

Motores de búsqueda

Elementos de un motor de búsqueda

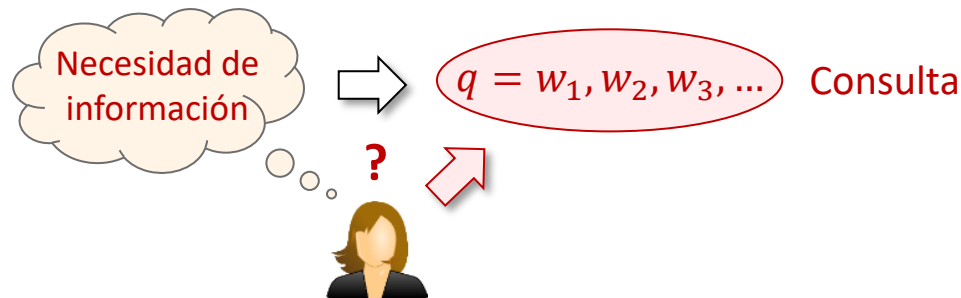
- ◆ Indexación del espacio de búsqueda
 - Extracción de “palabras” de los documentos (en el caso de texto)
 - Cómputo de ponderaciones palabra/documento
 - Creación de estructuras de índice
- ◆ Definición de una función de ránking
 - Búsqueda simple: funciones no supervisadas
 - Ránking avanzado: learning to rank supervisado
 - Típicamente multifase
- ◆ Docenas de refinamientos críticos
 - Comprensión de consultas: clasificación, erratas, sinonimia, NLP
 - Pre y post-filtros
 - Multilingualidad
 - Interfaz conversacional
 - Índices distribuidos, poda de resultados, cache
 - Etc.

Definición del problema de recuperación de información

- ♦ Dada una **necesidad de información** de un usuario
- ♦ Devolver al usuario la mayor cantidad posible de **información relevante** para su necesidad de información
- ♦ Y la menor cantidad posible de información no relevante
- ♦ Presentar la respuesta como **ranking** de opciones
para lidiar con el tamaño e incertidumbre del resultado

Necesidad de información

- ♦ Una información de la que el usuario precisa para realizar un objetivo
 - Precisa o vaga, estable o dinámica
- ♦ El usuario formula su necesidad en forma de consulta (o no)
 - Lista de palabras, preguntas, formulario, un ejemplo
 - No toda necesidad de información requiere un sistema IR!
- ♦ Entre necesidad y la consulta hay un salto
 - Expresividad del lenguaje de consulta del sistema
 - Consultas mejores y peores para una necesidad (relevance feedback)



Relevancia

- ◆ Un concepto central, abstracto, a veces escurridizo

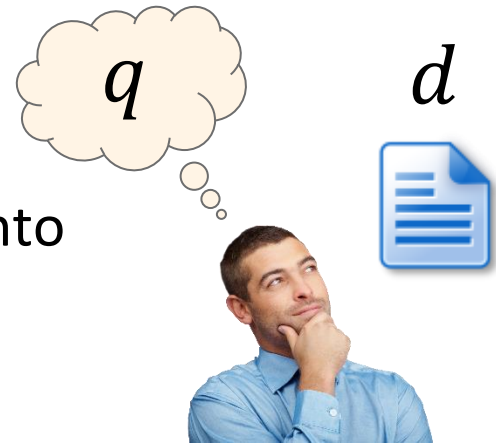
- Base de la evaluación offline de sistemas IR
- Base de algunos modelos probabilísticos

- ◆ Una propiedad del par necesidad / documento

- Acierto / utilidad / valor para el usuario

- ◆ Simplificaciones comunes

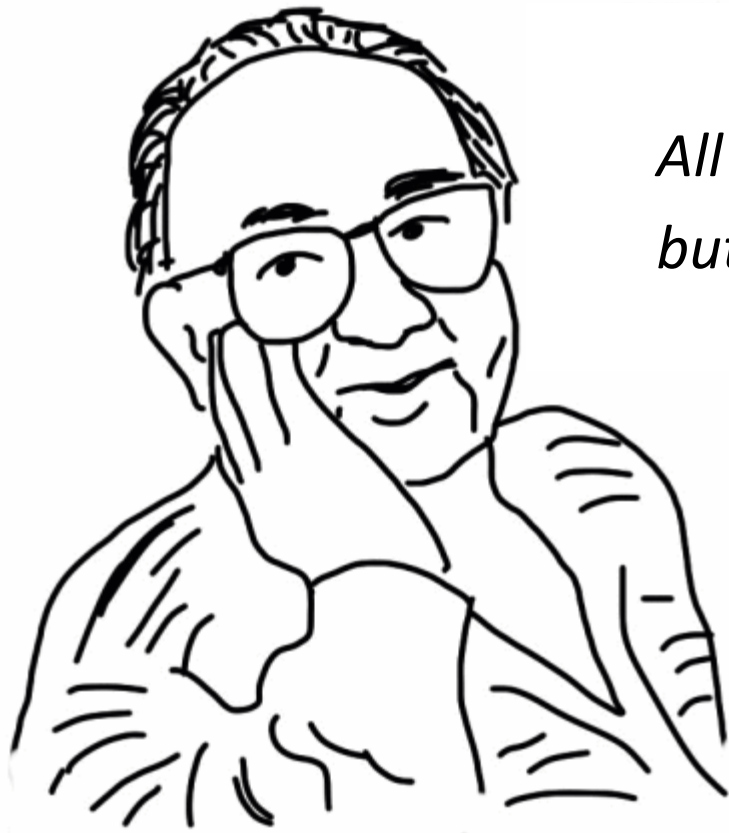
- Binaria (relevante / no relevante) → relevancia gradual
- Unidimensional → MOO
- Estable (independencia del tiempo) → modelos temporales
- Consistente (independencia de usuario y contexto) → personalización
- Independiente respecto a documentos ya vistos → diversidad



Ranking

- ♦ Incertidumbre dominante en un proceso IR
 - Es difícil tener certeza sobre el acierto
- ♦ Se presenta un orden del espacio de búsqueda
 - Para simplificar, orden lineal total
 - Otras presentaciones: ranking de estanterías, etc.
 - En la práctica se trunca un top N
- ♦ El usuario explora el ranking hasta encontrar lo que necesita
- ♦ Soluciones de ranking: **modelos de IR**
 - Supervisados vs. no supervisados

Modelos IR orientados a la búsqueda de texto



*All models are wrong
but some are useful*

George E. P. Box
(1919-2013, British statistician)

1. Modelos IR no supervisados

Modelos no supervisados

- ◆ Funciones de predicción
 - Inspiración heurística: modelo vectorial
 - Inspiración formal: modelos probabilísticos
- ◆ Representación sparse: bag of words
 - Basada en unigramas: palabras sueltas
 - Matriz de pesos término/documento
- ◆ Representación densa: embeddings
 - Se define un espacio de factores latentes de dimensión reducida
 - Se produce un vector denso para cada palabra, documento, consulta en el espacio común de factores latentes

Modelos sparse

- ◆ Previo al tiempo de consulta se crea un índice
 - Diccionario { palabra \rightarrow lista de documentos }
con información extra (frecuencias, pesos...)
 - Document map { documento \rightarrow lista de propiedades }
con cálculos globales: longitud del documento, PageRank...
 - Coste lineal en la suma de la longitud de los documentos
del espacio de búsqueda
- ◆ Calcular la función de scoring en tiempo de consulta
 - Coste lineal respecto al nº de documentos que contienen
las palabras de la consulta

Indexación

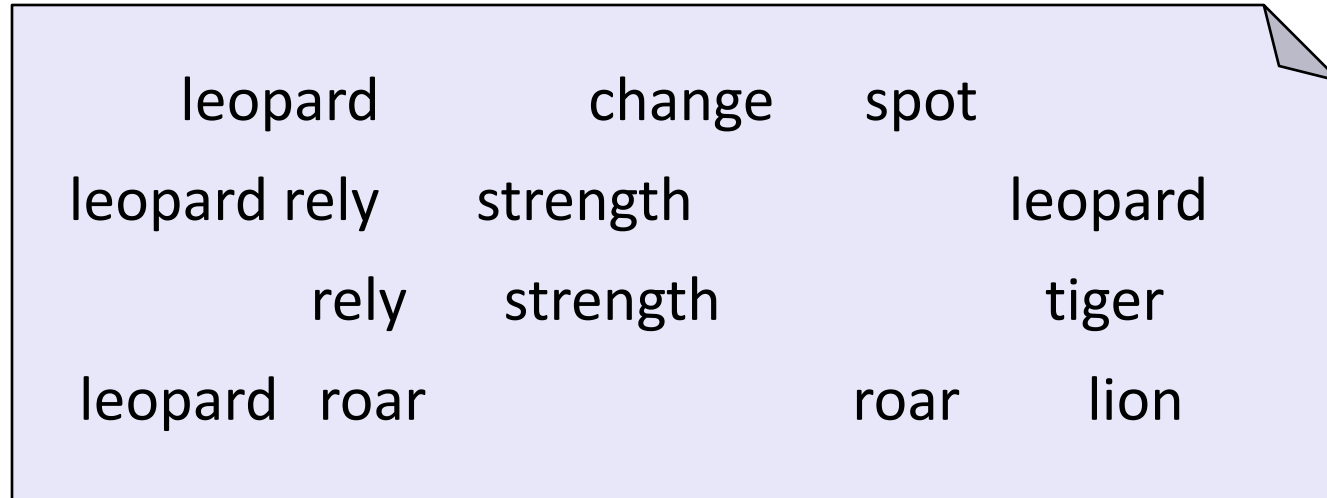
- ◆ Extracción de palabras clave
 - Tokenización, decodificación, eliminación de marcas...
- ◆ Filtrado y procesamiento de las palabras → términos (vocabulario)
 - Stopwords, normalización, stemming, grupos nominales...
- ◆ Construcción de índices optimizados
 - Estructuras, compresión, índices distribuidos

Representación sparse

The leopard cannot change its spots. Does the leopard rely on strength? Well no, a leopard does not rely on strength as does the tiger. Leopards roar but not like the roar of a lion.

Saco de palabras

- ◆ Los modelos dispersos simplemente extraen y **cuentan las palabras**
- ◆ Y construyen diferentes elaboraciones sobre el “saco” de palabras



leopard change spot
leopard rely strength leopard
rely strength tiger
leopard roar roar lion

Saco de palabras

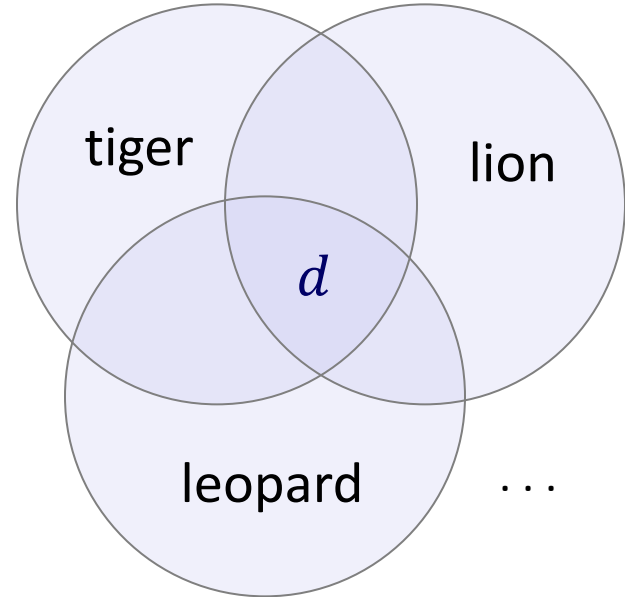
- ♦ Los modelos dispersos simplemente extraen y **cuentan las palabras**
- ♦ Y construyen diferentes elaboraciones sobre el “saco” de palabras

change	1
leopard	4
lion	1
rely	2
roar	2
spot	1
strength	2
tiger	1

Modelo booleano

- ♦ Las palabras se ven como **conjuntos de documentos** (los que contienen el término)
- ♦ Las consultas son **operaciones booleanas** (\cup , \cap , complemento) sobre palabras
- ♦ Un documento satisface la consulta si “**pertenece**” a ella

change	1	→	1
leopard	4	→	1
lion	1	→	1
rely	2	→	1
roar	2	→	1
spot	1	→	1
strength	2	→	1
tiger	1	→	1
...		→	0

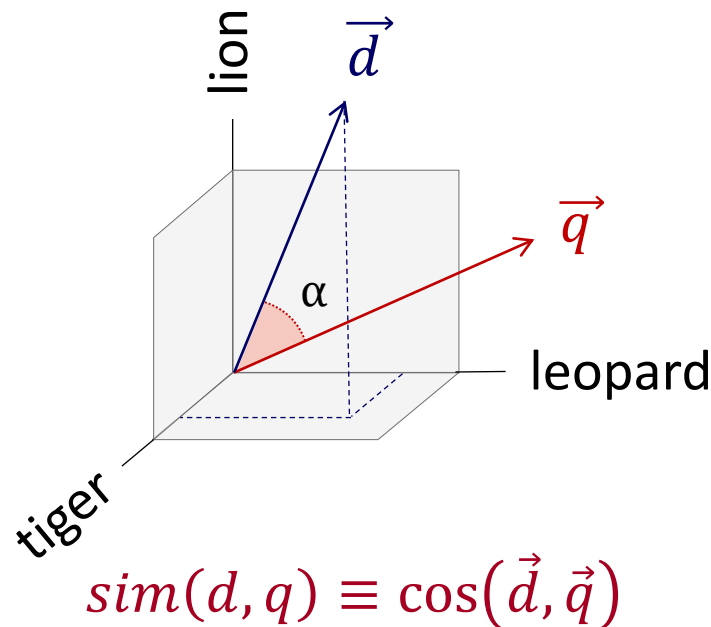


$$\text{sim}(d, q) \equiv [d \in q]$$

Modelo vectorial

- ♦ Las palabras son ejes de un **espacio vectorial**
- ♦ Los documentos y consultas son vectores en este espacio
- ♦ Un documento satisface la consulta \equiv el **ángulo** entre los vectores es pequeño

change	1	\rightarrow	4.3
leopard	4	\rightarrow	26.9
lion	1	\rightarrow	9.4
rely	2	\rightarrow	13.3
roar	2	\rightarrow	15.3
spot	1	\rightarrow	6.1
strength	2	\rightarrow	9.3
tiger	1	\rightarrow	9.0
...		\rightarrow	0

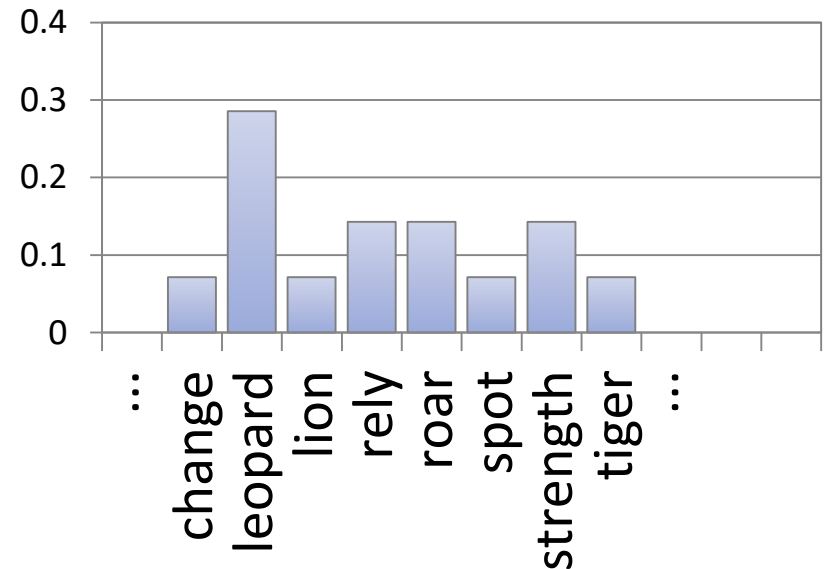


$$tf-idf(\text{leopard}, d) = (1 + \log_2 4) \log_2 \left(\frac{\# \text{ docs in collection}}{\# \text{ docs containing "leopard"}} \right)$$

Modelos probabilísticos

- ♦ Las palabras son **variables aleatorias**
- ♦ Los documentos y consultas “son” **distribuciones de palabras**
- ♦ Diferentes modelos para valorar la relación entre consultas y documentos

change	1	→	1/14
leopard	4	→	4/14
lion	1	→	1/14
rely	2	→	2/14
roar	2	→	2/14
spot	1	→	1/14
strength	2	→	1/14
tiger	1	→	1/14
...		→	0



$$P(\text{leopard}|d) \sim \frac{\text{\# occurrences of "leopard" in } d}{\text{\# words in } d}$$

$$\text{sim}(d, q) \equiv p(q|d)$$

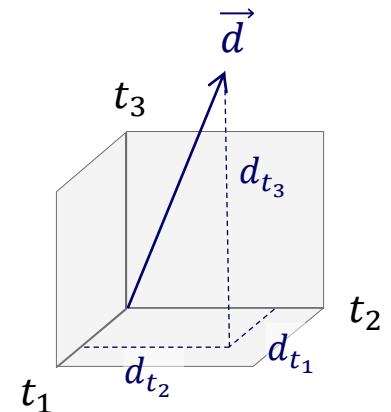
Modelo vectorial

Modelo vectorial (VSM)



Gerard Salton
(1927-1995)

- ♦ G. Salton, Harvard/Cornell University, 60-70's
- ♦ Se representan documentos y consultas en un espacio vectorial $\mathbb{R}^{|\mathcal{V}|}$, donde \mathcal{V} es el vocabulario
- ♦ La coordenada de los vectores de documento para cada $t \in \mathcal{V}$ son pesos $d_t = w(t, \vec{d})$ que se calculan con una fórmula, típicamente basada en frecuencias
- ♦ Definir una ponderación representativa
 - Que por un lado cuantifique cuán representativo es cada término en el documento
 - Que por otro matice entre términos muy comunes y otros más específicos (y por tanto significativos)

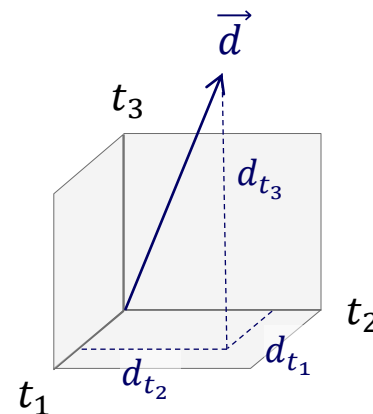


Modelo vectorial: esquema *tf-idf*

- ◆ El esquema típico de ponderación es *tf-idf*

$$d_t = tf(t, d) \cdot idf(t)$$

- *tf* mide la “importancia” de los términos en los documentos
- *idf* mide el poder de discriminación del término
- ◆ Existen diversas variantes para concretar las funciones *tf* e *idf*, en todas ellas:
 - *tf*(*t*, *d*) es creciente respecto a la frecuencia de *t* en *d*
 - *idf*(*t*) mide la especificidad de *t* por su frecuencia en la colección



El esquema *tf-idf*

$$tf(t, d) = \begin{cases} 1 + \log_2 frec(t, d) & \text{si } frec(t, d) > 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$idf(t) = \log \frac{|\mathcal{D}|}{|\mathcal{D}_t|}$$

\mathcal{D} = la colección de documentos
(espacio de búsqueda)

\mathcal{D}_t = documentos que contienen
el término t

- ♦ *tf* tiene que ver con la probabilidad del término en el documento
- ♦ E *idf* con la probabilidad en la colección

El esquema *tf-idf* (cont)

Otras variantes:

$$tf(t, d) = \frac{frec(t, d)}{\max_{t' \in \mathcal{V}} frec(t', d)}$$

- ♦ Pro: evita ventaja para documentos largos
- ♦ Contra: sensible a outliers

$$tf(t, d) = \lambda + (1 - \lambda) \frac{frec(t, d)}{\max_{t' \in \mathcal{V}} frec(t', d)} \quad \text{p. e. } \lambda = 0.5$$

$$idf(t) = \log \frac{|\mathcal{D}| + 1}{|\mathcal{D}_t| + 0.5}$$

...y unas cuantas más (tuning)

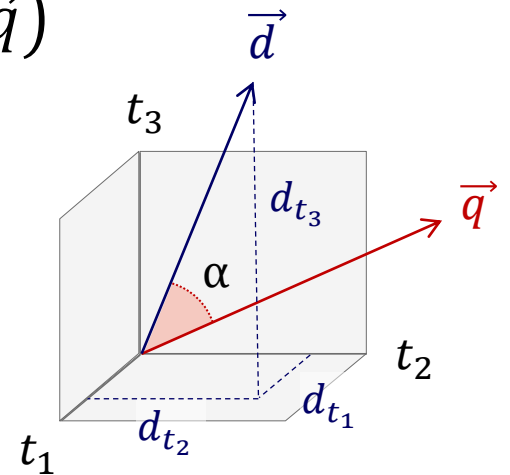
Modelo vectorial: función de ránking

Finalmente...

- ♦ Construimos \vec{q}
 - También por *tf-idf*, aunque no necesariamente con la misma variante
- ♦ Comparamos los vectores \vec{d} y \vec{q} en similitud por ángulo

$$f(d, q) = \text{sim}(d, q) = \text{angulo}(\vec{d}, \vec{q}) \propto \cos(\vec{d}, \vec{q})$$

$$\cos(\vec{d}, \vec{q}) = \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| |\vec{q}|} = \frac{\sum_t d_t q_t}{\sqrt{\sum_t d_t^2} \sqrt{\sum_t q_t^2}} \in [0, 1]$$



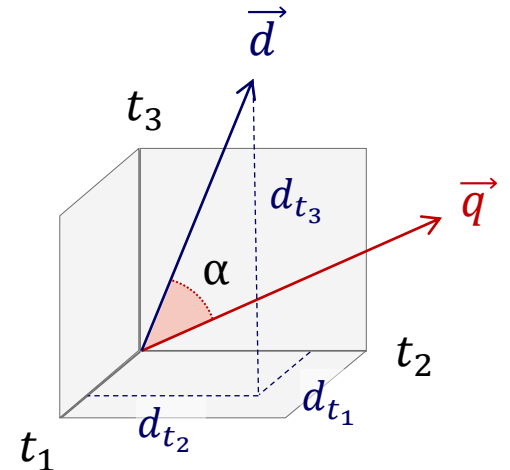
Modelo vectorial: coseno

Vector de consulta \vec{q}

- ♦ Se podría hacer tf binario
 - Salvo que la repetición de términos importe (p.e. consulta por ejemplos)
- ♦ *idf* penaliza doblemente los términos muy comunes (se puede omitir)
- ♦ Se puede omitir $|\vec{q}|$ en el denominador

Normalización longitud de documento \vec{d}

- ♦ El módulo del documento $|\vec{d}|$ en el denominador representa una normalización para evitar el sesgo a documentos largos



Ejemplo

q = “gold silver truck”

d_1 = “Shipment of gold damaged in a fire”

d_2 = “Delivery of silver arrived in a silver truck”

d_3 = “Shipment of gold arrived in a truck”

d_4 = “There was a fire at silver lake”

Ejemplo 2

$$1 + \log_2 \text{freq}(t, d)$$

$$\log \frac{|\mathcal{D}|}{|\mathcal{D}_t|}$$



$\text{freq}(t, d)$

$\text{tf}(t, d)$

$\text{tf-idf}(t, d)$

$d_1 \ d_2 \ d_3 \ d_4$

$d_1 \ d_2 \ d_3 \ d_4$

$\text{idf}(t)$

$d_1 \ d_2 \ d_3 \ d_4$

q

hielo

		4	1
--	--	---	---

0	0	3	1
---	---	---	---

1

0	0	3	1
---	---	---	---

--

hierba

1	4	2	1
---	---	---	---

1	3	2	1
---	---	---	---

0

0	0	0	0
---	---	---	---

--

hockey

4			
---	--	--	--

3	0	0	0
---	---	---	---

2

6	0	0	0
---	---	---	---

1

liga

	4	2	
--	---	---	--

0	3	2	0
---	---	---	---

1

0	3	2	0
---	---	---	---

1

street

		1	1
--	--	---	---

0	0	1	1
---	---	---	---

1

0	0	1	1
---	---	---	---

1

tenis

4		1	
---	--	---	--

3	0	1	0
---	---	---	---

1

3	0	1	0
---	---	---	---

--

$$|d| \begin{array}{|c|c|c|c|} \hline \sqrt{45} & 3 & \sqrt{15} & \sqrt{2} \\ \hline \end{array} \quad \begin{array}{|c|} \hline \sqrt{3} \\ \hline \end{array} |q|$$

q = "liga street hockey"

$$\frac{d \cdot q}{|d||q|}$$



$\cos(d, q)$

0.52	0.58	0.45	0.41
------	------	------	------