Fundamentos de Procesamiento de Lenguaje Natural

Alvaro J. Riascos Villegas
University of los Andes y Quantil

Junio, 2019

Contents

Un Modelo del Lenguaje

- El objetivo es calcular la probabilidad de una frase $W = (w_1, w_2, \ldots, w_n)$:

$$P(W) = P(w_1, w_2, \ldots, w_n) \tag{1}$$

- Está muy relacionado con el cálculo de la probabilidad de la "palabra siguiente":

$$P(w_5 | w_4, \ldots, w_1) \tag{2}$$

- Una distribución de probabilidad que nos de cualquiera de los dos valores es un **Modelo de Lenguaje.**

Regla de la cadena

- Recordemos la regla de Bayes:

$$P(A, B) = P(A|B)P(B) \tag{3}$$

- Para más variables:

$$P(A, B, C, D) = P(A)P(B|A)P(C|B, A)P(D|C, B, A) \tag{4}$$

- Entonces...

$$P(\text{FALTA}, \text{MUCHO}, \text{TIEMPO}) = \tag{5}$$

$$P(\text{FALTA})P(\text{MUCHO}|\text{FALTA})P(\text{TIEMPO}|\text{MUCHO},\text{FALTA}) \tag{6}$$

Probabilidad de una proposción

De acuerdo a lo anterior tenemos que:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_{i-1} \dots w_1) \tag{7}$$

- ¿Cómo podemos estimar esas probabilidades?
- ¿Podemos simplemente buscar las frecuencias empíricas?

$$P(\text{CLARA} | \text{EL AGUA ES MUY}) =$$

$$\frac{count(\text{EL AGUA ES MUY CLARA})}{count(\text{EL AGUA ES MUY})}$$

Calcular la probabilidad

- Lamentablemente, nunca tendríamos un **corpus** así de grande. Muchas combinaciones de 5 palabras (o **5−gramas**) nunca estarían en el **corpus**.
- Para poder estimar necesitamos hacer algunas simplificaciones.
- Vamos a suponer que las proposiciones se comportan como una **cadena de Markov**.

$$P(\text{CLARA}|\text{EL AGUA ES MUY}) = P(\text{CLARA}|\text{MUY}) \quad (8)$$

- La probabilidad de la palabra i, sólo está condicionada por la palabra anterior (¿o las dos anteriores?)

Unigramas

- El modelo más simple son los unigramas, donde:

$$P(w_n|w_{n-1},\dots w_1) \approx P(w_n) \tag{9}$$

- Otro menos simple son los bigramas, donde:

$$P(w_n|w_{n-1}\dots w_1) = P(w_n|w_{n-1}) \tag{10}$$

- Se puede extender a **trigramas**, $4-$gramas, $5-$gramas,etc.
- Cada vez son más costosos de estimar y siguen sin capturar las **dependencias largas** del lenguaje.
- "El *perro* que había estado merodeando la plaza me *mordió*"

Calculando bigramas

- Las probabilidades de los bigramas se pueden estimar fácilmente contando en el corpus de entrenamiento.

$$P(w_2|w_1) = \frac{count(w_1, w_2)}{count(w_1)} = \frac{c(w_1, w_2)}{c(w_1)} \tag{11}$$

- Ejemplo: **Esta clase es larga, muy larga y aburridora pero parece que algo enseñan de lo larga.**

Calculando bigramas

- Las probabilidades de los bigramas se pueden estimar fácilmente contando en el corpus de entrenamiento.

$$P(w_2|w_1) = \frac{count(w_1, w_2)}{count(w_1)} = \frac{c(w_1, w_2)}{c(w_1)} \tag{11}$$

- Ejemplo: **Esta clase es larga, muy larga y aburridora pero parece que algo enseñan de lo larga.**

Matriz de conteo

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Suavizamiento

- ¿Cuál es el problema con los bigramas que ocurren cero veces en el corpus?
- Para poder calcular las probabilidades en este caso y que todo no se vaya a cero, hacemos suavización.
- **Suavizamiento de Laplace:** Sumar 1 en todas las cuentas de bigramas.

$$P(w_2|w_1) = \frac{c(w_1, w_2) + 1}{c(w_1) + |V|} \tag{12}$$

Corpora

- Google N−gram release, 2006. Consta de más de **un trillón** de palabras, y da las probabilidades estimadas para más de un billón de 5−gramas que aparecen al menos 40 veces.
- Google books N−grams. Disponible en varios idiomas, incluyendo **español**.

Evaluar un modelo de lenguaje

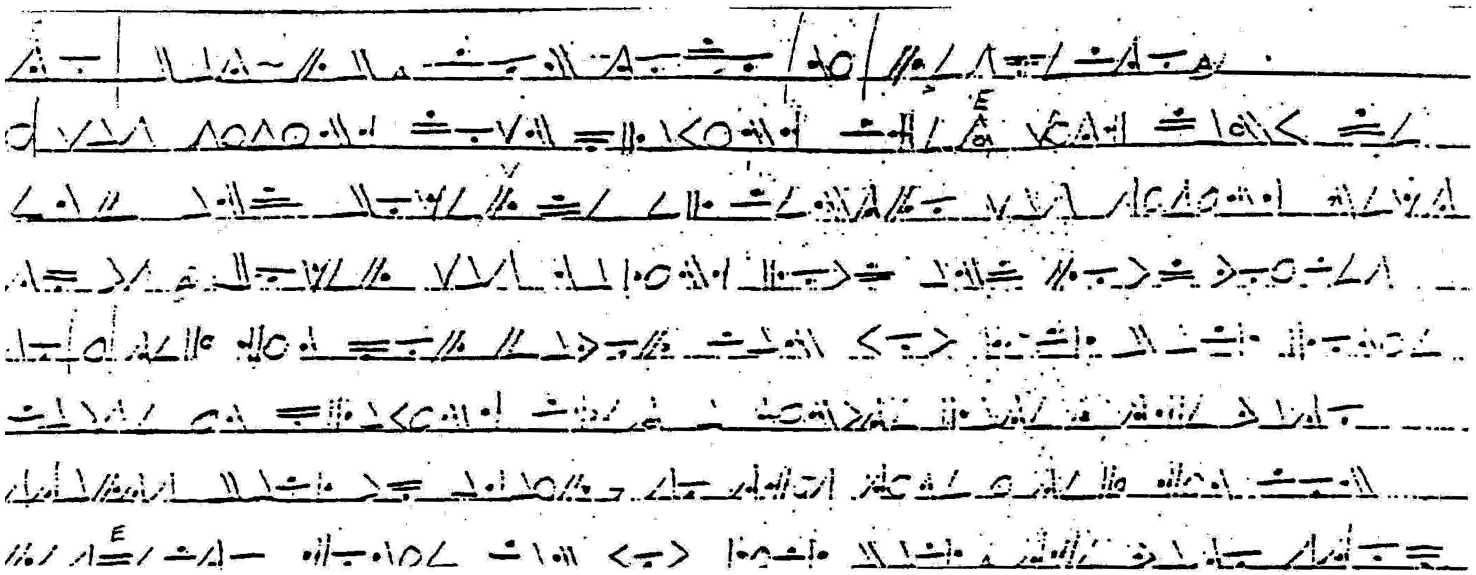
- ¿Cómo podemos evaluar un modelo de lenguaje?
- **Evaluación extrínseca:** Mirar el desempeño de ese modelo frente a otros modelos en tareas prácticas específicas.
- **Evaluación intrínseca:** Evaluar la verosimilitud del modelo en un corpus de prueba, que sea similar al de entrenamiento.
- Una medida común de verosimilitud es la **perplejidad**:

$$PP(W) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}} \tag{13}$$

Contents

Matemáticas en acción

- Un psicólogo de una prisión en California se acercó al Departamento de Estadística de Stanford con esto:



Matemáticas en acción

- Los estudiantes supusieron que el mensaje se podía decodificar de la forma:

$$f : \{\text{Símbolos del mensaje}\} \to \{\text{Alfabeto tradicional}\}$$

- El número de posibilidades para f es muy grande.
- Para juzgar la plausibilidad de que una f en particular podemos definir una medida de rendimiento.
- Tomaron un libro estándar (e.g. One Hundred Year of Solitude) y calcularon la frecuencia con la que después del símbolo x seguía el símbolo y.
- Interpretamos esto como la probabilidad de transición de x a y en el idioma Inglés (i.e. una cadena de Markov). Denotamos por $M(x, y)$ esta frecuencia.

Matemáticas en acción

- Los estudiantes supusieron que el mensaje se podía decodificar de la forma:

$$f : \{\text{Símbolos del mensaje}\} \to \{\text{Alfabeto tradicional}\}$$

- El número de posibilidades para f es muy grande.
- Para juzgar la plausibilidad de que una f en particular podemos definir una medida de rendimiento.
- Tomaron un libro estándar (e.g. One Hundred Year of Solitude) y calcularon la frecuencia con la que después del símbolo x seguía el símbolo y.
- Interpretamos esto como la probabilidad de transición de x a y en el idioma Inglés (i.e. una cadena de Markov). Denotamos por $M(x, y)$ esta frecuencia.

Matemáticas en acción

- Los estudiantes supusieron que el mensaje se podía decodificar de la forma:

$$f : \{\text{Símbolos del mensaje}\} \to \{\text{Alfabeto tradicional}\}$$

- El número de posibilidades para f es muy grande.
- Para juzgar la plausibilidad de que una f en particular podemos definir una medida de rendimiento.
- Tomaron un libro estándar (e.g. One Hundred Year of Solitude) y calcularon la frecuencia con la que después del símbolo x seguía el símbolo y.
- Interpretamos esto como la probabilidad de transición de x a y en el idioma Inglés (i.e. una cadena de Markov). Denotamos por $M(x, y)$ esta frecuencia.

Matemáticas en acción

- Los estudiantes supusieron que el mensaje se podía decodificar de la forma:

$$f : \{\text{Símbolos del mensaje}\} \to \{\text{Alfabeto tradicional}\}$$

- El número de posibilidades para f es muy grande.
- Para juzgar la plausibilidad de que una f en particular podemos definir una medida de rendimiento.
- Tomaron un libro estándar (e.g. One Hundred Year of Solitude) y calcularon la frecuencia con la que después del símbolo x seguía el símbolo y.
- Interpretamos esto como la probabilidad de transición de x a y en el idioma Inglés (i.e. una cadena de Markov). Denotamos por $M(x, y)$ esta frecuencia.

Matemáticas en acción

- Los estudiantes supusieron que el mensaje se podía decodificar de la forma:

$$f : \{\text{Símbolos del mensaje}\} \to \{\text{Alfabeto tradicional}\}$$

- El número de posibilidades para f es muy grande.
- Para juzgar la plausibilidad de que una f en particular podemos definir una medida de rendimiento.
- Tomaron un libro estándar (e.g. One Hundred Year of Solitude) y calcularon la frecuencia con la que después del símbolo x seguía el símbolo y.
- Interpretamos esto como la probabilidad de transición de x a y en el idioma Inglés (i.e. una cadena de Markov). Denotamos por $M(x, y)$ esta frecuencia.

- Ahora resolvemos este problema:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{14}$$

- Este problema se resuelve usando al algoritmo de Metropolis - Hasting (un truco muy creativo!)
- Ahora, por qué esto funcionaría? Para ganar un poco de confianza en el modelo veamos el resultado en un caso conocido.

Matemáticas en acción

- Ahora resolvemos este problema:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{14}$$

- Este problema se resuelve usando al algoritmo de Metropolis - Hasting (un truco muy creativo!)
- Ahora, por qué esto funcionaría? Para ganar un poco de confianza en el modelo veamos el resultado en un caso conocido.

- Ahora resolvemos este problema:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{14}$$

- Este problema se resuelve usando al algoritmo de Metropolis - Hasting (un truco muy creativo!)
- Ahora, por qué esto funcionaría? Para ganar un poco de confianza en el modelo veamos el resultado en un caso conocido.

Matemáticas en acción

- Revolver todos los símbolos de este párrafo de Hamlet:

```
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END
```

- Metropolis Hasting despues de n iteraciones:

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

- Resultado decodificación mensaje presos:

to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

Contents

Modelo documentos

Definition

Un vocabulario V es un conjunto finito de elementos diferentes. Los elementos del vocabulario son la unidades básicas de análisis (e.g., palabras).

Definition

Un documento sobre un vocabulario V es una sucesión de elementos $\mathbf{w} = (w_1, \ldots, w_n)$ con $w_i \in V$. N es la longuitud del documento $\mathbf{w}$.

Definition

Un Corpus $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ es un conjunto de documentos sobre un vocabulario V. M es el tamaño del Corpus .

Latent Dirichlet Allocation (LDA)

- LDA es un modelo probabilístico generativo de un corpus.
- La idea es que los documentos son representados como mezclas aleatorias sobre tópicos latentes, donde cada tópico es caracterizado como una distribución sobre palabras.

Latent Dirichlet Allocation (LDA)

- LDA asume el siguiente proceso generativo para cada documento: **w** en un corpus D (fijemos un número de tópicos k):
  1. Escoja $N \sim \text{Poisson}(\xi)$.
  2. Escoja $\theta \sim \text{Dirichlet}(\alpha)$.
  3. Para cada una de las N palabras w_n:
     - Escoja un tópico $z_n \sim \text{Multinomial}(\theta)$.
     - Escoja una palabra w_n de $p(w_n \mid z_n, \beta)$, una distribución de probabilidad multinomial condicionada en el tópico z_n, donde β es una matriz de dimensiones $k \times V$, donde $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
  4. El objetivo es estimar la distribución de las palabras condicionales al tópico (β).

Latent Dirichlet Allocation (LDA)

- Dados los parámetros α y β, la función de probabilidad conjunta de una mezcla de tópicos θ, un conjunto de k tópicos y un conjunto de N palabras está dado por:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \quad (15)$$

donde $p(z_n \mid \theta)$ es θ_i para el único i tal que $z_n^i = 1$.

Contents

Ejemplo Tópicos

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Validación: Perplejidad

- Una forma común de evaluar el rendimiento de un modelo de lenguaje es la perplejidad:

$$perplejidad(D) = \exp(-\frac{\sum_{d=1}^{M} log(p(\mathbf{w}_d))}{\sum_{d=1}^{M} N_d})$$

Aplicación: Modelación de Documentos

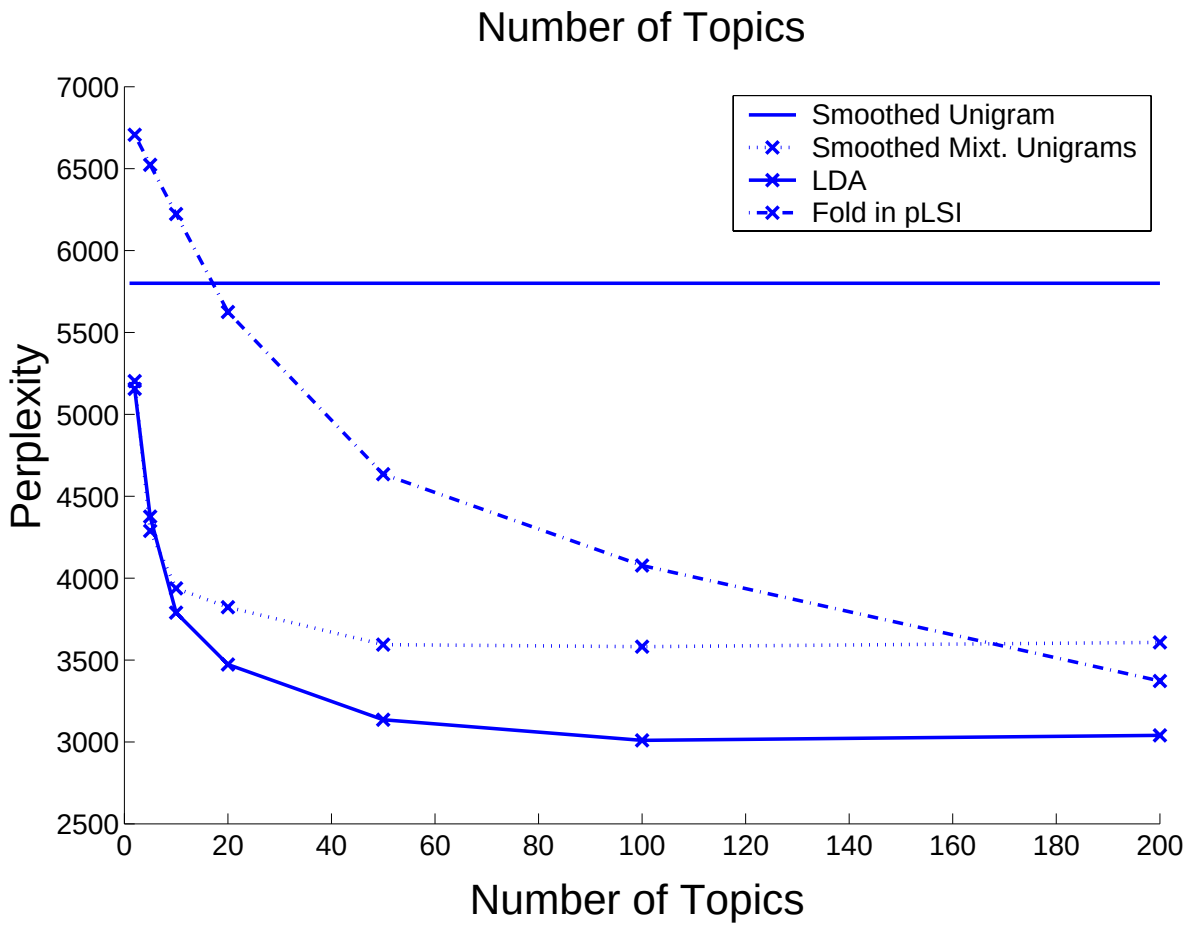
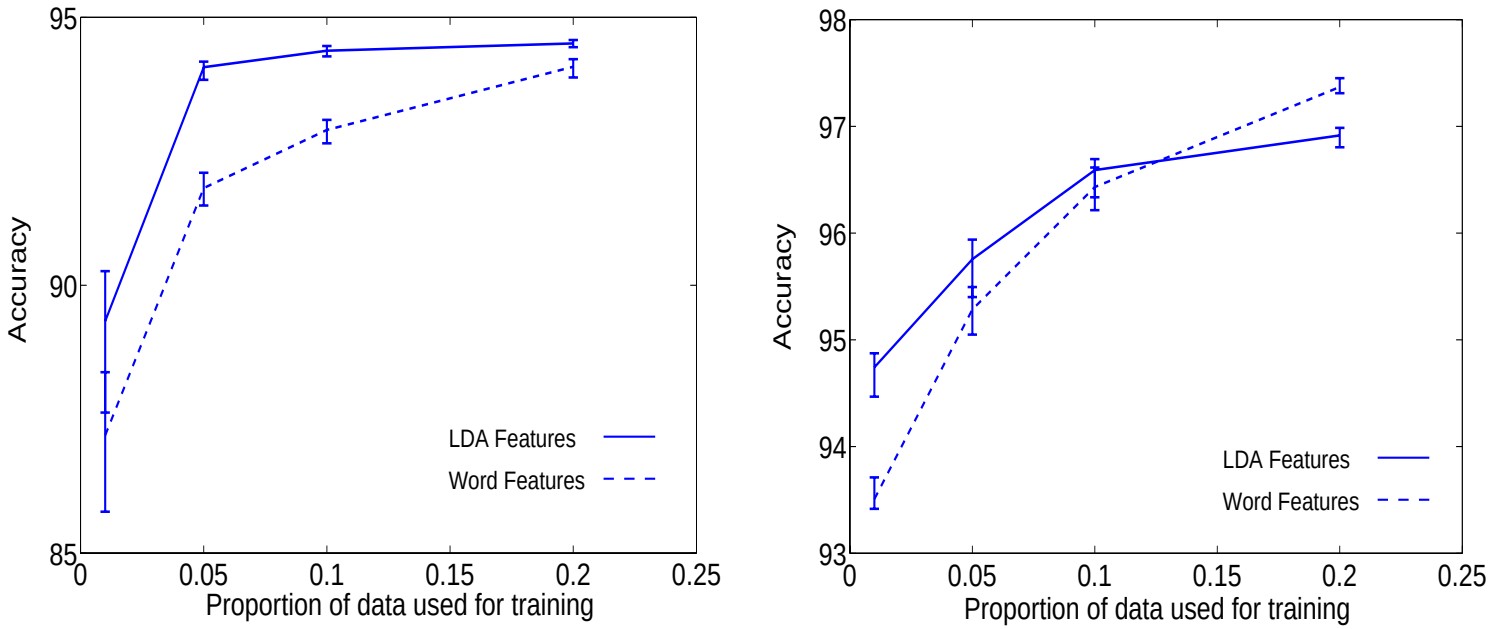


Figure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

Aplicación: Clasificación



(a) (b)

Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Aplicación: Filtros Colaborativo

- En esta aplicación un conjunto de usuarios califica una colección de películas. En este caso: Un usuario es un documento y las peliculas las palabras.

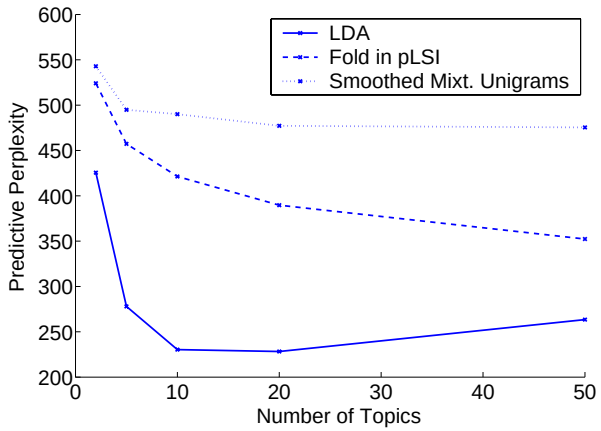


Figure 11: Results for collaborative filtering on the EachMovie data.