Conteo de palabras

IMPORTANTE

Seguramente en las diapositivas vienen tareas… por favor revíselas y **realícelas**

Datos numéricos

- De un documento además de obtener patrones lingüísticos podemos obtener datos numéricos.
- ¿Cómo cuales?|

Datos numéricos

- En la extracción automática se abordan 2 técnicas principales:
  - La técnica estadística que se basa principalmente en la frecuencia de aparición de una serie de combinaciones de palabras.
  - La técnica lingüística que se basa en detectar patrones de categorías morfológicas.

La ley de Zipf

- En todo texto escrito hay palabras que se repiten.
  - Contar cuántas veces aparece “de” y obtener un número.
- Si éste se divide entre el número total de palabras del texto, se obtiene su frecuencia.
  - Es la frecuencia de cada palabra que aparece en un texto.

La ley de Zipf

- Podemos hacer una lista
  - Colocar en 1er. lugar la palabra con mayor frecuencia
  - Después la palabra con segundo valor de frecuencia
  - Y así sucesivamente
- Al lugar que ocupa una palabra se llama rango de la palabra.

La ley de Zipf

- Supongamos que en un texto la palabra de más frecuencia es *"de"*, en la lista ocupará el 1er lugar y su rango será **¿?.**
- El artículo *"el"* tiene segundo valor de la frecuencia ocupará el segundo lugar en la lista y tendrá rango dos.
- Existe una relación entre la frecuencia de una palabra y su rango.

La ley de Zipf

- Mientras > sea el rango de una palabra, menor será la frecuencia con la que aparece.
- Mientras > sea su rango, más abajo estará la palabra en la lista, menor será su frecuencia.
- ¿Cómo es la dependencia de la frecuencia del rango?
  - Inversa (disminuye a medida que el rango aumenta) de la primera potencia del rango.

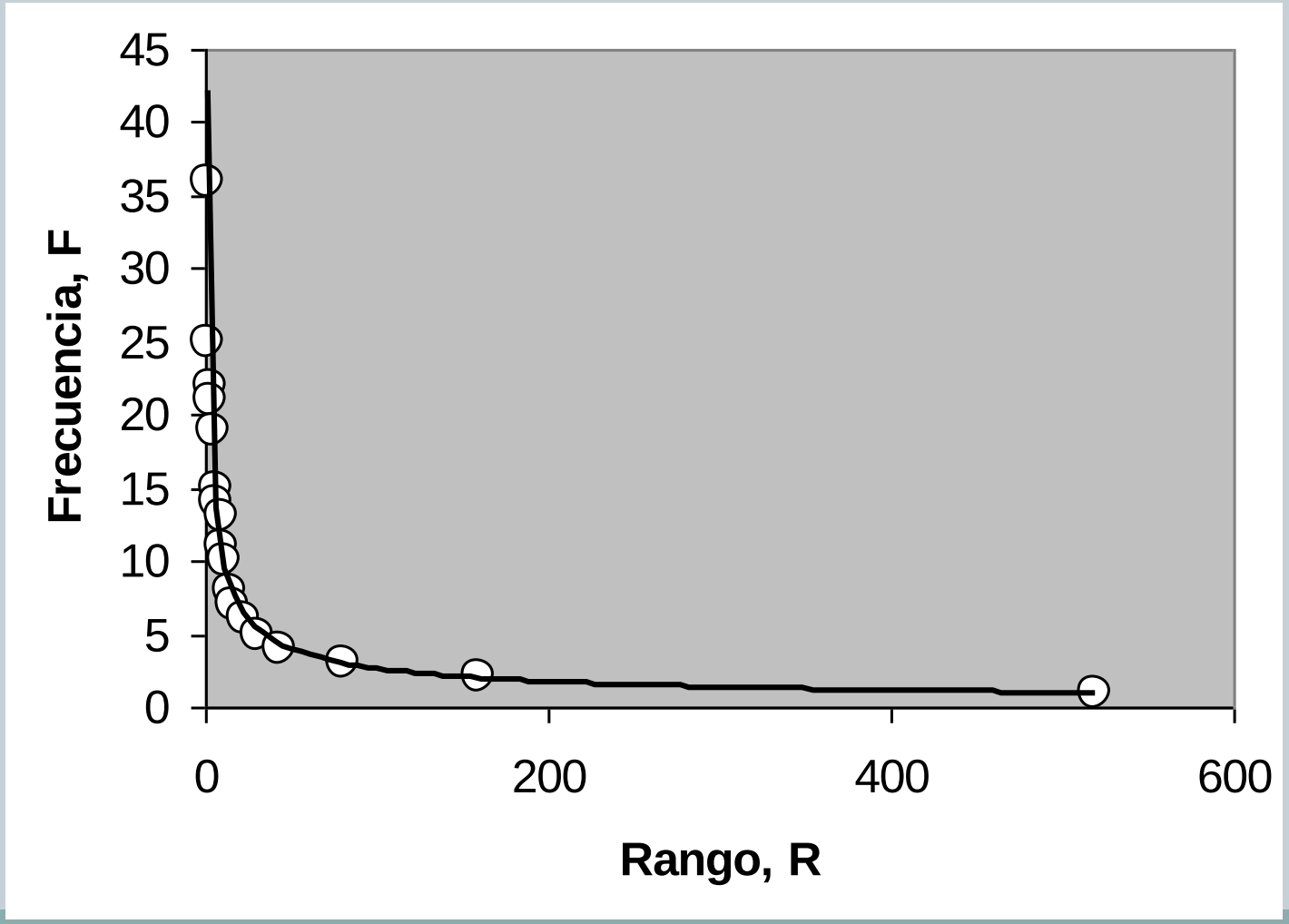
La ley de Zipf

- Sea *f* la frecuencia y *r* el rango.
  - La relación matemática es: f depende de r como (1/r)
- La Ley de Zipf se refiere a esto:
  - *"Un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran número de palabras son poco empleadas"*

La ley de Zipf

- En el español las palabras que encabezan este rango siempre son artículos y preposiciones.
  - Si *"el"* tiene rango 1, palabras como "oxímoron" o "escible" tendrán un rango altísimo.
- En el inglés, la palabra que casi siempre estará de inicio será *"the"*.

La ley de Zipf



La Ley de Zipf: Corpus LEXESP

Tabla#. Dos conjuntos de palabras. El primero incorpora las 20 palabras con mayor frecuencia de aparición en el castellano escrito. El segundo muestra 20 palabras de frecuencia mucho menor. Compárese el tipo de palabras (contenido vs función) y su longitud. Datos del corpus LEXESP (Sebastián-Gallés, et al., 2000) sobre un total de 5.020.930 unidades.

Palabra	Frecuencia
de	264.721
la	192.476
que	153.169
y	140.438
el	139.594
en	116.302
a	91.317
los	83.471
se	68.448
un	62.214
no	55.505
las	53.823
del	49.124
una	47.975
con	47.170
por	40.050
su	39.522
es	33.826
lo	31.481
para	27.646

Palabra	Frecuencia
crees	133
encontraban	133
entendido	133
explosión	133
frecuentes	133
guarda	133
hermanas	133
intelectuales	133
judíos	133
llegaban	133
monjas	133
moverse	133
occidente	133
sacado	133
seguían	133
sienten	133
sirvió	133
sospecha	133
sucedió	133
tela	133

Ley de Zipf

- Si aplicamos la ley de Zipf al universo de 5.020.930 unidades contenidas en el corpus LEXESP podemos calcular la frecuencia teórica esperable para cada rango.
  - Multiplicando la probabilidad de aparición por ese número.

Ley de Zipf

- En la tabla # se ven las frecuencias empíricas (datos reales).
- La 1ra. palabra *"de"* aparece en el corpus 264.721 veces, y según la ley debería aparecer 502.093.
- La palabra No. 20 *"para"* aparece 27.645 veces y la ley predice [(0,1 / 20) x 5.020.930)] = 25.105.
- La que ocupa el lugar 2.000 aparece 220 veces, y la ley predice 251, etc.

Tabla#. Frecuencias empíricas y frecuencias teóricas esperables según la ley de Zipf, para distintas palabras ordenadas por su uso. Las frecuencias corresponden a valores absolutos de un total de 5.020.930 unidades léxicas. Datos del corpus LEXESP (Sebastián-Gallés, Martí, Carreiras, Cuetos, 2000).

Nº orden	Frec. empírica	Frec. teórica	Nº orden	Frec. empírica	Frec. teórica	Nº orden	Frec. empírica	Frec. teórica
1	264.721	502.093	20	27.645	25.105	2000	220	251
2	192.476	251.047	40	7.933	12.552	2500	176	201
3	153.169	167.364	60	5.830	8.368	3000	148	167
4	140.438	125.523	80	4.157	6.276	3500	127	143
5	139.594	100.419	100	3.350	5.021	4000	110	126
6	116.302	83.682	250	1.467	2.008	10000	40	50
7	91.317	71.728	500	744	1.004	15000	24	33
8	83.471	62.762	750	521	669	20000	16	25
9	68.448	55.788	1000	405	502	25000	12	20
10	62.214	50.209	1500	289	335			

Ley de Zipf

- Se representan estos valores sobre ejes logarítmicos.
- Las frecuencias teóricas son la línea recta, de pendiente negativa
- Las frecuencias empíricas son la línea irregular
- Ajuste razonablemente bueno entre ambas

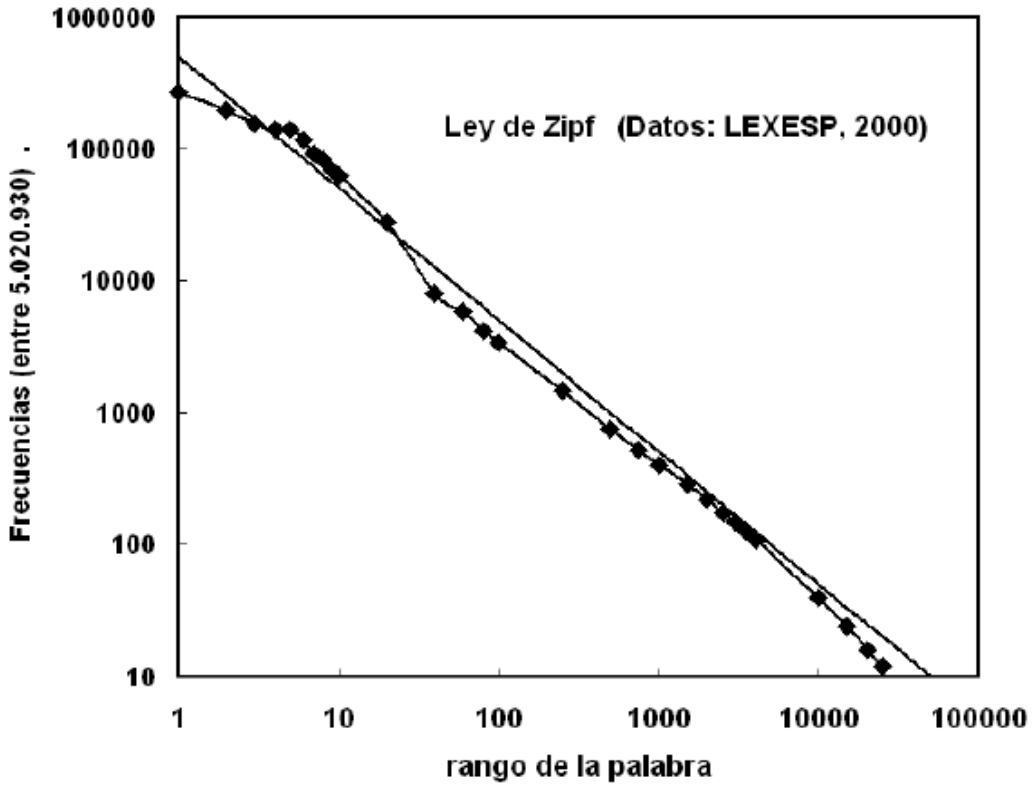


Figura #. Ley de Zipf aplicada al castellano (explicación en el texto). La abscisa corresponde al rango o puesto que ocupan las palabras ordenadas de mayor a menor frecuencia. La ordenada indica las frecuencias absolutas sobre un universo de 5.020.930 unidades. Los ejes son logarítmicos. Los datos empíricos pertenecen al copus LEXESP (Sebastián-Gallés, et al., 2000).

La ley de Zipf

- La ley de Zipf también da la dependencia de la frecuencia de ocurrencia de una palabra con respecto al número de palabras que se usen (amplitud del vocabulario utilizado).
  - Mientras < sea el vocabulario, > será la frecuencia de las palabras en los primeros rangos.
  - En un texto en español con un vocabulario de alrededor de 10 000 palabras, las frecuencias de las palabras de mayor rango, como *"de"*, *"el"*, *"y"*, son 0.11, 0.06, 0.33, respectivamente.

La ley de Zipf

- Mejora sustancial de la ley Zipf.
- En unas 10 lenguas, la previsibilidad de lo que una persona dice se ve más influenciado por la longitud de la palabra que por la frecuencia.
  - La longitud de una palabra es inversamente proporcional a la cantidad de información que contiene.

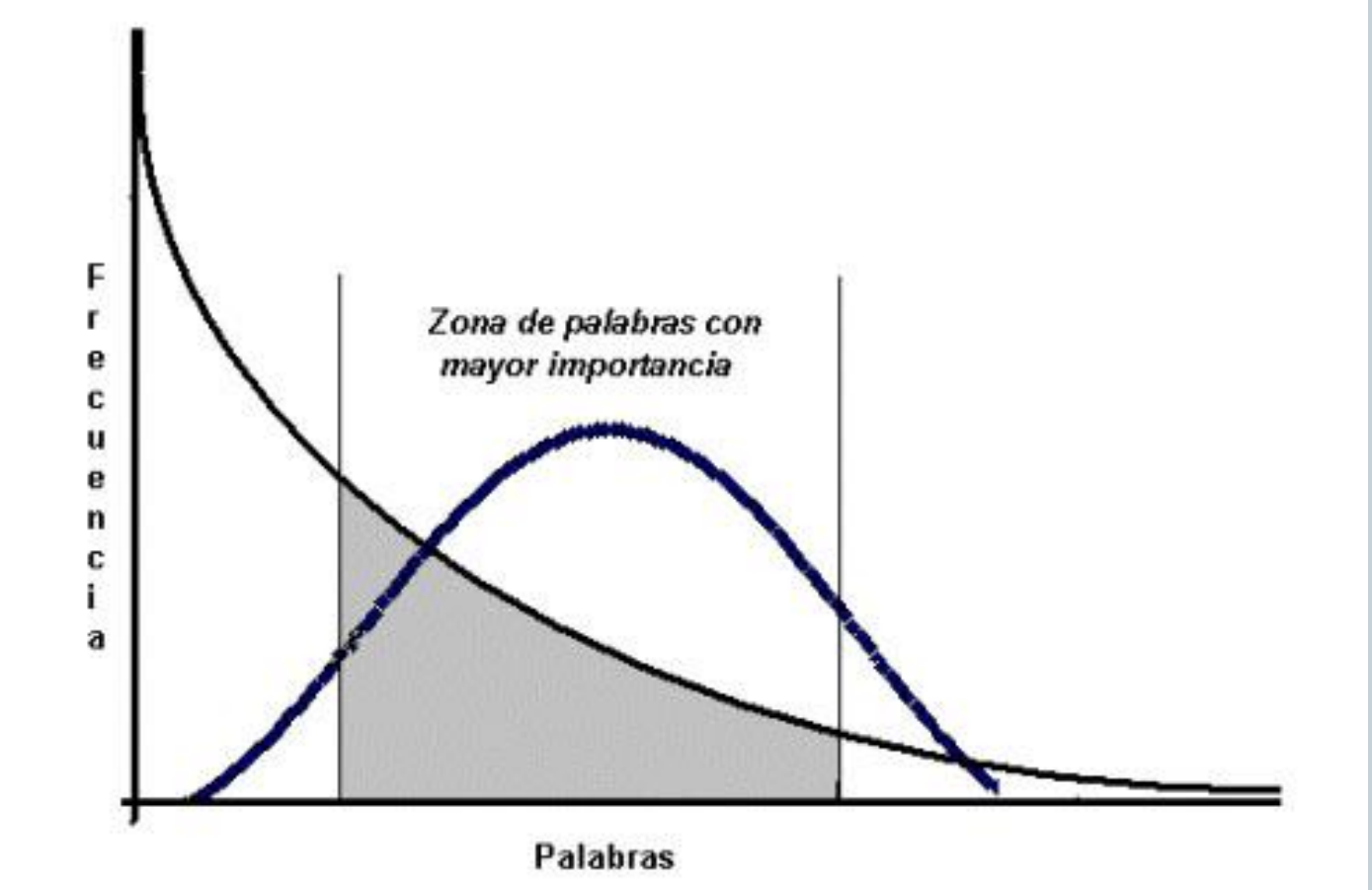
La ley de Zipf

- Propiedad universal del lenguaje humano:
  - *Las palabras más frecuentes tienden a ser cortas, ya que hacen la comunicación más eficiente que usando palabras largas.*
- Debido a la presión de la eficacia comunicativa.
  - Preguntar a todos en una cena si quieren un *plato de sopa* usando una palabra de 15 letras para la preposición “de”.

La ley de Zipf

- El comportamiento del vocabulario da indicios acerca de la utilidad de los términos.
  - Luhn sugirió: las palabras que describen de mejor forma el contenido se encuentran en una área comprendida entre las altamente frecuentes y las muy raras (baja frecuencia)

Ley de Zipf


Frecuencia
Zona de palabras con mayor importancia
Palabras

Ley de Zipf

- Las divisiones verticales definen una zona de transición entre las palabras de frecuencia muy alta y las de muy baja.
  - Están los términos con > contenido semántico.
- El límite superior indica las palabras vacías:
  - No se indexan por no tener poder de discriminación.

Paréntesis… Palabras vacías

- Palabras vacías o *stop-words*
- Son palabras que son filtradas en el preprocesamiento.
- Pueden ser artículos, conjunciones, preposiciones
- Palabras que por sí solas no transmiten carga temática.

A, acá, además, ahí, ahora, al, algo, algún, alguna, algunas, alguno, algunos, allá, allí, ante, antes, aparte, apenas, aquel, aquella, aquellas, aquello, aquellos, aquí, así, aun, aún, aunque, bajo, bastante, bien, bueno, cabe, casi, catorce, cerca, cien, ciento, cientos, cierto, cinco, cincuenta, como, con, conmigo, consigo, contigo, contra, cual, cuales, cualesquier, cualesquiera, cualquier, cualquiera, cuando, cuanta, cuantas, cuanto, cuantos, cuarenta, cuarto, cuatro, cuya, cuyas, cuyo, cuyos, de, debajo, décimo, del, delante, demasiado, dentro, desde, después, detrás, diecinueve, dieciocho, dieciséis, diecisiete, diez, doce, donde, dos durante, e, el, él, ella, ellas, ello, ellos, embargo, empero, en, encima, enseguida, entonces, entre, esa, esas, ese, eso, esos, esta, estas, este, esto, estos, frente, fuera, hacia, hasta, hoy, incluso, jamás, junto, justo, la, las, le, lejos, lo, los, luego, malo, mañana, manera, mas, me, menos mi, mía, mientras, mil, miles, millón, millones, mío, mis, misma, mismas, mismo, mismos, momento, mucha, muchas, mucho, muchos, muy, nadie, ni, ninguna, ningunas, ninguno, ningunos, no, nos, nosotras nosotros, noventa, nuestra, nuestras, nuestro, nuestros, nueve, nunca, o ochenta, ocho, octavo, once, ora, os, otra, otras, otro, otros, para, parte, peor, pero, pocas, poco, pocos, por, porque, primero, principio, pronto, pues, puesto, que, quien, quienes, quince, quinto, quizá, quizás, repente se, sea, según, segundo, seis, séptimo, sesenta, setenta, si, siempre, siete, sin, sino, so, sobre, solo, su, sus, suya, suyas, suyo, suyos, tal, tales, también, tampoco, tan, tanta, tantas, tanto, tantos, tarde, te, tercero, ti, toda, todas, todavía, todo, todos, tras, través, trece, treinta, tres, tu, tus, tuya, tuyas, tuyo, tuyos, un, una, unas, uno, unos, usted, ustedes, veinte, vez, vosotras, vosotros, vuestras, vuestro, vuestros, y, ya, yo.

Ley de Zipf

- El límite inferior indica el comienzo de las palabras más raras.
  - No se incluyen en el vocabulario, existe una baja probabilidad de que el usuario las use.

Ley de Zipf

- Las palabras de baja frecuencia denotan la riqueza y el estilo de vocabulario del autor o son errores de ortografía.
- Para la frecuencia límite se sugiere:
  - Eliminar aquellos términos que estén en 3 o menos documentos
  - Eliminar todas las palabras que ocurren una o dos veces.

Ley de Heaps

- Se plantea una relación entre el tamaño del texto (cantidad de palabras) y el crecimiento del vocabulario (cantidad de palabra únicas).

Ley de Heaps

- El tamaño del vocabulario (y su crecimiento) es una función del tamaño del texto.

$$\mathbf{V = K * N^{\beta}}$$

  - N: tamaño del documento (cantidad de palabras)
  - K: constante que depende del texto, entre 5 y 50.
  - β: constante que depende del texto, entre 0.4 y 0.6

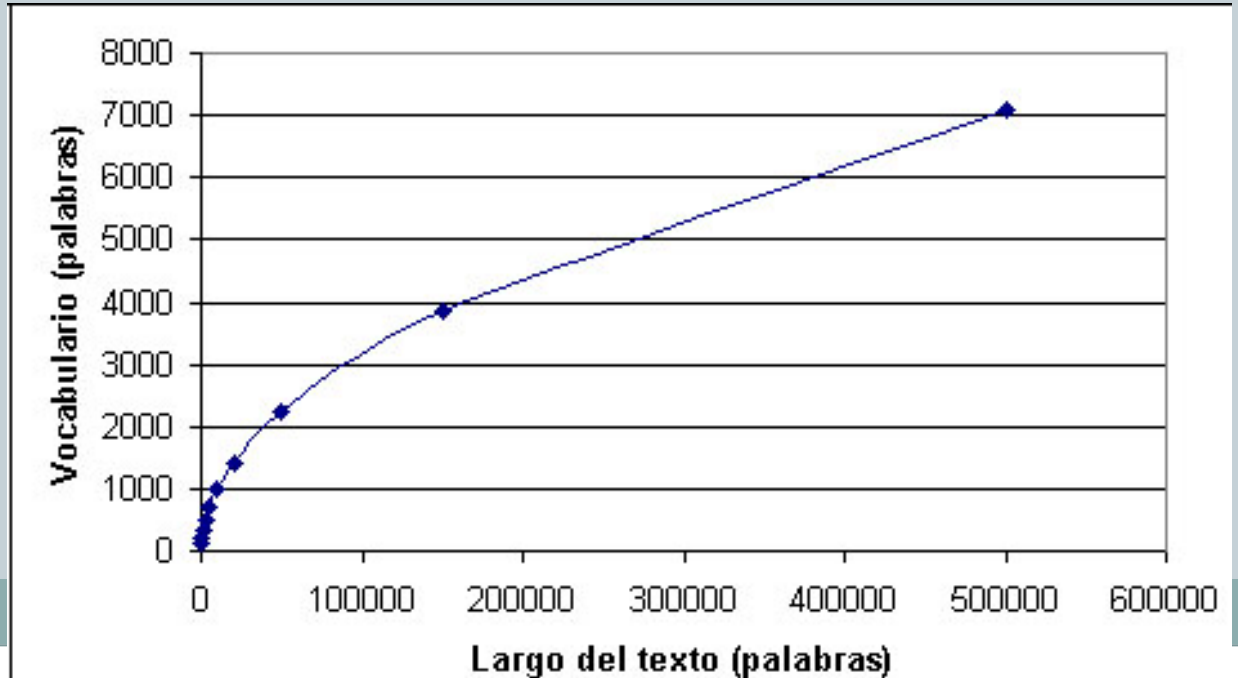
Ley de Heaps

- Si K=20 y β=0.5

N	V
100,000	6325
250,000	10000
400,000	12,649
800,000	17,889
1,000,000	20,000

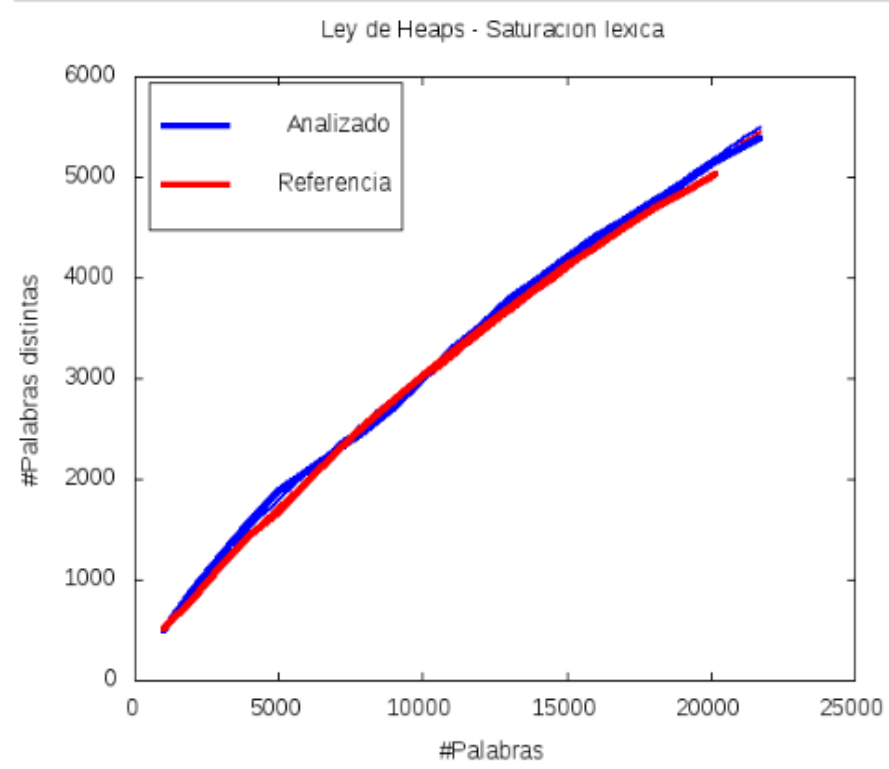
Ley de Heaps

- El tamaño del corpus creció 10 veces, el vocabulario superó las 3 veces su tamaño inicial.
- A medida que se agregan documentos a una colección, se descubrirán nuevos términos para el vocabulario.



Ley de Heaps

- Con la curvatura de la gráfica se puede saber si el texto es rico en vocabulario, la línea será más vertical
- Si está llegando a la saturación , la línea se curvará hasta llegar a la horizontalidad (no aparecen palabras nuevas).



TEXTO ANALIZADO		TEXTO DE REFERENCIA	
N° PALABRAS	N° PALABRAS DISTINTAS	N° PALABRAS	N° PALABRAS DISTINTAS
1000	497	1000	507
2000	895	2000	804
3000	1248	3000	1142
4000	1568	4000	1425
5000	1883	5000	1658
6000	2079	6000	1975
7000	2275	7000	2269
8000	2477	8000	2533
9000	2712	9000	2784
10000	2980	10000	3021
11000	3284	11000	3239
12000	3510	12000	3476
13000	3786	13000	3692
14000	3984	14000	3913

¡Predicción de palabras!

Predicción de palabras

- Las acciones se…

Predicción de palabras

- Las acciones se derrumbaron esta…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva Federal, mientras Wall…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva Federal, mientras Wall Street volvió a operar por primera…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva Federal, mientras Wall Street volvió a operar por primera vez desde los ataques…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva Federal, mientras Wall Street volvió a operar por primera vez desde los ataques terroristas del…

Predicción de palabras

- Las acciones se derrumbaron esta mañana, pese a la baja en las tasas de interés por parte de la Reserva Federal, mientras Wall Street volvió a operar por primera vez desde los ataques terroristas del 11 de septiembre.

Predicción de palabras

- Y ahora estas:
  - Como es de público…
  - Quedamos a la espera de…
  - Solicitamos a los amables…
  - Tómalo con…
- ¿Puede venir cualquier palabra? ¿con cualquier categorías gramatical?

Predicción de palabras

- En alguna medida, es posible predecir palabras futuras en una oración… ¿Como?
  - Conocimiento del dominio.
    - *baja en las tasas de interés*
  - Conocimiento sintáctico.
    - *el <sustantivo>, se <verbo>*
  - Conocimiento léxico.
    - *ataques terroristas, Reserva Federal*

Predicción de palabras

- En Google :
  - “Como es de público” -> 12,800,000
  - “Como es de público conocimiento” -> 1,320,000
- También podría ser:
  - “Como es de publico y notorio”
  - “Como es de publico entendimiento”
  - “Como es de publico dominio”
  - “Como es de publico reconocimiento”
  - **Tarea : leer http://searchenginelnd.com/how-google-instant-autocomplete-suggestions-work-62592**

Predicción de palabras

- Enfoque simbólico: reglas que “adivinen” la próxima palabra.

  if(entrada=“como es de publico”)

  then (próxima palabra=“conocimiento”)

  - ¿Problemas de este enfoque?

- Enfoque estadístico: modelos de N-gramas

N-gramas

- El PLN estadístico busca hacer inferencia estadística en el campo del LN.
- La inferencia estadística consiste en tomar algún dato (generado con alguna distribución de probabilidad) y hacer algunas inferencias acerca de esta distribución.

N-gramas

- Adivinar la siguiente palabra o predicción de palabra es una sub-tarea esencial:
  - De reconocimiento del habla, reconocimiento de escritura, detección de errores de deletreo.
  - La identificación de palabras es difícil por la entrada ruidosa y ambigua.
- *Ver* palabras previas puede darnos una idea acerca de cuales serán las palabras siguientes.

N-gramas

Traducción automática

落实义务教育法，关心教师的今天和明天

(a) Implementación de la Ley de Educación, en vista de...
(b) Implementación a la Educación Ley, vista de...
(c) Implementación Ley Educación, en vista de...

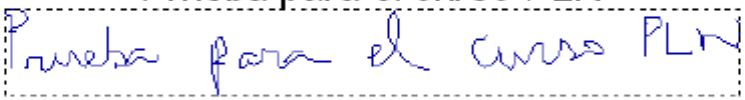
Corrección de errores

Sus **hordas** de Botero, sus hoteles de paso.

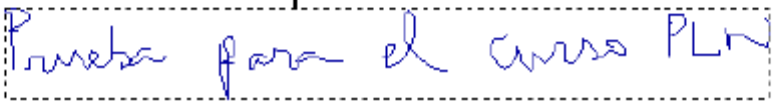
Reconocimiento del habla.

Reconocimiento de escritura.

Prneba para el cnrso PLN



Prueba para el curso PLN



N-gramas

- Adivinar la siguiente palabra está fuertemente relacionado a otro problema:
  - Calcular la probabilidad de una secuencia de palabras.
- Algoritmos que asignan una probabilidad a una sentencia pueden usarse para asignar una probabilidad a la siguiente palabra en una sentencia incompleta.

N-gramas

- Este modelo de predicción de palabras es el N-grama.
- Un modelo N-grama utiliza las N-1 palabras previas para predecir la siguiente.
- Los *n-gramas* de palabras son combinaciones de *n* palabras consecutivas.

N-gramas

- En la frase:
  - *Nuestro sistema de gestión empresarial incluye un programa de facturación y una base de datos de recursos humanos.*
- Los 1-*gramas* presentes en el texto son:
  - *Nuestro, sistema, de, gestión, empresarial, incluye, un, programa, facturación, y, una, base, datos, recursos, humanos.*

N-gramas

- Los 2-*gramas* son:
  - *Nuestro sistema, sistema de, de gestión, gestión empresarial, empresarial incluye, incluye un, un programa, programa de, de facturación, facturación y, y una, una base, base de, de datos, datos de, de recursos, recursos humanos.*
- Los 3-*gramas* son:
  - *Nuestro sistema de, sistema de gestión, de gestión empresarial, gestión empresarial incluye, empresarial incluye un, incluye un programa, un programa de, programa de facturación, de facturación y, facturación y una, y una base, una base de, base de datos, de datos de, datos de recursos, de recursos humanos.*
- Así sucesivamente hasta el orden *n* deseado.

N-gramas

- Un modelo de N-gramas (modelo de lenguaje) intenta predecir la próxima palabra de una oración a partir de las N-1 anteriores.
- El orden importa:
  - Programación de lenguaje.
  - Lenguaje de programación.
  - ¿Otro ejemplo?

N-gramas

- Del orden correcto nos podemos dar cuenta por la sintaxis y la semántica.
  - Pero los N-gramas se basan en probabilidades.
- Los N-gramas pueden usarse para asignar la probabilidad de una oración completa.

N-gramas

- ¿Que elementos vamos a contar para modelar el lenguaje?
  - *Con alivio, con humillación, con terror, comprendió que él también era una apariencia, que otro estaba soñándolo.*
- ¿14, 17 o 22 palabras?
- Fragmentos, rellenos, repetición de palabras, mayúsculas, formas flexionadas…

N-gramas simples

- Necesario considerar la forma de asignar probabilidades a cadenas de palabras:
  - Para calcular la probabilidad de una sentencia entera o para dar una predicción probabilística de cual será la siguiente palabra en una secuencia.
- Modelo más simple: cualquier palabra puede seguir a cualquier otra.

N-gramas

- Cada palabra tiene una probabilidad igual de seguir otra palabra.
  - ¿Cuántas palabras tiene el español?
    - Pocas, muchas, chorrocientas.

> R. El diccionario de la RAE contiene 88.000 palabras. El de americanismos 70.000; pero en este último aparecen muchas variantes que en el diccionario académico ocuparían una sola entrada, como guaira, huaira, huayra, waira, wayra, guayra. Se suele estimar el léxico de una lengua añadiendo un 30% al de los diccionarios. En cuanto a la posición del español en número de palabras, solo puede responderse con respecto a las que aparecen en los diccionarios y para ello basta con comparar las 150.000 de nuestro *Diccionario histórico* con las 350.000 del Oxford.

  - La probabilidad de cualquier palabra después de otra palabra sería 1/100,000 o 0.00001

N-gramas

- Otro modelo: cualquier palabra podría seguir cualquier otra, *pero* la siguiente debería aparecer con su frecuencia normal de ocurrencia.
  - Ejemplo: la palabra *el* tiene una frecuencia relativa alta, ocurre 69,971 veces en el corpus Brown de 1,000,000 (7% de las palabras en este corpus son “el”).
  - La palabra conejo ocurre sólo 11 veces.

N-gramas

- Utilizar frecuencias para asignar una distribución de probabilidad a través de las siguientes palabras.
  - Si hemos visto la cadena "*entonces*", podemos utilizar la probabilidad 0.07 para *el* y 0.00001 para *conejo* para adivinar la siguiente palabra.
- Suponga que hemos visto:
  - *Entonces el conejo*

N-gramas

- En este contexto *blanco* parece ser una palabra más razonable de seguir a conejo que la propia *el*.
  - Buscar en la probabilidad condicional de una palabra dadas las palabras previas.
  - La probabilidad de tener *blanco* dado que hemos visto *conejo* (blanco | conejo) es más alta que la probabilidad de blanco en otro caso.

N-gramas

- Como calcular la probabilidad de una cadena completa de palabras $(w_1,\ldots,w_n$ o $w_1^n)$
- Consideramos cada palabra ocurriendo en su ubicación correcta como un evento independiente.
- Se representa la probabilidad como:

$$P(w_1, w_2,\ldots,w_{n-1}, w_n)$$

N-gramas

- Descomponiendo la probabilidad:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

- Se deben calcular las probabilidades como:

$$P(w_n|w_1^{n-1})$$

N-gramas

- Es decir:
  - ¿Cómo calcular la probabilidad de una palabra, dada una secuencia larga de palabras precedentes?
- Simplificación útil:
  - Aproximar la probabilidad de una palabra dadas todas las palabras previas.
  - Aproximación muy simple: la probabilidad de la palabra dada sólo la palabra previa.

N-gramas

- Mientras más larga sea la secuencia, es menos probable que la encontremos en un conjunto de entrenamiento:

  Encontrar: P(La mayoría de biólogos y especialistas en folklore creen, de hecho, que los cuernos del unicornio mítico derivaron del)

N-gramas

- El modelo bigrama aproxima la probabilidad de una palabra dadas todas las palabras previas $P(w_n | w_1^{n-1})$ por la probabilidad condicional de la palabra precedente $P(w_n | w_{n-1})$
  - En lugar de calcular P(conejo | apenas el otro día vi un).
  - Se aproxima con P(conejo | un)

N-gramas

- Presunción de Markov: la probabilidad de una palabra depende sólo de la palabra previa.
- Modelos de Markov: modelos probabilísticos, podemos predecir la probabilidad de alguna unidad futura sin ver muy lejos en el pasado.

N-gramas

- Podemos generalizar el bigrama (busca una palabra en el pasado) al trigrama (busca 2 palabras en el pasado) y así hasta el n-grama (busca N-1 palabras en el pasado).
  - A un valor de n más alto, mayor cantidad de datos se necesitan para entrenar los modelos.

N-gramas

- La aproximación N-grama a la probabilidad condicional de la siguiente palabra en una secuencia es:

$$P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-N+1}^{n-1})$$

- La probabilidad de una palabra w_n dadas todas las palabras previas pueden ser aproximadas por la probabilidad dadas sólo las N palabras previas.

N-gramas

- Para un enfoque con bigrama, se calcula la probabilidad de una cadena completa al sustituir las ecuaciones:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k | w_{k-1})$$

Ejemplo

- La probabilidad de "El gato está en el tapete":

  P(el gato está en el tapete) = P(el | <s>)

  P(gato | <s> el)

  P(está | <s> el gato)

  P(en | <s> el gato está)

  P(el | <s> el gato está en el)

  P(tapete | <s> el gato está en el)

  P(</s> | <s> el gato está en el tapete)

- <s> y </s> indican el inicio y final de la sentencia.

N-gramas

- No es una solución práctica.
- Tomar sólo los dos tokens previos
  - P(el gato está en el tapete) = P(el | <s>)
    - P(gato | <s> el)
    - P(está| el gato)
    - P(en | gato está)
    - P(el| está en)
    - P(tapete | en el)
    - P(</s> | el tapete)

N-gramas

- Ejemplo de un sistema de entendimiento del habla.
- Los usuarios hacen preguntas sobre restaurantes y el sistema despliega información apropiada de una BD de restaurantes locales.

N-gramas

- Ejemplos de consultas:
  - Estoy buscando por comida cantonesa.
  - Me gustaría cenar en un lugar cercano.
  - Dime acerca Chez Panisse
  - ¿Podría darme una lista de los tipos de comida que están disponibles?
  - Estoy buscando un buen lugar para desayunar.
  - Definitivamente no quiero comer comida china.
  - No quiero caminar más de 10 minutos

N-gramas

- Se tiene una muestra de las probabilidades de bigramas para alguna de las palabras que pueden seguir la palabra *eat*, tomadas de oraciones reales habladas por usuarios.

N-gramas

Eat on	.16	Eat Thai	.03
Eat some	.06	Eat breakfast	.03
Eat lunch	.06	Eat in	.02
Eat dinner	.05	Eat chinese	.02
Eat at	.04	Eat mexican	.02
Eat a	.04	Eat tomorrow	.01
Eat indian	.04	Eat dessert	.007
Eat today	.03	Eat British	.001

N-gramas

- Asuma que además de las probabilidades, también se incluyen las probabilidades de bigramas.

<s> I .25	I want .32	want to .65	to eat .26	British food .60
<s> I ‘d .06	I would .29	want a .05	to have .14	British restaurant .15
<s> Tell .04	I don’t .08	want some .04	to spend .09	British cuisine .01
<s> I’m .02	I have .04	want thai .01	to be .02	British lunch .01

N-gramas

- Un ejemplo conocido

Estimar la probabilidad de la oración:

- *I want to eat Chinese food.*

$$P(\text{I want to eat Chinese food}) =$$
$$P(\text{I} \mid \langle\text{start}\rangle)\ P(\text{want} \mid \text{I})\ P(\text{to} \mid \text{want})\ P(\text{eat} \mid \text{to})$$
$$P(\text{Chinese} \mid \text{eat})\ P(\text{food} \mid \text{Chinese})\ P(\langle\text{end}\rangle \mid \text{food})$$

¿Qué necesitamos para estos cálculos?

- Probabilidad $P(w_m|w_{m-1})$ para cada par de palabras.
- Pre-calculadas de un corpus grande.

N-gramas

- Calcular la probabilidad de sentencias:
  - *I want to eat British food*
  - *I want to eat Chinese food*
  - P(I want to eat British food) = P(I|<s>) P(want|I) P(to|want) P(eat|to) P(British|eat) P(food|British)

    = .25 * .32 * .65 * .26 * .002 * .60

    = .000016

N-gramas

- Problema:
  - Por definición las probabilidades son menores a 1, el producto de muchas probabilidades se vuelve menor conforme más probabilidades se multipliquen.

N-gramas

- En un modelo trigrama se condiciona bajo las 2 palabras previas
  - P(food|eat British) en lugar de P(food|British)
- Para calcular las probabilidades de trigramas al inicio, se utilizan las etiquetas:
  - P(I|<start1><start2>)

N-gramas

- Los modelos N-gramas pueden ser entrenados al contar y normalizar:
  - Normalizar : dividir por algún conteo total para que las probabilidades caigan legalmente entre 0 y 1.

N-gramas

- Se toma un corpus de entrenamiento.
- Se toma la cuenta de un bigrama particular.
- Se divide la cuenta por la suma de todos los bigramas que comparten la misma primera palabra.

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

N-gramas

- Simplificación: la suma de todos los conteos de bigramas que empiezan con una palabra dada w_{n-1} debe ser igual al conteo de unigramas para esa palabra w_{n-1}.

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

N-gramas

- Caso general de estimación de parámetros N-Grama:

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}$$

  - Estima la probabilidad al dividir la frecuencia observada de una secuencia particular por la frecuencia de un precedente.
- Este radio es llamado una frecuencia relativa.
  - Para estimación de probabilidades, es un ejemplo de Maximum Likelihood Estimation(MLE).

N-gramas

- El conjunto de parámetros resultante es uno de los cuales la probabilidad del conjunto de entrenamiento T dado el modelo M
  - P(T|M) es maximizada
- Suponga que la palabra Chinese ocurre 400 veces en un corpus de un millón de palabras.
- ¿Cual es la probabilidad de que ocurra en algún otro texto de un millón de palabras?
  - La estimación MLE es 400/1000000 o 0.0004

N-gramas

- .0004 no es la mejor estimación posible de la probabilidad de que Chinese ocurra en todas las situaciones.
- Es la probabilidad que hace “más probable” que aparezca Chinese 400 veces en un corpus de un millón de palabras.

N-gramas

- Se muestran los conteos de bigramas.

BERP (Berkeley Restaurant Project)

- Consultas de usuarios a un sistema de diálogo hablado.

	I	*Want*	*To*	*Eat*	*Chinese*	*Food*	*lunch*
I	8	1087	0	13	0	0	0
Want	3	0	786	0	6	8	6
To	3	0	10	860	3	0	12
Eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
Food	19	0	17	0	0	0	0
Lunch	4	0	0	0	0	1	0

- $P(want \mid I) = \#(I\ want) / \#(I) = 1087 / 3437 = 0.32$

- En la siguiente tabla, se muestan las probabilidades de bigramas después de la normalización:
  - Dividir cada fila por los conteos de unigramas.
  - I 3437
  - want 1215
  - to 3256
  - eat 938
  - Chinese 213
  - food 1506
  - lunch 459

N-gramas

¿Qué cosas captura este modelo?

- *P(want | I)* = .32
- *P(to | want)* = .65
- *P(eat | to)* = .26
- *P(food | Chinese)* = .56
- *P(lunch | eat)* = .055
- *P(I | I)* = .0023 *I I I I want*
- *P(I | want)* = .0025 *I want I want*
- *P(I | food)* = .013 *the kind of food I want is...*

N-gramas

- Se muestran las probabilidades de bigramas después de la normalización

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Actividad

- Leer *More on N-grams and their sensitivity to the training corpus.*

Tarea

- ¡La tarea!:
  - Buscar que es Google books Ngram Viewer y su funcionamiento, hacer un ejemplo.
  - Revisar las herramienta goldfish http://www.equiposcreativos.com/blog/redaccion-de-contenidos/herramienta-para-investigacion-de-palabras-claves-en-tiempo-real-gofish y http://www.seomoz.org/blog/using-social-media-to-get-ahead-of-search-demand

Proyecto 5

- En equipos construir un corpus de documentos (500 documentos), y dividirlos en clases.
  - Pueden ser sólo dos clases.
  - Descargar e instalar Weka

Por favor avancen en sus temas de tesis.