Regresión Logística y Naïve Bayes

Denis Parra

Curso Exploratorio de Computación – IIC 1005

2018

PLAN SEMESTRAL

Week	Fecha semana	Clase Martes	Clase Jueves	Presentador	ayudante	Ayudantía	Control	Tarea
I	6 - 8 Mar	Introduccion+terminal	Github+Jupyter					
II	13 - 15 Mar	Leng. Prog + Jupyter 2	Visualizacion + HCI		Dan + Vi	Jupyter Pandas		
III	20 - 22 Mar	Tecn Web HTML + CSS	Tecn Web JS		Vi + Dan	Jupyter Plots		tc1 Git+Shell
IV	27 - 29 Mar	Arquitectura	SO+Redes	HL + CR	Dal + FI	Web		
V	3 - 5 Abr	BD	BD	AS	FI + Dal +ViD	Web		TG1 Jupyter + We
VI	10 - 12 Abr	Algoritmos	Ingenieria de Sotware	YE + JN	--	--	I1: 13Abr Web/HCI	
VII	17 - 19 Abr	ML	ML		Dal + FI	SQL		tc2 BD (SQL)
VIII	24 - 26 Abr	ML	ML		Antonio	ML		
IX	3 may.	FERIADO	Guest: DL	HL	Antonio	Sala de Ayuda ML		TG2 ML
X	8 - 10 May	Computabilidad	Complejidad	CRi	*No hay*	*No hay*		
XI	15 - 17 May	Prog Logica	Prog Logica	JB	Vicho	Turing	I2: 16May IngSoft	
XII	22 - 24 Ma	BPM	BPM	MS	Vicho	Sala de Ayuda TC3		tc3 Maq de Turing
XIII	29 - 31 Ma	Guest: Criptomonedas	Guest: TBA	CR - ??	Dal +	BPM		
XIV	5 - 7 Jun	Guest: CSCW	Guest: MOOC	VH - Mar	Dal +	BPM + Prolog		tc4 BPM
XV	12 -14 Jun	Guest: Miguel Nussb.	Guest: Mobile & Cloud	MN - AN		Prolog	I3: 14Jun ML+IA	
XVI	19 - 21 Jun	Resumen Final						

Resumen

- Inteligencia de Máquina: **Aprender de los Datos**
- Revisemos de modo conceptual un ejemplo: Construir para un banco un sistema que automáticamente apruebe o niegue crédito

Applicant information:

Rechazar o Aprobar credito??

age	23 years
gender	male
annual salary	\$30,000
years in residence	1 year
years in job	1 year
current debt	\$15,000
...	...

Formalización del Problema de Aprendizaje

- Encontrar la formula de aprobación *g* que se aproxime lo más posible a la formula ideal *f*

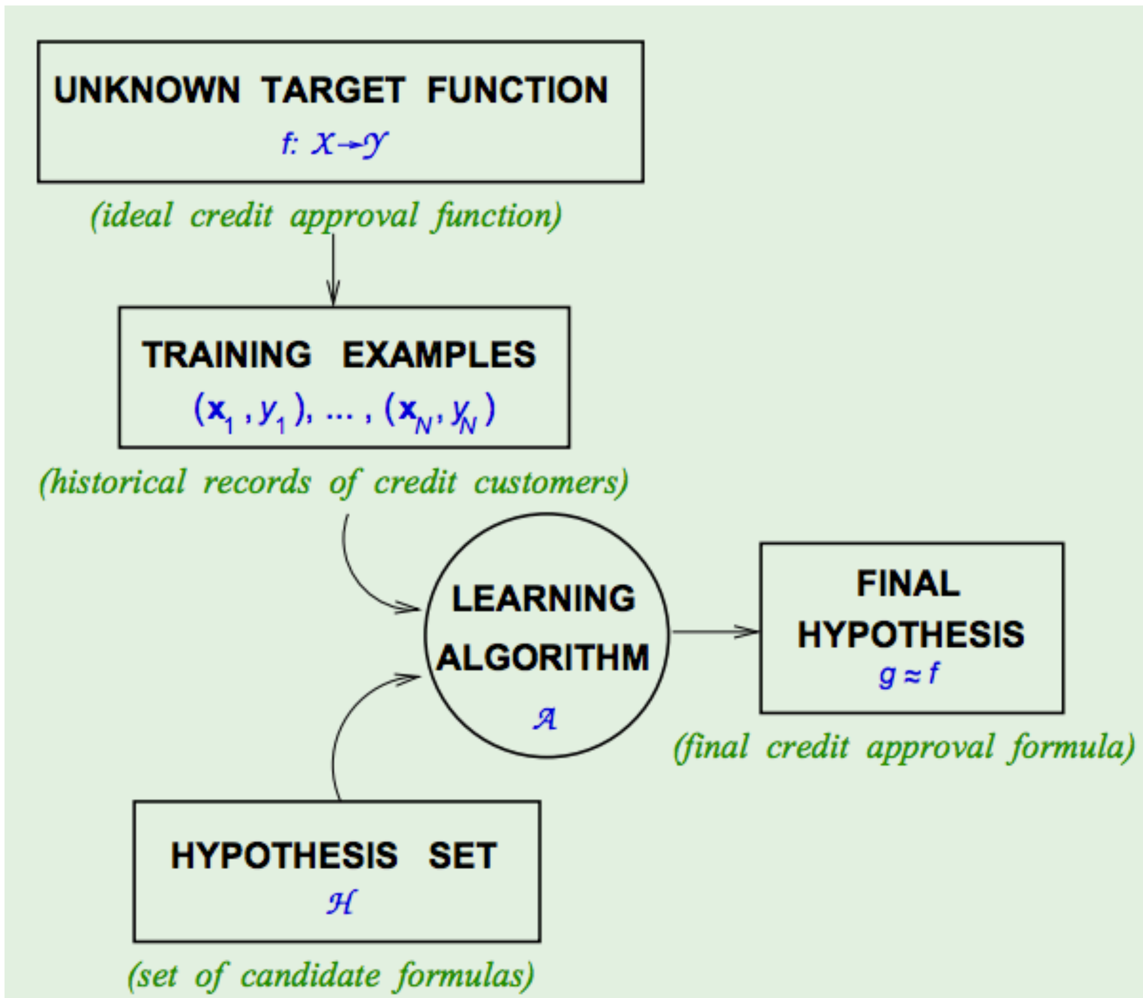
- Input: $\mathbf{x}$ *(customer application)*
- Output: y *(good/bad customer?)*
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ *(ideal credit approval formula)*
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ *(historical records)*

$$\downarrow \quad \downarrow \quad \downarrow$$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ *(formula to be used)*

Formalizacion del Problema de Aprendizaje

- El conjunto de hipótesis H



Motivación

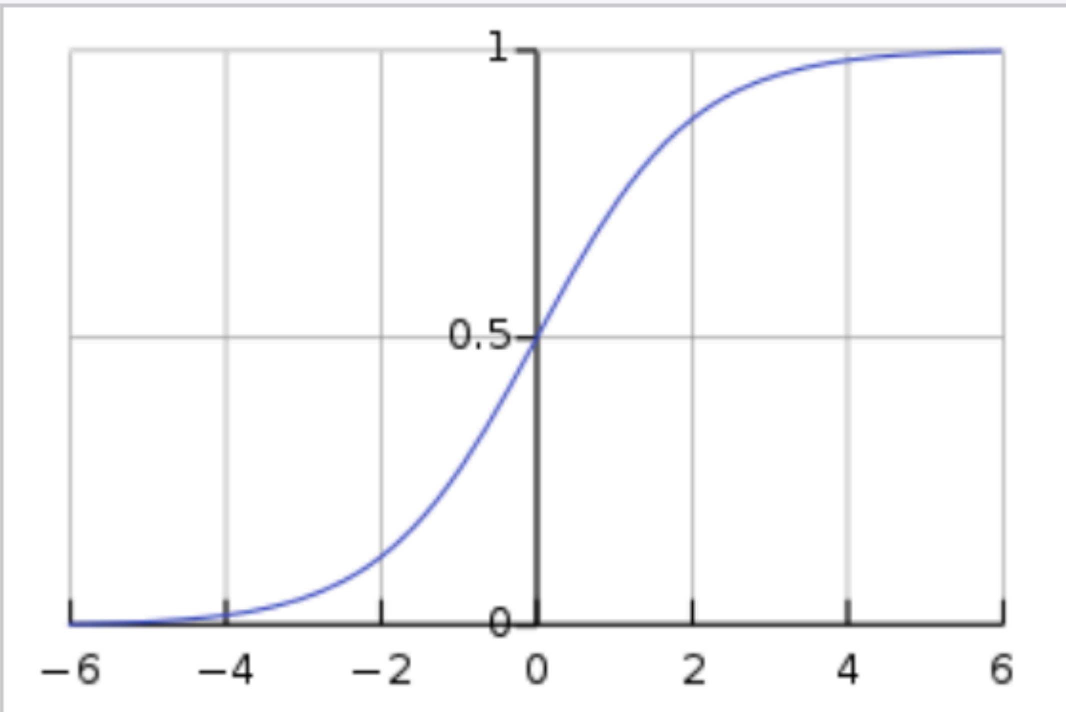
- Nos interesa usar un modelo de regresión en que la variable dependiente es binaria (1 ó 0) o multinomial (espacio finito de valores).
- Para la tarea, Y = 1 significa que el usuario es vendedor

$$P(Y=1)=\beta_0+\beta_1 x_1+\ldots+\beta_k x_k$$

Motivación

- Sin embargo, considerando que en una regresión lineal múltiple la variable dependiente “Y” podría tomar cualquier valor en un rango continuo, hacemos uso de la **función logística** para hacer una transformación

Función logística



Función logística con $\beta_0 + \beta_1 x + e$ en el eje horizontal y $\pi(x)$ en el eje vertical.

$$y = \frac{1}{1 + e^{-f(X)}}$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

Modelo de regresión logística

- De esta forma, el modelo logístico nos queda:

$$\log\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$$

- donde P = P(Y = 1 | x_1, .., x_p), es decir, la probabilidad de que el evento Y ocurra (dadas las covariables x_1, .., x_p). La expresión es equivalente a:

$$\Pr(Y=1 \mid x_1, x_2, \cdots, x_p) = \frac{1}{1 + \exp(-\alpha - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p)}$$

Tarea 2: Regresión Logística para Predicción

- Con el objetivo de simplificar algunos temas como los supuestos del modelo y cómo obtener los valores de β, usaremos el software scikit-learn para "obtener" los valores de β. Quedan pendientes:
  - ¿Cómo interpretar estos valores?
  - ¿Cómo evaluar qué tan buena es la predicción del modelo?

Interpretar valores de β

- β_j es la cantidad de cambio en logit por cada unidad que cambia X_j
- A diferencia de regresión lineal, los β_j no se interpretan directamente, sino que $\exp(\beta_j)$
- $\exp(\beta_j)$: odds ratio (razón de disparidad)
  - $\exp(\beta_j) = 1$, no hay cambio en los odds ratio
  - $\exp(\beta_j) < 1$ odds ratio decrece
  - $\exp(\beta_j) > 1$ odds ratio crece
  - $\exp(\beta_0)$: baseline

Interpretar valores de β

- Ejemplo: en la tarea, si $\exp(\beta_j) = 1.2$ y X_j es "largo de la pagina en letras", por cada letra adicional mi chance (odds) de ser *phishy* es 20% mayor.

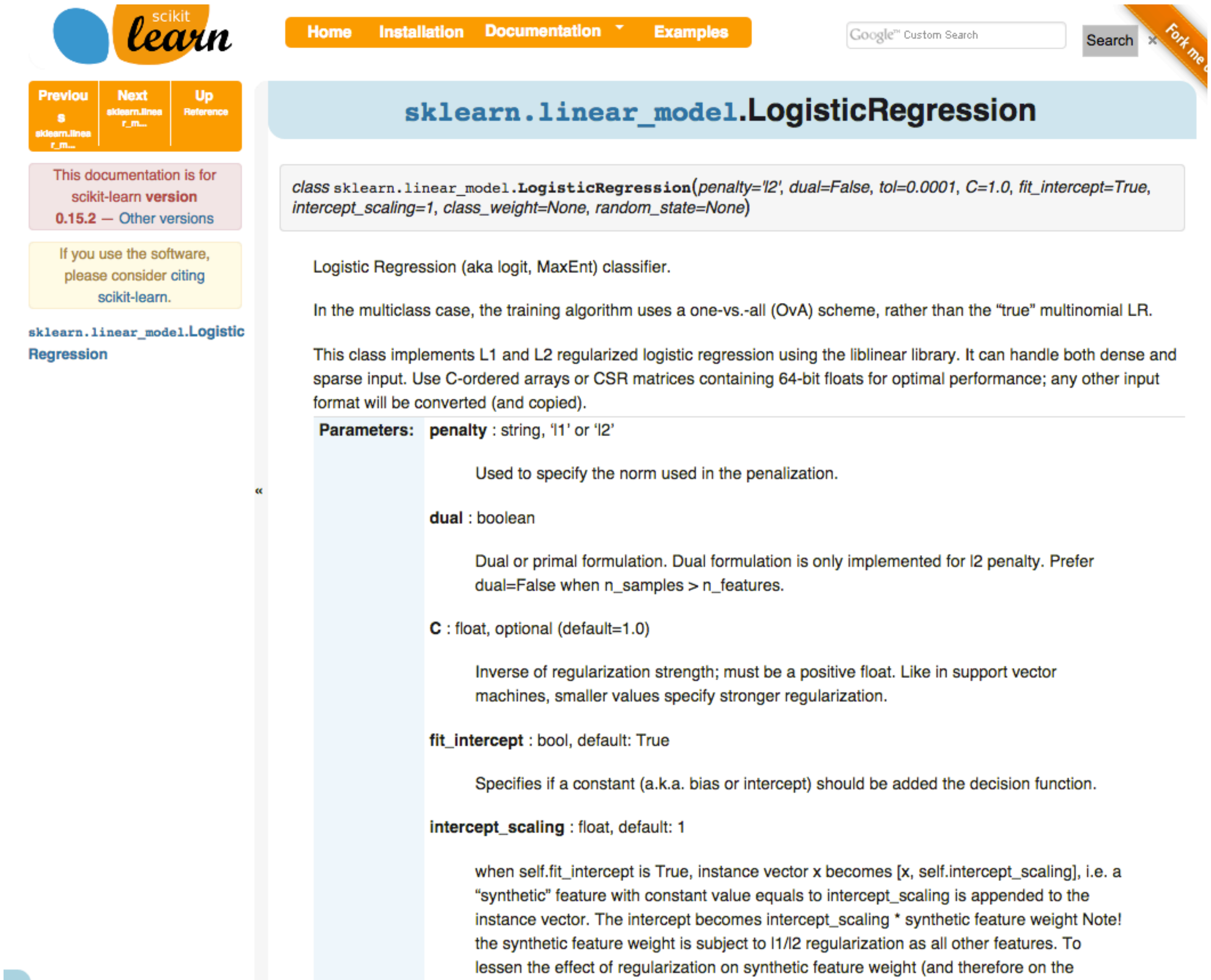
Si tomamos un valor de ejemplo, digamos $p(50) = 2/3$, entonces

$$\frac{p(50)}{1-p(50)} = \frac{\frac{2}{3}}{1-\frac{2}{3}} = 2.$$

- Recomiendo revisar: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm

sklearn

- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html



sklearn.linear_model.LogisticRegression

class `sklearn.linear_model.LogisticRegression`(*penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None*)

Logistic Regression (aka logit, MaxEnt) classifier.

In the multiclass case, the training algorithm uses a one-vs.-all (OvA) scheme, rather than the "true" multinomial LR.

This class implements L1 and L2 regularized logistic regression using the liblinear library. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied).

Parameters:

penalty : string, 'l1' or 'l2'

Used to specify the norm used in the penalization.

dual : boolean

Dual or primal formulation. Dual formulation is only implemented for l2 penalty. Prefer dual=False when n_samples > n_features.

C : float, optional (default=1.0)

Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.

fit_intercept : bool, default: True

Specifies if a constant (a.k.a. bias or intercept) should be added the decision function.

intercept_scaling : float, default: 1

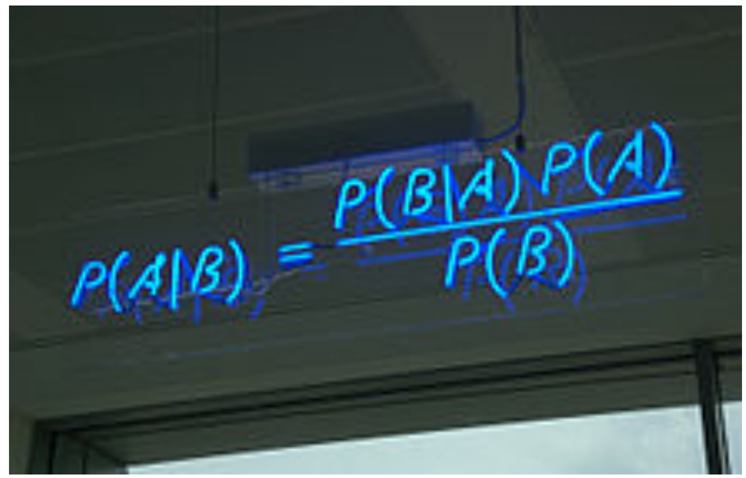
when self.fit_intercept is True, instance vector x becomes [x, self.intercept_scaling], i.e. a "synthetic" feature with constant value equals to intercept_scaling is appended to the instance vector. The intercept becomes intercept_scaling * synthetic feature weight Note! the synthetic feature weight is subject to l1/l2 regularization as all other features. To lessen the effect of regularization on synthetic feature weight (and therefore on the

Clasificador Bayesiano “Naive”

- Este modelo se basa en Probabilidades, particularmente usa el teorema de Bayes y más precisamente el supuesto “Naïve” de Independencia Condicional

Teorema de Bayes

Thomas Bayes (1701 – 1761)



Teorema de Bayes

- P(A = sí): Probabilidad del evento A sea "sí"
- P(A=sí|B=sí): Probabilidad de que el evento A sea "sí" DADO QUE el evento B fue "sí"
- Por simplicidad, usamos P(A) = P(A="sí")

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(A \mid B) = \frac{P(B|A) * P(A)}{P(B)}$$

Noción del Teorema de Bayes

- << La riqueza hace la felicidad >>
- ¿Son felices los ricos? P(feliz = sí | rico = sí)
- … yo sé que de la gente feliz, 20% es rica.

Supongamos:

A: gente feliz = 40% de la población

B: gente rica = 10% de la población

C: P(rico | feliz) = 20%

Noción del Teorema de Bayes

- << La riqueza hace la felicidad >>
- ¿Son felices los ricos? P(feliz = sí | rico = sí)
- … yo sé que de la gente feliz, 20% es rica.

Supongamos:

A: gente feliz = 40% de la población

B: gente rica = 10% de la población

C: P(rico | feliz) = 20%

$$P(feliz \mid rico) = \frac{P(rico|feliz) * P(feliz)}{P(rico)}$$

$$P(feliz \mid rico) = \frac{0.2 \ * 0.4}{0.1}$$

= 0.8 = **80%**

¿Y por qué se llama “Naïve” ?

- Naïve significa “**ingenuo**”
- Es “ingenuo” por que asume independencia de los eventos*
- * en realidad, asume independencia condicional

Independencia de Eventos

- Supongamos que lanzo una moneda al aire y quiero saber la probabilidad de que ocurra "cara"

$$P(Cara = 1) = P(Cara)$$

- ¿Y si quiero calcular la probabilidad de que ocurra dos veces seguidas? Si son Independientes...

$$\Rightarrow P(Cara=1, Cara=1) = P(Cara = 1) * P(Cara = 1)$$

- Si no asumo independencia:

$$=> P(Cara\_2=1, Cara\_1=1) = P(Cara\_2=1 | Cara\_1=1) * P(Cara\_1=1)$$

El segundo lanzamiento depende del primero

Independencia Condicional

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

- ¿Probabilidad de oír un trueno dado que llueve y ya hubo un rayo?
- …en realidad, basta con ver el rayo para saber si habrá un trueno

P(trueno = Si | lluvia = si, rayo=si) = P(trueno =Si | rayo = si)

- No necesito “llueve=sí”

Independencia Condicional

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

- ¿Probabilidad de oír un trueno dado que llueve y ya hubo un rayo?
- …en realidad, basta con ver el rayo para saber si habrá un trueno

P(trueno = Si | lluvia = si, rayo=si) = P(trueno =Si | rayo = si)

- No necesito “llueve=sí”

$$\begin{aligned} P(X|Y) &= P(X_1, X_2|Y) \\ &= P(X_1|X_2, Y) P(X_2|Y) \\ &= P(X_1|Y) P(X_2|Y) \end{aligned}$$

OK, pero ¿¿por qué es “Naïve”??

- Ejemplo: Clasificar una frase como SPAM
- P(spam=Si | <frase>) , y <frase> es “viagra feliz”. Usando Bayes

$$\frac{P(viagra=SI \mid spam=si) * P(feliz=SI \mid spam=si) * P(spam=si)}{P(viagra=SI, feliz=si)}$$

De la misma forma, podemos calcular

P(spam = No | <frase>)

Volviendo: Ejemplo de Clasificación

- Consideremos un auto SUV, color rojo, doméstico. ¿La probabilidad de que la roben es mayor o menor de que no la roben?

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Latent Dirichlet Allocation (LDA)

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.
- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Representaciones Distribucionales

- Podemos asumir un modelo en el cual la distribución de contextos en los cuales las palabras (consideradas átomos) aparecen, nos dice algo sobre esa palabra y sobre esos contextos.
- Contexto: puede ser palabras adyacentes, trozos de discursos, o simplemente … Documentos.
- Si dos palabras aparecen en el mismo contexto (documento), podrían pertenecer a una misma clase semántica.

Latent Dirichlet Allocation

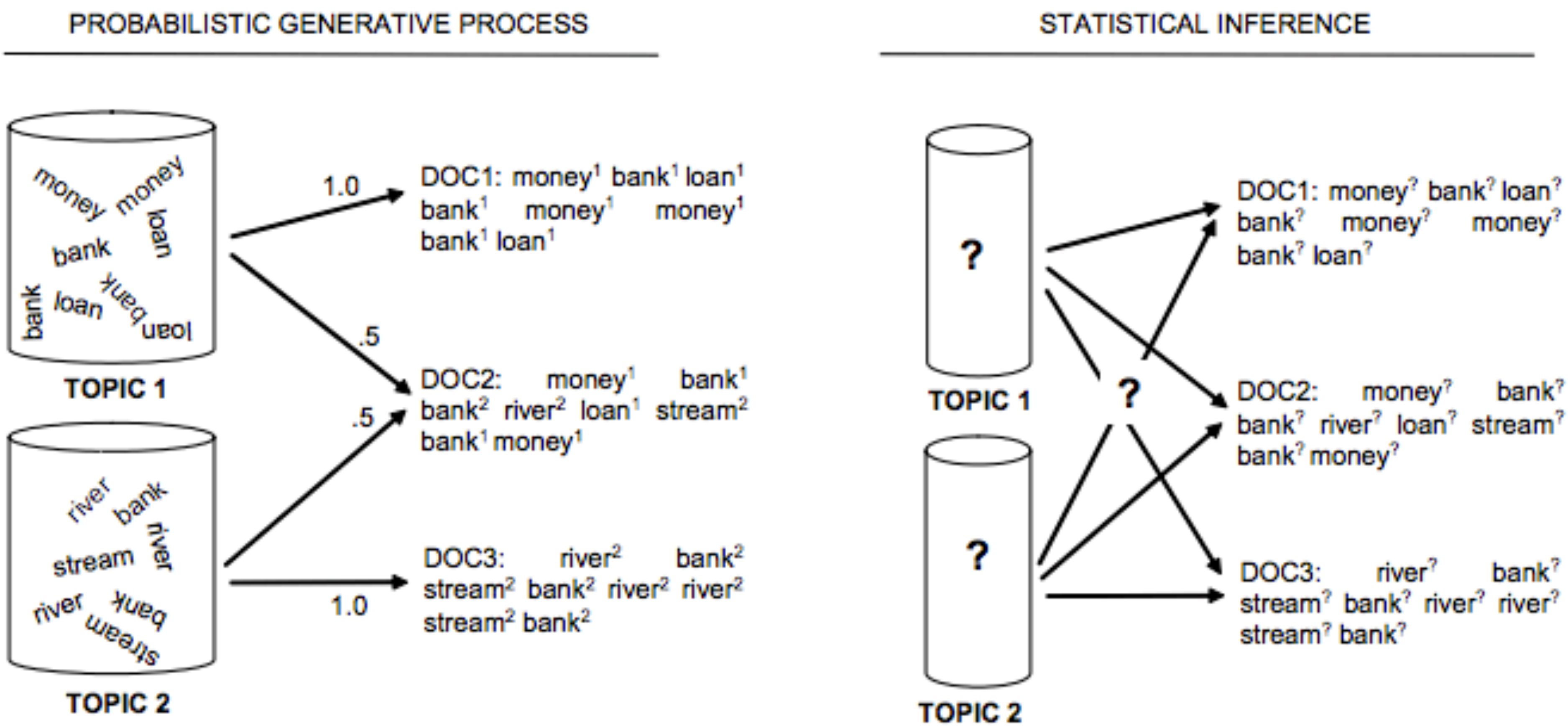
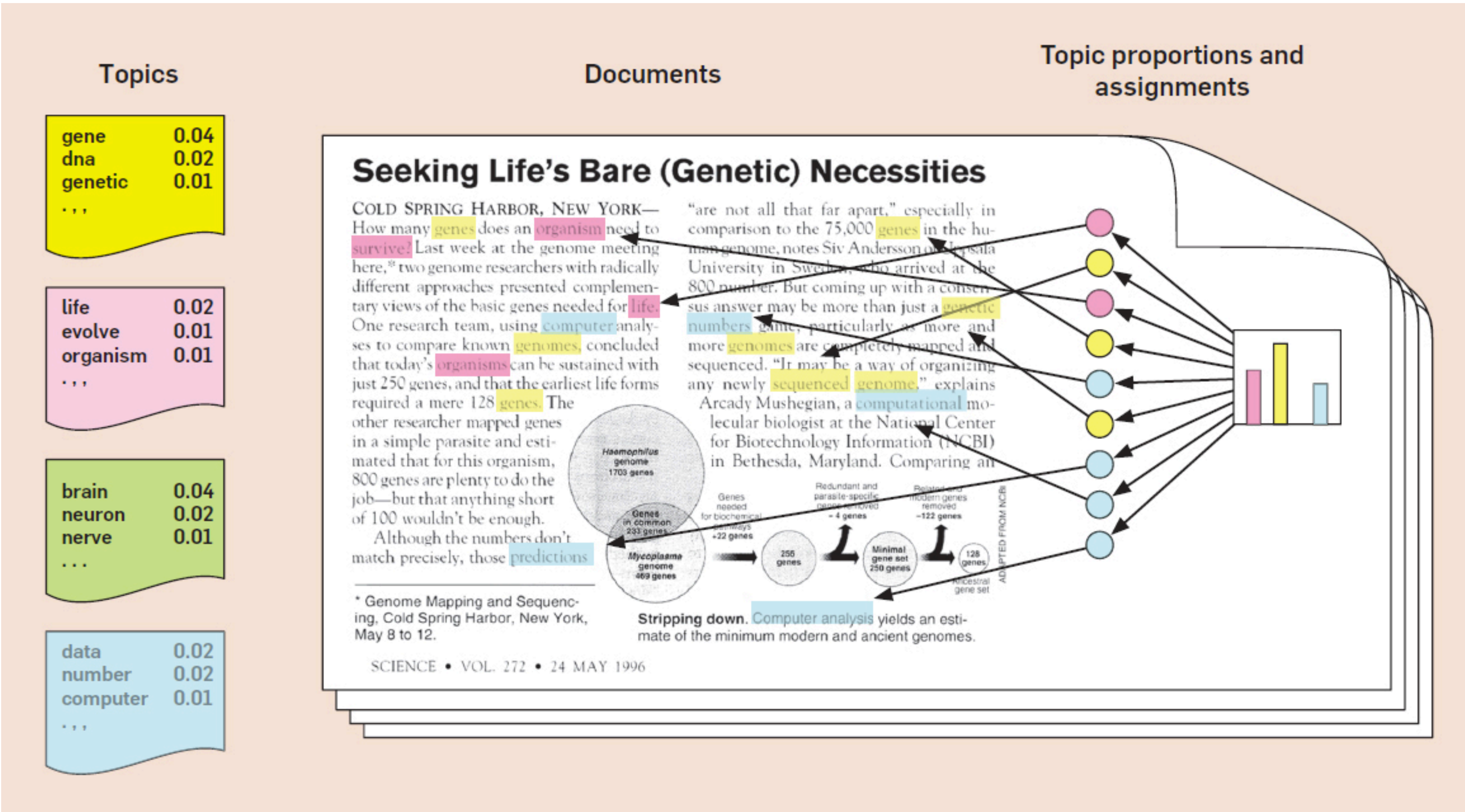


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

Steyvers & Griffiths, 2007

LDA (Blei)



Blei, 2012

LDA III

- Plate notation

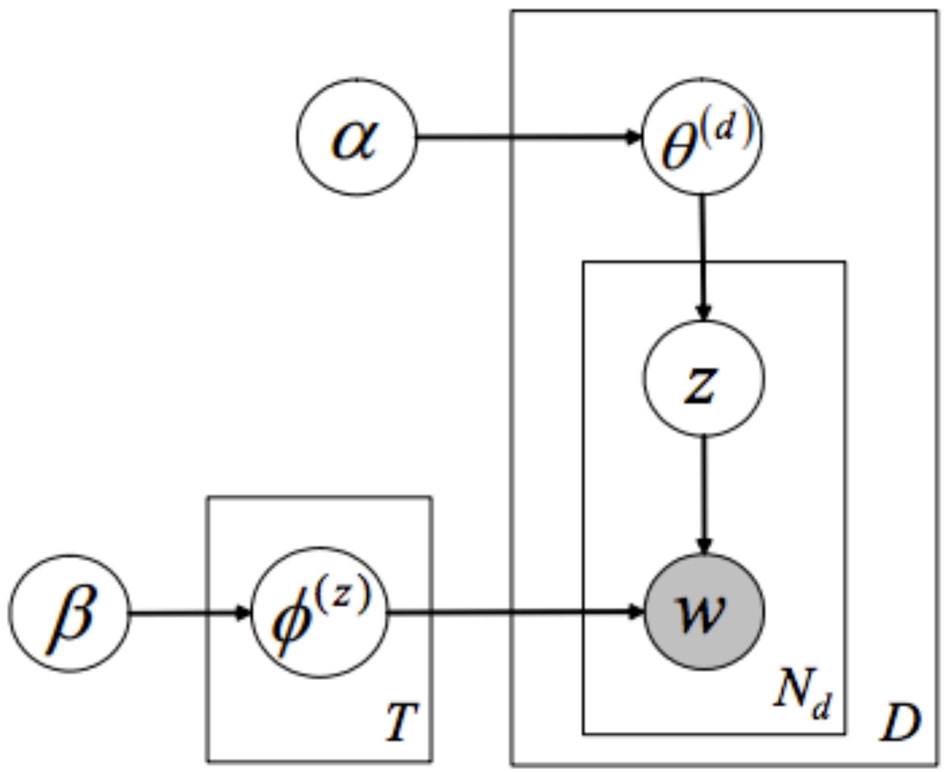
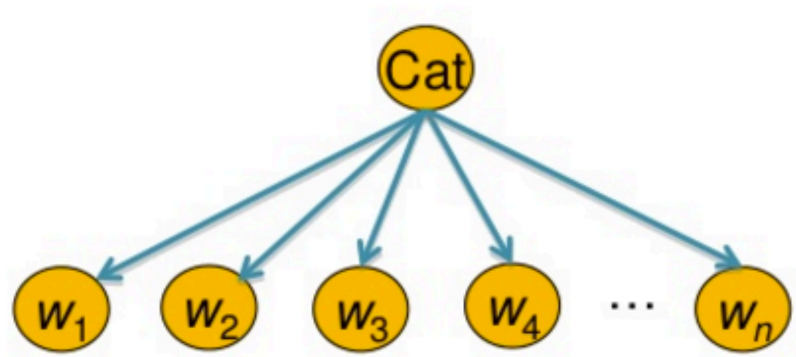


Figure 4. The graphical model for the topic model using plate notation.

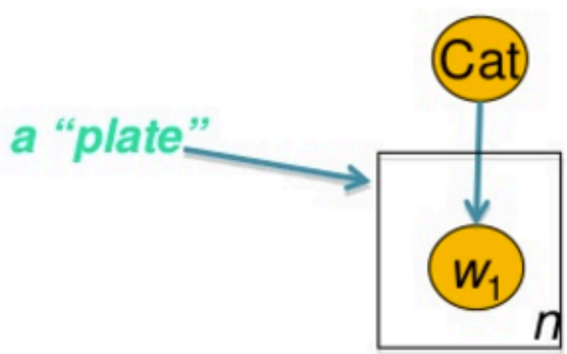
Steyvers & Griffiths, 2007

Modelos Gráficos

Graphical model representations



Compact notation:

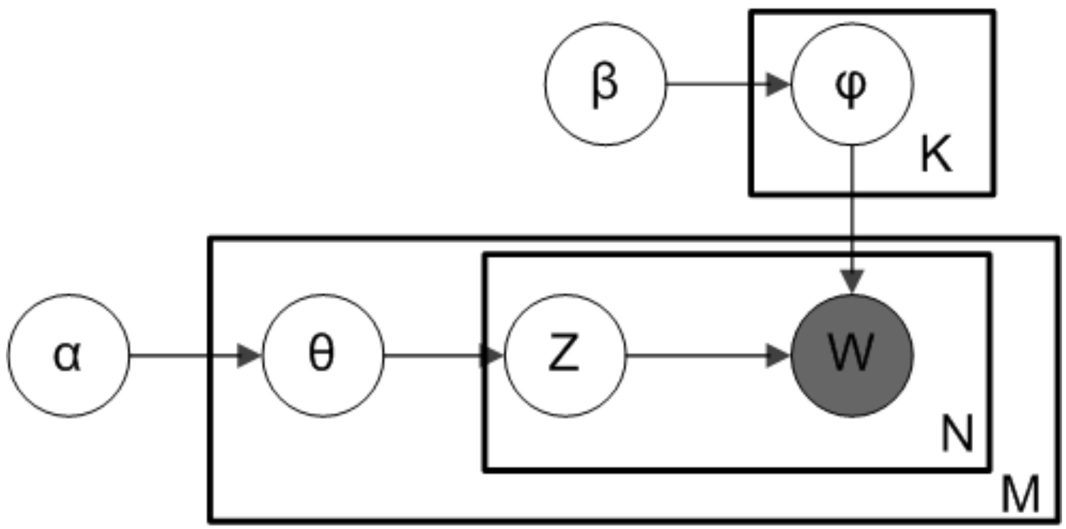


"generate a word from Cat *n* times"

*Original slide by Roger Levy, UCSD

Documents in Latent Space - LDA

- Latent Dirichlet Allocation: extends pLSI by adding two Dirichlet priors.



α is the parameter of the Dirichlet prior on the per-document topic distributions.
β is the parameter of the Dirichlet prior on the per-topic word distribution.
θ_i is the topic distribution for document i,
ϕ_k is the word distribution for topic k,
z_{ij} is the topic for the jth word in document i, and
wij is a specific word in the document

LDA II

Document #29795

Bix beiderbecke, at age^{060} fifteen^{207}, sat^{174} on the slope^{071} of a bluff^{055} overlooking^{027} the mississippi^{137} river^{137}. He was listening^{077} to music^{077} coming^{009} from a passing^{043} riverboat. The music^{077} had already captured^{006} his heart^{157} as well as his ear^{119}. It was jazz^{077}. Bix beiderbecke had already had music^{077} lessons^{077}. He showed^{002} promise^{134} on the piano^{077}, and his parents^{035} hoped^{268} he might consider^{118} becoming a concert^{077} pianist^{077}. But bix was interested^{268} in another kind^{050} of music^{077}. He wanted^{268} to play^{077} the cornet. And he wanted^{268} to play^{077} jazz^{077}...

Document #1883

There is a simple^{050} reason^{106} why there are so few periods^{078} of really great theater^{082} in our whole western^{046} world. Too many things^{300} have to come right at the very same time. The dramatists must have the right actors^{082}, the actors^{082} must have the right playhouses, the playhouses must have the right audiences^{082}. We must remember^{288} that plays^{082} exist^{143} to be performed^{077}, not merely^{050} to be read^{254}. (even when you read^{254} a play^{082} to yourself, try^{288} to perform^{062} it, to put^{174} it on a stage^{078}, as you go along.) as soon^{028} as a play^{082} has to be performed^{082}, then some kind^{126} of theatrical^{082}...

Document #21359

Jim^{296} has a game^{166} book^{254}. Jim^{296} reads^{254} the book^{254}. Jim^{296} sees^{081} a game^{166} for one. Jim^{296} plays^{166} the game^{166}. Jim^{296} likes^{081} the game^{166} for one. The game^{166} book^{254} helps^{081} jim^{296}. Don^{180} comes^{040} into the house^{038}. Don^{180} and jim^{296} read^{254} the game^{166} book^{254}. The boys^{020} see a game^{166} for two. The two boys^{020} play^{166} the game^{166}. The boys^{020} play^{166} the game^{166} for two. The boys^{020} like the game^{166}. Meg^{282} comes^{040} into the house^{282}. Meg^{282} and don^{180} and jim^{296} read^{254} the book^{254}. They see a game^{166} for three. Meg^{282} and don^{180} and jim^{296} play^{166} the game^{166}. They play^{166}...

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

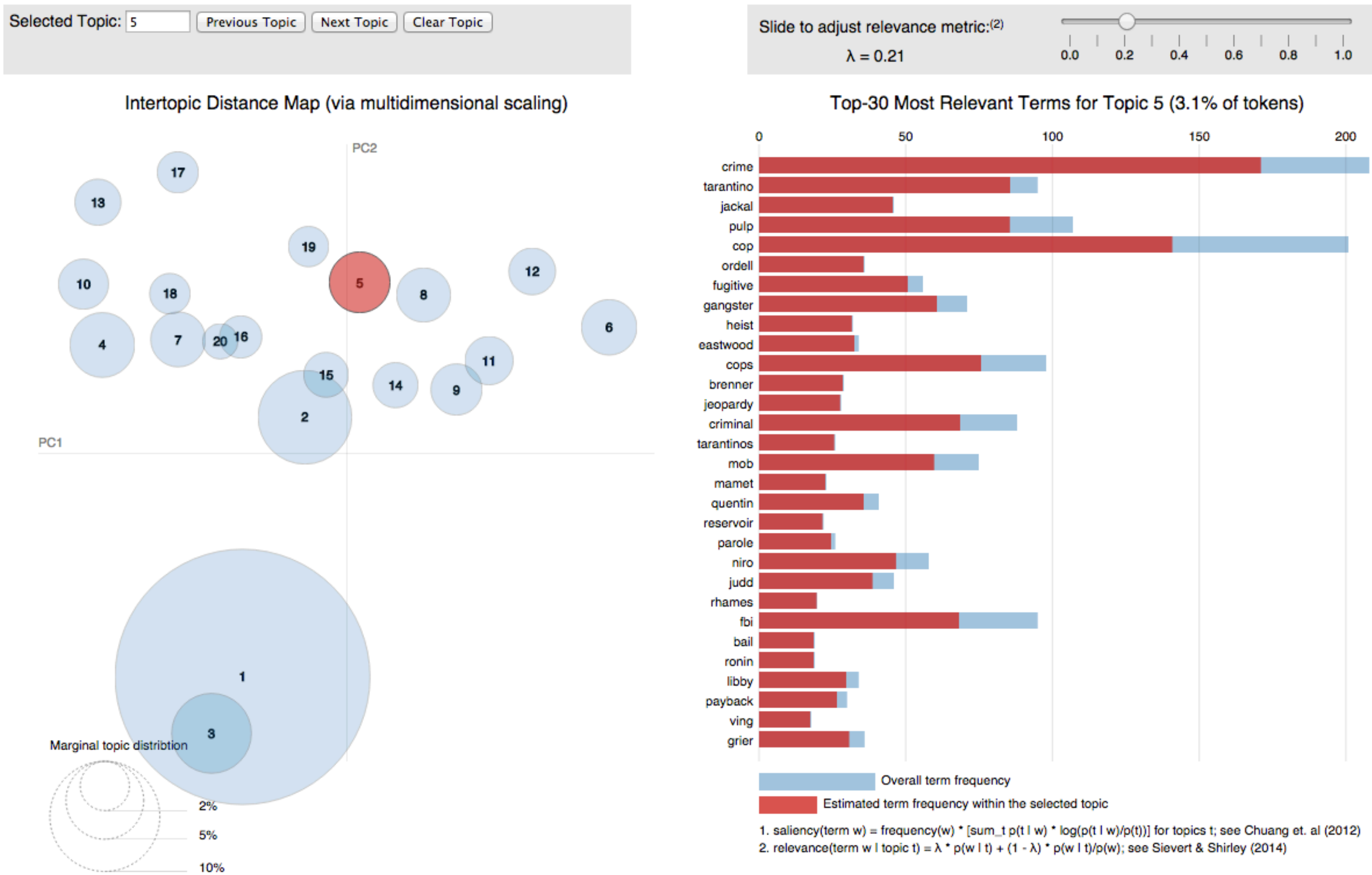
word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

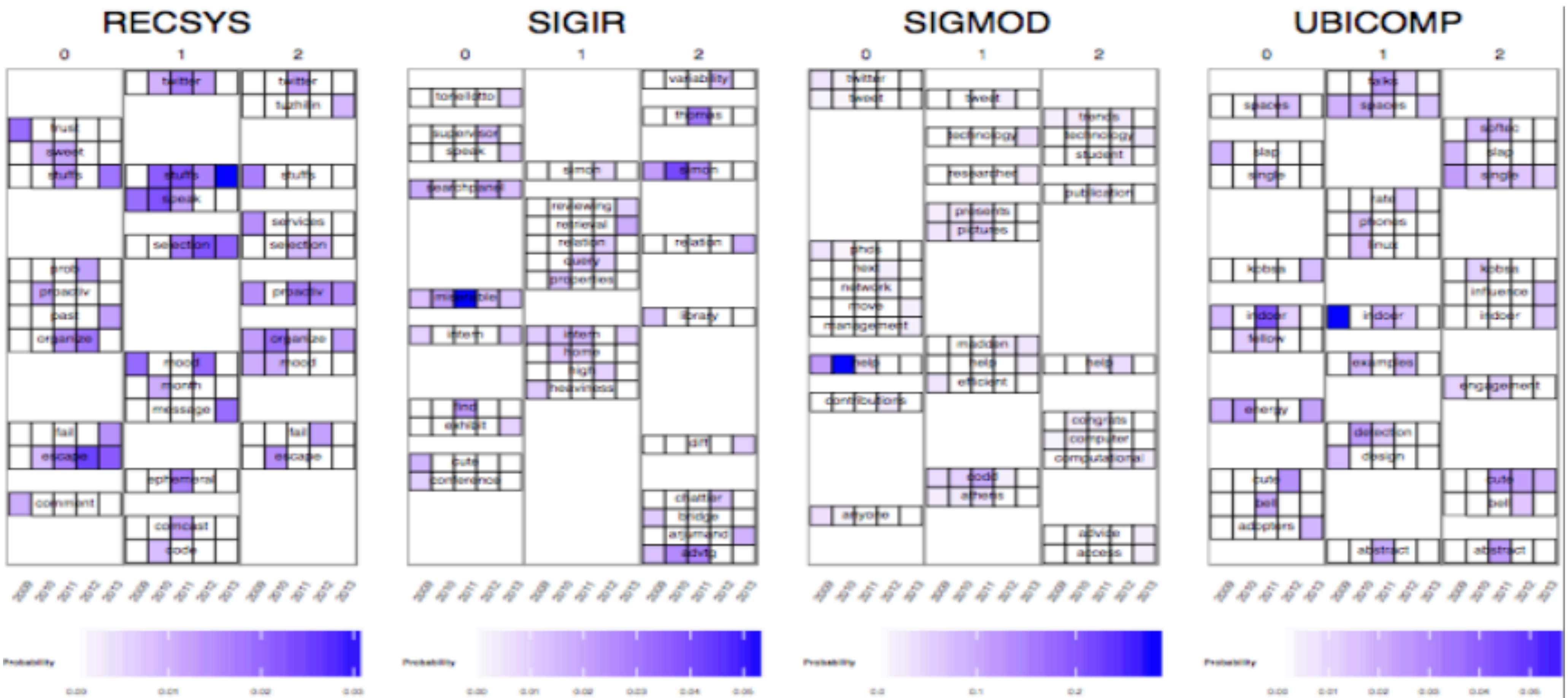
Steyvers & Griffiths, 2007

LDAvis


Selected Topic: 5
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.21
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Marginal topic distribtion
2%
5%
10%
Top-30 Most Relevant Terms for Topic 5 (3.1% of tokens)
0
50
100
150
200
crime
tarantino
jackal
pulp
cop
ordell
fugitive
gangster
heist
eastwood
cops
brenner
jeopardy
criminal
tarantinos
mob
mamet
quentin
reservoir
parole
niro
judd
rhames
fbi
bail
ronin
libby
payback
ving
grier
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Otras extensiones

- Tópicos a lo largo del tiempo (DTM)



http://socialcomputing.ing.puc.cl/uploads/DynamicTopicModellingTutorial.pdf

Gracias!

- Consultas a:

Denis Parra

dparra@ing.puc.cl