DIPARTIMENTO DI
MATEMATICA E INFORMATICA

UNIVERSITÀ
degli STUDI
di CATANIA

Social Media
Data Analysis
2023/2024

Information Retrieval

Francesco Ragusa

francesco.ragusa@unict.it

https://iplab.dmi.unict.it/ragusa/

IMAGE PROCESSING LABORATORY

https://iplab.dmi.unict.it/fpv/

Definition

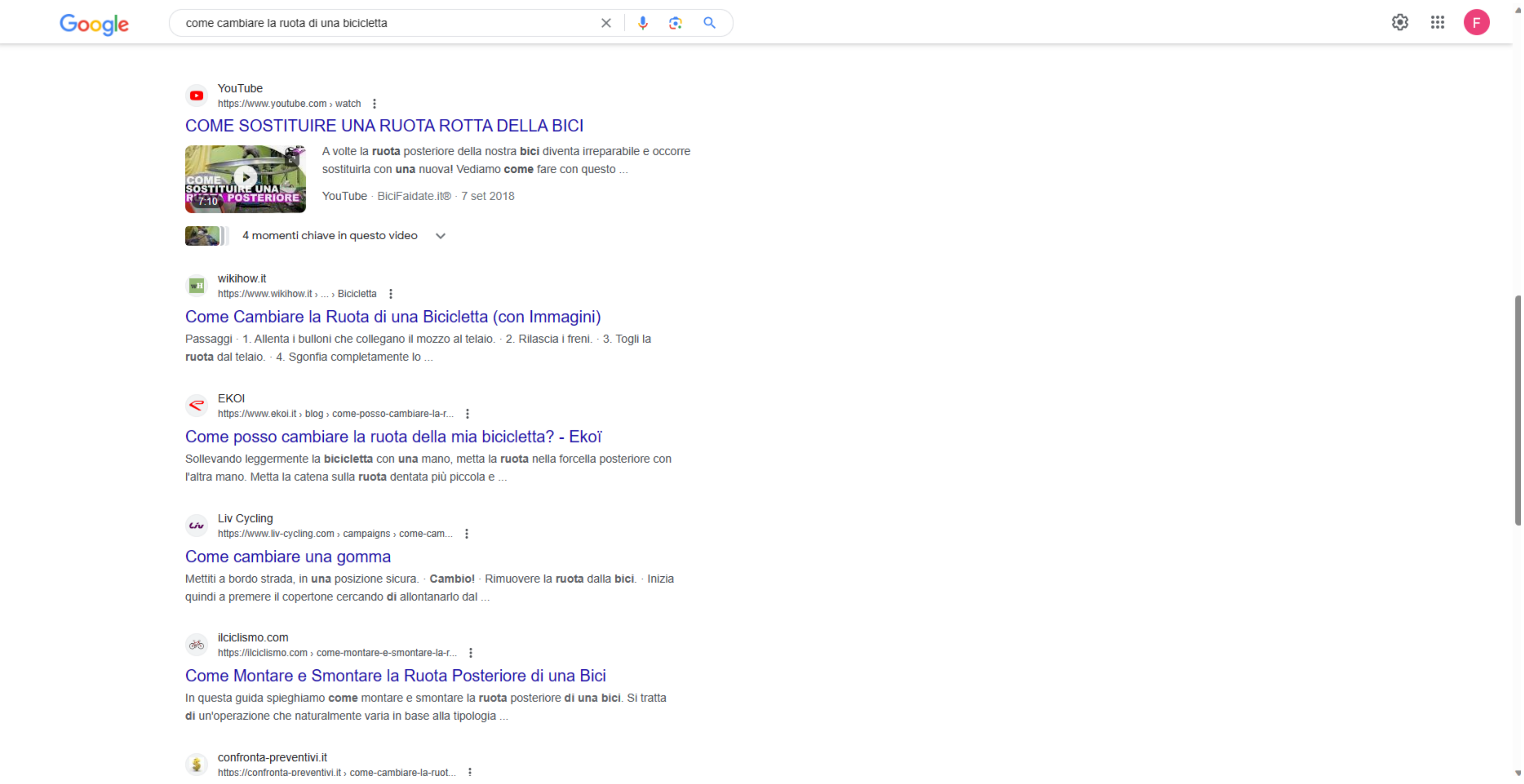
Information retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an information need from within **large collections** (usually stored on computers)

Hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email

Definition

The **IR system** assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies regarding the existence and location of documents that might consist of the required information.

An IR system has the ability to represent, store, organize, and access information items. A set of **keywords** are required to search



come cambiare la ruota di una bicicletta

YouTube
https://www.youtube.com › watch

COME SOSTITUIRE UNA RUOTA ROTTA DELLA BICI

A volte la **ruota** posteriore della nostra **bici** diventa irreparabile e occorre sostituirla con **una** nuova! Vediamo **come** fare con questo ...

YouTube · BiciFaidate.it® · 7 set 2018

4 momenti chiave in questo video

wikihow.it
https://www.wikihow.it › ... › Bicicletta

Come Cambiare la Ruota di una Bicicletta (con Immagini)

Passaggi · 1. Allenta i bulloni che collegano il mozzo al telaio. · 2. Rilascia i freni. · 3. Togli la **ruota** dal telaio. · 4. Sgonfia completamente lo ...

EKOI
https://www.ekoi.it › blog › come-posso-cambiare-la-r...

Come posso cambiare la ruota della mia bicicletta? - Ekoï

Sollevando leggermente la **bicicletta** con **una** mano, metta la **ruota** nella forcella posteriore con l'altra mano. Metta la catena sulla **ruota** dentata più piccola e ...

Liv Cycling
https://www.liv-cycling.com › campaigns › come-cam...

Come cambiare una gomma

Mettiti a bordo strada, in **una** posizione sicura. · **Cambio!** · Rimuovere la **ruota** dalla **bici**. · Inizia quindi a premere il copertone cercando **di** allontanarlo dal ...

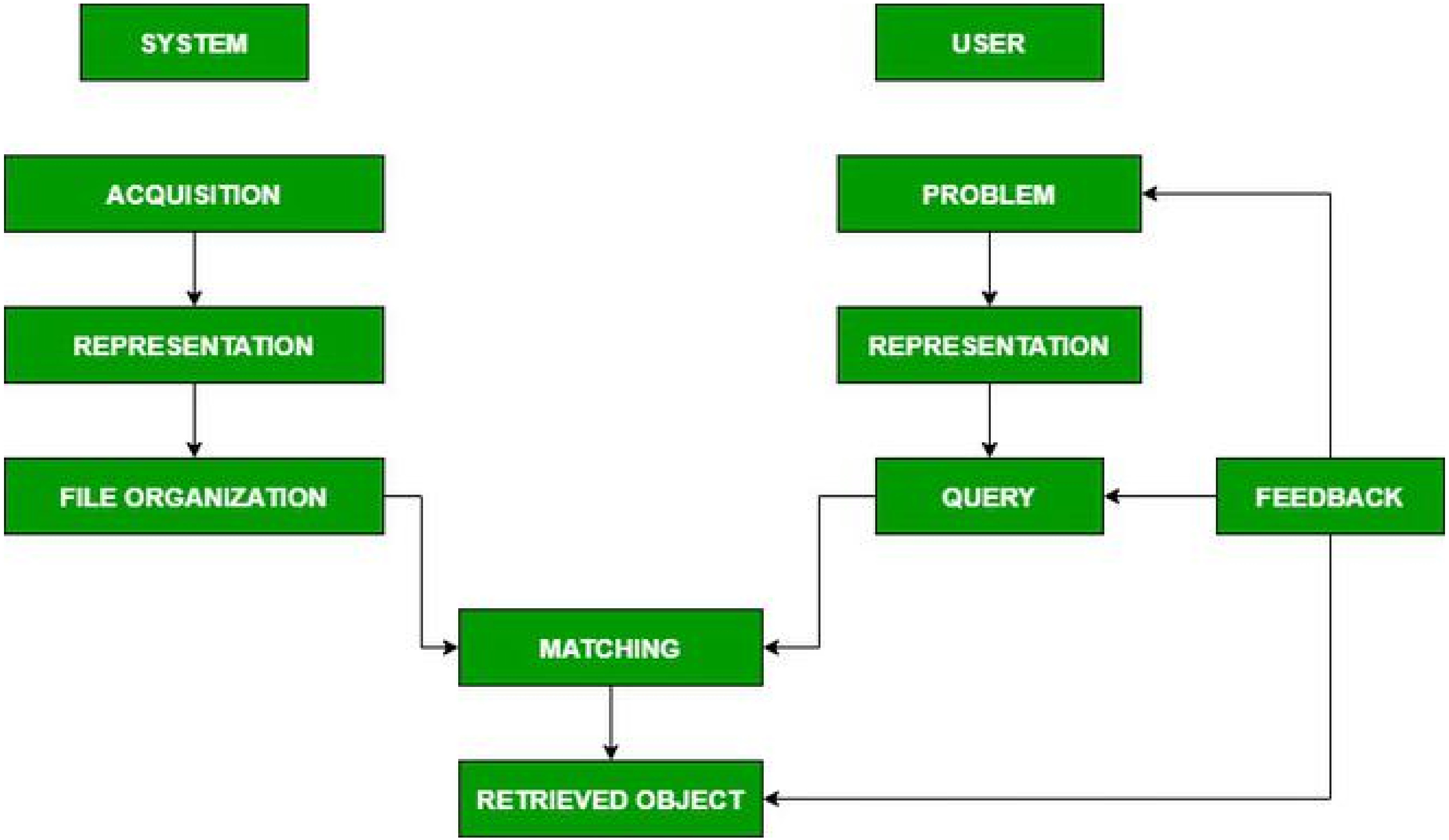
ilciclismo.com
https://ilciclismo.com › come-montare-e-smontare-la-r...

Come Montare e Smontare la Ruota Posteriore di una Bici

In questa guida spieghiamo **come** montare e smontare la **ruota** posteriore **di una bici**. Si tratta **di** un'operazione che naturalmente varia in base alla tipologia ...

confronta-preventivi.it
https://confronta-preventivi.it › come-cambiare-la-ruot...

IR Model


SYSTEM
USER
ACQUISITION
PROBLEM
REPRESENTATION
REPRESENTATION
FILE ORGANIZATION
QUERY
FEEDBACK
MATCHING
RETRIEVED OBJECT

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	It is a process of identifying and retrieving the data from the database, based on the query provided by user or application.

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	It is a process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	It is a process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
Unstructured	Structured

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
Unstructured	Structured
Ambiguous	Well defined semantics

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
Unstructured	Structured
Ambiguous	Well defined semantics
The results obtained are approximate matches	The results obtained are exact matches

Information vs. Data Retrieval

Information Retrieval	Data Retrieval
Unstructured	Structured
Ambiguous	Well defined semantics
The results obtained are approximate matches	The results obtained are exact matches
Results are ordered by relevance	Results are unordered by relevance

Formally..

$\mathbf{R'(q)}$

Vocabulary $V = \{w_1, w_2, \dots, w_N\}$

Query $Q = q_1, q_2, \dots, q_m \quad where\ q_i \in V$

Document $d_k = \{d_{k1}, d_{k2}, \dots, d_{km}\} \quad where\ d_{ki} \in V$

Collection $\mathrm{C} = \{d_1, d_2, \dots, d_n\}$

Set of relevant documents $\mathrm{R(q)} \subseteq C$

The Boolean Model

The boolean model is a **deterministic** model that uses logical operators to combine the terms of a query and return the documents that satisfy the logical condition.

AND **OR** **NOT**

The Boolean model

cat AND dog

cat OR dog

cat NOT dog

(cat AND dog) OR bird

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

Q = (cat AND dog) OR bird

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

[1 1 1 1 1 0 0 0 0 0 0 0 0 0]

[0 0 1 1 0 1 1 1 0 0 0 0 0 0]

Q = (cat AND dog) OR bird

[0

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

[1 1 1 1 1 0 0 0 0 0 0 0 0 0]

[0 0 1 1 0 1 1 1 0 0 0 0 0 0]

Q = (cat AND dog) OR bird

[0 0

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

[1 1 1 1 1 0 0 0 0 0 0 0 0 0]

[0 0 1 1 0 1 1 1 0 0 0 0 0 0]

Q = (cat AND dog) OR bird

[0 0 (1 AND 1) OR 1

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

[1 1 1 1 1 0 0 0 0 0 0 0 0 0]

[0 0 1 1 0 1 1 1 0 0 0 0 0 0]

Q = (cat AND dog) OR bird

[0 0 (1 AND 1) OR 1 0

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

D1: I like cats and dogs.
D2: Cats are cute and fluffy.
D3: Dogs are loyal and friendly.
D4: Birds are colorful and sing beautifully.

[1 1 1 1 1 0 0 0 0 0 0 0 0 0]

[0 0 1 1 0 1 1 1 0 0 0 0 0 0]

Q = (cat AND dog) OR bird

[0 0 (1 AND 1) OR 1 0 (1 AND 1) OR 1 …

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

How it works

Q = (cat AND dog) OR bird

[0 0 **(1 AND 1)** OR 1 0 (1 AND 1) OR 1 …

[0 0 (1) OR 1 0 (1 AND 1) OR 1 …

How it works

Q = (cat AND dog) OR bird

[0 0 **(1 AND 1)** OR 1 0 (1 AND 1) OR 1 …

[0 0 **(1) OR (1)** 0 (1 AND 1) OR 1 …

[0 0 (1) 0 (1 AND 1) OR 1 …

How it works

Q = (cat AND dog) OR bird

[0 0 **(1 AND 1)** OR 1 0 (1 AND 1) OR 1 …

[0 0 **(1) OR (1)** 0 (1 AND 1) OR 1 …

[0 0 **(1)** 0 (1 AND 1) OR 1 …

…

[0 0 1 0 1 0 0 0 0 0 1 0 0 0]

Similarity

Q = [0 0 1 0 1 0 0 0 0 0 1 0 0 0]

D1: [1 1 1 1 1 0 0 0 0 0 0 0 0 0] [2]
D2: [0 0 1 1 0 1 1 1 0 0 0 0 0 0] **[1]**
D3: [0 0 0 1 1 1 0 0 1 1 0 0 0 0] **[1]**
D4: [0 0 0 0 0 1 0 0 0 0 1 1 1 1] **[1]**

S (D1, Q) = [(0*1) + (0*1) + (1*1) + …] = [2]

Limitations

They are very rigid and difficult to express complex user requests.

They are difficult to control the number and quality of documents retrieved.

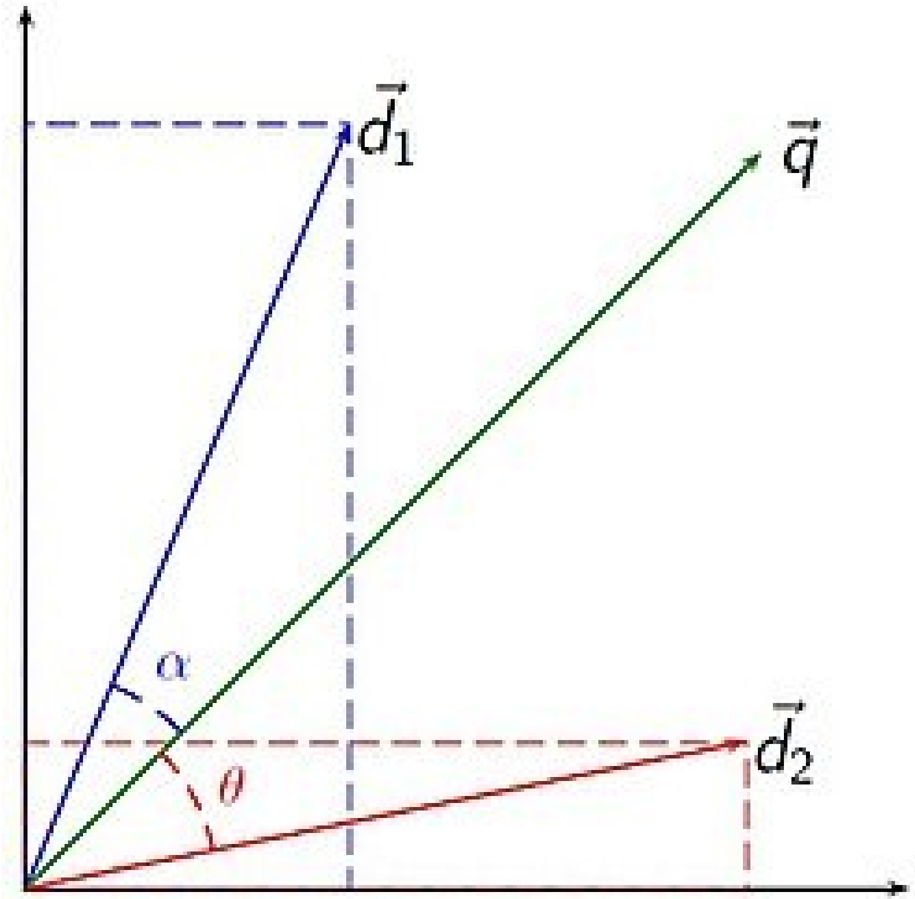
They are difficult to perform relevance feedback.

The Vector Space model

The representation of a set of documents as vectors in a common vector space

The (documents) and the queries are considered as vectors embedded in a high dimensional Euclidean space

Cosine similarity



Vector Space

V is a vector space over a field F (for example, the field of real or of complex numbers) if:

an operation vector addition defined in V, denoted v + w (where v, w ∈ V)

an operation, scalar multiplication in V, denoted a * v (where v ∈ V and a ∈ F)

the following properties hold for all a, b ∈ F and u, v, and w ∈ V:

v + w belongs to V

u + (v + w) = (u + v) + w

There exists a neutral element 0 in V, such that for all elements v in V, v + 0 = v

Vector Space

For all v in V, there exists an element w in V, such that $v + w = 0$

$v + w = w + v$

$a * v$ belongs to V

$a * (b * v) = (ab) * v$

If 1 denotes the multiplicative identity of the field F, then $1 * v = v$

$a * (v + w) = a * v + a * w$

$(a + b) * v = a * v + b * v$

The Vector Space model vs. Boolean model

Partial Matching

Ranking

Weighting schemes

Distance

We can compute the **norm.**

L2-**norm** (Euclidean norm)

L2-norm

x $= \{x_1, x_2, x_3\}$

y $= \{y_1, y_2, y_3\}$

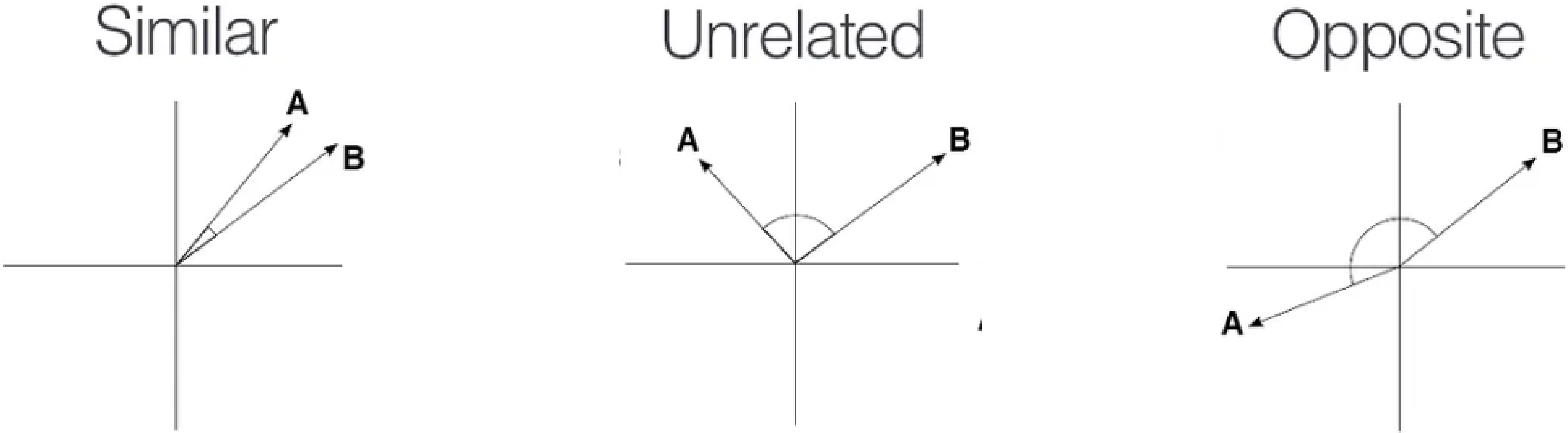
x-y $= (x_1 - y_1, x_2 - y_2, x_3 - y_3)$

L2-norm = $\sqrt{(x_1{-}y_1)^2 + (x_2{-}y_2)^2 + (x_3 - y_3)^2}$

Cosine Similarity

It is a metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space

$$A = \{x_1, x_2, x_3\} \qquad B = \{y_1, y_2, y_3\}$$



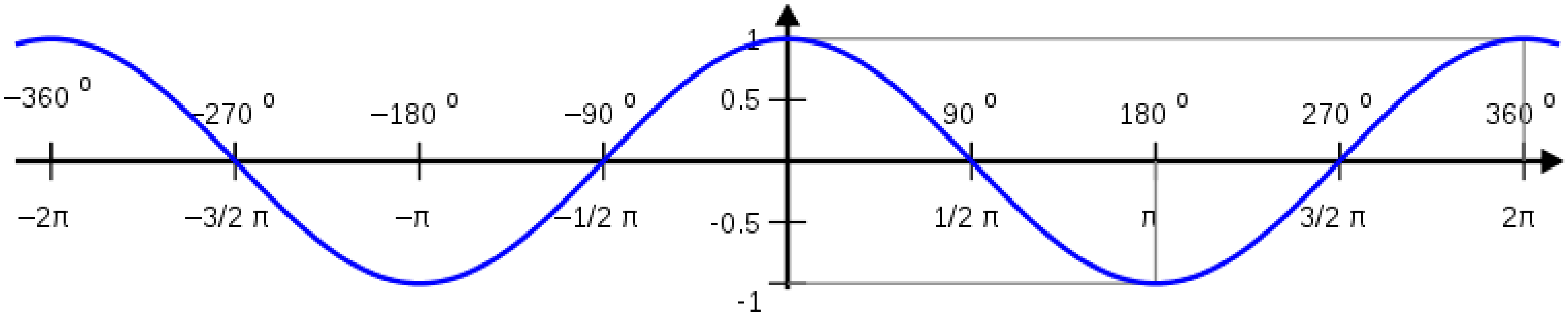
Similar

Unrelated

Opposite

Cosine Similarity

It is a metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space



Cosine Similarity

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$$

-1 = strongly opposite vectors → no similarity

0 = orthogonal vectors → independent

1 = overlap →high similarity

Cosine Distance

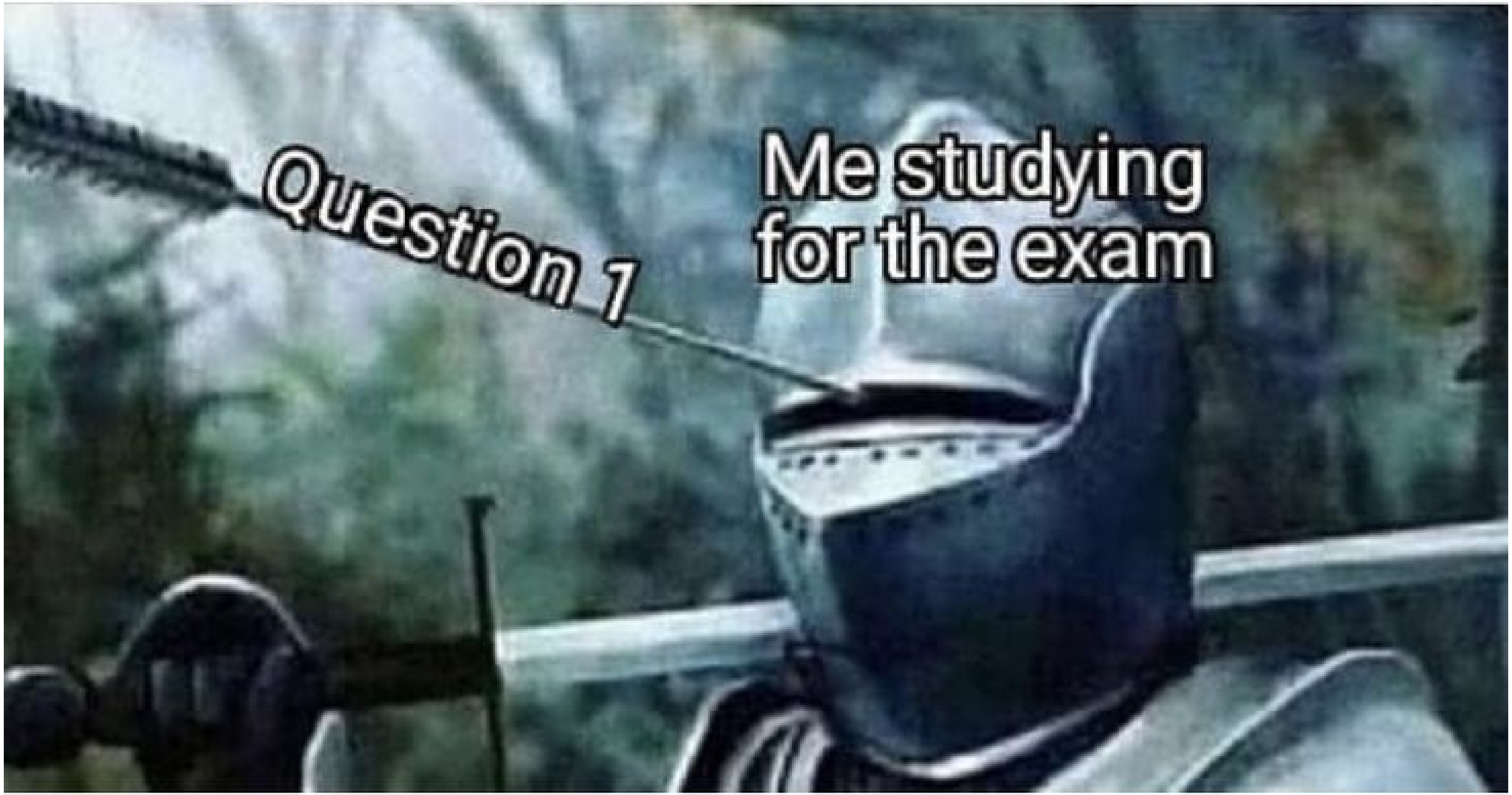
Cosine distance = 1 – cosine similarity

If 2 vectors are perfectly the same:

Cosine similarity = 1 (angle = 0, $cos(\theta)$=1)

Cosine distance = 1 – cosine similarity = 0

How can we represent documents as vectors?



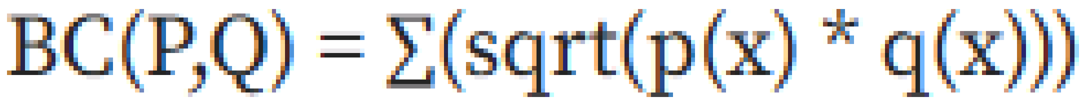
Bhattacharya Distance

It is a measure of the similarity between two probability distributions.
It tells us how much overlap there is between the two distributions, and can help us determine how similar or dissimilar they are.

$$BD(P,Q) = -\ln(BC(P,Q))$$

P, Q are two probability distributions
BC(P,Q) is the Bhattacharyya coefficient

Bhattacharya Distance

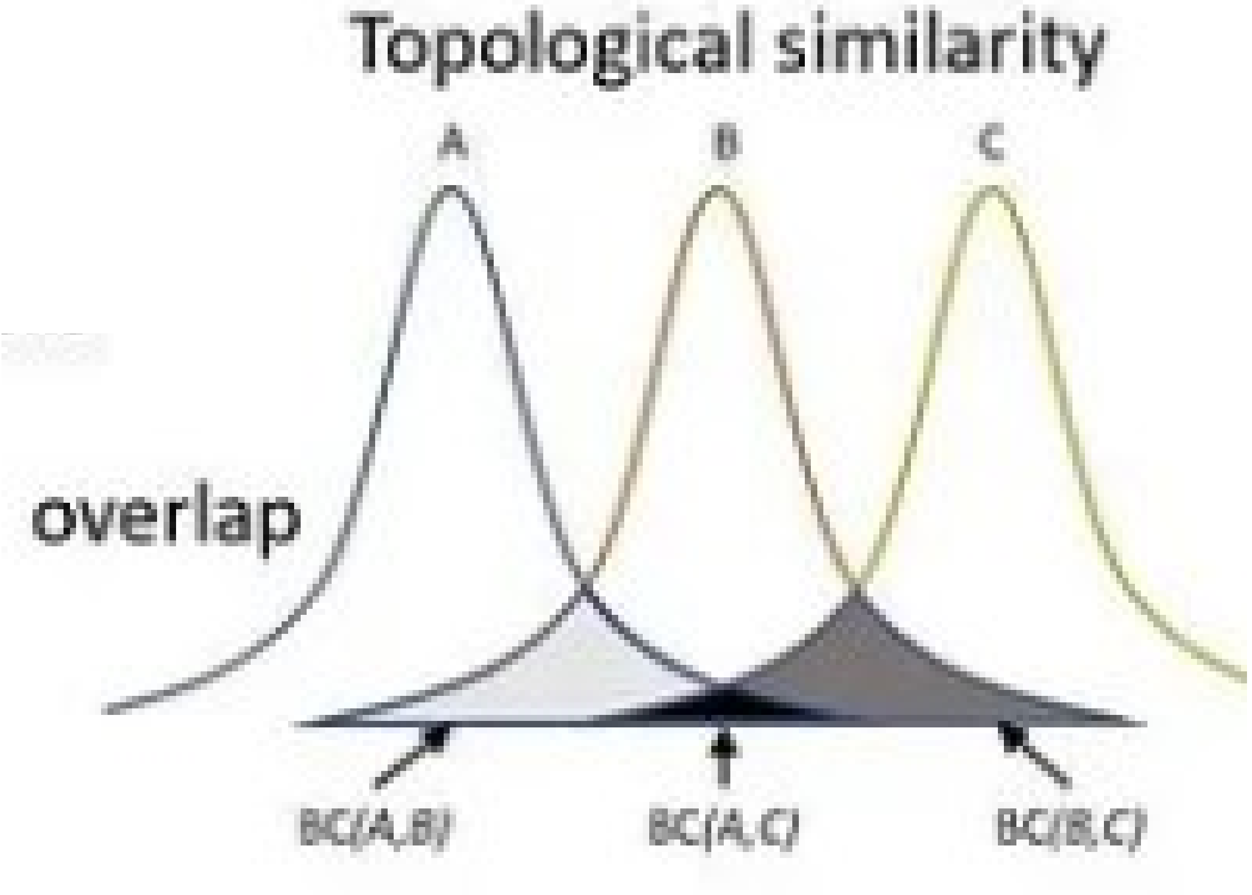


$$BC(P,Q) = \Sigma(\mathrm{sqrt}(p(x) * q(x)))$$

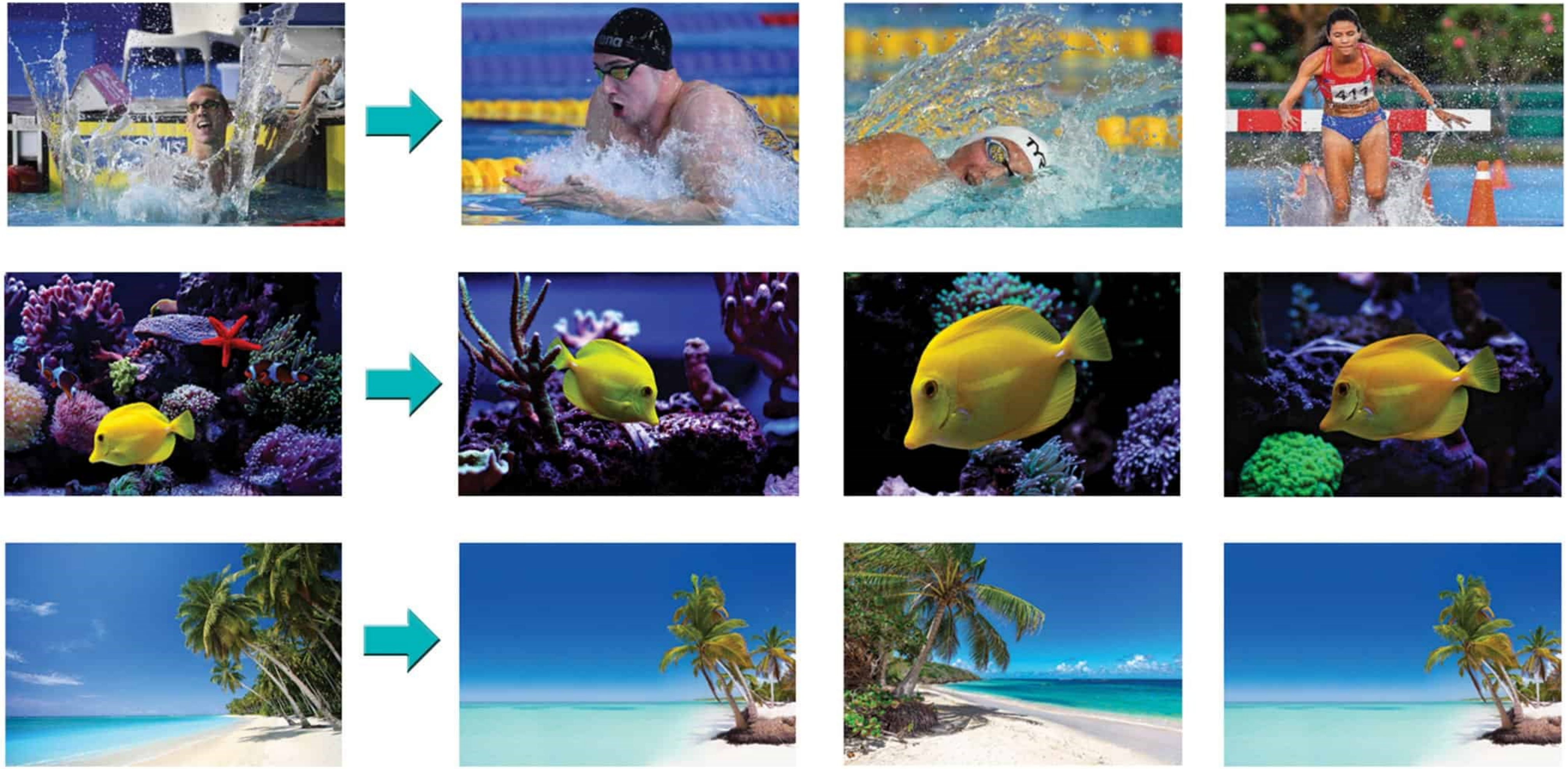
the summation is taken over all possible outcomes or events x

p(x), q(x) the probabilities of occurrence of event x in distributions P and Q

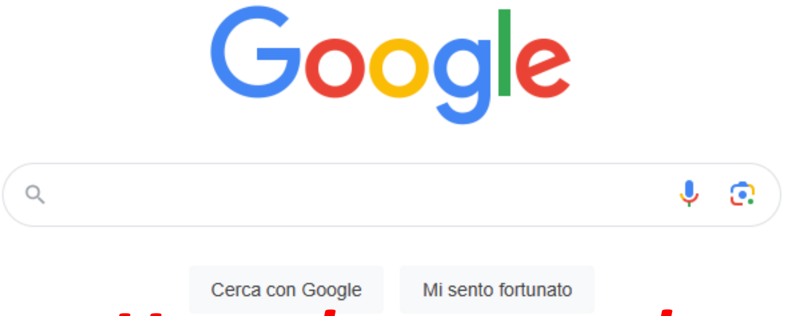
Bhattacharya Distance



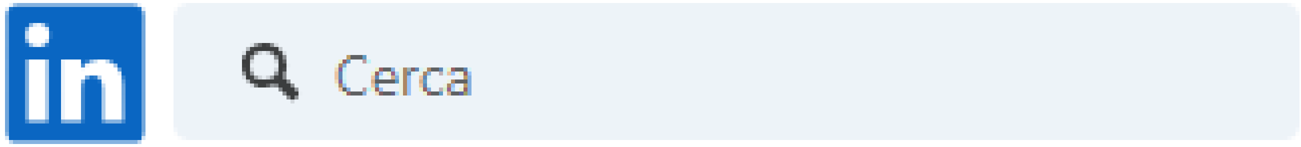
Bhattacharya Distance



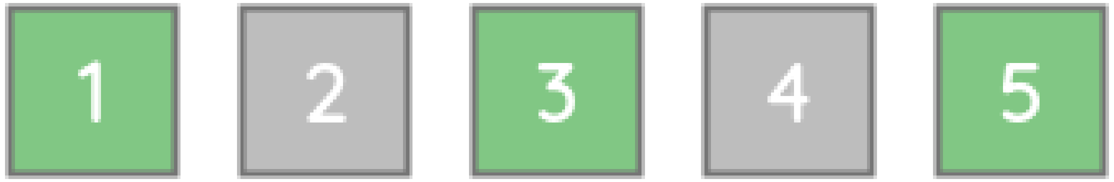
Evaluation Measures



How do we evaluate how good the top-N results are?



Evaluation Measures: Binary Relevance



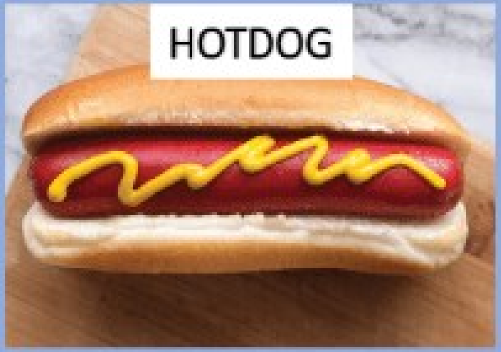
ground-truth annotation

Order-Unaware Metrics

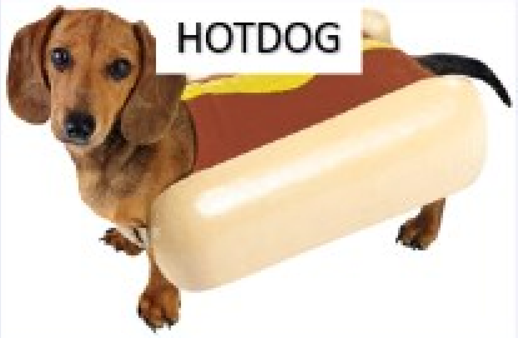
Order Aware Metrics

True/False Positives/Negatives

True positive

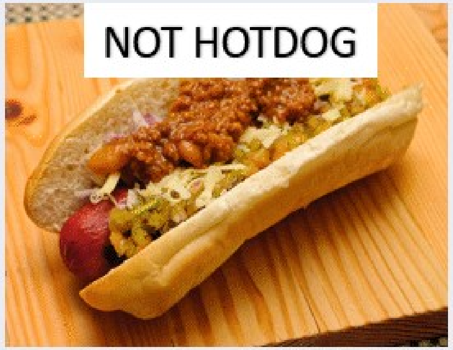


False positive

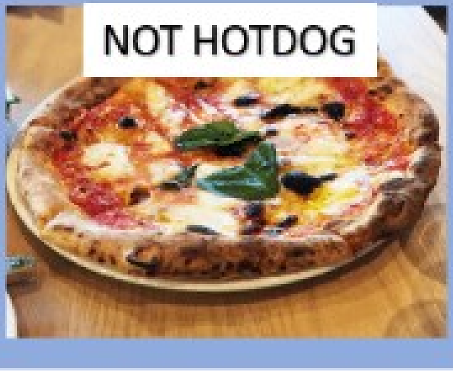


Hot Dog

False negative

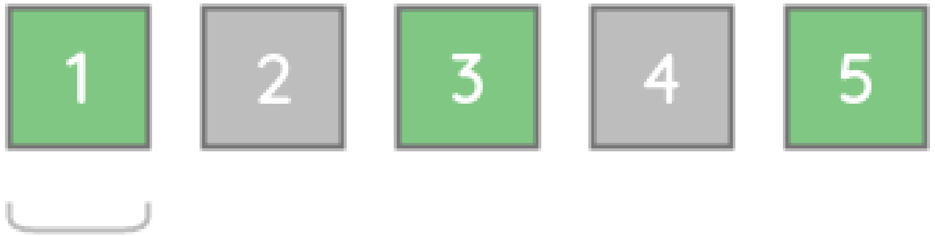


True negative

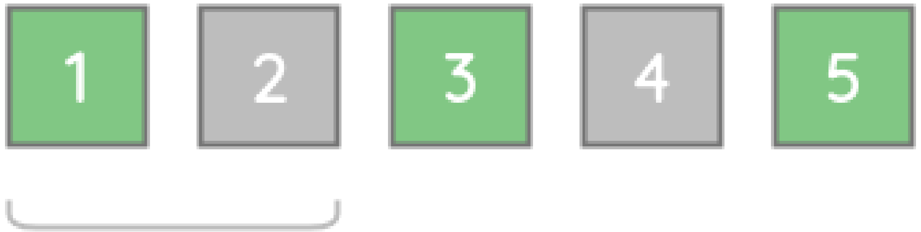


Order-Unaware Metrics: Precision

$$Precision@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ positives@k)}$$



Precision@1 = 1/1 = 1

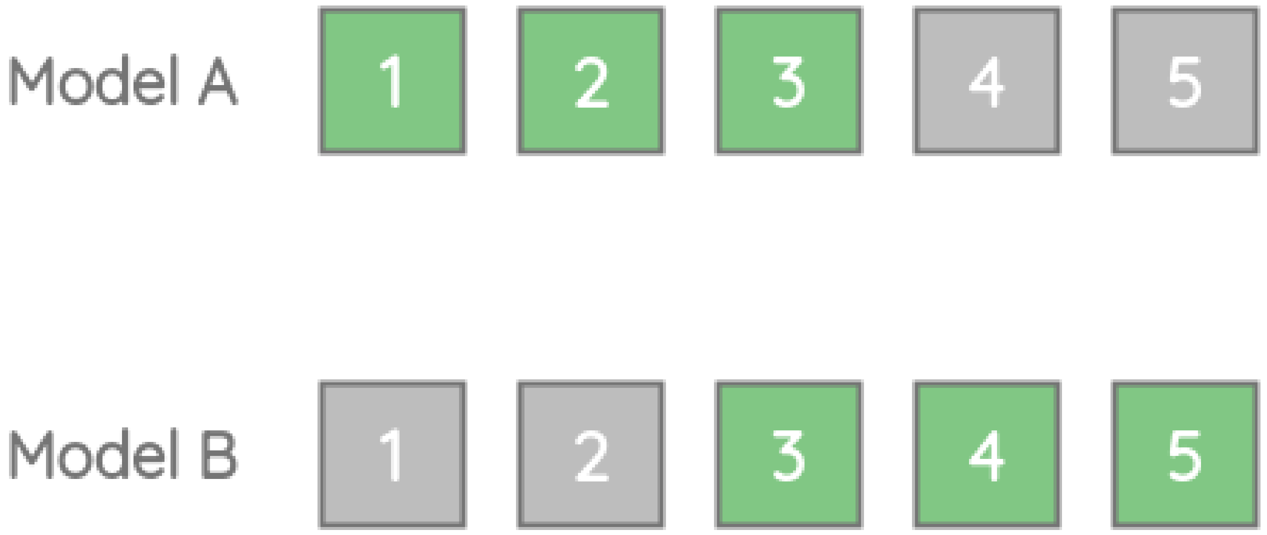


Precision@2 = 1/(1+1) = 1/2 = 0.5

k	1	2	3	4	5
Precision@k	$\frac{1}{1}=1$	$\frac{1}{2}=0.5$	$\frac{2}{3}=0.67$	$\frac{2}{4}=0.5$	$\frac{3}{5}=0.6$

Order-Unaware Metrics: Precision

It doesn’t consider the position of the relevant items



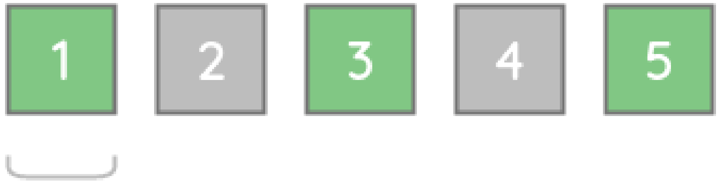
Precision@5 = 3/(3+2) = 3/5 = 0.6

Precision@5 = 3/(2+3) = 3/5 = 0.6

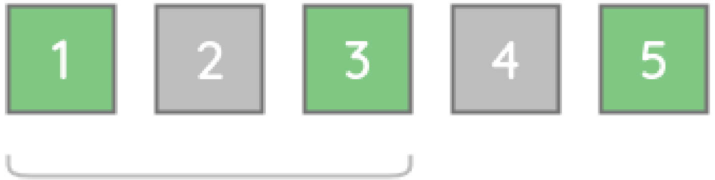
Order-Unaware Metrics: Recall

It gives how many actual relevant results were shown out of all actual relevant results for the query.

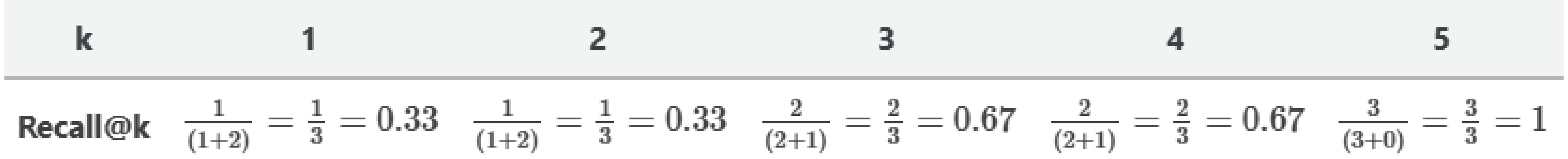
$$Recall@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ negatives@k)}$$



Recall@1 = 1/3 = 0.33



Recall@3 = 2/(2+1) = 2/3 = 0.67

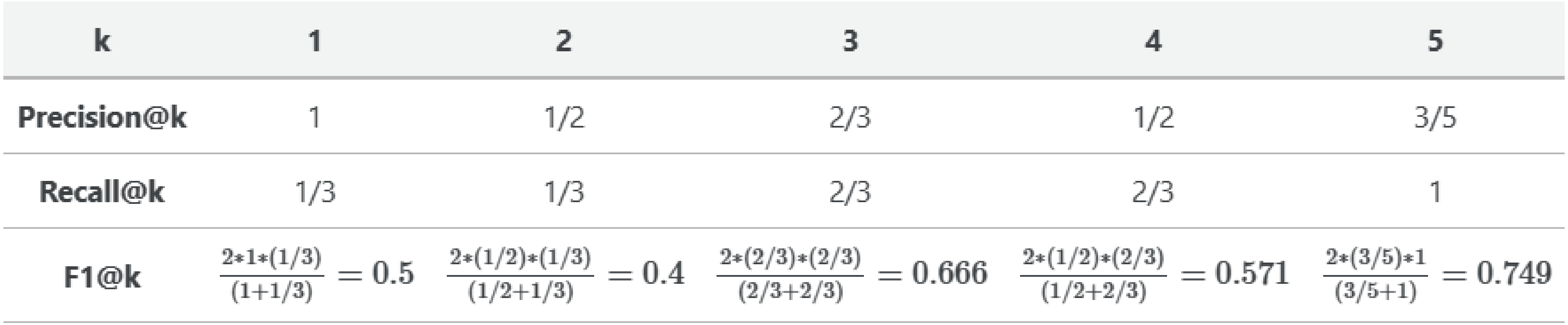


k	**1**	**2**	**3**	**4**	**5**
Recall@k	$\frac{1}{(1+2)} = \frac{1}{3} = 0.33$	$\frac{1}{(1+2)} = \frac{1}{3} = 0.33$	$\frac{2}{(2+1)} = \frac{2}{3} = 0.67$	$\frac{2}{(2+1)} = \frac{2}{3} = 0.67$	$\frac{3}{(3+0)} = \frac{3}{3} = 1$

Order-Unaware Metrics: F1-score

This is a combined metric that incorporates both Precision@k and Recall@k by taking their harmonic mean

$$F1@k = \frac{2 * (Precision@k) * (Recall@k)}{(Precision@k) + (Recall@k)}$$



k	1	2	3	4	5
Precision@k	1	1/2	2/3	1/2	3/5
Recall@k	1/3	1/3	2/3	2/3	1
F1@k	$\frac{2*1*(1/3)}{(1+1/3)} = 0.5$	$\frac{2*(1/2)*(1/3)}{(1/2+1/3)} = 0.4$	$\frac{2*(2/3)*(2/3)}{(2/3+2/3)} = 0.666$	$\frac{2*(1/2)*(2/3)}{(1/2+2/3)} = 0.571$	$\frac{2*(3/5)*1}{(3/5+1)} = 0.749$

Order Aware Metrics: Mean Reciprocal Rank (MRR)

This metric is useful when we want our system to return the best relevant item and want that item to be at a higher position.

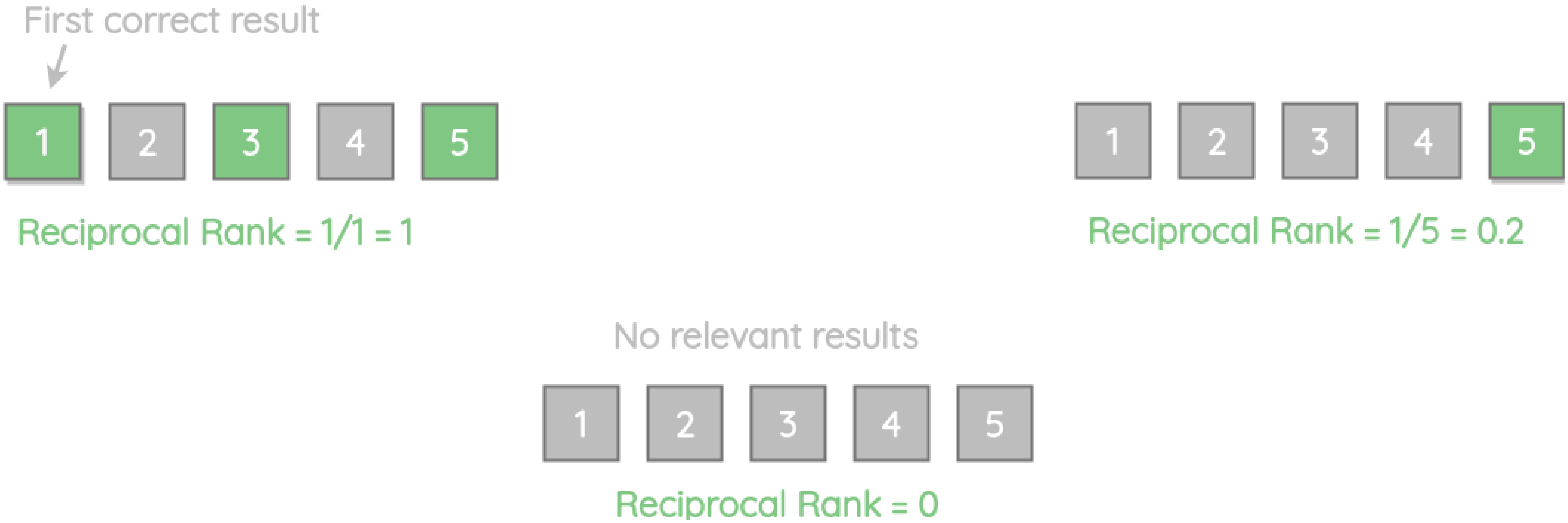
$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

|Q| denotes the total number of queries

$rank_1$ denotes the rank of the first relevan result

Order Aware Metrics: Mean Reciprocal Rank (MRR)

Reciprocal rank = the reciprocal of the rank of the first correct relevant result and the value ranges from 0 to 1.



First correct result

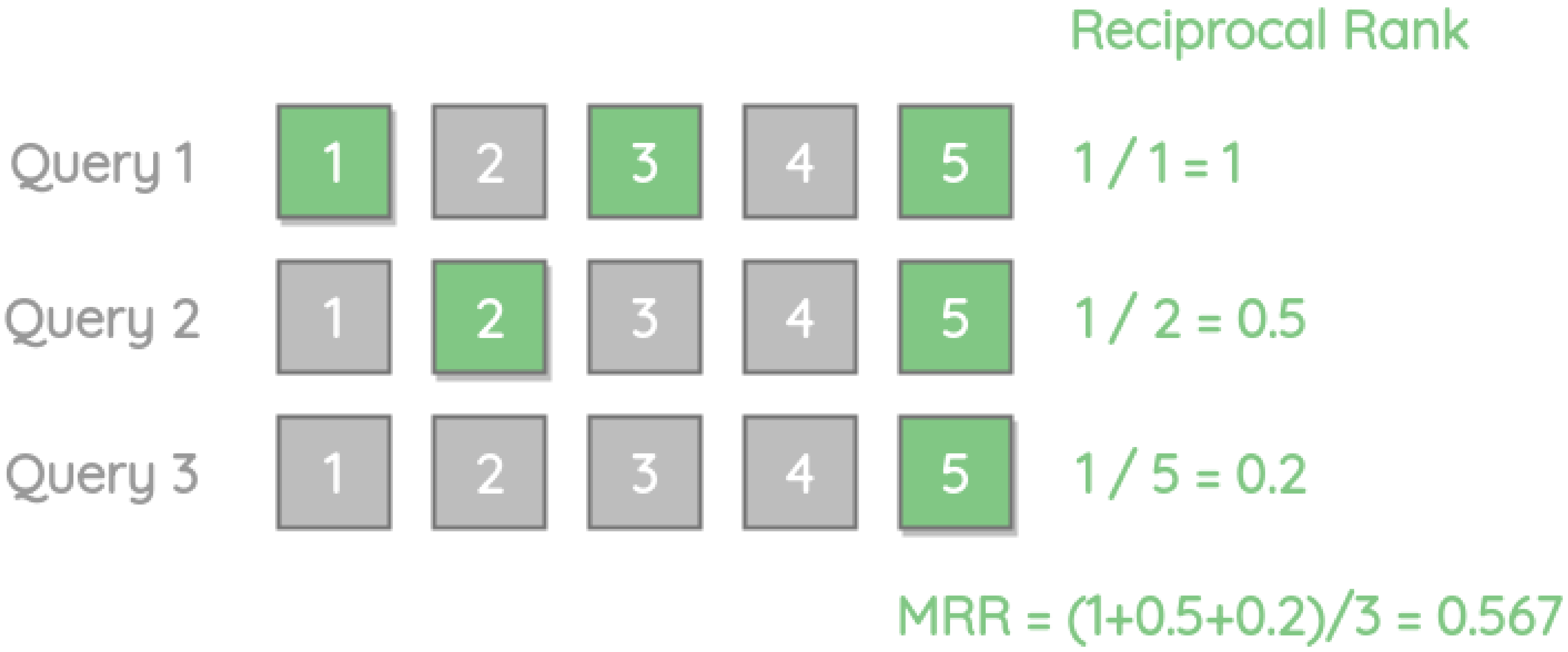
Reciprocal Rank = 1/1 = 1

Reciprocal Rank = 1/5 = 0.2

No relevant results

Reciprocal Rank = 0

Order Aware Metrics: Mean Reciprocal Rank (MRR)



						Reciprocal Rank
Query 1	🟩 1	2	🟩 3	4	🟩 5	1 / 1 = 1
Query 2	1	🟩 2	3	4	🟩 5	1 / 2 = 0.5
Query 3	1	2	3	4	🟩 5	1 / 5 = 0.2

MRR = (1+0.5+0.2)/3 = 0.567

Order Aware Metrics: Average Precision (AP)

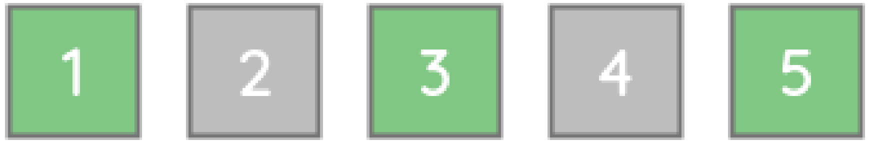
It is a metric that evaluates whether all of the ground-truth relevant items selected by the model are ranked higher or not.

$$AP = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{number\ of\ relevant\ items}$$

rel(k) is an indicator function which is 1 when the item at rank K is relevant

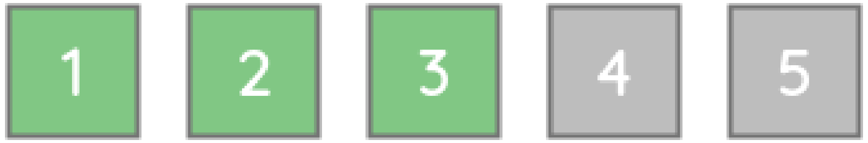
P(k) is the Precision@k metric

Order Aware Metrics: Average Precision (AP)



Precision@K	1	1/2	2/3	2/4	3/5

$$AP = \frac{(1 + 2/3 + 3/5)}{3} = 0.7555$$



Precision@K	1	1	1	3/4	3/5

$$AP = \frac{(1 + 1 + 1)}{3} = 1$$

Order Aware Metrics: Mean Average Precision (MAP)

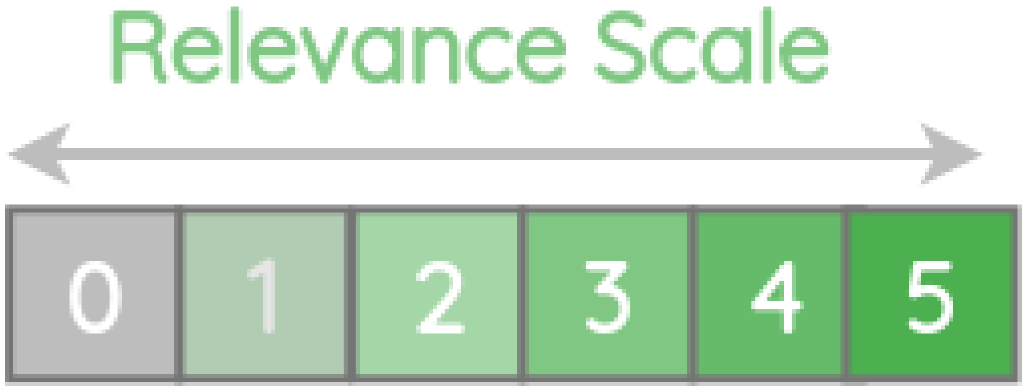
It is simply the mean of the average precision for all queries

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q)$$

Q is the total number of queries

AP(q) is the average precision for query q

Graded Relevance



Relevance Scale

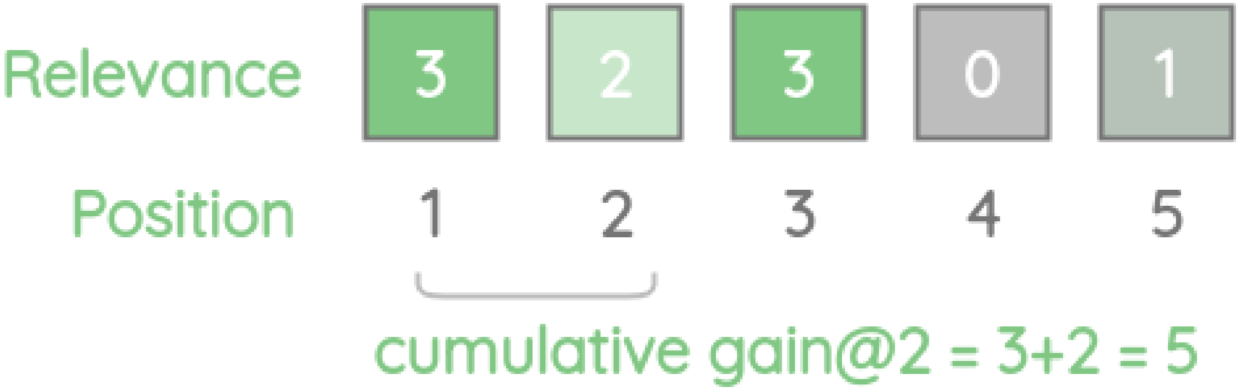
0 1 2 3 4 5

Relevance	3	2	3	0	1
Position	1	2	3	4	5

Order-Unaware Metrics: Cumulative Gain

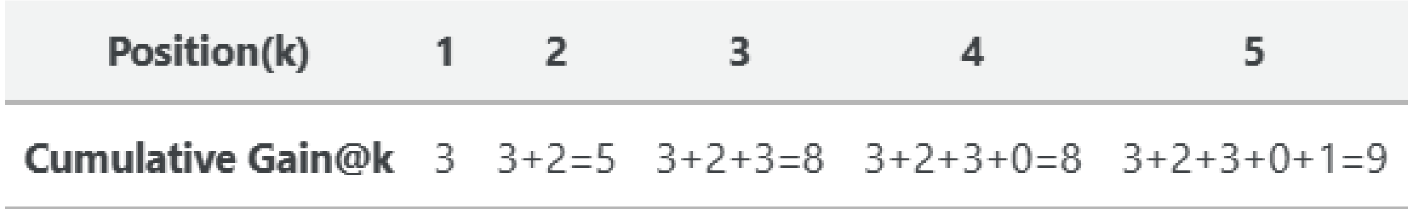
This metric uses a simple idea to just sum up the relevance scores for top-K items. The total score is called cumulative gain.

$$CG@k = \sum_{1}^{k} rel_i$$



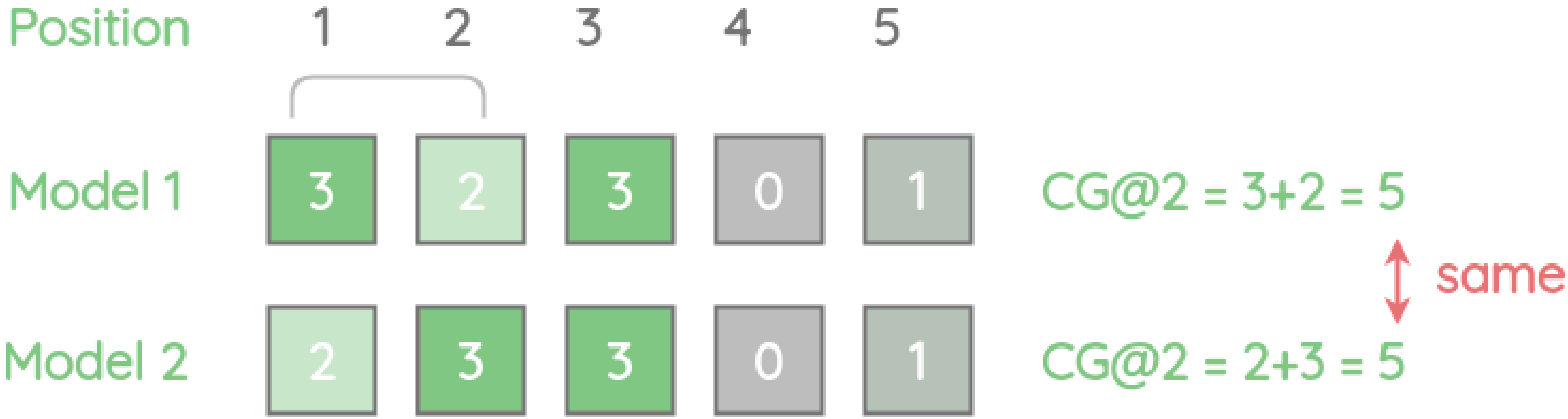
Relevance	3	2	3	0	1
Position	1	2	3	4	5

cumulative gain@2 = 3+2 = 5



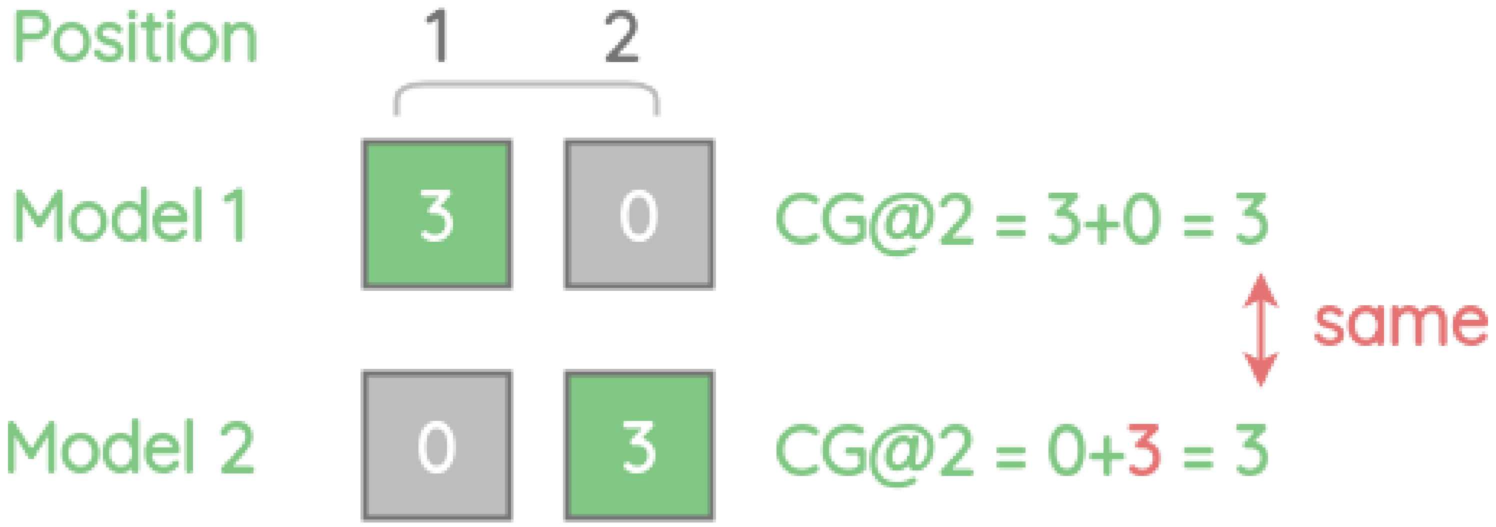
Position(k)	1	2	3	4	5
Cumulative Gain@k	3	3+2=5	3+2+3=8	3+2+3+0=8	3+2+3+0+1=9

Order-Unaware Metrics: Cumulative Gain



Position	1	2	3	4	5	
Model 1	3	2	3	0	1	CG@2 = 3+2 = 5
Model 2	2	3	3	0	1	CG@2 = 2+3 = 5

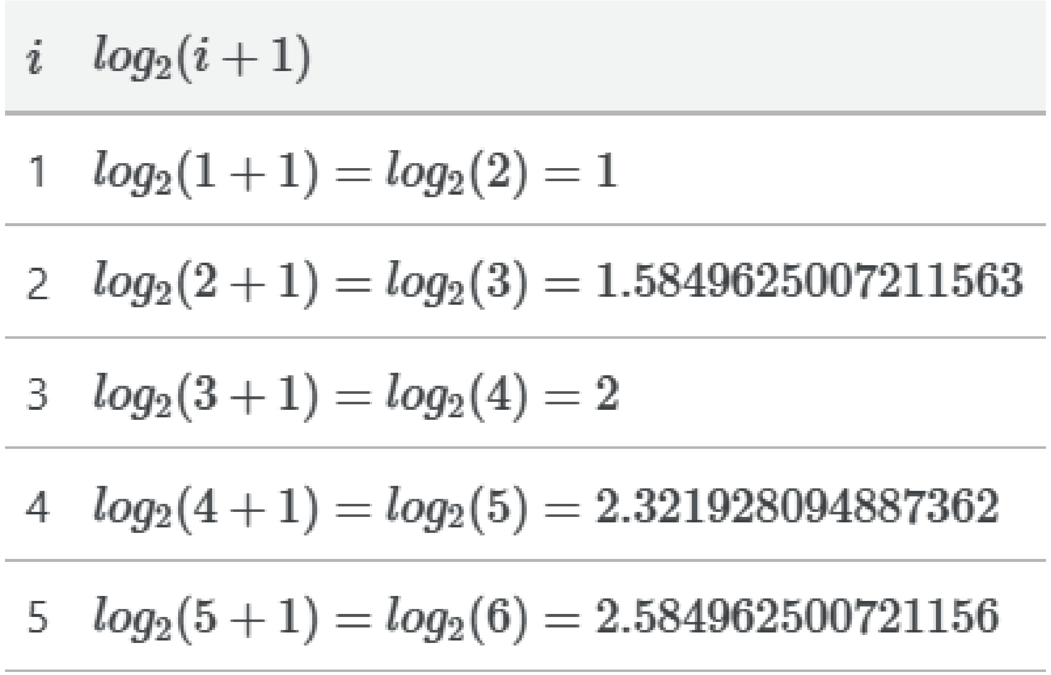
Order Aware Metrics: Discounted Cumulative Gain



An item with a relevance score of 3 at position 1 is better than the same item with relevance score 3 at position 2.

Order Aware Metrics: Discounted Cumulative Gain

DCG introduces a log-based penalty function to reduce the relevance score at each position.



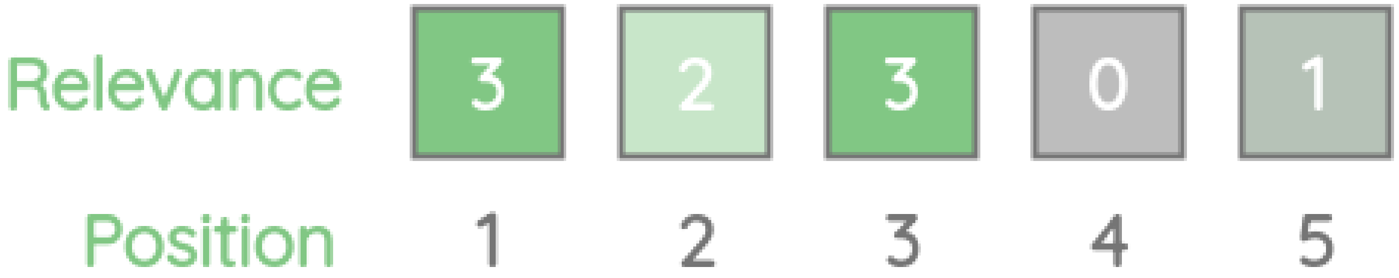
i	$log_2(i+1)$
1	$log_2(1+1) = log_2(2) = 1$
2	$log_2(2+1) = log_2(3) = 1.5849625007211563$
3	$log_2(3+1) = log_2(4) = 2$
4	$log_2(4+1) = log_2(5) = 2.321928094887362$
5	$log_2(5+1) = log_2(6) = 2.584962500721156$

Order Aware Metrics: Discounted Cumulative Gain

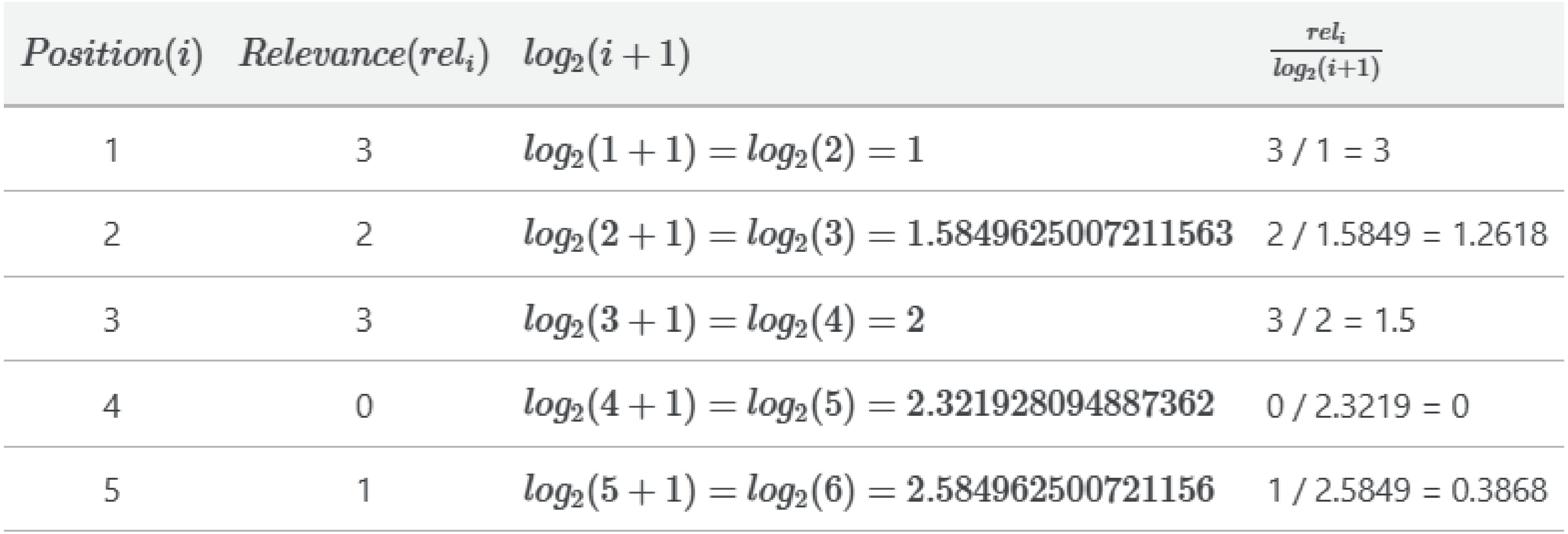
The discounted cumulative gain simply takes the sum of the relevance score normalized by the penalty.

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$$

Order Aware Metrics: Discounted Cumulative Gain

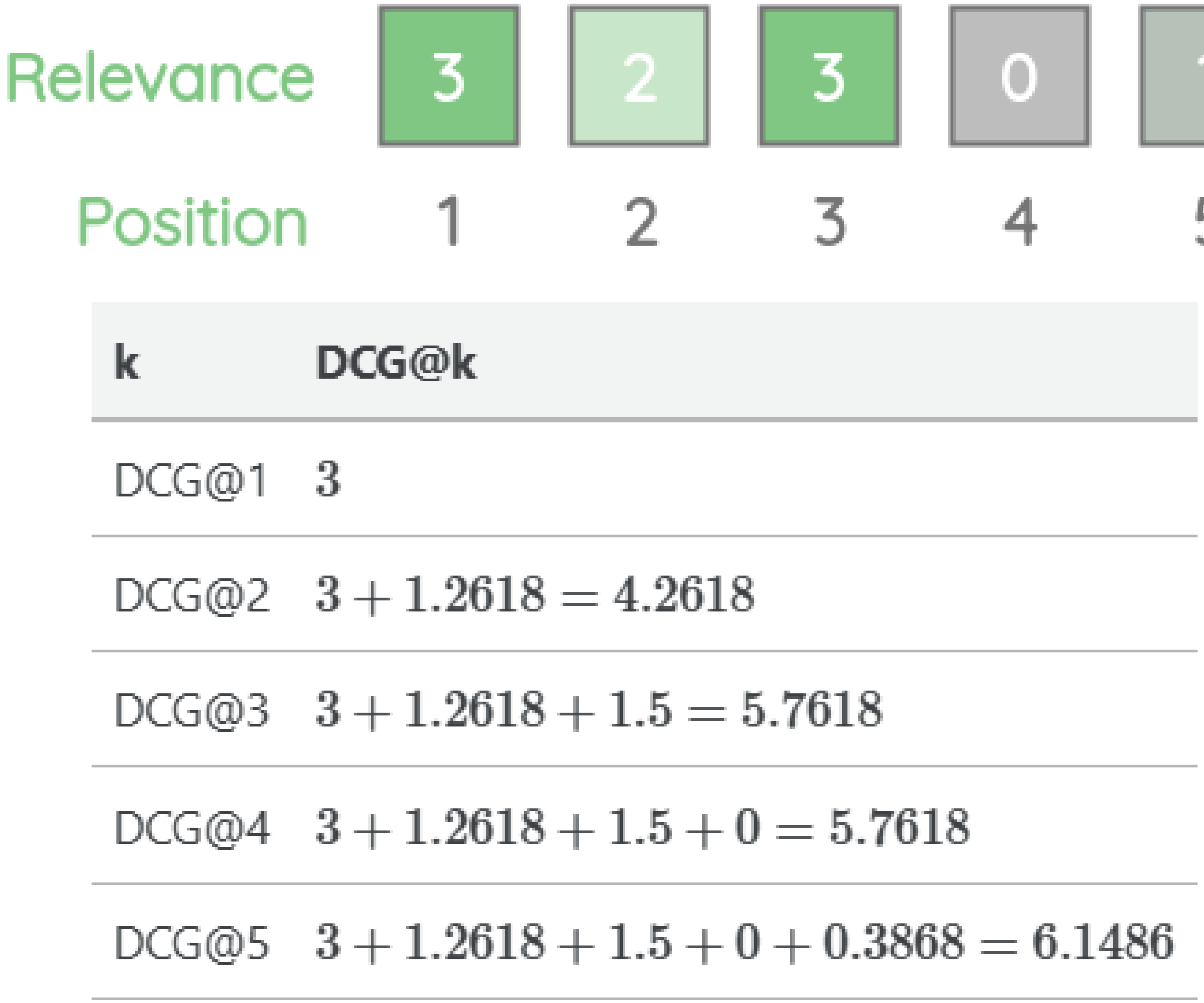


Relevance	3	2	3	0	1
Position	1	2	3	4	5



$Position(i)$	$Relevance(rel_i)$	$log_2(i+1)$	$\frac{rel_i}{log_2(i+1)}$
1	3	$log_2(1+1) = log_2(2) = 1$	3 / 1 = 3
2	2	$log_2(2+1) = log_2(3) = 1.5849625007211563$	2 / 1.5849 = 1.2618
3	3	$log_2(3+1) = log_2(4) = 2$	3 / 2 = 1.5
4	0	$log_2(4+1) = log_2(5) = 2.321928094887362$	0 / 2.3219 = 0
5	1	$log_2(5+1) = log_2(6) = 2.584962500721156$	1 / 2.5849 = 0.3868

Order Aware Metrics: Discounted Cumulative Gain

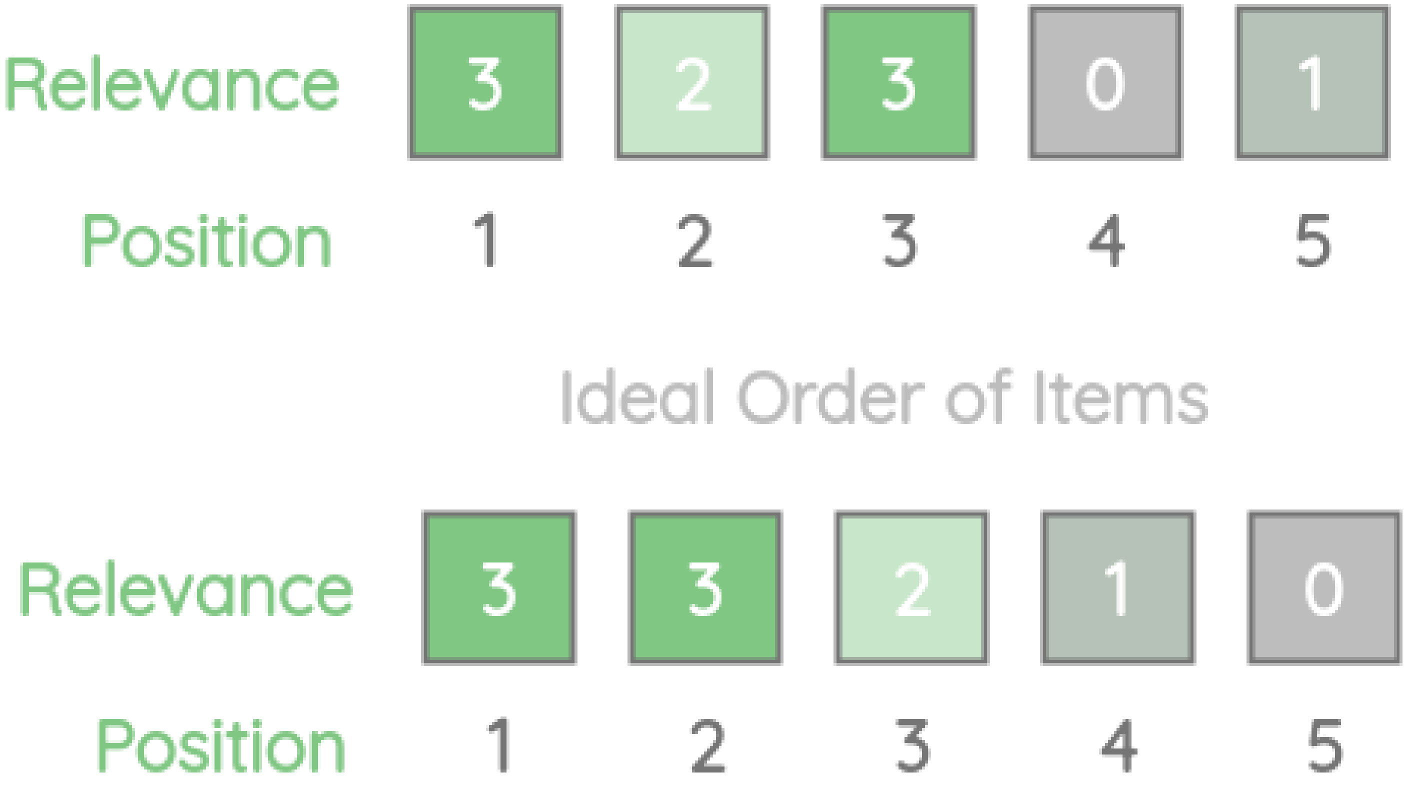


Relevance	3	2	3	0	1
Position	1	2	3	4	5

k	DCG@k
DCG@1	3
DCG@2	$3 + 1.2618 = 4.2618$
DCG@3	$3 + 1.2618 + 1.5 = 5.7618$
DCG@4	$3 + 1.2618 + 1.5 + 0 = 5.7618$
DCG@5	$3 + 1.2618 + 1.5 + 0 + 0.3868 = 6.1486$

Order Aware Metrics: Normalized Discounted Cumulative Gain

It normalizes the DCG values using the ideal order of the relevant items.

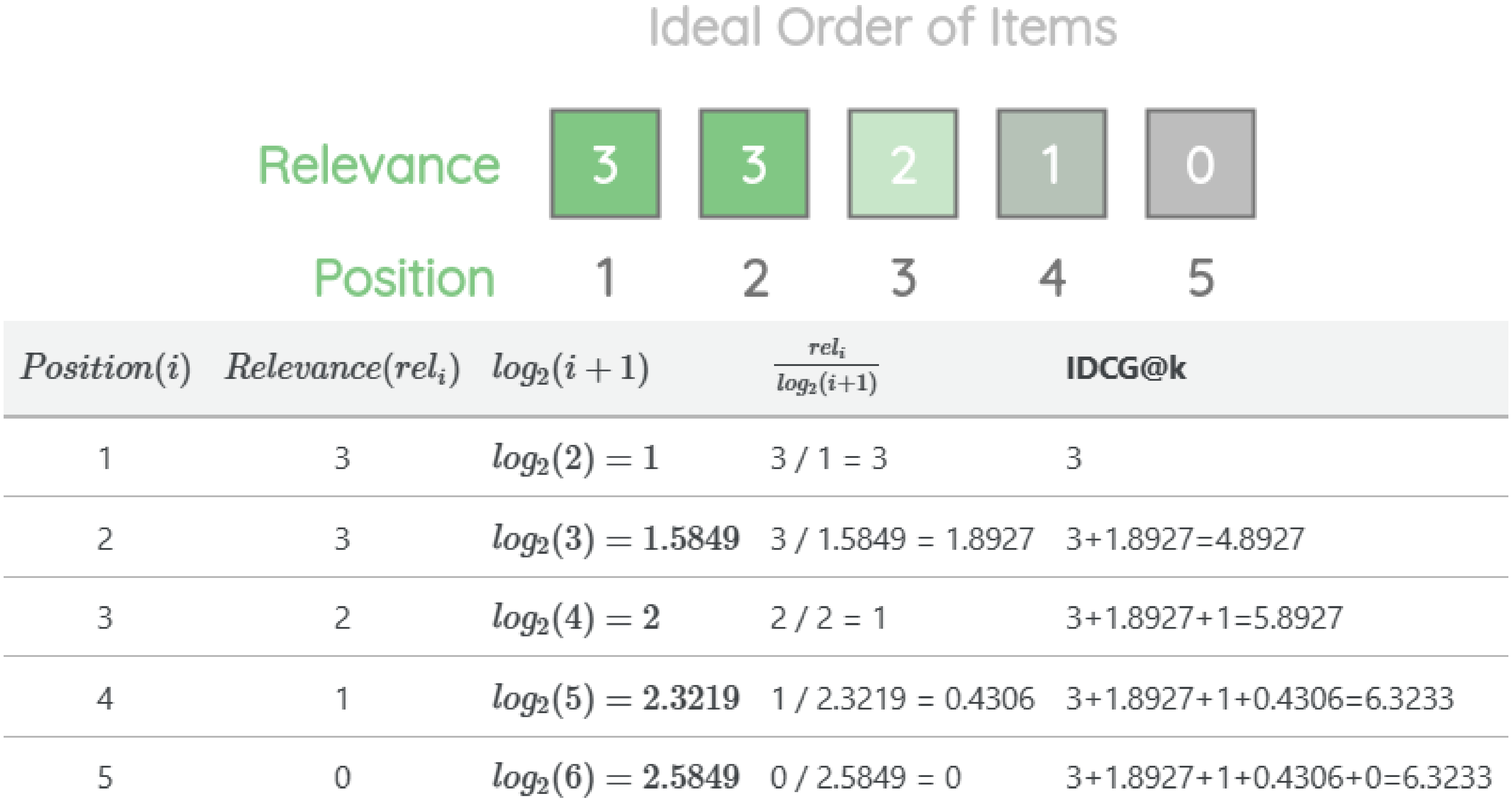


Relevance	3	2	3	0	1
Position	1	2	3	4	5

Ideal Order of Items

Relevance	3	3	2	1	0
Position	1	2	3	4	5

Order Aware Metrics: Normalized Discounted Cumulative Gain



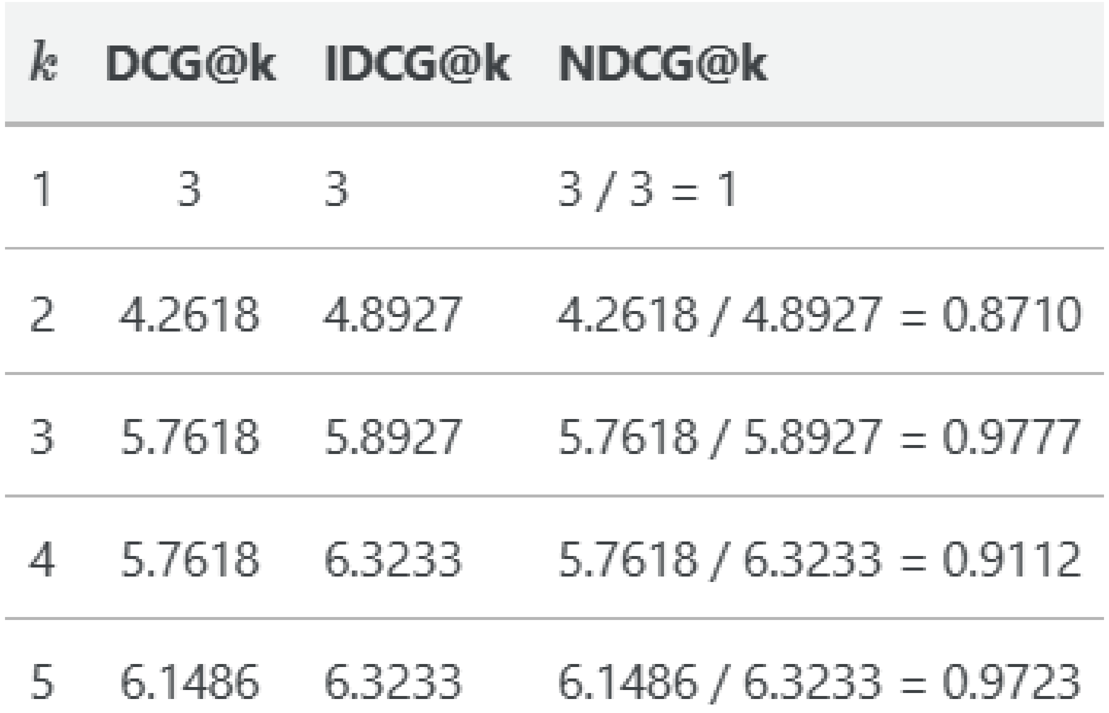
Ideal Order of Items

Relevance	3	3	2	1	0
Position	1	2	3	4	5

$Position(i)$	$Relevance(rel_i)$	$log_2(i+1)$	$\frac{rel_i}{log_2(i+1)}$	**IDCG@k**
1	3	$log_2(2) = 1$	3 / 1 = 3	3
2	3	$log_2(3) = \mathbf{1.5849}$	3 / 1.5849 = 1.8927	3+1.8927=4.8927
3	2	$log_2(4) = \mathbf{2}$	2 / 2 = 1	3+1.8927+1=5.8927
4	1	$log_2(5) = \mathbf{2.3219}$	1 / 2.3219 = 0.4306	3+1.8927+1+0.4306=6.3233
5	0	$log_2(6) = \mathbf{2.5849}$	0 / 2.5849 = 0	3+1.8927+1+0.4306+0=6.3233

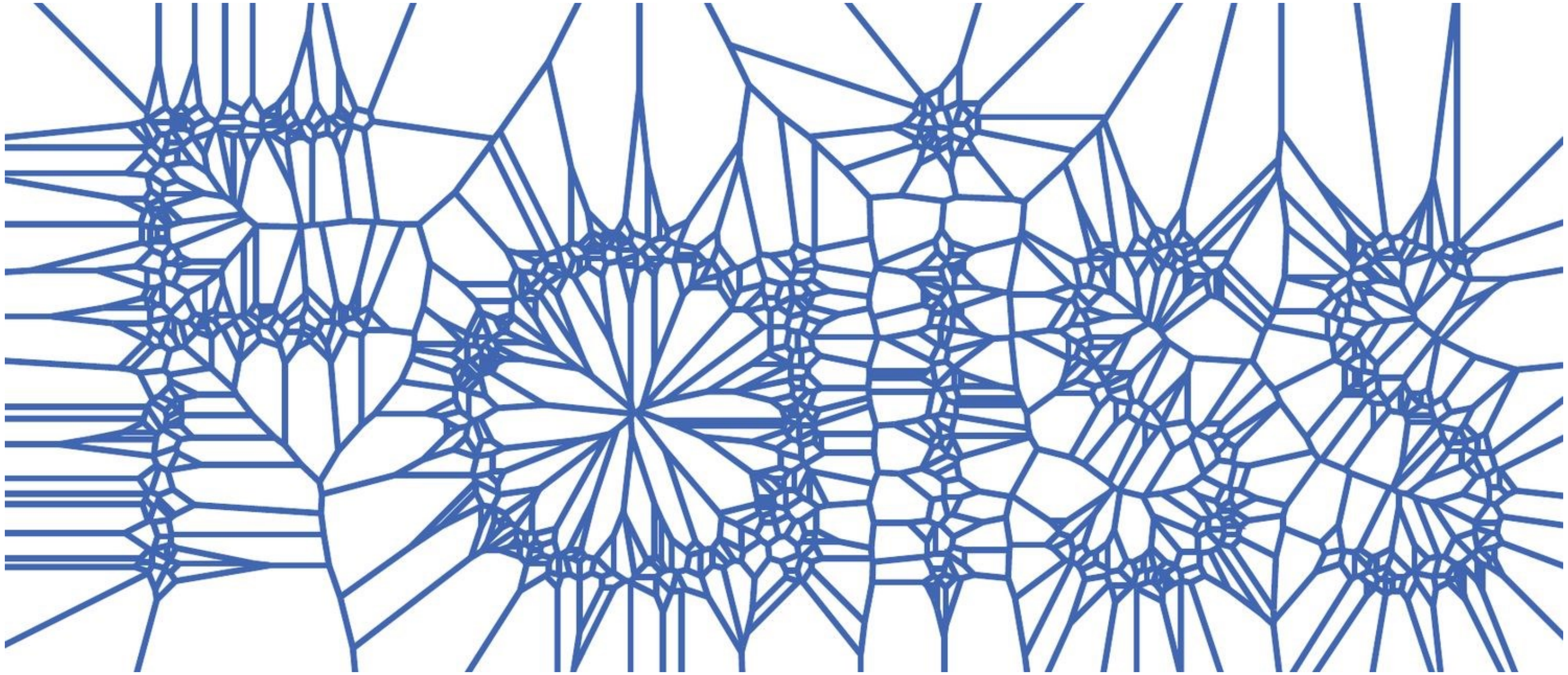
Order Aware Metrics: Normalized Discounted Cumulative Gain

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$



k	DCG@k	IDCG@k	NDCG@k
1	3	3	3 / 3 = 1
2	4.2618	4.8927	4.2618 / 4.8927 = 0.8710
3	5.7618	5.8927	5.7618 / 5.8927 = 0.9777
4	5.7618	6.3233	5.7618 / 6.3233 = 0.9112
5	6.1486	6.3233	6.1486 / 6.3233 = 0.9723

Meta



facebookresearch/faiss: A library for efficient similarity search and clustering of dense vectors. (github.com)

References

Introduction to Information Retrieval (stanford.edu)

Introduction to Information Retrieval | Kaggle

Intro-to-NLP-IR.pdf (unitn.it)

What is Information Retrieval (IR) in Machine Learning? | Simplilearn

A Gentle Introduction to Vector Space Models - MachineLearningMastery.com

Evaluation Metrics For Information Retrieval (amitness.com)