UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

ST7003 Procesamiento Natural del Lenguaje

Lecture 02

Contenido

1. Minería de tópicos / Temas
2. Modelos de lenguaje n-gram
3. LDA
4. Categorización de texto

Contenido

1. Minería de tópicos / Temas
2. Modelos de lenguaje n-gram
3. LDA
4. Categorización de texto

UNIVERSIDAD EAFIT

Escuela de
Ciencias Aplicadas
e Ingeniería

1. Minería de tópicos / temas

Primeras intuiciones a la generación de documentos y contexto

Minería de tópicos

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Minería de conocimiento sobre el lenguaje, descubrimiento de asociaciones entre palabras (paradigmáticas o sintagmáticas en la visión clásica, o de contexto en la visión moderna).
- Descubrimiento de conocimiento sobre los temas principales del texto -> minería y análisis.
- Que es un tema?
  - Es la idea principal que se analiza en los datos de texto.
  - De lo que está 'hablando' un texto.
  - Hay diferentes granularidades: un conjunto de documentos, a nivel de docto, párrafo o sentencia.
- Ejemplo:
  - De que hablan los usuarios de X hoy en día en Colombia?. Hablan de deportes? De política? De economía? Salud?, etc..
  - O mejor aún, se hablan de diferentes temas, pero en que proporción?
  - Tambien nos interesa saber que le gusta y no le gusta de un producto a la gente.
  - O saber cuales fueron los temas principales, durante el discurso de posesión de Donald Trump.

Tópicos / temas

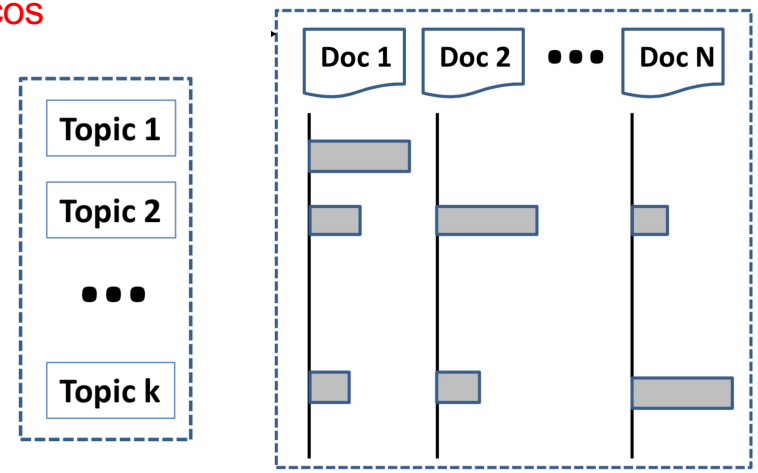
- Podemos ver un tema como un conocimiento sobre el mundo
- A partir de los datos textuales esperamos descubrir una serie de temas
- Estos temas proporcionan una descripción del mundo
- Se puede contar con otros tipos de datos adicionales al texto: ubicación, autores, fuentes, tiempo (serie), etc.
- Cual es la tarea de mineria: descubrir muchos temas (k temas), Tambien nos interesa que temas se tratan en un documento y en que proporción.

Tareas de Minería y Análisis de TÓPICOS

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Tarea1: descubrir k tópicos



Tarea 2: Determinar qué documentos cubren qué temas

Tareas

- Por ejemplo, en el documento uno, podemos ver que el tema 1 se trata mucho, y el tema 2 y el tema k se tratan en una pequeña parte no estén cubiertos.
- El documento dos, por otro lado, cubrió muy bien el tema 2, pero no cubrió el tema 1 en absoluto, y también cubre el tema k hasta cierto punto, etc.
- Dos tareas:
  1. descubrir k temas de un conjunto de documentos.
  2. en un documento, que temas cubre y en que proporción.
- A continuación la definición formal:

Definición formal de la minería y el análisis de tópicos

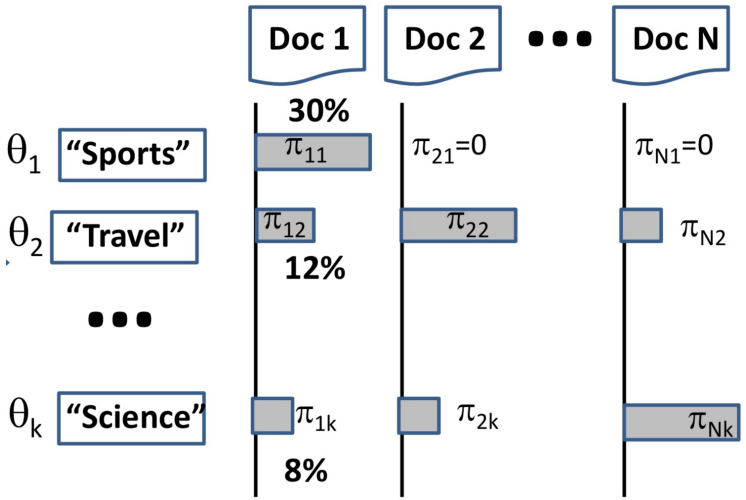
UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Entrada:
  - Una colección de N documentos C={d1, ..., dN}
  - Número de tópicos: k
- Salida:
  - k tópicos: { θ_1, ..., θ_k }
  - Cobertura de tópicos en cada d_i: {π_{i1}, ..., π_{ik} }
  - π_{ij} = probabilidad de d_i de cubrir el tópico j

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

- Como definir θ_i?

Idea inicial: Término = Tópico

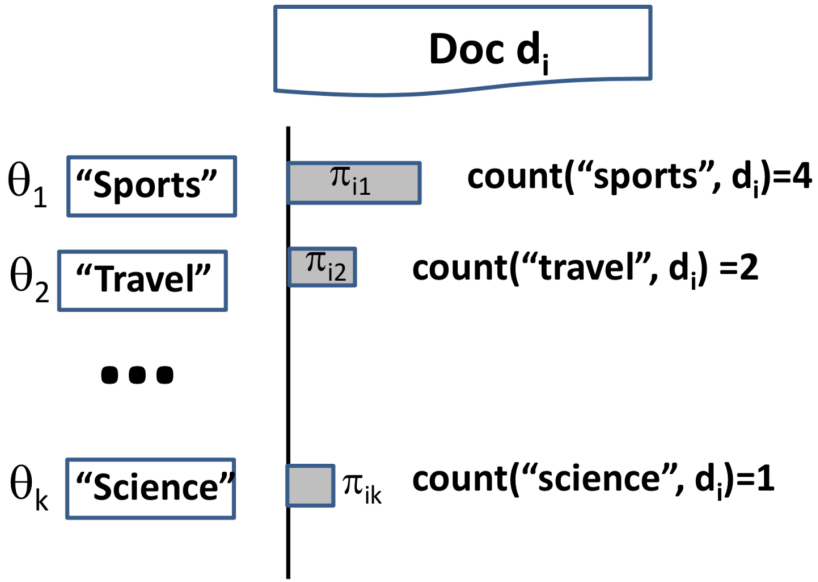


Minería k términos tópicos de la colección C

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Analice el texto en C para obtener términos candidatos (por ejemplo, término = palabra).
- Diseñe una función de puntaje para medir qué tan bueno es cada término como tópico.
  - Favorecer un término representativo (se favorece la alta frecuencia)
  - Evite las palabras que son demasiado frecuentes (por ejemplo, "el", "a").
  - La ponderación TF-IDF de la recuperación puede ser muy útil.
  - La heurística específica del dominio es posible (por ejemplo, favorecer palabras de título, hashtags en tweets).
- Escoja los k términos con las puntuaciones más altas, pero trate de minimizar la redundancia.
  - Si los términos múltiples son muy similares o están estrechamente relacionados, escoja sólo uno de ellos e ignorar a los demás.

Calcular la cobertura de tópicos: π_{ij}



$$\pi_{ij} = \frac{count(\theta_j, d_i)}{\sum_{L=1}^{k} count(\theta_L, d_i)}$$

Problemas con "El término como tópico"

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

- Falta de poder expresivo
  - Solo palabras simples/generales
  - No puede representar tópicos complejos
- Incompletitud en la cobertura del vocabulario
- Ambigüedad en el sentido de la palabra

Idea Mejorada: Tópico = Distribución de palabras

Θ_1 = sports	Θ_2 = travel	Θ_k = science
$P(w\|\Theta_1)$	$P(w\|\Theta_2)$	$P(w\|\Theta_k)$
sports 0.02 game 0.01 basketball 0.005 football 0.004 play 0.003 star 0.003 ... nba 0.001 ... travel 0.0005 ...	travel 0.05 attraction 0.03 trip 0.01 flight 0.004 hotel 0.003 island 0.003 ... culture 0.001 ... play 0.0002 ...	science 0.04 scientist 0.03 spaceship 0.006 telescope 0.004 genomics 0.004 star 0.002 ... genetics 0.001 ... travel 0.00001 ...

$$\sum_{w \in V} p(w \mid \theta_i) = 1$$

Vocabulario V = {w1, w2, ….}

Minería y análisis de TÓPICOS probabilísticos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

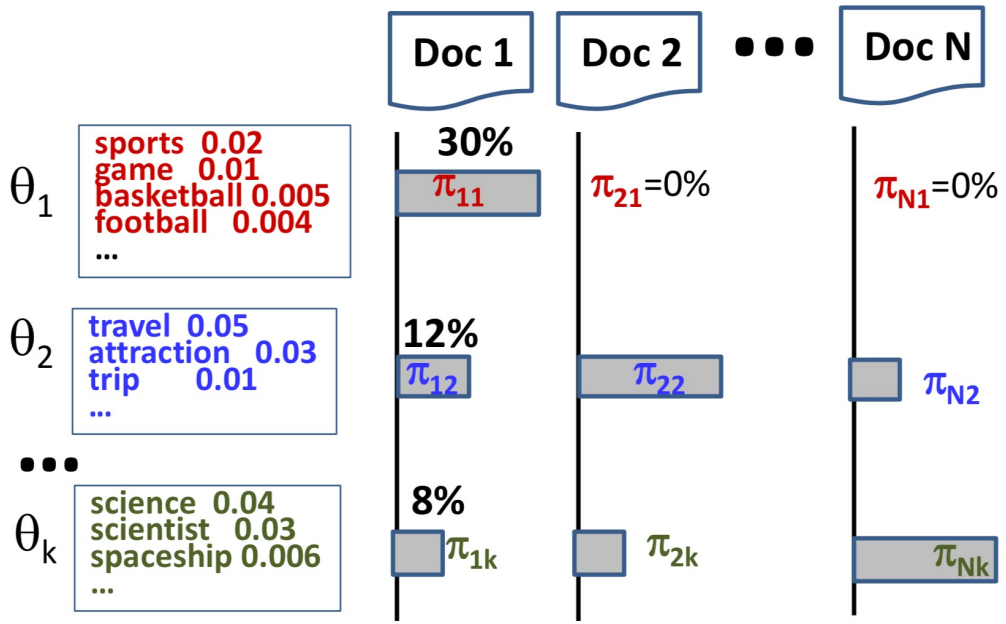
- Entrada:
  - Una colección de N documentos C={d1, ..., dN}
  - Conjunto de vocabulario: V={w1, ..., wM}
  - Número de tópicos: k
- Salida:
  - k tópicos, cada distribución de palabras: { θ_1, ..., θ_k }
  - Cobertura de tópicos en cada d_i: {π_{i1}, ..., π_{ik} }
  - π_{ij} = probabilidad de d_i de cubrir el tópico j

$$\sum_{w \in V} p(w \mid \theta_i) = 1$$

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

La tarea de cálculo

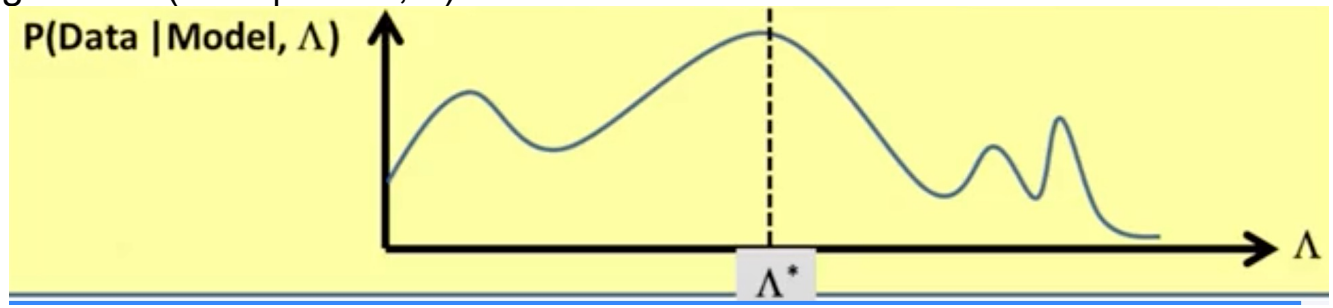
- Entrada: C,k,V
- Salida: { θ_1, ..., θ_k } y {π_{i1}, ..., π_{ik} }



Modelo Generativo para minería de texto

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Input: C, k, V Output: { θ_1, ..., θ_k } y {π_{i1}, ..., π_{ik} }
- Diseñar un modelo probabilistico
- P (Data | Model, Λ)
- Λ = ({θ_1, ..., θ_k }, {π_{11}, ..., π_{1k} },…,{π_{N1}, ..., π_{Nk} })
- Cuantos parámetros hay en Total?
  - En θ_i -> |V| x k
  - En π_{ij} -> N x k
- Función de optimización: obtener los parámetros hasta que consigamos el conjunto de datos de probabilidad máxima
  - Λ* = argmax Λ P(Data | Model, Λ)



Resumen

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Idea mejorada: un topico como una distribución de palabras
- Tareas:
  - Entrada: C,k,V
  - Salida: conjunto de tópicos, cada uno con una distribución de palabras; cobertura de tópicos en cada documento
- $\Lambda = (\{\theta_1, ..., \theta_k\}, \{\pi_{11}, ..., \pi_{1k}\}, \ldots, \{\pi_{N1}, ..., \pi_{Nk}\})$

$$\forall j \in [1,k], \sum_{w \in V} p(w \mid \theta_j) = 1$$

$$\forall i \in [1,N], \sum_{j=1}^{k} \pi_{ij} = 1$$

retricciones

Resumen

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Modelo generativo
  - Modelo de generación de datos con un modelo de probabilidad:
    - P (Data | Model, Λ)
  - Inferir los parámetos Λ* para un dataset particular:
    - Λ* = argmax Λ P(Data | Model, Λ)
    - Λ* como el 'conocimiento' a ser obtenido por la minería de texto.
    - Ajuste el diseño del modelo para descubrir diferente conocimiento.

Contenido

1. Minería de tópicos / Temas
2. Modelos de lenguaje n-gram
3. LDA
4. Categorización de texto

UNIVERSIDAD EAFIT

Escuela de
Ciencias Aplicadas
e Ingeniería

2. Modelos de lenguaje n-gram

Modelo de lenguaje

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Un modelo de lenguaje es un modelo de aprendizaje que predice palabras.
- Un modelo de lenguaje asigna una probabilidad a cada posible próxima palabra o equivalente dando una probabilidad de distribución sobre posibles próximas palabras.

Prediciendo ‘palabras’

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- *El agua de Manantiales es hermosamente …*

azul
gris
clara

*nevera
*que

Modelos de Lenguaje

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Sistemas que pueden predecir palabras adelante
  - Asignar una probabilidad a cada potencial próxima palabra
  - Asignar una probabilidad a una ORACIÓN COMPLETA (muy retador)

Porque predecir palabras?

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Varias tareas

- Corrección gramatical

  Their are two midterms ~~Their~~ There are two midterms

  Everything has improve Everything has ~~improve~~ improved

- Speech recognition

  I will be back soonish I will be bassoon dish

Modelo de lenguaje estadístico

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Una probabilidad de distribución sobre una secuencia de palabras
  - p('hoy es martes') = 0.01
  - p('hoy martes es') = 0.0000001
  - p('hoy es banana') = 0.0000000001

- Depende del contexto
- Tambien puede ser visto como un mecanismo probabilistico para generar texto, modelo generativo.

Modelo de lenguaje: Unigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Generar texto Generando cada palabra independientemente.
- $p(w_1,w_2,\ldots, w_n) = p(w_1)p(w_2)\ldots p(w_n)$
- Parametos: $\{p(w_i)\}$ -> $p(w_1) + p(w_2)+\ldots+p(w_N) = 1$ (N tamaño vocab.)

- Texto -> muestra extraída de acuerdo con una 'distribución palabras'
- Por ejemplo, podemos intentar juntar palabras
- Así podemos calcular la probabilidad de una secuencia de palabras solo con las probabilidades de las palabras
  - p('hoy es martes) = p('hoy')p('es')p('martes') =

    0.0002 x 0.001 x 0.0000015

  Problema: ignora el orden de las palabras-> p('hoy es martes) es igual a p('martes hoy es').

  PERO puede ser usado en ANALISIS DE TÓPICOS, porque?

El modelo de lenguaje unigram es pieza fundamental en modelos de detección de tópicos como LDA

Generación de texto con LM Unigram

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

	Unigram LM p(w|θ)	Muestreo	documento d p(d|θ)=?
Topico 1: Text Mining	… text 0.2 mining 0.1 clustering 0.02 … food 0.00001 …	→	Text Mining paper
Topico 2: health	… food 0.25 nutrition 0.1 … mining 0.0002 …	→	Heath nutrition paper

Estimación con LM Unigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Unigram LM p(w|θ)=?

Estimación

Text Mining paper d
 # palabras = 100

…	…
10/100	text ?
5/100	mining ?
3/100	clustering ?
…	…

text 10
mining 5
clustering 3

Estimación de máxima Verosimilitud (Maximum Likelihood Estimate)

Heath nutrition paper

Cual es la mejor estimación?
Como definir 'mejor'?

Maximum likelihood vs Bayesian

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Maximum likelihood estimation
  - Problema: muestra pequeña

$$\hat{\theta} = \arg \max_{\theta} P(X \mid \theta)$$

- Estimacion bayesiana

Bayes Rule

$$p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$$

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid X) = \arg \max_{\theta} P(X \mid \theta) P(\theta)$$

  - Problema: como definir a priori?

LM setup on unigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

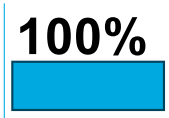
- Input: C={d}, V Output:{θ}

-datos texto-
“Ipsun larum
Pepe el toro
Hola mundo
…”

θ

P(w| θ)
Ipsun ?
larum ?
Pepe ?
el ?
toro ?
Hola ?
mundo ?
…

Doc d



Mining One Topic

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Data: Documento d= x1x2…x|d|, xi pertenece a V = {w1,w1,…wM}

- Modelo: Unigram LM θ (=topic) : {θi=p(wi | θ)}, i=1, …, M; θ1+…+ θM=1
- Función likelihood: p(d | θ) = $p(x_1| \theta)\ p(x_2 | \theta) \ldots p(x_{|d|} | \theta)$
  = $p(w_1 | \theta)^{c(w_1,d)}$ x … x $p(w_M | \theta)^{c(w_M,d)}$
  = $\prod_{i=1}^{M} p(wi | \theta)^{c(wi,d)} = \prod_{i=1}^{M} \theta_i^{c(wi,d)}$

- ML estimado:

$$(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d \mid \theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$$

Finalmente: $\hat{\theta}_i = p(w_i \mid \hat{\theta}) = \frac{c(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \frac{c(w_i,d)}{|d|}$ ← **Contador normalizado**

LM setup on unigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

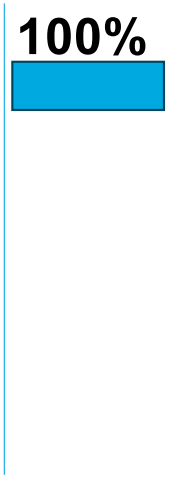
- Input: C={d}, V Output:{θ}

-datos texto-
“Ipsun larum
Pepe el toro
Hola mundo
…”

θ

P(w| θ)
Ipsun 0.01
larum 0.023
Pepe 0.001
el 0.1
toro 0.02
Hola 0.03
mundo 0.03
…

Doc d



Modelo de Lenguaje (LM) más formalmente

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Meta: CALCULAR LA PROBABILIDAD de una oración (muy retador) o secuencia de palabras W:

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- Tarea relacionada: probabilidad de una palabra siguente:

$$P(w_5|w_1,w_2,w_3,w_4) \text{ or } P(w_n|w_1,w_2 \ldots w_{n-1})$$

- Expresado como:

$$P(W) \quad \text{or} \quad P(w_n|w_1,w_2 \ldots w_{n-1})$$

Como estimar la probabilidad?

- Contar y Dividir

P(azul | ‘El agua de Manantiales es hermosamente’) =

C (‘El agua de Manantiales es hermosamente azul’)

C (‘El agua de Manantiales es hermosamente’)

- No! Demasiadas posibles oraciones.
- No hay sufientes datos para estimar esto.

Calcular P(W) como palabras

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Probabilidad conjunta de P(W)

  P(el, agua, de, Mantiales, es, hermosamente, azul)

- Regla de cadena de probabilidad

Regla de la cadena

- Definición de probabilidades condicionales

$$P(B|A) = P(A,B)/P(A) \quad \text{->} \quad P(A,B) = P(A)\,P(B|A) = P(B)\,P(A|B)$$

- Más variables:

$$P(A,B,C,D) = P(A)\,P(B|A)\,P(C|A,B)\,P(D|A,B,C)$$

- Regla de la cadena en general

$$P(x_1,x_2,x_3,\ldots,x_n) =$$
$$P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

Regla de la cadena aplicado a la probabilidad conjunta de palabras en una oración

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2})\ldots P(w_n|w_{1:n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_{1:k-1})$$

P("El agua de Manantiales") =

P(El) × P(agua|El) × P(de|El agua) × P(Mantiales|El agua de)

Todavía problemas

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Probabilidad condicional de oraciones.
- Solución:
  - Supuesto de Markov

P (azul | El agua de Manantiales es hermosamente)

≈ P (azul | hermosamente)

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$$

Supuesto de Markov Bigram

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(w_k | w_{k-1})$$

- En vez de: $\prod_{k=1}^{n} P(w_k | w_{1:k-1})$

Queda así: $P(w_n | w_{1:n-1}) \approx P(w_n | w_{n-N+1:n-1})$

Modelo Unigram

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Modelo Bigram

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

Ejercicios

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

1. Dado el siguiente modelo de lenguaje θ sobre un vocabulario V compuesto de solo 4 palabras: “the”, “machine”, “learning”, “data”. La distribución de θ esta dada por la siguiente tabla:

w	P(w|θ)
machine	0.1
learning	0.2
data	0.3
the	0.4

Calcular y seleccionar la respuesta correcta: P(“machine learning”|θ) =

- 0.004
- 0.02
- 0.2
- 0.3

Ejercicios

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

2. Asumiendo el mismo modelo de lenguaje de la pregunta 1:

Calcular y seleccionar la respuesta correcta: P(“learning machine”|θ) =

- 0.004
- 0.02
- 0.2
- 0.3

Ejercicios

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

3. Asumiendo el mismo modelo de lenguaje de la pregunta 1:

Calcular y seleccionar la respuesta correcta: P(“learning machine learning”|θ) =

- 0.004
- 0.02
- 0.2
- 0.3

Ejercicios

4. Asumiendo que las palabras son generadas por una mezcla de 2 modelos de lenguaje θ_1 y θ_2, donde $P(\theta_1)=0.5$ y $P(\theta_2)=0.5$. Observe la siguiente distribución de palabras:

w	$P(w\|\theta_1)$	$P(P(w\|\theta_2)$
the	0.4	0.15
and	0.4	0.15
genes	0.05	0.3
biology	0.15	0.4

Calcular y seleccionar la respuesta correcta: P(“biology”) =

- 0.55
- 0.15
- 0.175
- 0.275

Ejercicios

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

5. Asumiendo el mismo modelo de la pregunta 4.

Calcular y seleccionar la respuesta correcta: P(“the biology”) =

- 1
- 0.275
- 0.275625
- 0.075625

Ejercicios

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

6. Dado el siguiente modelo de lenguaje θ sobre un vocabulario V compuesto de solo 4 palabras: “the”, “global”, “warming”, “efffects”. La distribución de θ esta dada por la siguiente tabla:

w	P(w|θ)
the	0.3
global	0.2
warming	0.2
effects	X

Calcular y seleccionar la respuesta correcta: P(“effects” | θ) =

- 0.1
- 0.2
- 0.3
- 0

Ejercicios

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

7. Asumiendo el mismo modelo de lenguaje de la pregunta 6:

w	P(w|θ)
the	0.3
global	0.2
warming	0.2
effects	X

Cual de las siguientes sentencias es falsa:

- P(“the global warming effects”|θ)<P(“global warming effects”|θ)
- P(“global warming”|θ)>P(“warming global”|θ)
- P(“text mining”|θ)=0
- P(“global warming”|θ)=0.04

Ejercicios

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

8. Asumiendo que las palabras son generadas por una mezcla de 2 modelos de lenguaje θ_1 y θ_2, donde $P(\theta_1)=0.5$ y $P(\theta_2)=0.5$. Observe la siguiente distribución de palabras :

w	$P(w\|\theta_1)$	$P(w\|\theta_2)$
sports	0.35	0.05
basketball	0.2	0.05
fast	0.3	0.3
computer	0.1	0.4
smartphone	0.05	0.2

Cual es la probabilidad de observar “computer” desde el modelo mixto:

- 0.45
- 0.4
- 0.05
- 0.25

Ejercicios

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

w	$P(w\|\theta_1)$	$P(w\|\theta_2)$
sports	0.35	0.05
basketball	0.2	0.05
fast	0.3	0.3
computer	0.1	0.4
smartphone	0.05	0.2

9. Asumiendo el mismo modelo de la pregunta 9. Queremos inferir cual de las 2 distribuciones de palabras, θ1 y θ2, ha sido usada para generar “computer” y quisieramos calcula la probabilidad que este ha sido generado usando θ1 y θ2. Es decir: $P(\theta_1 |\text{“computer"})$ y $P(\theta_2 |\text{“computer"})$, respectivamente, luego los valores de $P(\theta_1 |\text{“computer”})$ y $P(\theta_2 |\text{“computer"})$ son:

Clave: aplicar regla de bayes

- 0.9 y 0.1
- 0.1 y 0.9
- 0.8 y 0.2
- 0.2 y 0.8

Problemas con modelos N-gram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- N-grams no puede manejar **long-distance dependencies**:

"**The soups** that I made from that new cookbook I bought yesterday **were** amazingly delicious."

- N-grams no trabaja bien modelando nuevas secuencias con significado similar.

Solución: **Large language models**

- Contexto más largo
- Espacio de embeddings en vez de palabras.
- Puede manera sinonimia mejor

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Modelos de lenguaje n-gram

Estimando probabilidades n-gram

Estimando probabilidades bigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Estimando Maxima Verosimilitud

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Ejemplo

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

$$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$$

$$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5 \qquad P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$$

Otro ejemplo:
oraciones del Proyecto: Berkeley Restaurant

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

can you tell me about any good cantonese restaurants close by

tell me about chez panisse

i’m looking for a good place to eat breakfast

when is caffe venezia open during the day

Conteo bigram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- 9222 oraciones

	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

Probabilidades bigram

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

- unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Resultado:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Bigram estimates of sentence probabilities

P(<s> I want english food </s>) =

P(I | <s>)

× P(want | I)

× P(english | want)

× P(food | english)

× P(</s> | food)

= .000031

Que clase de conocimiento representa n-grams?

- P(english | want) = .0011
- P(chinese | want) = .0065
- P(to | want) = .66
- P(eat | to) = .28
- P(food | to) = 0
- P(want | spend) = 0
- P (i | <s>) = .25

grandes ngrams

- 4-grams, 5-grams
- Large datasets of large n-grams have been released
  - N-grams from Corpus of Contemporary American English (COCA) 1 billion words (Davies 2020)
  - Google Web 5-grams (Franz and Brants 2006) 1 trillion words)
  - Efficiency: quantize probabilities to 4-8 bits instead of 8-byte float

Newest model: infini-grams (∞-grams) (Liu et al 2024)

- No precomputing! Instead, store 5 trillion words of web text in **suffix arrays**. Can compute n-gram probabilities with any n!

N-gram LM Toolkits

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- SRILM
  - http://www.speech.sri.com/projects/srilm/
- KenLM
  - https://kheafield.com/code/kenlm/
  - (demo e interpretación)

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Modelo de Lenguaje

Evaluación y perplejidad

Como evaluar modelos de lenguaje n-gram

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- ”Evaluación extrinseca (en-vivo)"

Comparar modelos A y B

1. Coloque cada modelo en una tarea real
   - Machine Translation, speech recognition, etc.
2. Ejecutar la tarea, score de A y B
   - Cuantas palabras traduce bien?
   - Cuantas palabras transcribe bien?
   - Compare la precision de A y B

Intrinsic (in-vitro) evaluation

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Evaluación extrinseca no siempre es possible
  - Caro y consumidor de tiempo
  - No siempre generaliza a otras aplicaciones
- Evaluación intrinseca: **perplexity**
  - Directamente mide el rendimiento del modelo de lenguaje prediciendo palabras
  - No necesariamente corresponde con rendimiento de aplicaciones reales
  - Pero nos da una métrica generica simple para modelos de lenguaje
  - Util para large language models (LLMs) así como n-grams

Conjunto de entrenamiento y prueba

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Entrenamos nuetro modelo con **training set**.

Evaluamos el modelo con datos no vistos (**test set**)

- El **test set** son datos con los que no fue entrenado el modelo.
  - Intuición: medir la generalización con datos no vistos.
- Una **evaluation metric** (como **perplexity**) nos dice que tan bien esta Nuestro modelo con los datos de prueba.

Contenido

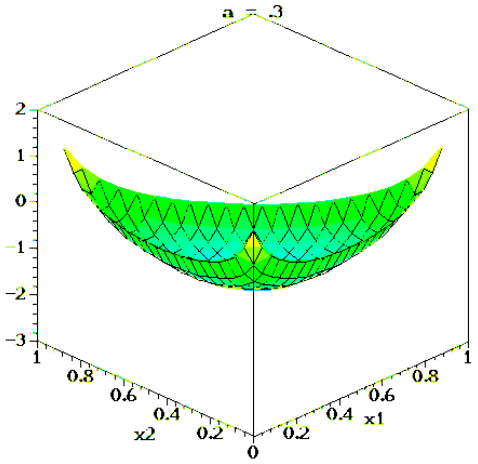
1. Minería de tópicos / Temas
2. Modelos de lenguaje n-gram
3. LDA
4. Categorización de texto

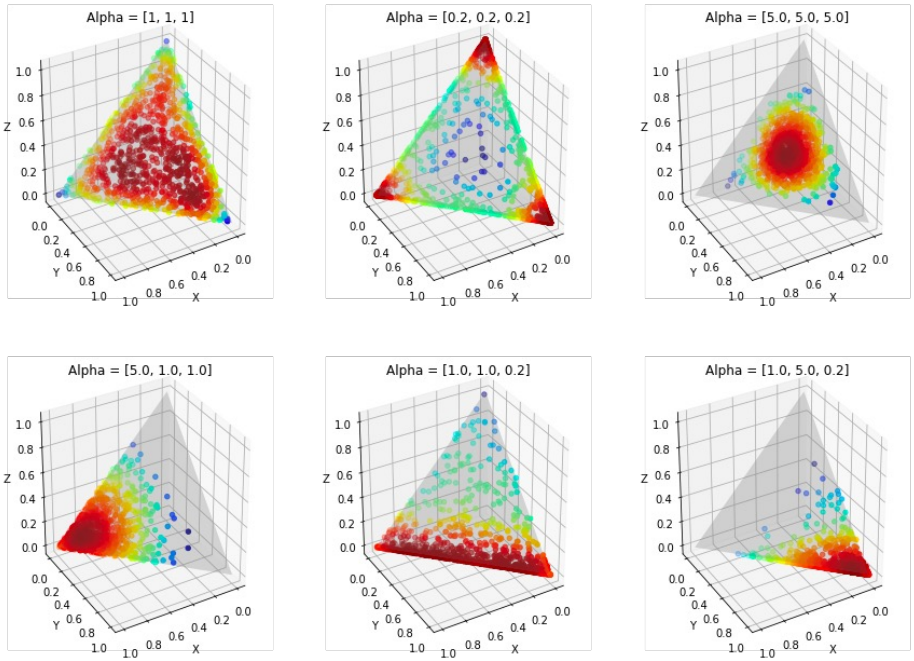
3. Latent Dirichlet Allocation (LDA)

Preliminares: Distribución de Direchlet

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

$$f(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$





https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_Dirichlet

Modelos de Temas Probabilísticos

- Los algoritmos de modelado de temas no requieren anotaciones previas ni etiquetado de los documentos
- **Latent Dirichlet Allocation (LDA)**
  - LDA es un modelo estadístico de colecciones de documentos que intenta captar este proceso generativo de intuición
  - Definimos formalmente un tema para ser una distribución sobre un vocabulario fijo
- el objetivo del modelado de temas es descubrir automáticamente los temas de una colección de documentos.
- El problema computacional central para el modelado de temas es usar los documentos observados para inferir la estructura de temas ocultos.

LDA

- LDA es un modelo probabilístico. modelado probabilístico generativo.
  - tratamos los datos como si fueran el resultado de un proceso generativo que incluye variables ocultas.
  - Este proceso generativo define una distribución de probabilidad conjunta entre las variables aleatorias observadas y las ocultas.
  - utilizando esa distribución conjunta para calcular la distribución condicional de las variables ocultas dadas las variables observadas.
  - Esta distribución condicional también se denomina distribución posterior.
- Las variables observadas son las PALABRAS de los documentos;
- Las variables ocultas son la estructura de TÓPICOS.

Distribución de dirichlet

- Una variable aleatoria de Dirichlet k-dimensional puede tomar valores en el (k-1)-simplex, y tiene la siguiente densidad de probabilidad en este simplex:

$$f(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i-1},$$

$$\sum_{i=1}^{K} x_i = 1$$

$$x_1,\ldots,x_K > 0$$

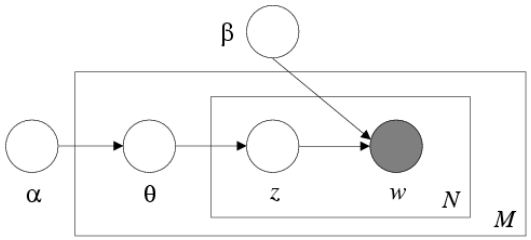
$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_K)}$$

donde $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)$.

Diagrama LDA

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería



Representación del diagrama LDA. Las cajas son "placas" que representan réplicas.

La placa exterior representa documentos

mientras que la placa interior representa la elección repetida de temas y palabras dentro de un documento.

Probabilidad de un documento

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Temas Palabras Temas Palabras

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

Distribuciones Direchlet

Distribuciones Multinomiales

Un ejemplo

"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.

Contenido

1. Minería de tópicos / Temas
2. Modelos de lenguaje n-gram
3. LDA
4. Categorización de texto

UNIVERSIDAD EAFIT

Escuela de
Ciencias Aplicadas
e Ingeniería

4. Categorización de texto

Clasificación de textos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- La categorización de texto es una tarea de asignar una o más categorías predefinidas al documento analizado, en función de su contenido.
- Sea d_i un documento de un conjunto D y $\{c_1, c_2, c_3, \ldots, c_n\}$ un conjunto de todos las etiquetas (labels) a ser asignados. Un Clasificador de Texto asigna uno o más labels c_j a un documento d_i dependiendo del contenido. (single-label vs multi-label)
- Usaremos técnicas de ML supervisadas y no supervisadas, predictivas:
  - Recorderis: ML Supervisadas: Datos {(xi,yi)}, Ni (Ni conjuntos de entrenamiento), yi puede ser una variable categórica, ej: {c1, c2, …. Cn}, de esta forma Text Classification va por esta línea. Ej: análisis de sentimiento
  - Recorderis: ML no supervisada: Datos {xi} Ni, donde encontramos info y conocimiento de los mismos datos, ej: clustering, detección de tópicos, etc.

Evaluación de Clasificación de textos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Conjunto aleatorio de los documentos con label (test set), diferente al training set
- Se clasifica el test set con el Clasificador
- Se compara los labels estimados vs labels reales
- La porción de documentos correctamente clasificados es llamado ‘accuracy’
  - Las métricas comunes son:
    - Precision, recall, F1-score
    - "la precisión es la fracción de las instancias correctas entre las instancias positivas identificadas. Recall es el porcentaje de instancias correctas entre todas las instancias positivas. Y F-1 score es la media geométrica de precisión y recuperación".

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

UNIVERSIDAD EAFIT

Escuela de
Ciencias Aplicadas
e Ingeniería

Técnicas de clasificación

Clasificador Bayesiano

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Es un clasificador NO iterativo de texto, Clasificador Probabilístico
- Basado en el teorema de Bayes
  - Las características (features) son independientes
- El clasificador Naive Bayes esta basado en un modelo de probabilidad condicional.
  - Dado D = (a1,a2,a3, …., an}. -> ai nro de características del documento
  - => p(ck || a1,a2,a3, …, an }

Naïve Bayes Fundamentos

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Modelo probabilistico
- Tenemos probabilidades *a priori* basadas en el conjunto de entranamiento (por eso es supervisada)
- Generamos probabilidades a *posteriori*.
- Formula general:

$$Pr(y \mid X) = \frac{Pr(y) \times Pr(X \mid y)}{Pr(X)}$$

- y son las clases, X los documentos o términos

Ejemplo sencillo

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- Tres clases {Y} = { ‘computer science - CS’, ‘zoology’, ’entertaiment’} = {y1,y2,y3}
- Un conjunto de documento D = {X1, X2, …., Xn}

- Hago dos querys:
  - Query1: ‘python’
  - Query2: ‘python download’

- Un caso: $P(y=CS \mid \text{‘python’}) = \frac{(P(y=CS) \times P(\text{‘python’} \mid y = CS))}{P(\text{‘python’})}$

Otros casos:

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- $P(y=\text{Zoology} \mid \text{'python'}) = \frac{(P(y=\text{Zoology}) \times P(\text{'python'} \mid y=\text{Zoology}))}{P(\text{'python'})}$
- $P(y=\text{Entert} \mid \text{'python'}) = \frac{(P(y=\text{Entert}) \times P(\text{'python'} \mid y=\text{Entert}))}{P(\text{'python'})}$

Clasificador Naïve Bayes

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- y* = MAX(yj) de P(yj|X) = MAX(yj) P(yj) x P(X | yj)
- $P(X \mid yj) = \prod_{i=1}^{n} P(xi \mid yj)$
- y* = MAX(yj) de P(yj|X) = MAX(yj) P(yj) x $\prod_{i=1}^{n} P(xi \mid yj)$
- P(yj) = n / N, n #doctos con la clase y ∈ N
- P(xi | yi) for xi ∈ X, yj ∈ Y -> cuantas veces aparece xi en las instancias con label yj
- Si hay p instancia de la clase yj y xi aparece en k de ellos ->
  - P(xi|yj) = k / p
- Problemas:
  - ? P(xi|y) = 0 ???? Sln -> función de suavización (k+1) / (p+n). n nro de características. O eliminas la xi no vista.
  - Desbalanceo de clases?

NB -> 2 propuestas concretas

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Modelo de bernoulli (clasificador binario de clases)
- Modelo Multinomial (pesos en características, tipo tf, tfidf, etc)

Naïve Bayes – sintesis más formal

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

- El modelo comprende componentes $c_j \in C = \{c_1, c_2, ..., c_k\}$.
- Cada documento $d_i = \{w_1, w_2, ..., w_{ni}\}$ se genera seleccionando primero un componente de acuerdo con los antecedentes, $P(c_j | \theta)$
- luego utiliza el componente para crear el documento de acuerdo con sus propios parámetros, $P(d_i | c_j; \theta)$.
- podemos calcular la probabilidad de un documento utilizando la suma de probabilidades en todos los componentes:

$$P(d_i|\theta) = \sum_{j=1}^{k} P(c_j|\theta) P(d_i|c_j;\theta)$$

Bayes

- Asumimos una correspondencia uno a uno entre
  - clases L = {l1, l2, . . . , lk } y components cj
- por lo tanto cj indica tanto el componente j^{th} como la clase c^{th}.
- En consecuencia, dado un conjunto de ejemplos de entrenamiento etiquetados, D = $\{d_1, d_2, ..., d|D|\}$
- primero aprendemos (estimamos) los parámetros del modelo de clasificación probabilística, θ , y luego usando las estimaciones de estos parámetros, realizamos la clasificación de los documentos de prueba calculando las probabilidades posteriores de cada clase c_j, dado el documento de prueba, y seleccionar la clase más probable (clase con la mayor probabilidad).

$$P(c_j|d_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})P(d_i|c_j;\hat{\theta}_j)}{P(d_i|\hat{\theta})}$$

$$= \frac{P(c_j|\hat{\theta})P(w_1, w_2, \ldots, w_{n_i}|c_j;\hat{\theta}_j)}{\sum_{c\in C} P(w_1, w_2, \ldots, w_{n_i}|c;\hat{\theta}_c)P(c|\hat{\theta})}$$

$$P(w_1, w_2, \ldots, w_{n_i}|c_j;\hat{\theta}_j) = \prod_{i=1}^{n_i} P(w_i|c_j;\hat{\theta}_j)$$

Otras técnicas de clasificación de texto

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Arboles de decisión
- KNN
- Máquinas de soporte vectorial
- Random Forest
- Redes neuronales

Arboles de decisión

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- Técnica supervisada
- El conjunto de entranamiento forma una estructura de arbol
  - Nodos internos denotan una prueba sobre un atributo
  - Una rama representa un resultado del test
  - Las hojas se encuentran las categorias Ci.
  - El arbol genera reglas

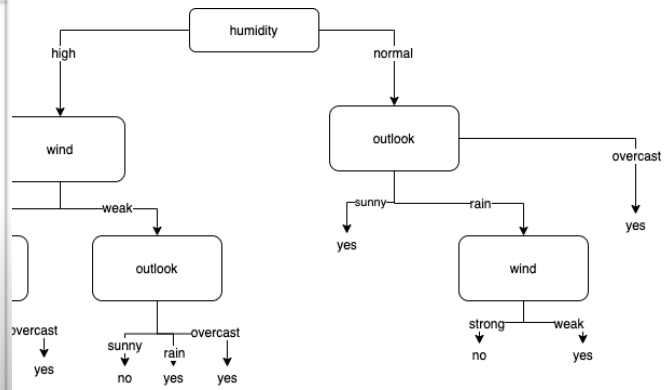
Arboles de decisión: un ejemplo

UNIVERSIDAD EAFIT

Escuela de **Ciencias Aplicadas e Ingeniería**

Dataset de entrada

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



KNN (k nearest neighbour)

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- El algorithmo asume que es posible clasificar documentos en un espacio euclidiano como puntos / distancia entre 2 puntos en un plano con coordenadas p(x,y) y q(a,b) calculado como:
  - $d(p,q) = d(q,p) = \sqrt{((x-a)^2 + (y-b)^2)}$
  - Dado un documento x, ya instanciado en vsm:
  - [w1(x),w2(x),w3(x), ……. wn(x)]

  wj(x) es el peso del termino j, de acuerdo a diferentes métricas (bit-vector,tf, etc)

referencias

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

- [ASG] Anubhav Aggarwal, Jasmeet Singh, Dr Kapil Gupta. A Review of Different Text Categorization Techniques, DOI: 10.14419/ijet.v7i3.8.15210
- [R.Jindal] R. Jindal, R. Malhotra, A. Jain (2015), “Techniques for text classification: Literature review and current trends”, Webology, Volume 12, Number 2.