Clasificación de texto

Reconocimiento de patrones

Prof. Mauricio Cerda
mauriciocerda@med.uchile.cl
http://www.scian.cl

Programa de Anatomía y Biología del Desarrollo
I.C.B.M., Facultad de Medicina, Universidad de Chile

- Santiago 13/11/2017 -

Outline

- Pendientes semana pasada
- Clasificadores Bayesianos (BoW)
- Vectorización de texto y otras técnicas

Motivación

- Clasificación de texto: ¿es spam o no?, ¿opinión positiva o negativa? ¿tema de un artículo?
- ¿Cómo representar texto en un espacio vectorial?

Motivación

¿Es este correo spam?

13920 mauriciocerda Papelera x

Equipo de Gmail <mail-noreply@google.com> 16:24 (Hace 6 horas.)

para mí

El mensaje "13920 mauriciocerda" de jonas.karlsson@sparktrade.se contenía un virus o un archivo adjunto sospechoso. Por lo tanto, no se pudo obtener desde tu cuenta mauriciocerda@med.uchile.cl y fue dejado en el servidor.

ID del mensaje: <149409603130.31457.2930160597399828074@muravlenko.ru>

Si deseas escribirle a esta persona, selecciona responder y envía un mensaje.

Gracias.

El equipo de Gmail

Motivación

¿Comentarios positivos o negativos de una película?

- Increíblemente desilusionante

- Llena de caracteres coloridos, sátira inteligente, y tramas sorprendentes.

- La mejor comedia para niños nunca filmada

- Fue patética. La peor parte fueron las escenas de acción.

Motivación

¿Tema de un artículo?

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Categorización MeSH

?

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology

Definiciones

- Entrada:
  - Un documento d
  - Un conjunto de clases posibles $C = \{c_1, ..., c_n\}$

- Salida: una predicción de la clase c_i

Aproximación basada en reglas

- Basadas en reglas sobre palabras o conjuntos de ellas (expresiones regulares)
  - *Ejemplos, ¿para el caso de spam?*
- Puede tener un buen desempeño, si se realiza por expertos.
- Construir y mantener estas reglas puede ser costoso

Aproximación basada en métodos de clasificación supervisados

- Entrada:
  - Un documento d
  - Un conjunto de n clases posibles $C = \{c_1, ..., c_n\}$
  - Un conjunto de m documentos ya clasificados $(d_1, c_1), ..., (d_m, c_m)$
- Salida: un clasificador γ entrenado (con una predicción de la clase c_i).

Metodos supervisados

- Cualquier de los métodos supervisados estudiados se podría usar (MLP, SVM, Random Forest)
- Uno de los primeros propuestos son los métodos probabilísticos basado en la regla de Bayes.

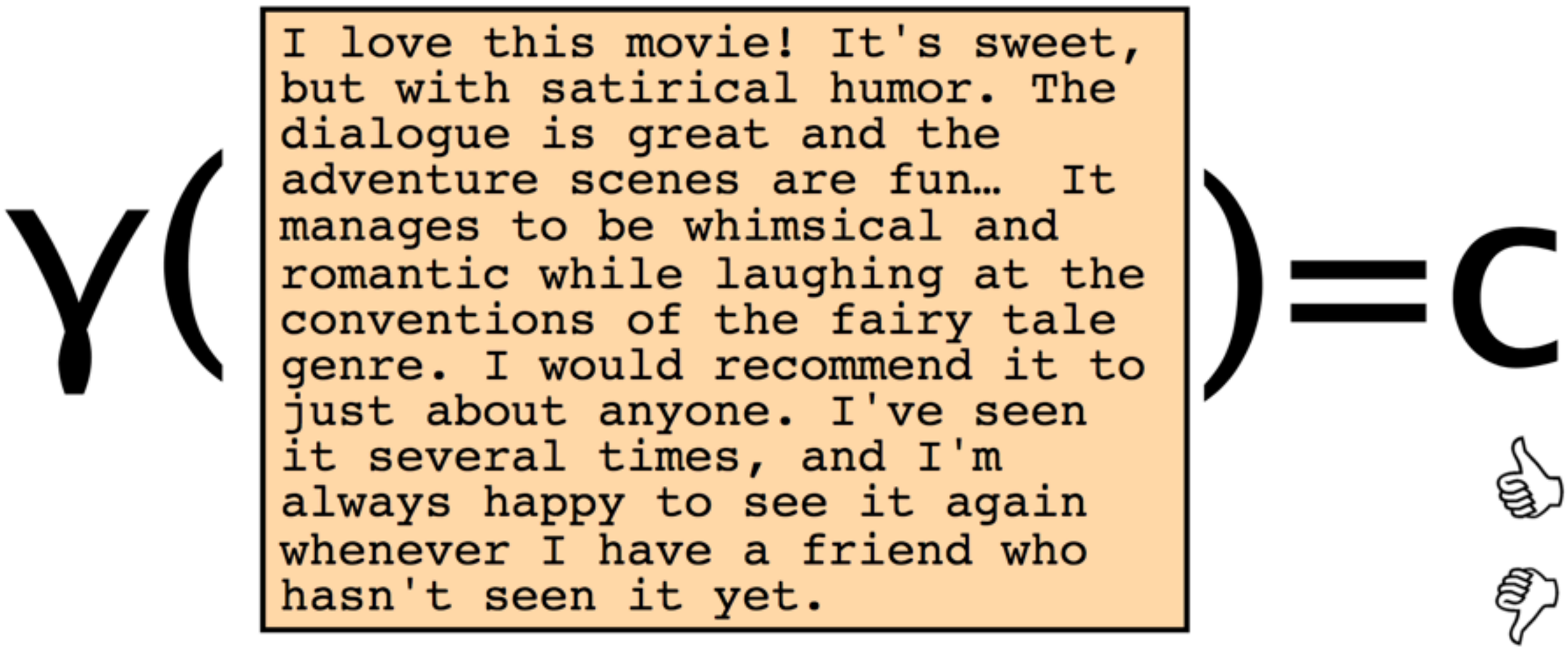
Clasificador de Bayes Naïve

- Clasificador simple (naïve) basada en la regla de Bayes.

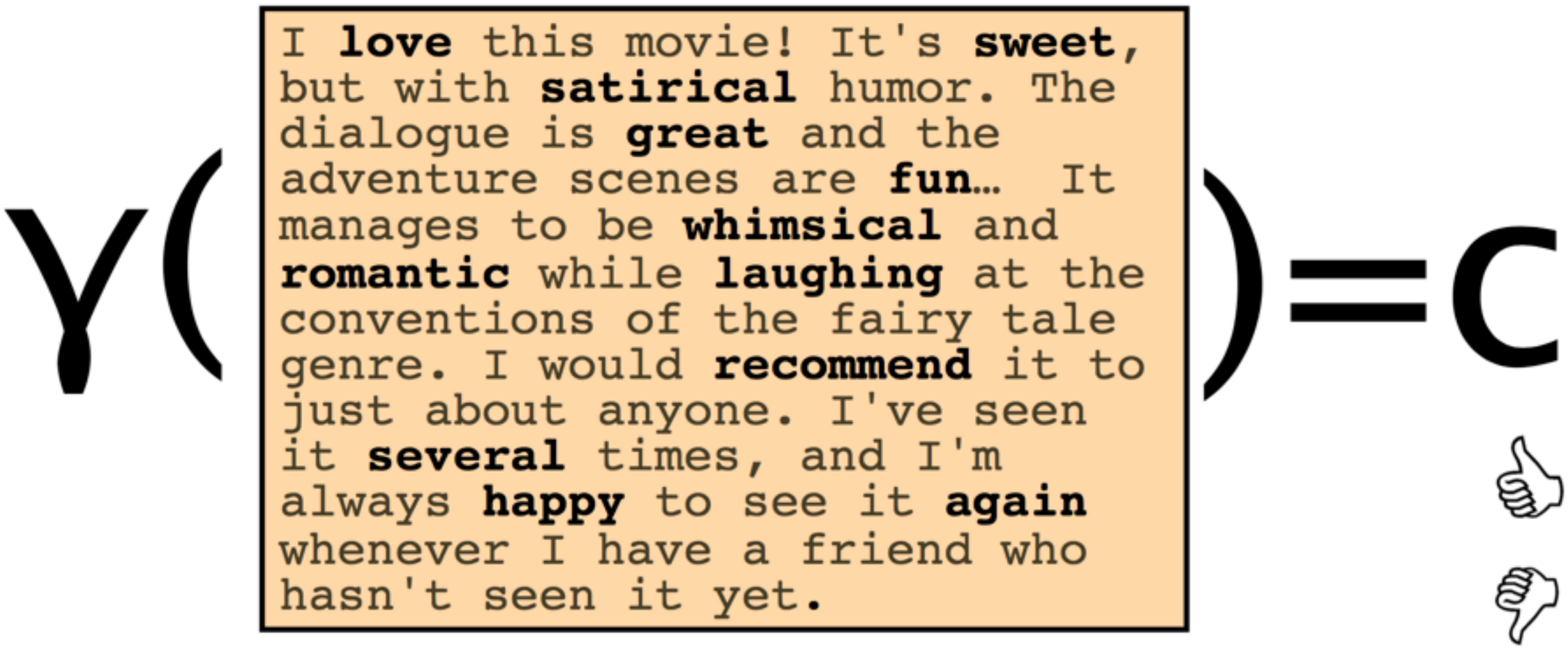
- Se basa en una representación “bag of words” de un documento.

- Asume independencia de las características. Ej. no puede entender que me gustan las películas con arnold schwarzenegger y dani de vito, pero no sólo con arnold.

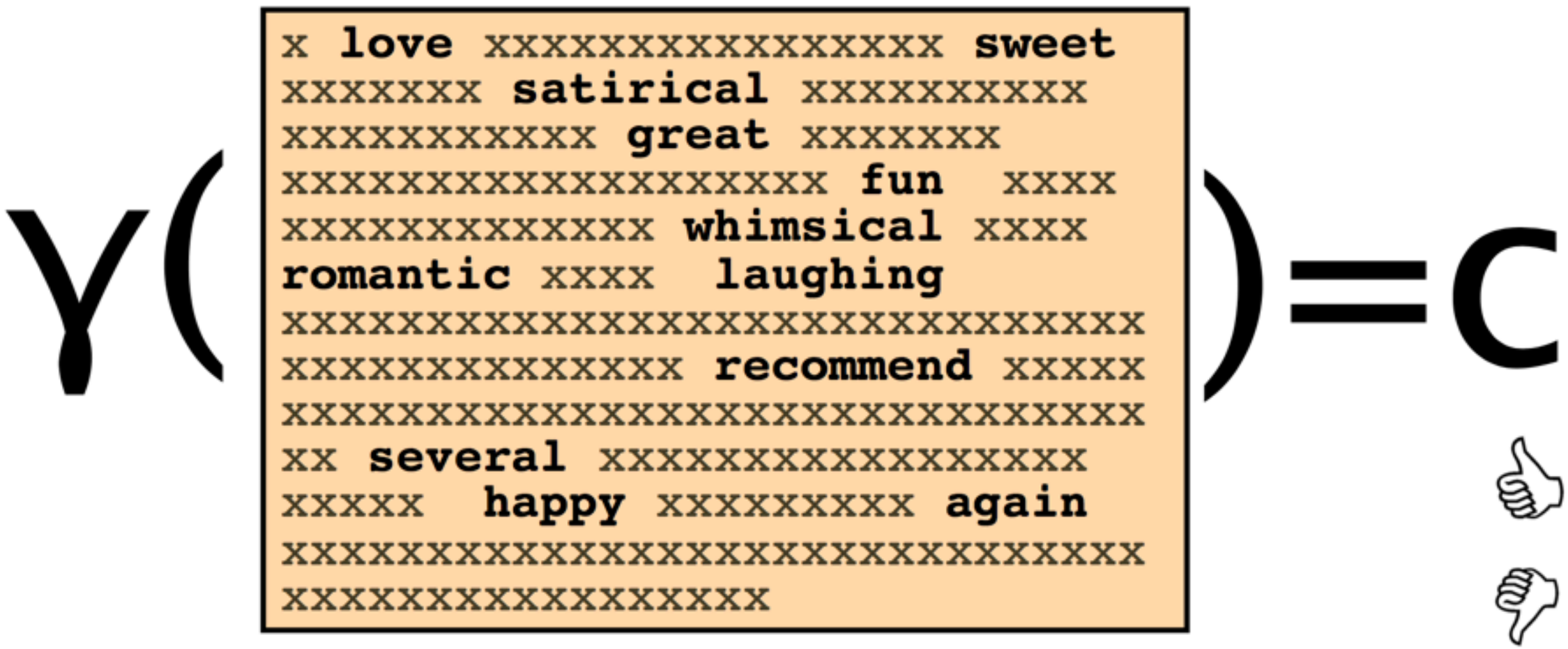
Clasificador de Bayes Naïve



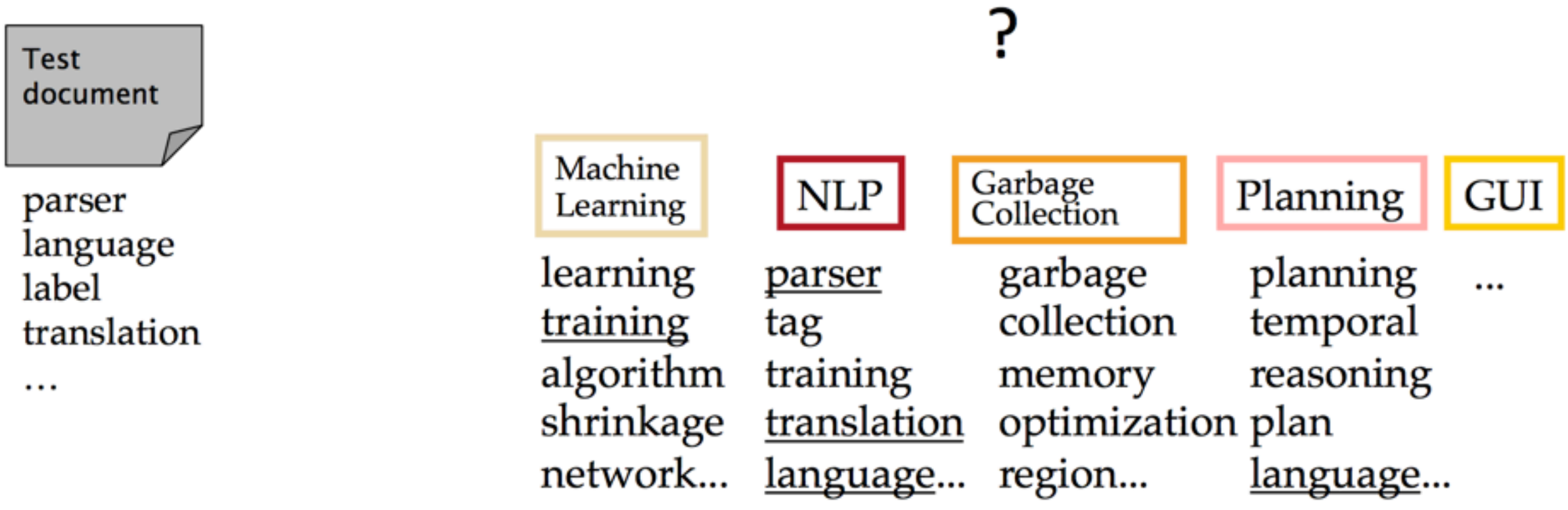
Clasificador de Bayes Naïve



Clasificador de Bayes Naïve



Clasificador de Bayes Naïve



Formalización

- Para un documento d y una clase c:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Clasificador Bayesiano

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$

Clase más probable
(maximum a posteriori)

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d \mid c)P(c)}{P(d)}$$

Regla de Bayes

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

Descartando
denominador

Clasificador Bayesiano

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Documento d representado como n características.

Clasificador Bayesiano

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$O(X^n C)$ ejemplos

Requiere cantidad de ejemplos masiva

Probabilidad de una clase

Frecuencia relativa en la BD

Supuesto de independencia

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Supuesto Bag o Words:** la posición en el documento no importa.

- **Independencia condicional:** Las características son independientes.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Clasificador Bayesiano Naïve

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Bayes: aprendizaje

- Primera aproximación: MLE, contando frecuencias.

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Bayes: aprendizaje

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

- fracción del nro de veces que la palabra ω_i aparece en los documentos de clase c_j

- crear un gran documento de la clase c_j

Bayes: aprendizaje

- Problema con MLE: ¿Qué pasa si no hemos visto una palabra con una cierta clasificación?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Las probabilidades 0 no se pueden eliminar, sin importar el resto de la evidencia.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

Bayes: aprendizaje

- Suavizado de Laplace para Bayes.

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Aprendizaje pseudocódigo:

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
  - For each c_j in C do
    $docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*
    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Bayes: ejemplo

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Bayes: ejemplo

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

Bayes: evaluación reuters

- Most (over)used data set, 21,578 docs (each 90 types, 200 toknens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Bayes: evaluación reuters

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Bayes: resumen

- Clasificador Naive no es tan Naive.
- Rápido y fácil de implementar.
- Robusto a características irrelevantes.
- Óptimo si es supuesto de independencia se verifica.
- Es un clasificador de comparación (baseline)
- Cuando la cantidad de ejemplos es masiva es una buena opción (SpamAssasin).

Bayes: extensiones

- Manejo de negaciones
- N-gramas

Table 1.RESULTS TIMELINE

Feature Added	Accuracy on test set
Original Naive Bayes algorithm with Laplacian Smoothing	73.77%
Handling negations	82.80%
Bernoulli Naive Bayes	83.66%
Bigrams and trigrams	85.20%
Feature Selection	88.80%

Narayanan et al, “Fast and accurate sentiment classification using an

”

Más alla de BoW…

- Uno de los principales problemas de los métodos bayesianos, es que el orden de las palabras es irrelevante.

- Especialmente importante en textos cortos.

- Entendemos que “suave” y “fuerte”, si fueran vectores, deberían estar más cercan que “Chile”.

Vectorización de texto

- Desde el 2006 se han propuesto aproximaciones para vectorizar texto [1][2].

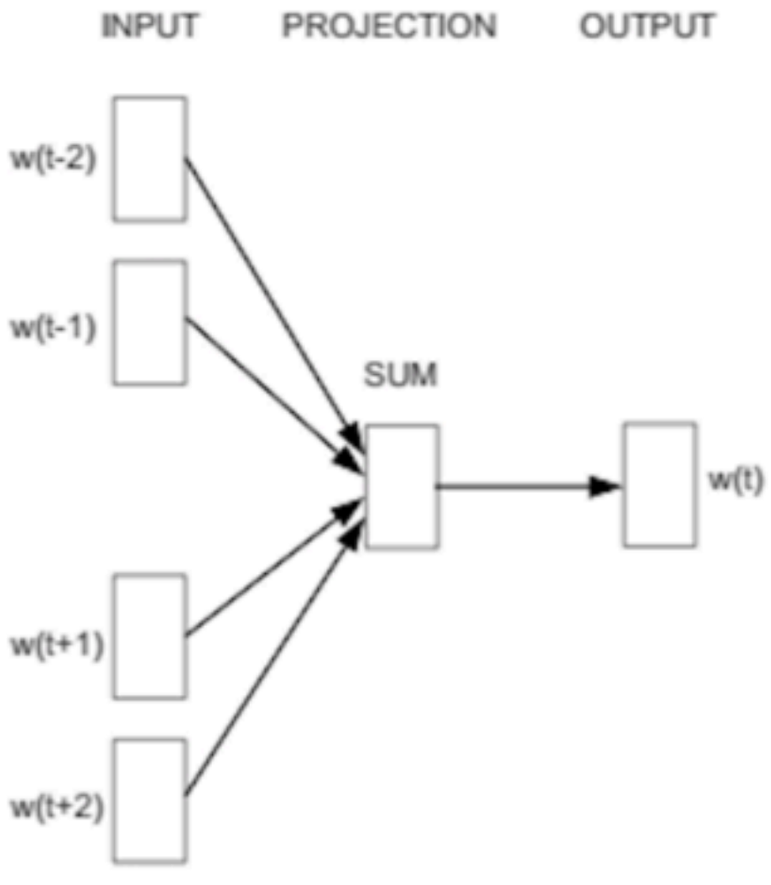
- Principalmente motivadas por problemas de predicción de palabras, y traducción.

- Métodos supervisados para aprender el contexto de las palabras.

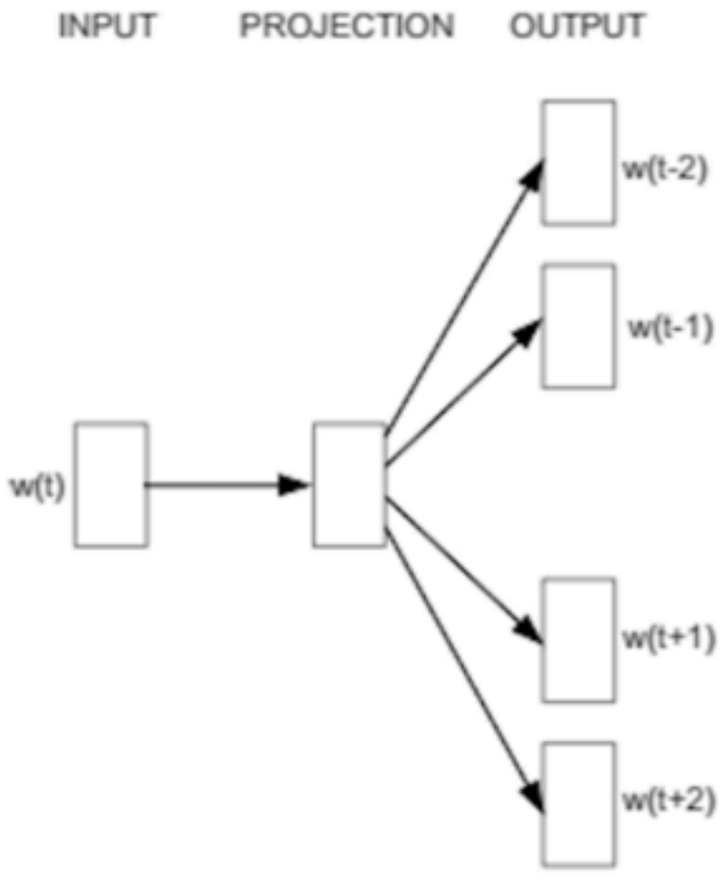
1. Bengio et al. Neural probabilistic language models. In Innovations in Machine Learning, pp. 137–186. Springe 2006.
2. Mikolov, Tomas. Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.

Vectorización de texto

- Entrenamiento en dos variantes:

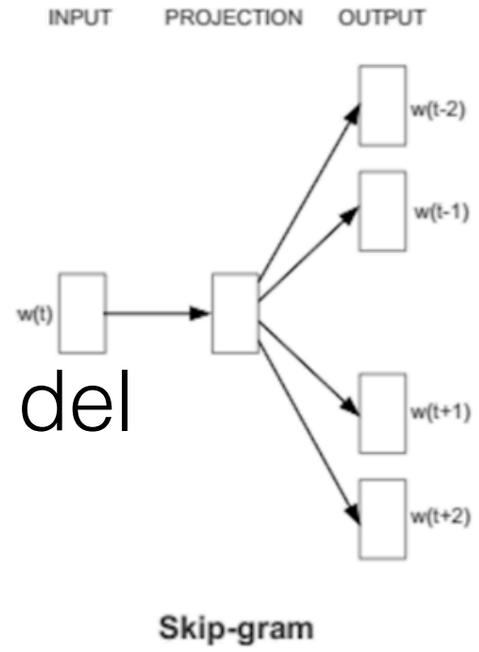


CBOW



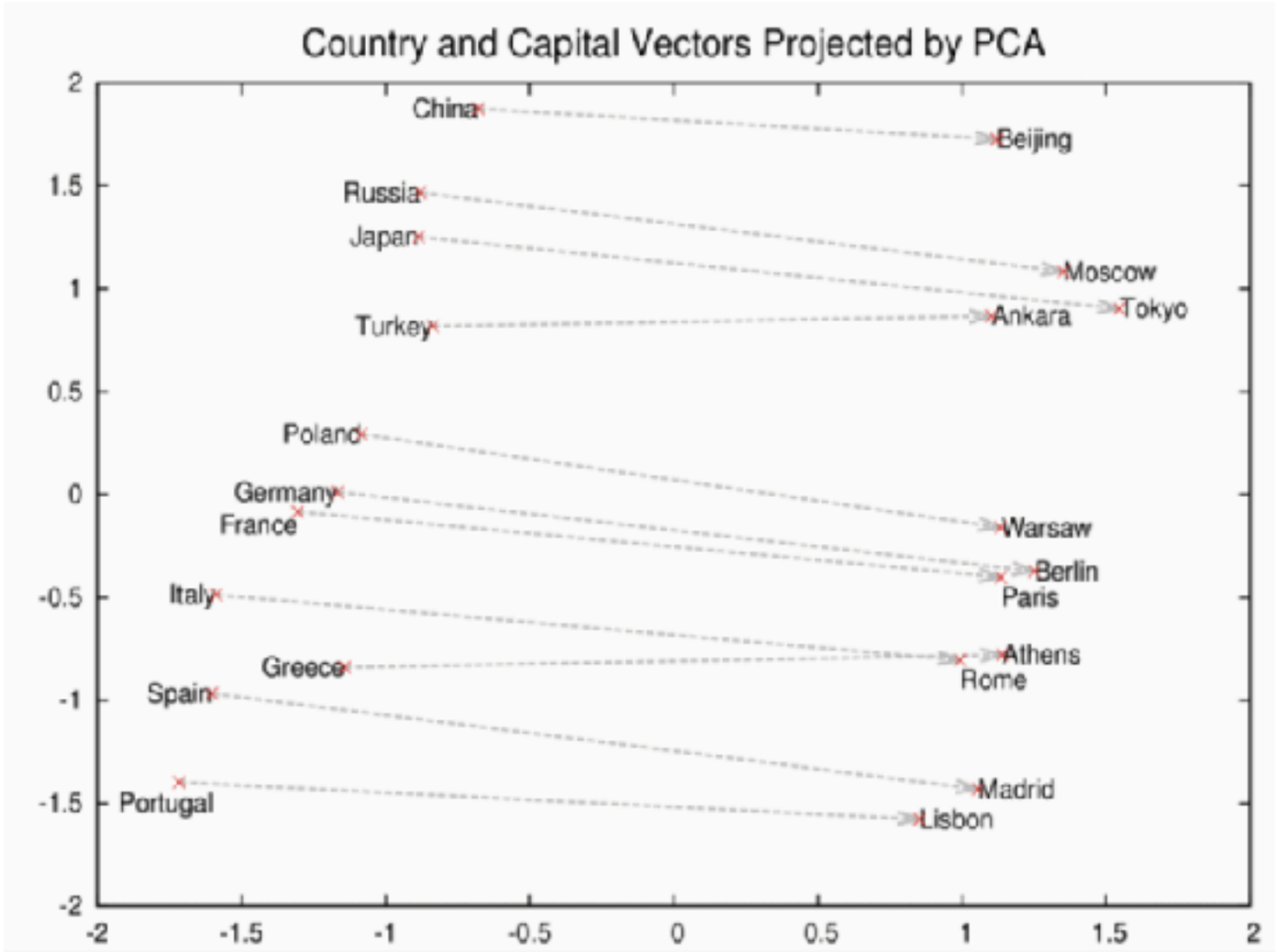
Skip-gram

Aprendizaje skip-gram



- Suponer una palabra w(t) de entrada.
- Se buscar realizar una predicción de palabras del contexto w(t-2), w(t-1), w(t+1), w(t+2).
- Si el espacio es de dimensión 400
- Las primeras 100 dimensiones podrían predecir w(t-2)
- Si no la predicen correctamente se cambia la proyección.

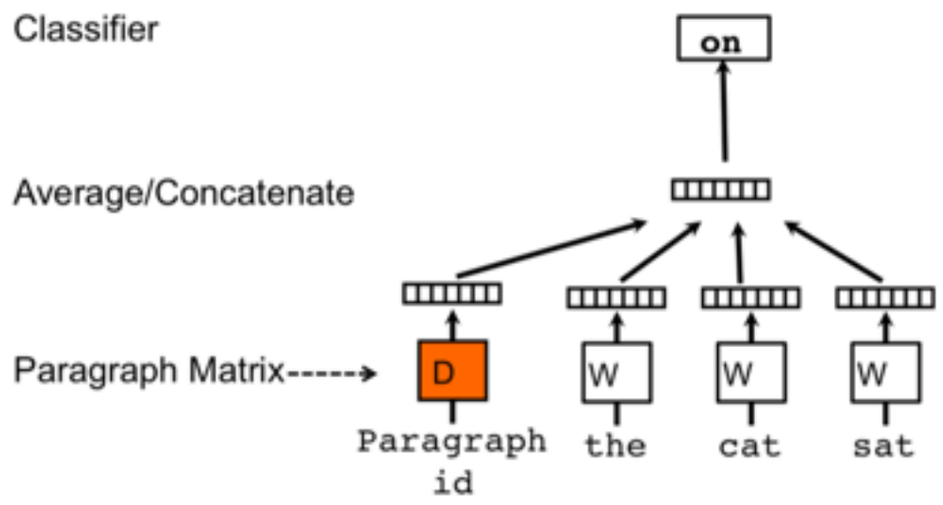
word2vec funcionando



- Permite usar algebra entre palabras, ejemplos:
  - distancia entre términos: Rome - Italy = Beijing - China
  - inferencia: Rome -Italy + China

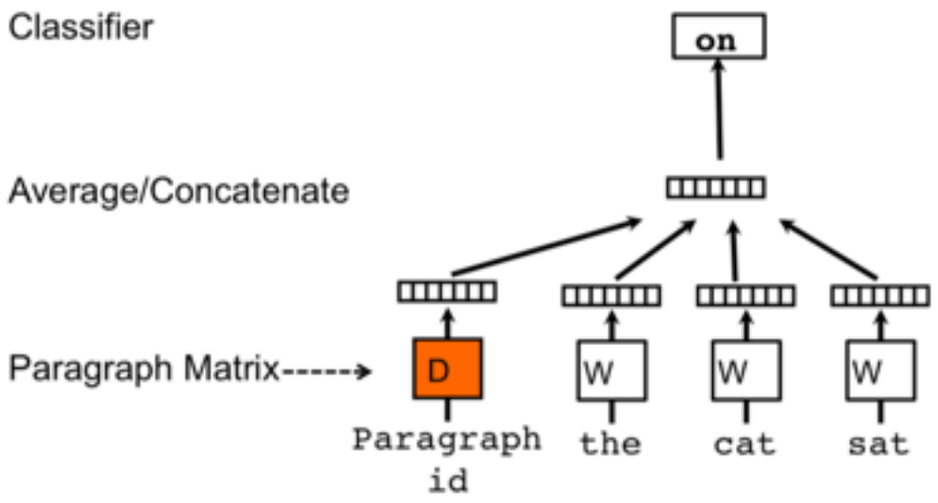
word2vec para clasificar

- word2vec no es una manera directa de clasificar texto.
- Se han propuesto extensiones que extienden esta idea para clasificar texto [Lee & Mikolov 2013].



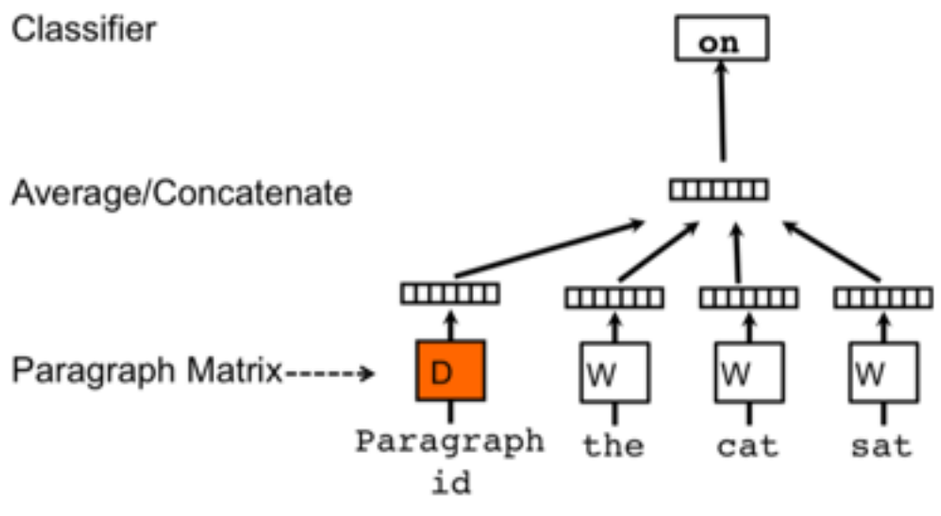
text2vec entrenamiento

- Se agrega una “palabra” extra, común a todas las ventanas de un párrafo (D).
- Entrenamiento en la tarea de predicción de la palabra siguiente.



text2vec inferencia

- En la parte de inferencia, se dejan fijos los pesos de las palabras y se entrena nuevamente para tener una representación de D.



text2vec evaluación

- Clasificación de texto 11855 comentarios de películas de Rotten Tomatoes.
- 8544 párrafos para entrenamiento, 2210 test, 1101 para validación.
- Parrafos tienen una clasificación de muy mala (0) a muy buena (1.0)
- Luego de aprender la representación vectorial de los párrafos se entrega un regresión logística.

text2vec resultados

Model	Error rate (Positive/ Negative)	Error rate (Fine-grained)
Naïve Bayes (Socher et al., 2013b)	18.2 %	59.0%
SVMs (Socher et al., 2013b)	20.6%	59.3%
Bigram Naïve Bayes (Socher et al., 2013b)	16.9%	58.1%
Word Vector Averaging (Socher et al., 2013b)	19.9%	67.3%
Recursive Neural Network (Socher et al., 2013b)	17.6%	56.8%
Matrix Vector-RNN (Socher et al., 2013b)	17.1%	55.6%
Recursive Neural Tensor Network (Socher et al., 2013b)	14.6%	54.3%
Paragraph Vector	**12.2%**	**51.3%**

[Lee & Mikolov 2013]

text2vec resultados IMBD

Table 2. The performance of Paragraph Vector compared to other approaches on the IMDB dataset. The error rates of other methods are reported in (Wang & Manning, 2012).

Model	Error rate
BoW (bnc) (Maas et al., 2011)	12.20 %
BoW (bΔt'c) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full+BoW (Maas et al., 2011)	11.67%
Full+Unlabeled+BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

[Lee & Mikolov 2013]

links y tutoriales

https://deeplearning4j.org/word2vec#anatomy

https://analyzecore.com/2017/02/08/twitter-sentiment-analysis-doc2vec/