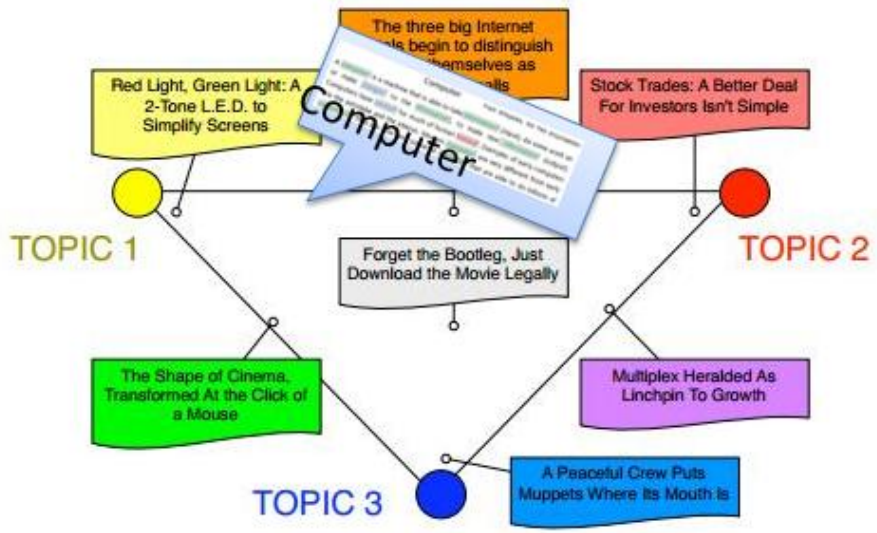


(b) Document Assignments to Topics

Topic Models : Documents exhibit multiple topics



(a) Topics



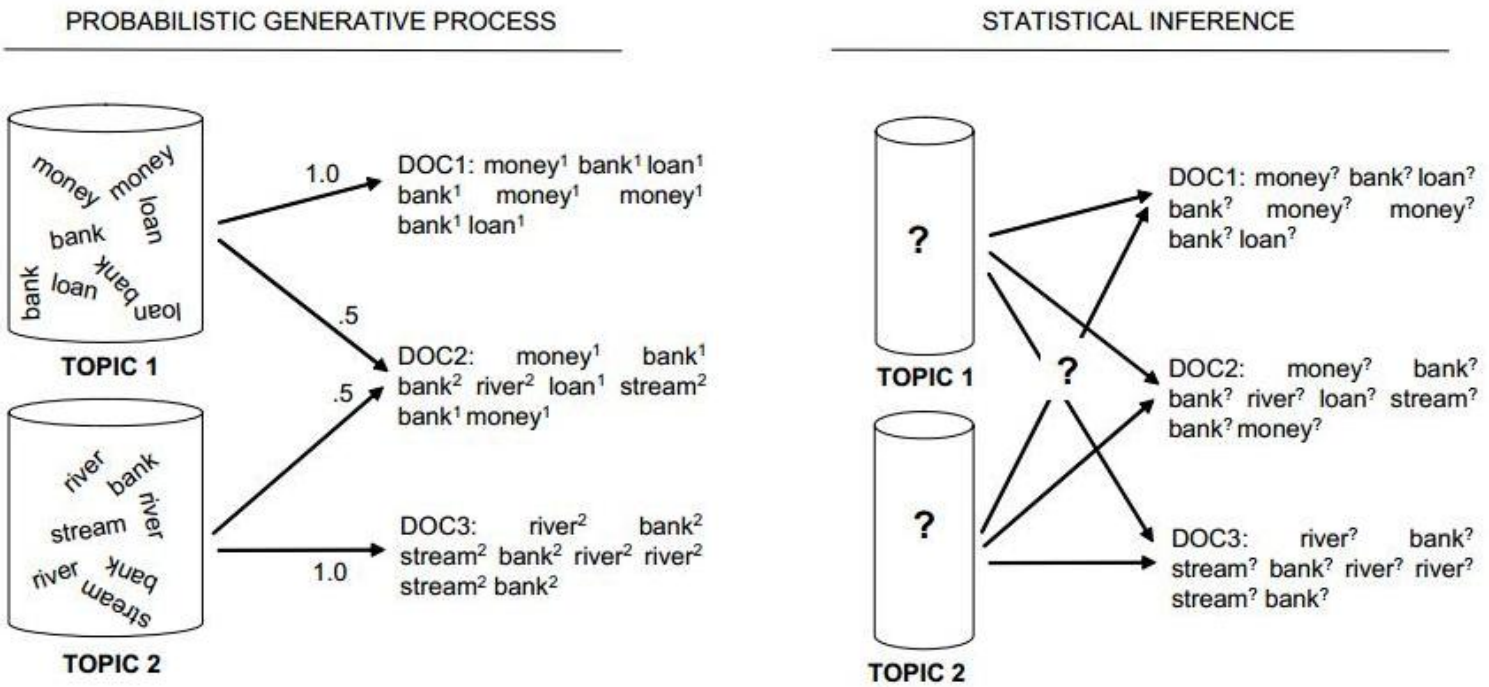
(b) Document Assignments to Topics

Introduction to Topic Models :Latent Dirichlet Allocation

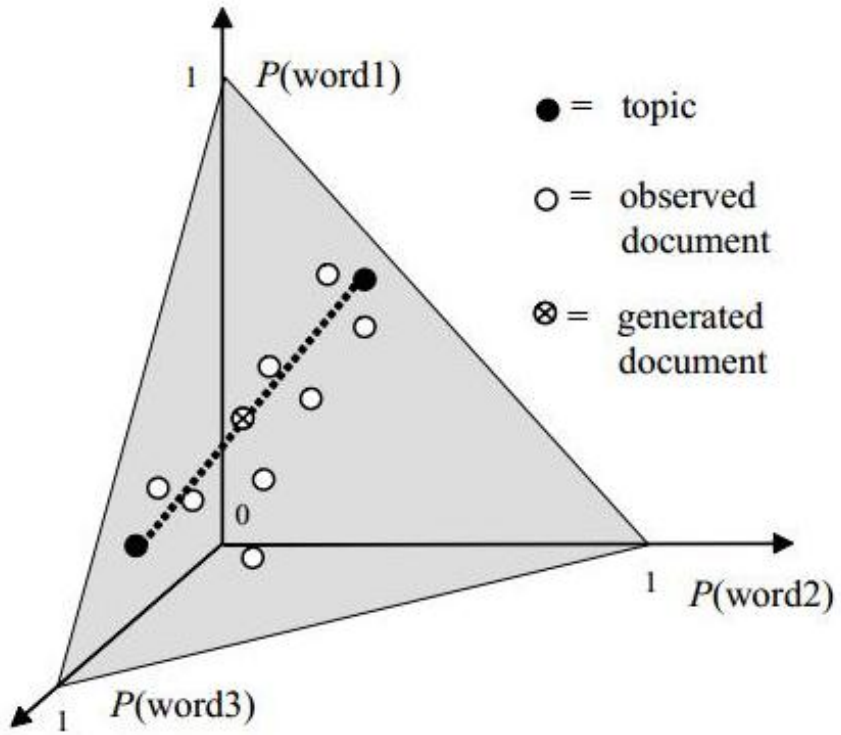
Latent Dirichlet Allocation (Blei et al, 2003)

- Each **topic** is a distribution over words.
- Each **document** is a mixture of multiple topics.
- Each **word** is sampled from one of those topics.

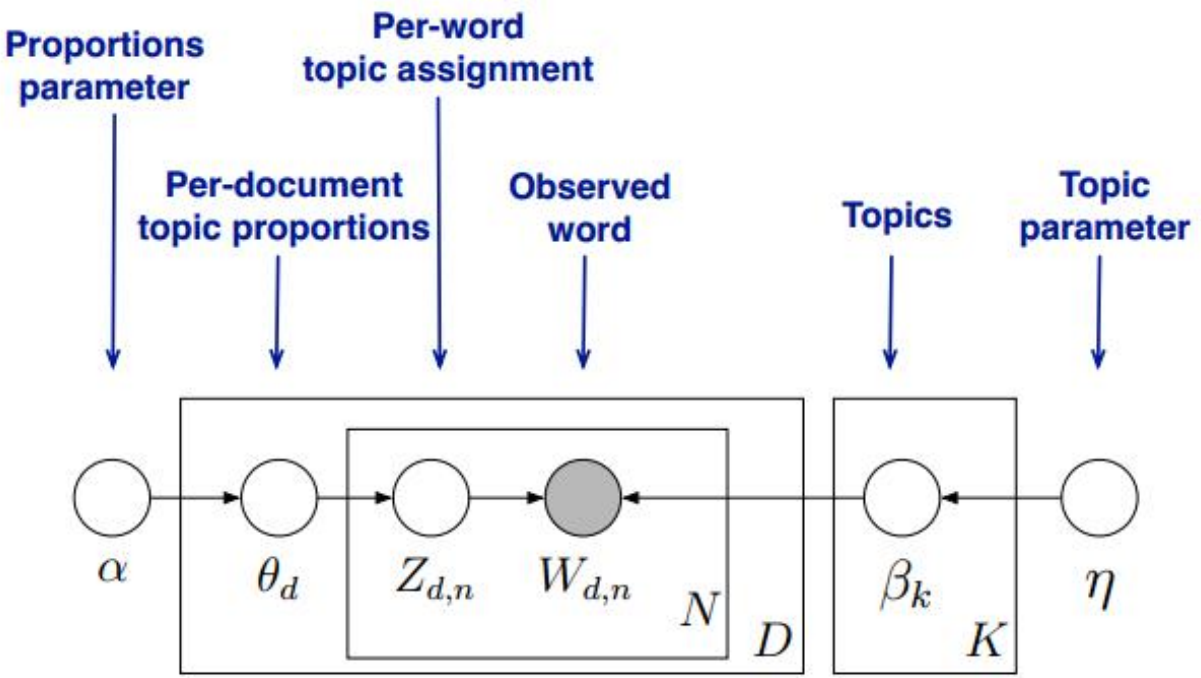
LDA - Generative Models



LDA - Generative Models



Graphical model for Latent Dirichlet Allocation



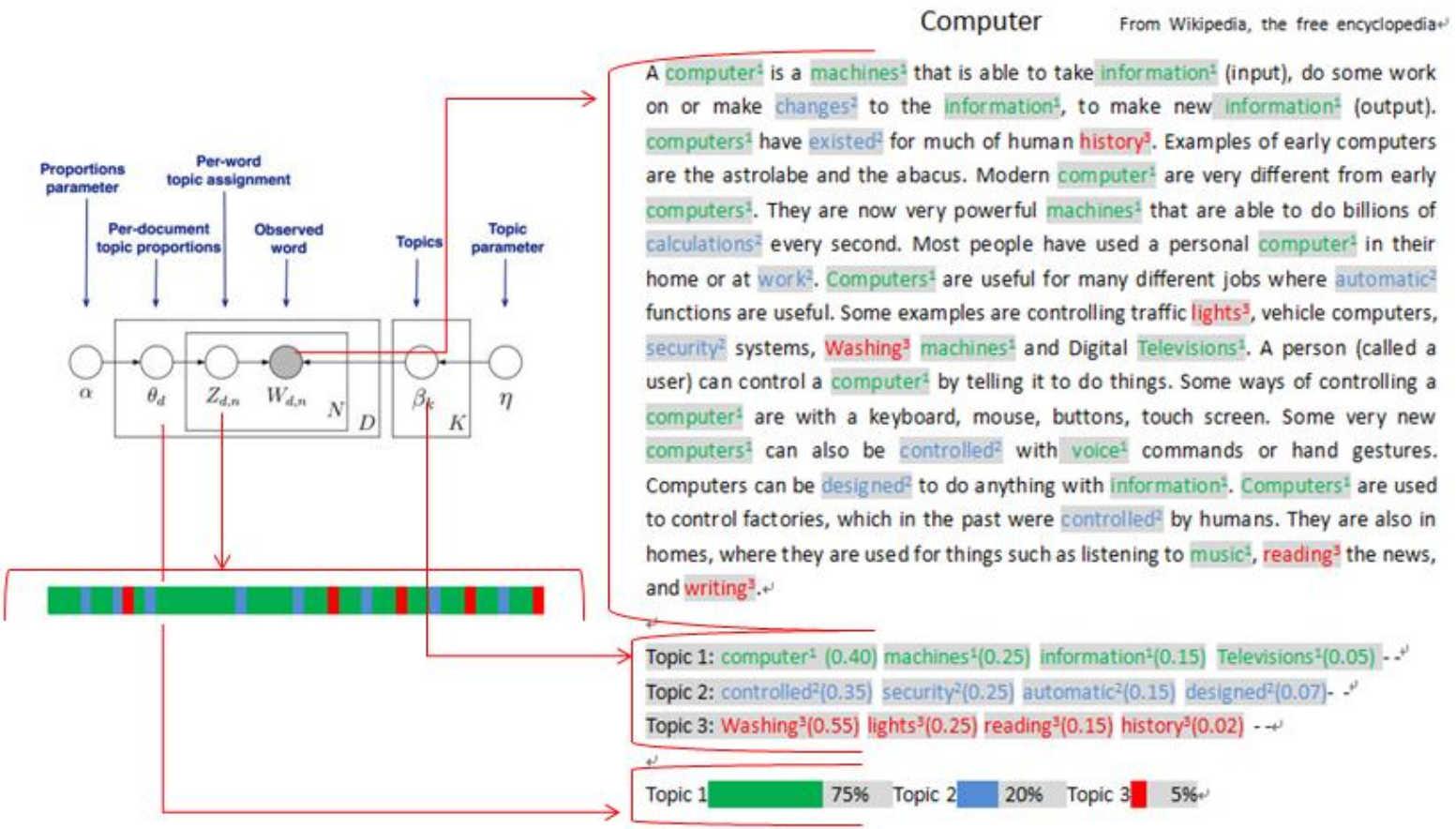
Topic labels for the Document

Computer

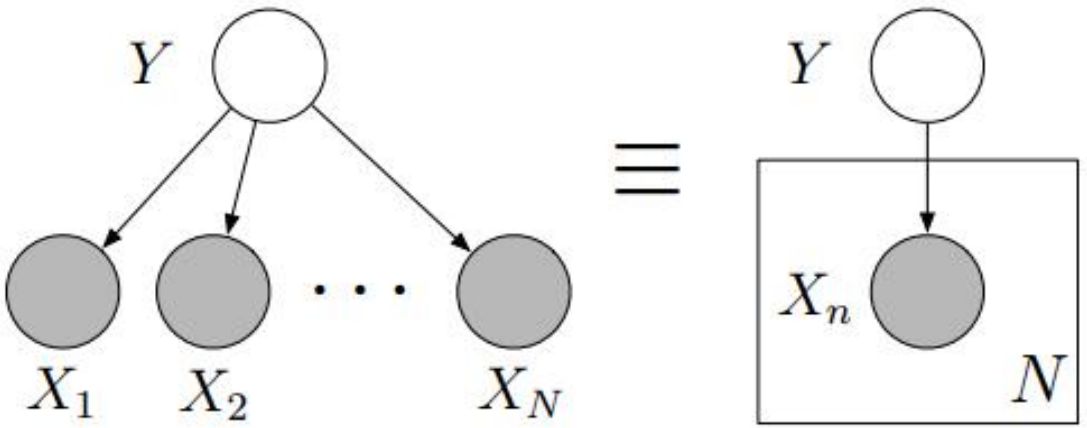
From Wikipedia, the free encyclopedia

A computer[1] is a machines[1] that is able to take information[1] (input), do some work on or make changes[2] to the information[1], to make new information[1] (output). Computers have existed[2] for much of human history[3]. Examples of early computers are the astrolabe and the abacus. Modern computer[1] are very different from early computers[1]. They are now very powerful machines[1] that are able to do billions of calculations[2] every second. Most people have used a personal computer[1] in their home or at work[2]. Computers[1] are useful for many different jobs where automatic[2] functions are useful. Some examples are controlling traffic lights[3], vehicle computers, security[2] systems, Washing[3] machines[1] and Digital Televisions[1]. A person (called a user) can control a computer[1] by telling it to do things. Some ways of controlling a computer[1] are with a keyboard, mouse, buttons, touch screen. Some very new computers[1] can also be controlled[2] with voice[1] commands or hand gestures. Computers can be designed[2] to do anything with information[1]. Computers[1] are used to control factories, which in the past were controlled[2] by humans. They are also in homes, where they are used for things such as listening to music[1], reading[3] the news, and writing[3].

Graphical model and the Document

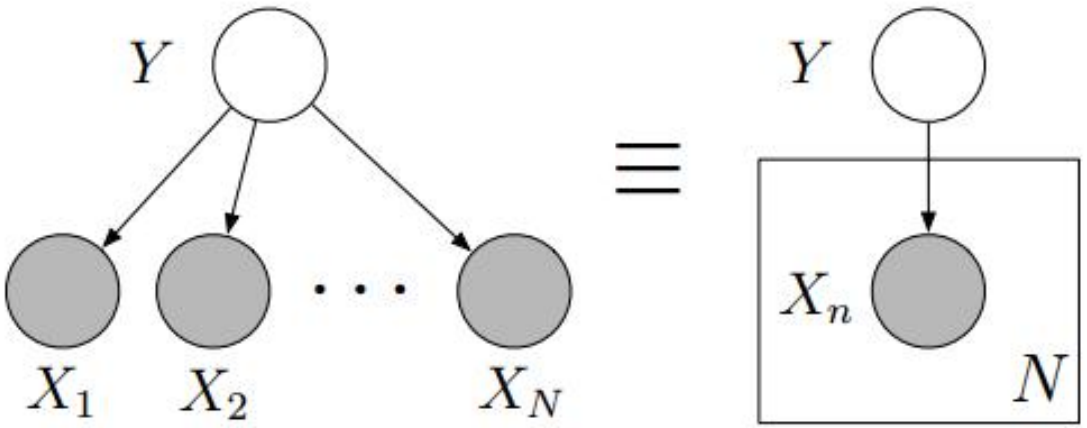


Introduction : Graphical model



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

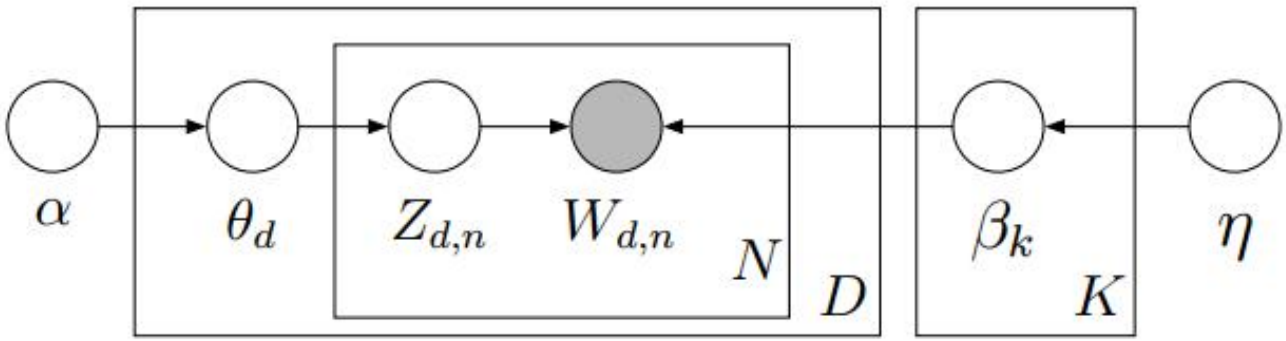
Introduction : Graphical model



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

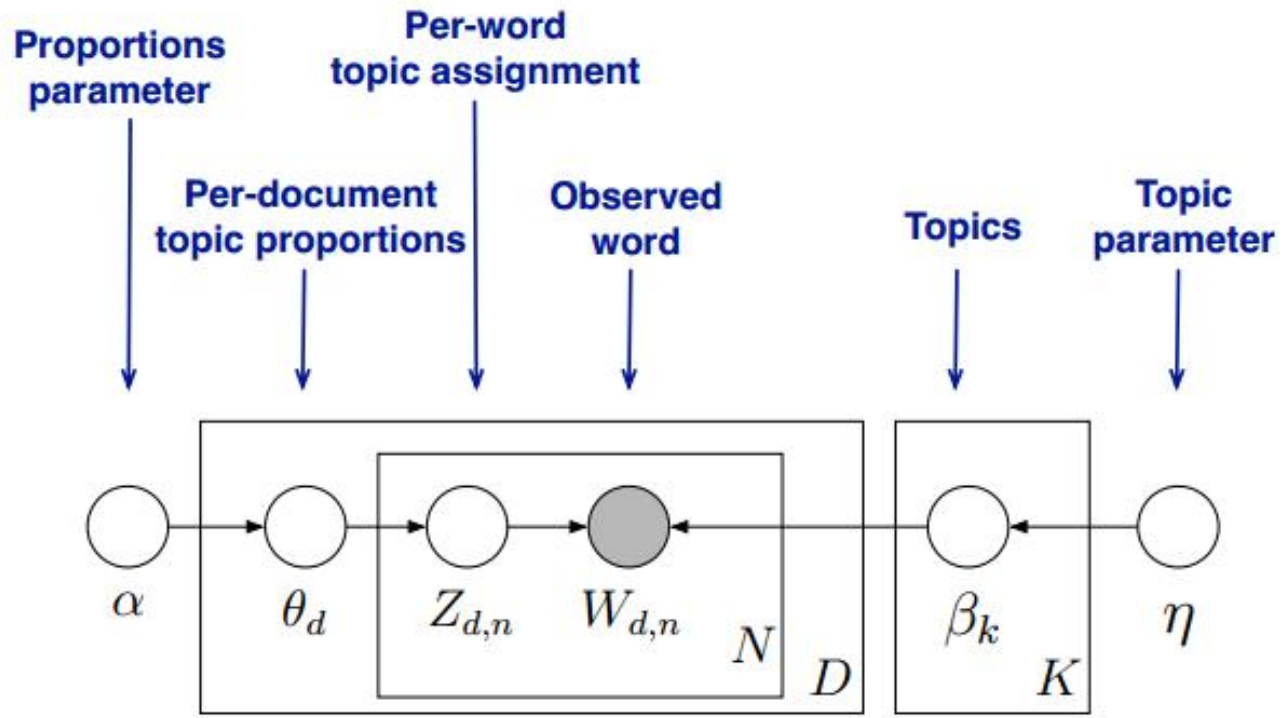
$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

The Joint Distribution



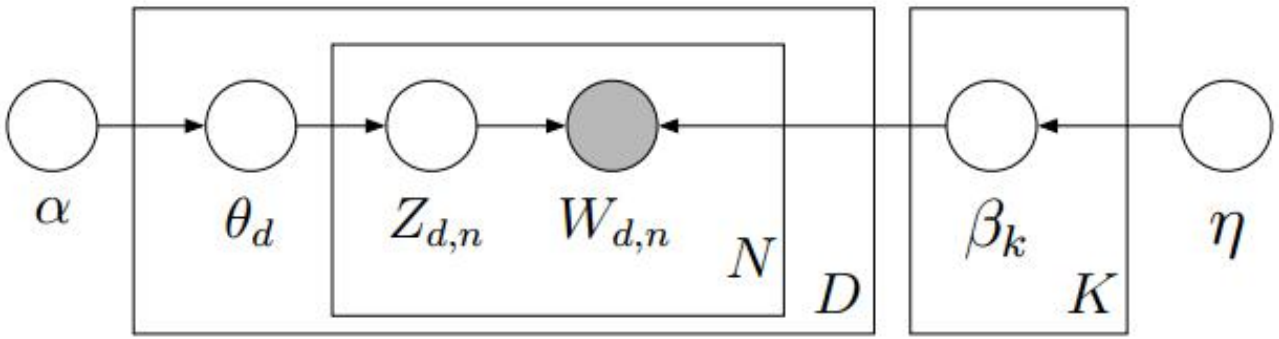
1. Draw each topic $\beta_i \sim \text{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \text{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$.

The Joint Distribution



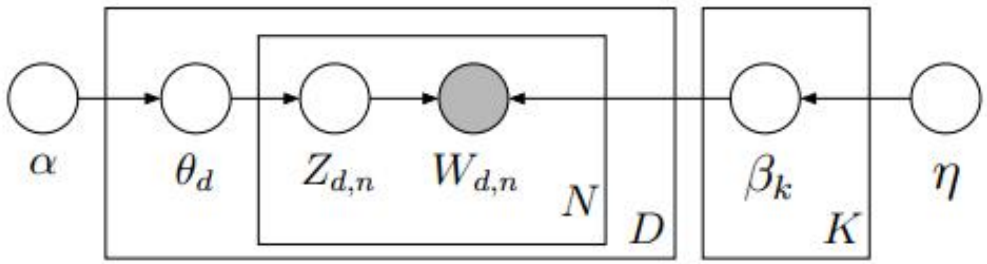
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

The Joint Distribution



- This joint defines a posterior.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k

The Posterior inference for LDA



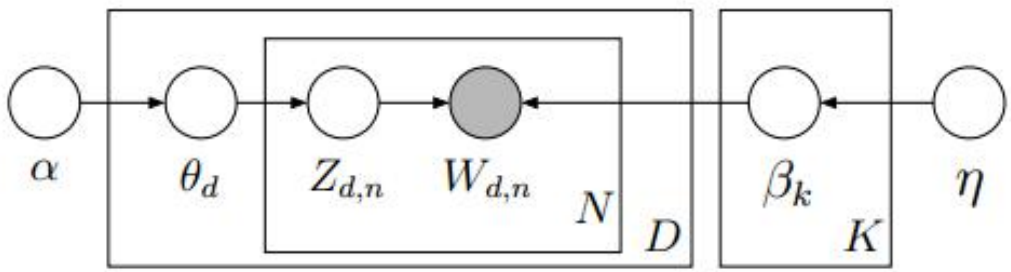
- The joint distribution of the latent variables and documents is

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right).$$

- The posterior of the latent variables given the documents is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N} \mid w_{1:D,1:N}).$$

The Posterior inference for LDA

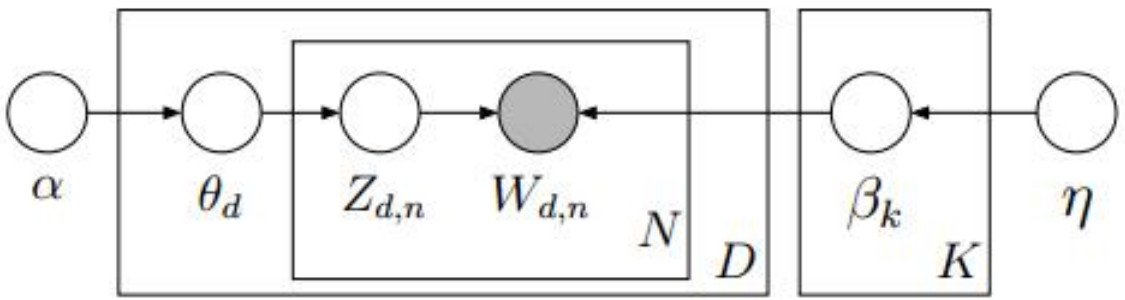


- This is equal to

$$\frac{p(\beta_{1:K}, \theta_{1:D}, \mathbf{z}_{1:D}, \mathbf{w}_{1:D})}{\int_{\beta_{1:K}} \int_{\theta_{1:D}} \sum_{\mathbf{z}_{1:D}} p(\beta_{1:K}, \theta_{1:D}, \mathbf{z}_{1:D}, \mathbf{w}_{1:D})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w}_{1:D})$.

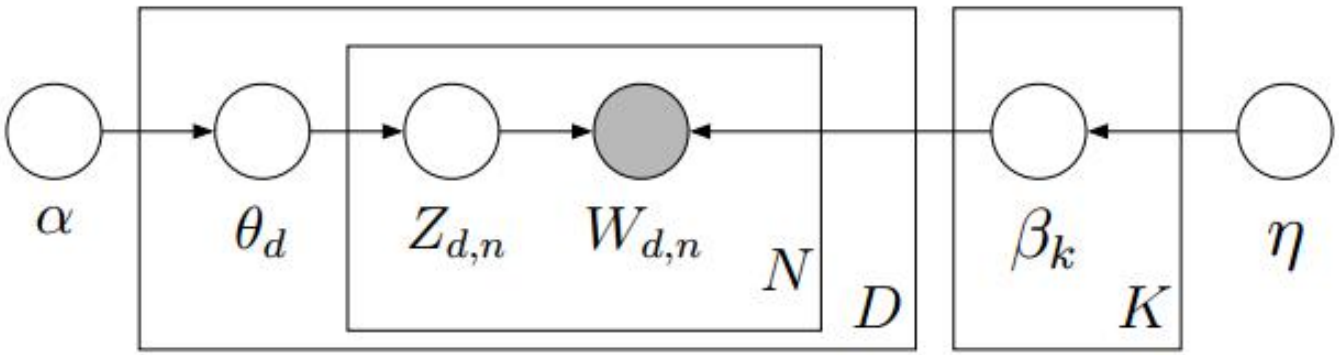
The Posterior inference for LDA



- There is a large literature on approximating the posterior.

- We will focus on Gibbs Sampling.

The Posterior inference for LDA



Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)

Also see Mukherjee and Blei (2009) and Asuncion et al. (2009).

The Gibbs Sampling algorithm

zero all count variables NWZ, NZM, NZ ;
foreach *document* $m \in [1, D]$ **do**
  foreach *word* $n \in [1, N_m]$ *in document* m **do**
    sample topic index $z_{m,n} \sim \text{Mult}(1/K)$ for word $w_{m,n}$;
    increment document-topic count: NZM$[z_{m,n}, m]$++ ;
    increment topic-term count: NWZ$[w_{m,n}, z_{m,n}]$++ ;
    increment topic-term sum: NZ$[z_{m,n}]$++ ;
  end
end
while *not finished* **do**
  foreach *document* $m \in [1, D]$ **do**
    foreach *word* $n \in [1, N_m]$ *in document* m **do**
      NWZ$[w_{m,n}, z_{m,n}]$--, NZ$[z_{m,n}]$--, NZM$[z_{m,n}, m]$-- ;
      sample topic index $\tilde{z}_{m,n}$ according to A
      NWZ$[w_{m,n}, \tilde{z}_{m,n}]$++, NZ$[\tilde{z}_{m,n}]$++, NZM$[\tilde{z}_{m,n}, m]$++ ;
    end
  end
  if *converged and L sampling iterations since last read out* **then**
    read out parameter set $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ according to B
  end
end

$$p(z_i) = \frac{\Psi_{z_i,w_i} + \beta_{w_i}}{\left[\sum_{v=1}^{V} \Psi_{k,v} + \beta_t\right]} \cdot \left[\Omega_{d_i,z_i} + \alpha_{z_i}\right] .$$

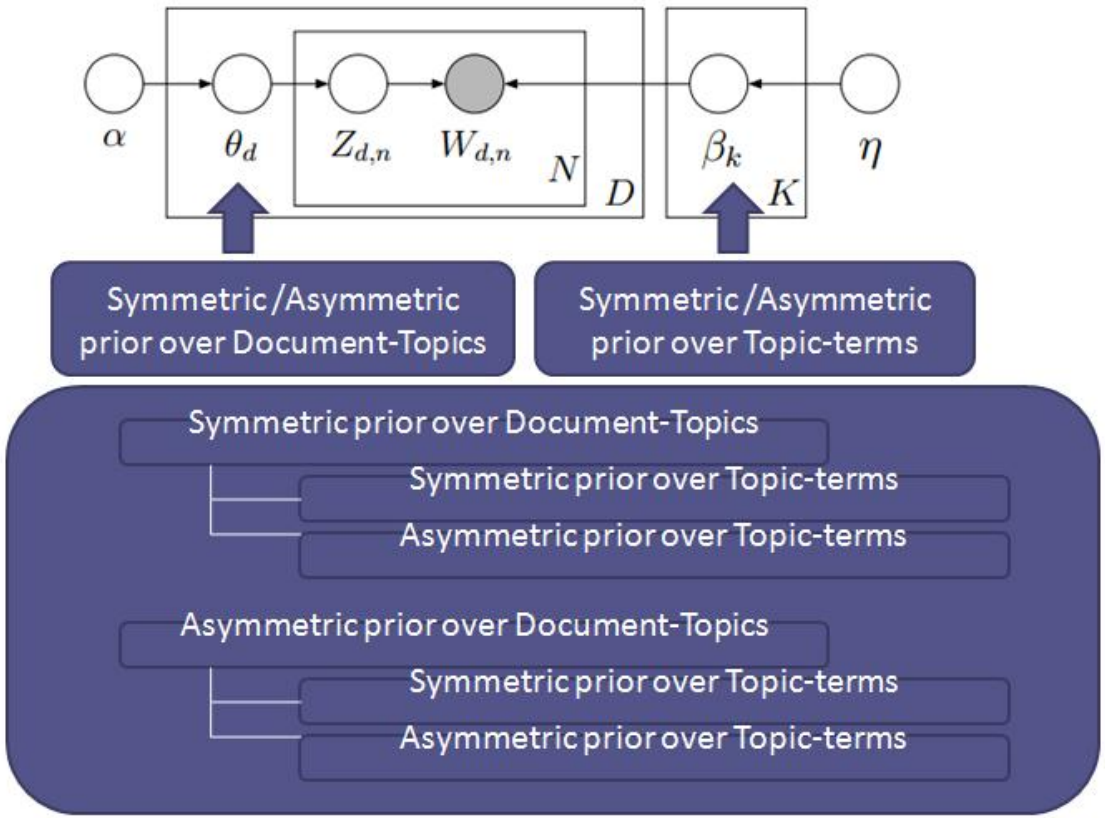
Parameter Estimation By the definition of $\phi_{k,v}$ and $\theta_{d,k}$, we have

$$\phi_{k,v} = \frac{\Psi_{k,v} + \beta_t}{\left(\sum_{v'=1}^{V} \Psi_{k,v'} + \beta_{v'}\right)} ,$$

$$\theta_{m,k} = \frac{\Omega_{d,k} + \alpha_k}{\left(\sum_{k=1}^{K} \Omega_{d,k} + \alpha_z\right)} .$$

Introduction to Dirichlet Distribution

Priors for LDA [4]



[4] Y.H., Lin , 2013, The choice of prior to LDA for fitting topic models, 22nd SSC, NKU

Dirichlet Distributions

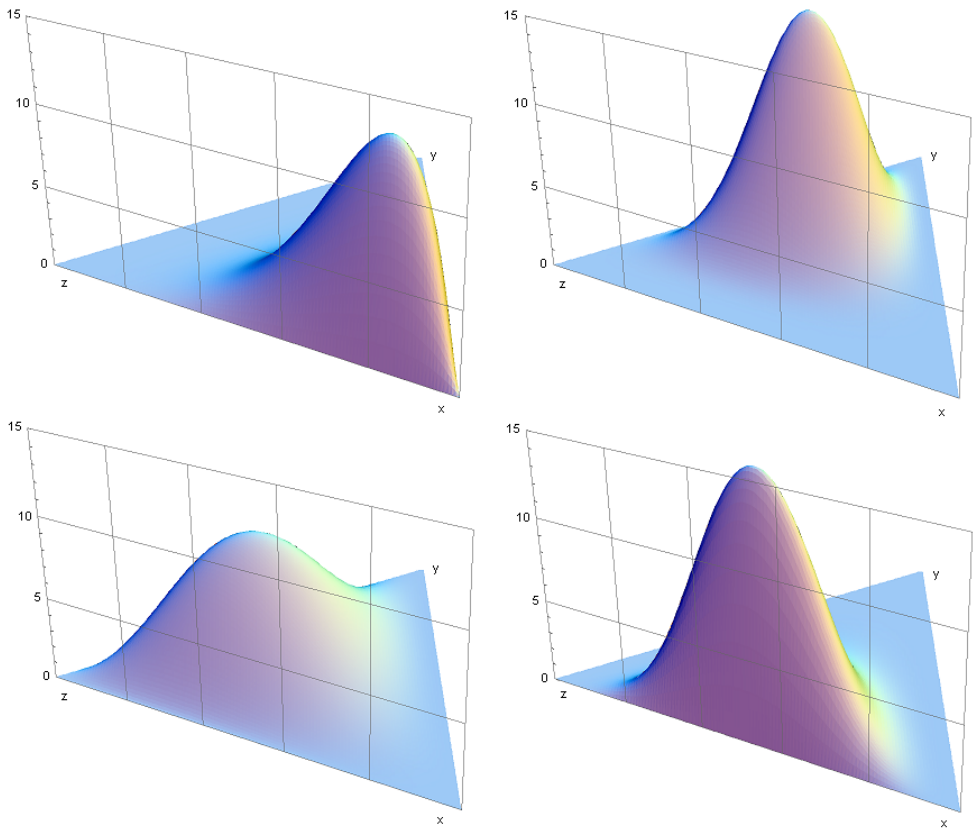
LDA makes central use of the Dirichlet distribution, the exponential family distribution over the simplex of positive vectors that sum to one. The Dirichlet has density

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}. \tag{1}$$

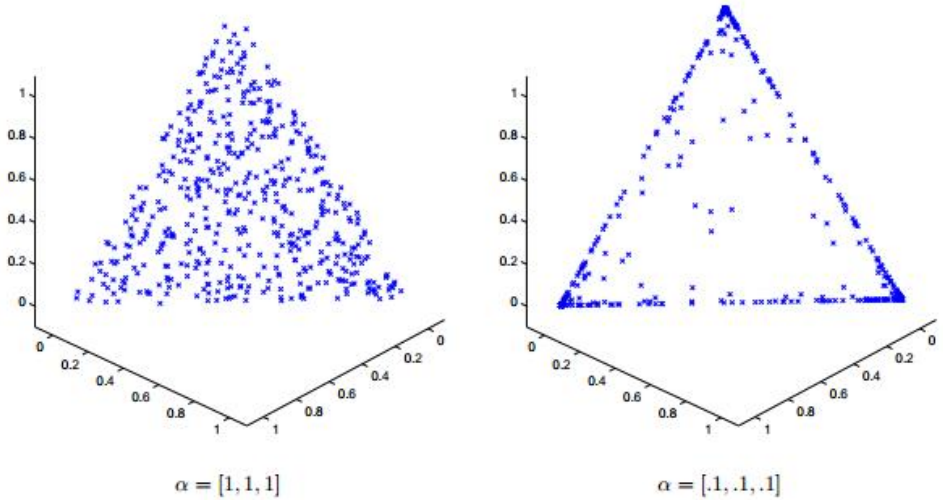
The parameter $\vec{\alpha}$ is a positive K-vector, and Γ denotes the Gamma function, which can be thought of as a real-valued extension of the factorial function. A *symmetric Dirichlet* is a Dirichlet where each component of the parameter is equal to the same value. The Dirichlet is used as a distribution over discrete distributions; each component in the random vector is the probability of drawing the item associated with that component.

Dirichlet Distributions : Prob. Density plots, K=3

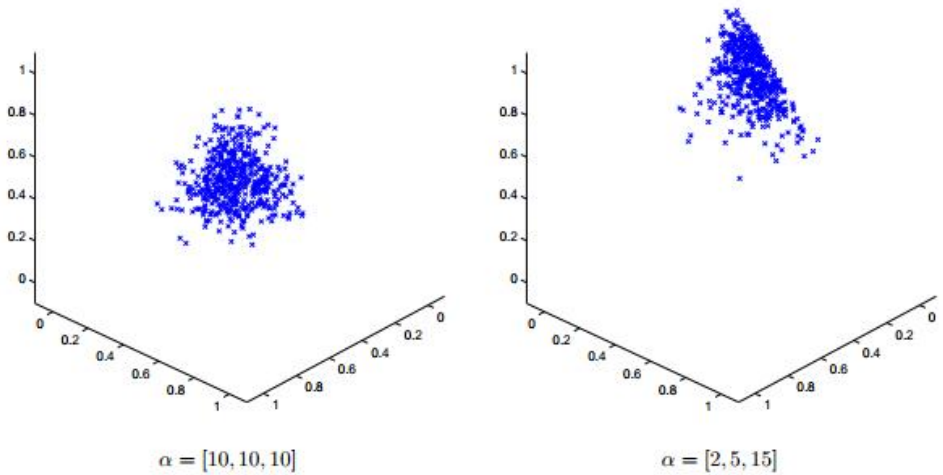
- Clockwise from top left: $\alpha = (6, 2, 2)(3, 7, 5)(6, 2, 6)(2, 3, 4)$.



Samples drawn from Dirichlet distributions : K=3

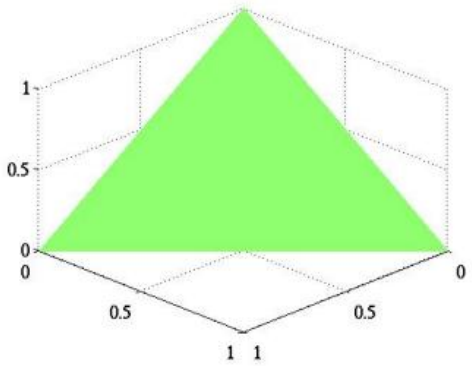


$\alpha = [1, 1, 1]$ $\alpha = [.1, .1, .1]$

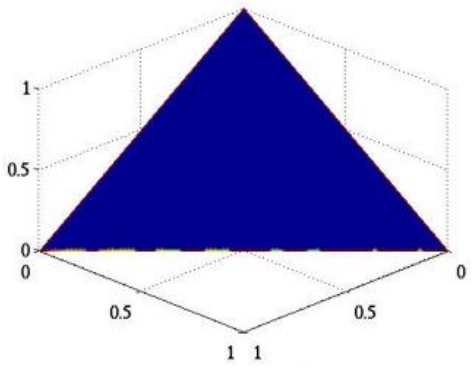


$\alpha = [10, 10, 10]$ $\alpha = [2, 5, 15]$

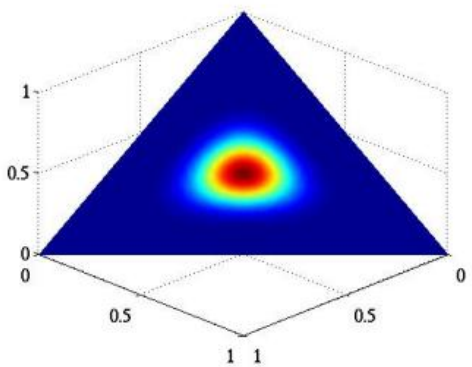
Density plots (blue = low, red = high): K=3



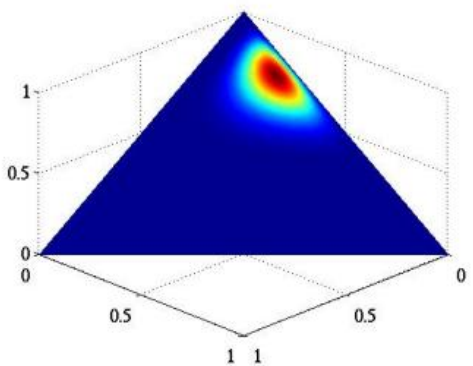
$\alpha = [1, 1, 1]$



$\alpha = [.1, .1, .1]$

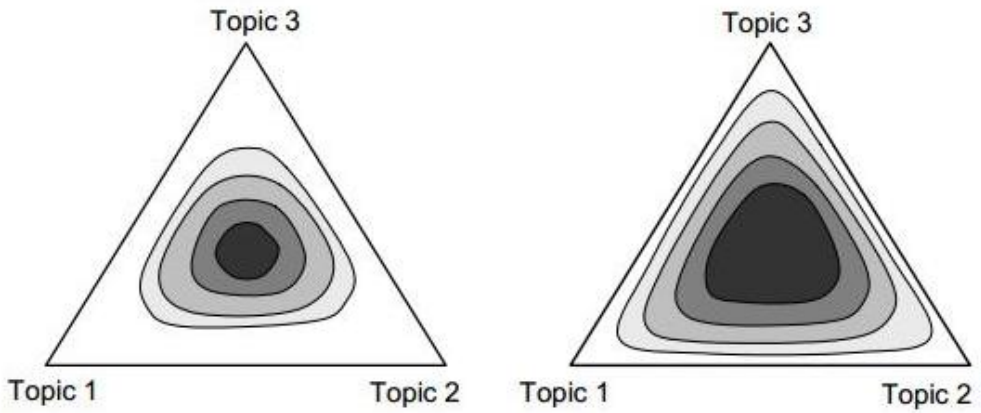


$\alpha = [10, 10, 10]$



$\alpha = [2, 5, 15]$

Symmetric Dirichlet Prior : K=3

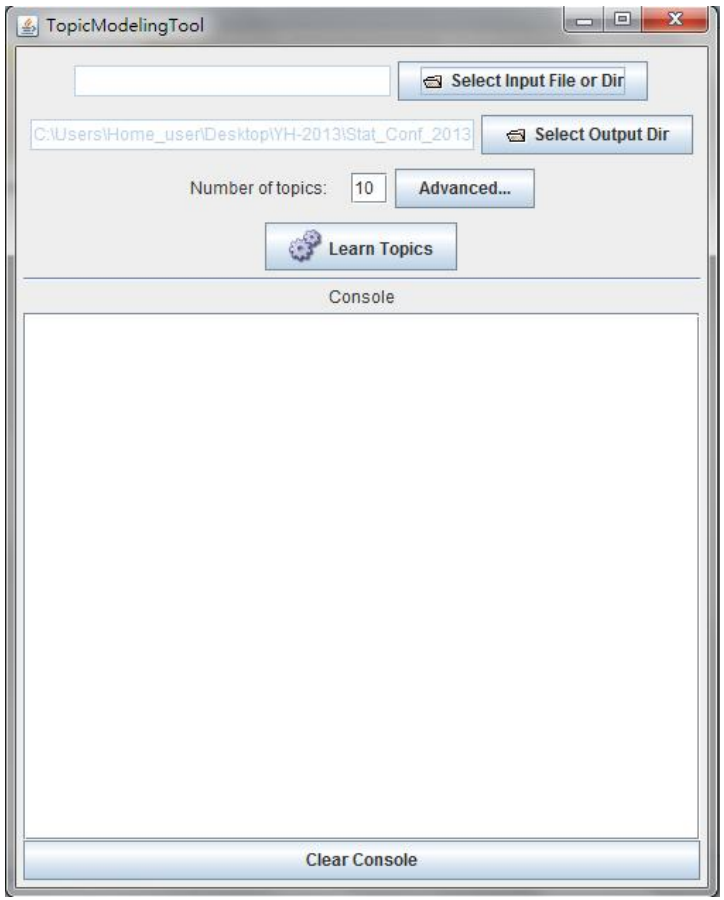


- Symmetric Dirichlet Distributions : left $\alpha = 4$, right: $\alpha = 2$

Machine Learning application for fitting topic models: MALLET

Topic Modeling Tool

- A Java-based package for fitting topic models.



LDA Output:

Four (out of 300) topics extracted from TASA corpus :

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure shows four example topics that were derived from the TASA corpus. The words in these topics relate to drug use, colors, memory and the mind, and doctor visits. Documents with different content can be generated by choosing different distributions over topics. For example, by giving equal probability to the first two topics, one could construct a document about a person that has taken too many drugs, and how that affected color perception. By giving equal probability to the last two topics, one could construct a document about a person who experienced a loss of memory, which required a visit to the doctor.

EX: 2014 Travellers' Chioce

Data : The file of UGC on TripAdvisor.com for Taiwan Top 25 Hotels includes 25 documents, 218K words, 1.1K unique terms (stop words and rare words removed).

Topic Models : Using example for illustration (0)

研究方法及步驟：R 軟體程式設計將文件資料轉換建立用於 LSI 之字詞-文件矩陣。

研究方法：使用 R 軟體相關套件，及撰寫耙梳線上數位資料程式碼。

研究步驟：今假設先將待檢索的文件集中放置程式根目錄，透過本文所設計之程式進行讀檔輸出（程式含關鍵詞判斷、stop word 判斷、一詞多義、多詞一義）自動輸出字詞-文件矩陣。如下圖：

輸出為一字詞-文件矩陣，後續透過文字探勘與文字雲圖將可得專題研究所需訊息。

Topic Models : Using example for illustration (1)

Topic Models : Using example for illustration (2)

A、以中文字分析預期成果（Hotel 中譯為飯店或酒店，星級飯店大多以酒店命名。 ）

(1)關鍵字關係分析（預期:檢視與**酒店**有關的其他關鍵字，如:**房間**關係達 0.56）

	酒店
房間	0.56
大堂	0.53
前台	0.51
美味	0.46
浴室	0.45

表一：與【酒店】關鍵字關係大於 0.40 之字詞組

由中文字詞矩陣中發現與【酒店】相關程度達 0.40 以上的關鍵字有：房間、大堂、前台、美味、浴室等。可見旅客的用後評論資料中對於酒店最在意的話題在於【房間】的議題，其次為【大堂】、【前台】指的就是大廳與接待櫃台，換句話說為在意剛進入酒店時的感覺，包含被服務的感覺及大廳的設計等。

Topic Models : Using example for illustration (2)

(2)文字雲:關鍵字關係可視圖

圖一：旅客用後評論資料之文字雲圖

由圖一藉由視覺化了解，整個旅客用後評論資料以【酒店】、【房間】為主要關鍵字，其中伴隨有討論：飯店內設備(馬桶、泳池、電視)、景觀、網路及交通等。議題。

Topic Models : Using example for illustration (2)

台灣 Top 25 最佳飯店旅客用後評論資料 List of Topics

Topic 1	服務員 笑容 心情 貴賓 熱忱 品牌 態度 ...
Topic 2	飲料 美食 菜色 主菜 牛肉 蛋糕 零食...
Topic 3	味道 房間 異味 細節 套房 備品 設備...
Topic 4	美景 風景 風格 景觀 環境 養生 隱私...
Topic 5	出租車 車子 地鐵 司機 車站 機場 計程車...
Topic 6	色調 噪音 綠色 擺設 空調 冰箱 拖鞋...
Topic 7	婚禮 婚宴 生日 時尚 氛圍 美容 SPA ...
Topic 8	旅客 媽媽 全家 小孩 家人 朋友 長輩...
Topic 9	健身房 泳池 窗戶 窗簾 電梯 陽台 酒吧...
Topic 10	電影 電話 書桌 網路 電視 電腦 音響...
Topic 11	浴室 浴衣 廁所 浴缸 馬桶 毛巾 沙發...
Topic 12	機場 車站 溫泉 便利商店 地點印象 夜市 夜景...

Reference

1. David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.
2. Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In Neural Information Processing Systems.
3. David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In Proceedings of International Conference of Machine Learning.
4. Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In Proceedings of Emperical Methods in Natural Language Processing.
5. Jonathan Chang. 2010. Not-so-latent dirichlet allocation: Collapsed gibbs sampling using human judgments. In NAACL Workshop: Creating Speech and Language Data With Amazon'ss Mechanical Turk.