CSCE604135 | Perolehan Informasi (Information Retrieval)

An Intro to Information Retrieval

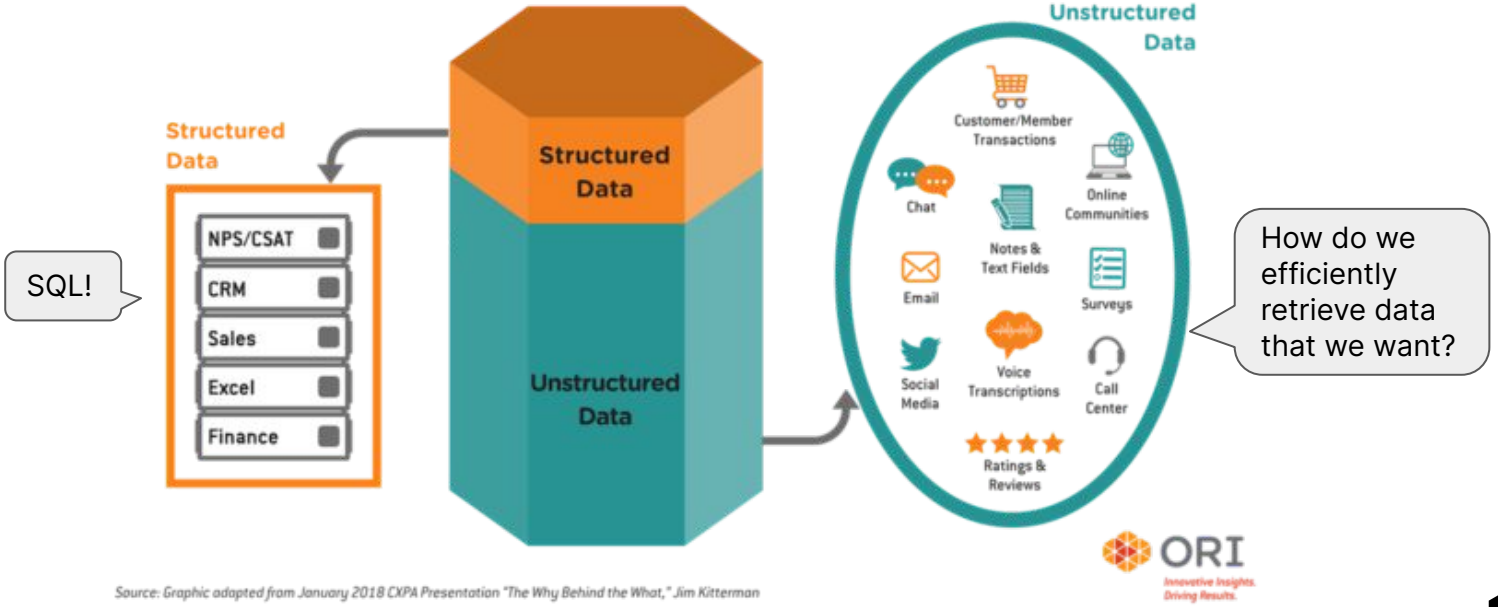
Radityo Eko Prasojo, PhD

Contents

- Why IR?
- What is IR? What are the components?

Why IR?

In 2021: unstructured data becomes more and more common



Source: Graphic adapted from January 2018 CXPA Presentation "The Why Behind the What," Jim Kitterman

A trick: metadata - but what if there are none?

Medium

1K 11

Climate Change | Global Warming | Oceans | Marine Life | Society

https://pixabay.com/photos/maine-coon-cat-cat-s-eyes-black-cat-694730/

maine coon | cat | cat's eyes | black cat | animal portrait | zoom background | kitty | pet | feline | domestic cat | cat portrait | photos

No metadata - just content

How do we obtain relevant information effectively?

Here comes Information Retrieval (IR)!

In essence: finding relevant materials (e.g. documents) of an unstructured nature (e.g. texts), according to some criteria (e.g. query, or keywords), from a large collection of resources (e.g. millions of documents stored in a computer).

**applies to many kinds of unstructured data, but in this course we focus on text*

Contemporarily, the most typical use case of an IR is a **search engine,** e.g.:

- On your email
- On your PC folder
- On your company/university shared repository
- Web search engine

IR application: Search Engines

Example

Google Scholar

information retrieval

Articles

About 4,410,000 results (**0.04** sec)

Any time
Since 2021
Since 2020
Since 2017
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

[BOOK] Modern **information retrieval**
R Baeza-Yates, B Ribeiro-Neto - 1999 - people.ischool.berkeley.edu
Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out of date …
Cited by 18991 Related articles All 25 versions

[BOOK] Introduction to modern **information retrieval**
GG Chowdhury - 2010 - books.google.com
An **information retrieval** (IR) system is designed to analyse, process and store sources of **information** and retrieve those that match a particular user's requirements. A bewildering range of techniques is now available to the **information** professional attempting to …
Cited by 877 Related articles All 6 versions

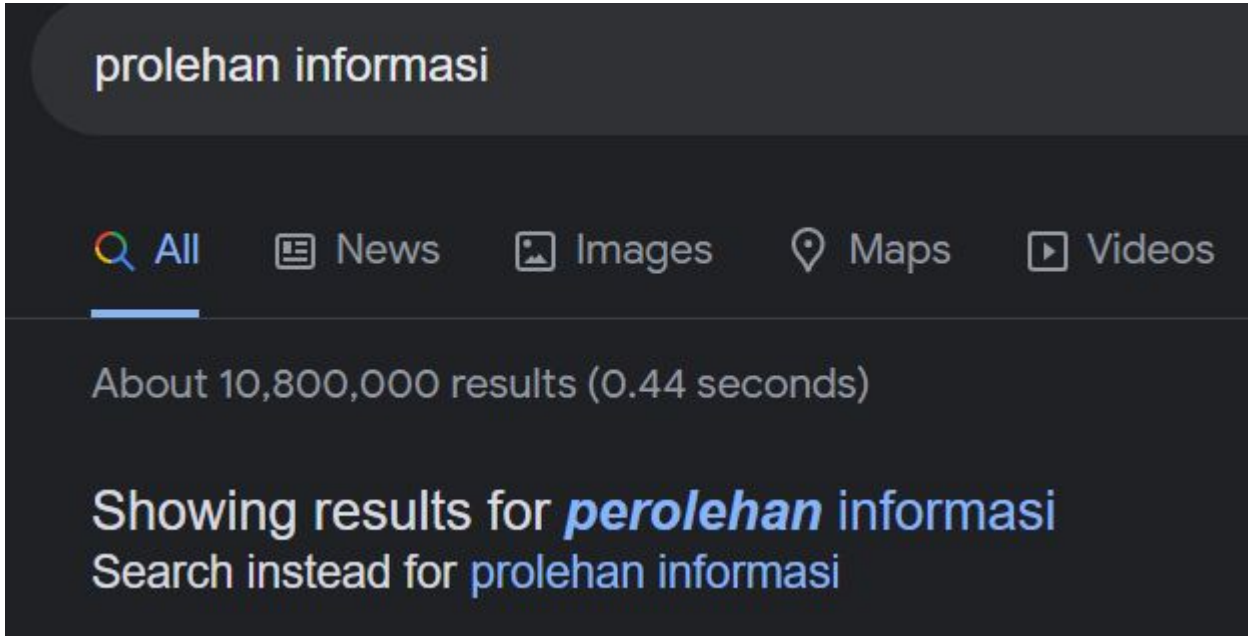
[BOOK] Readings in **information retrieval**
KS Jones, P Willett - 1997 - books.google.com
This is a marvelous collection. The papers have been selected with care, and provide an unprecedented concentration of knowledge about **information retrieval**. The introductory material, by two of the leading researchers in IR, is itself a valuable reference to the history …
Cited by 857 Related articles All 2 versions

It's just string matching right? What's the big deal

```
# criteria like year, etc.
def process(query,criteria):
    relevantdocs = []
    for doc in docs:
        # check if doc contains q
        if query in doc
        and doc matches criteria:
            relevantdocs.add(doc)
    return relevantdocs
# Boom! I have an IR system?
```

- **Syntax**/textual problem: typo, word order, etc.: "Efektivitas vaksin" vs "vaksin efektifitas"
- **Semantic** problem: polysemy, homonymy: "milk", "hak", "bunga"
- **Relevancy** order: which doc comes first?
- **Performance**: what if there are billions of documents?

Google’s typo detection



Semantic problem: I meant the animal!

Google

jaguar

All | Images | Videos | News | Maps | More | Settings | Tools

About 443,000,000 results (0.92 seconds)

www.jaguar.com › market-selector

Market Selector | Jaguar | View the site in your preferred ...

... Rover - Above and Beyond. Land Rover. Faded out winding mountain pass. VIEW **JAGUAR** IN YOUR MARKET. ASIA PACIFIC. AUSTRALIA · BANGLADESH.

Build · Jaguar XJ · New Jaguar F-PACE · I-Pace

www.jaguar.com

Jaguar: Luxury Saloons, Performance SUVs & Sports Cars

The official website of **Jaguar**. Discover our range of vehicles from the XE, XF, XJ, F-Type, F-PACE, E-PACE & our all-electric sports car I-PACE.

Images for jaguar

Jaguar Cars

Luxury vehicles company

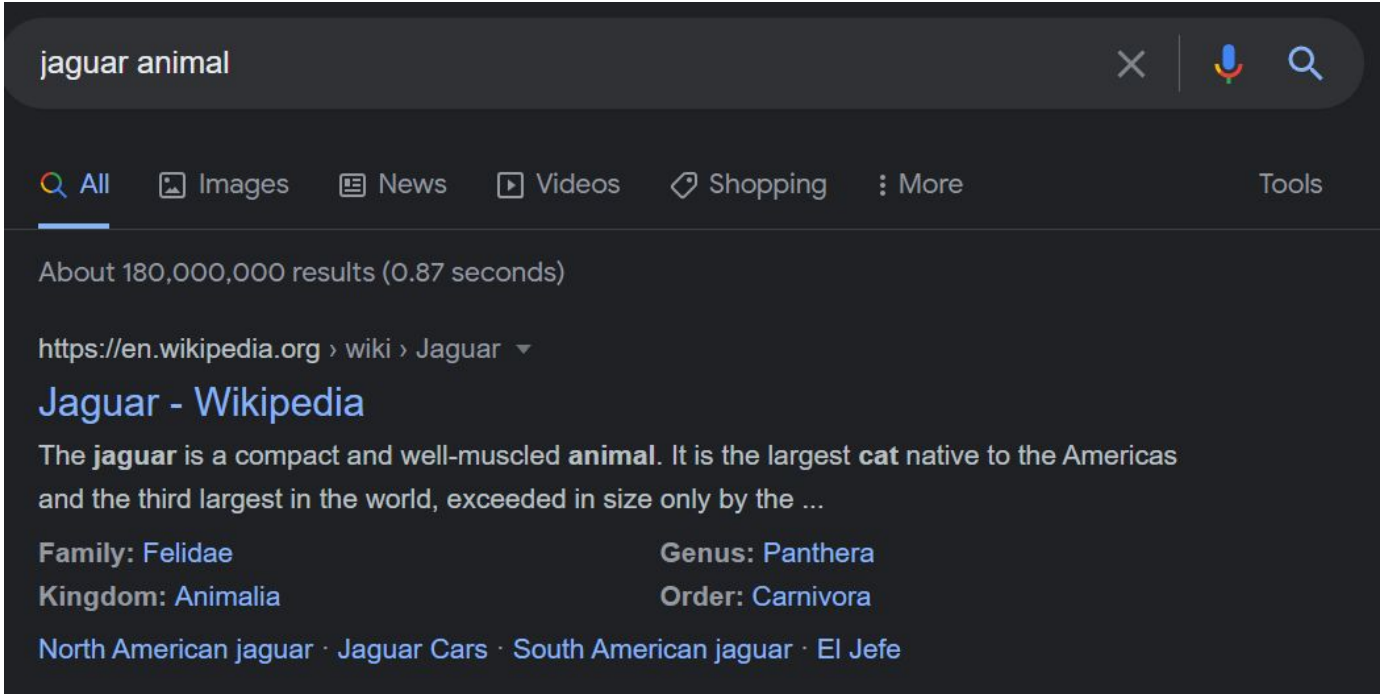
JAGUAR

jaguarlandrover.com

Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England.

Wikipedia

As an expert googler, I can do:



Synonyms

siapa anak jokowi yang jualan terang bulan

All | Images | Maps | News | Videos | More | Settings | Tools

About 6 results (0.61 seconds)

Ini satu-satunya **martabak manis** yang tampilannya menarik, rasanya juga berkualitas. **Gibran** Rakabuming Raka, **anak** Presiden Joko Widodo, mempromosikan Markobar (Martabak Kota Barat) di outlet ke-21 di Kota Malang, tepatnya di Monopoli Cafe, Jl Soekarno Hatta, Sabtu (28/1/2017). Jan 29, 2017

surabaya.tribunnews.com › Travel › Kuliner

Anak Jokowi Buka Outlet Martabak di Kota Malang, Berikutnya ...

About Featured Snippets | Feedback

Why in this **order**, Google?

About 180,000,000 results (0.87 seconds)

https://en.wikipedia.org › wiki › Jaguar

Jaguar - Wikipedia

The **jaguar** is a compact and well-muscled **animal**. It is the largest **cat** native to the Americas and the third largest in the world, exceeded in size only by the ...

Family: Felidae
Genus: Panthera
Kingdom: Animalia
Order: Carnivora

North American jaguar · Jaguar Cars · South American jaguar · El Jefe

https://www.wwf.org.uk › learn › fascinating-facts › jag...

Top 10 facts about Jaguars | WWF

3. They're on the chunky side. The **jaguar** is the third biggest **cat** in the world - after the tiger and the lion - and is ...

And how is it so blazing fast?

About 180,000,000 results (0.87 seconds)

https://en.wikipedia.org › wiki › Jaguar

Jaguar - Wikipedia

The **jaguar** is a compact and well-muscled **animal**. It is the largest **cat** native to the Americas and the third largest in the world, exceeded in size only by the ...

Family: Felidae
Genus: Panthera
Kingdom: Animalia
Order: Carnivora

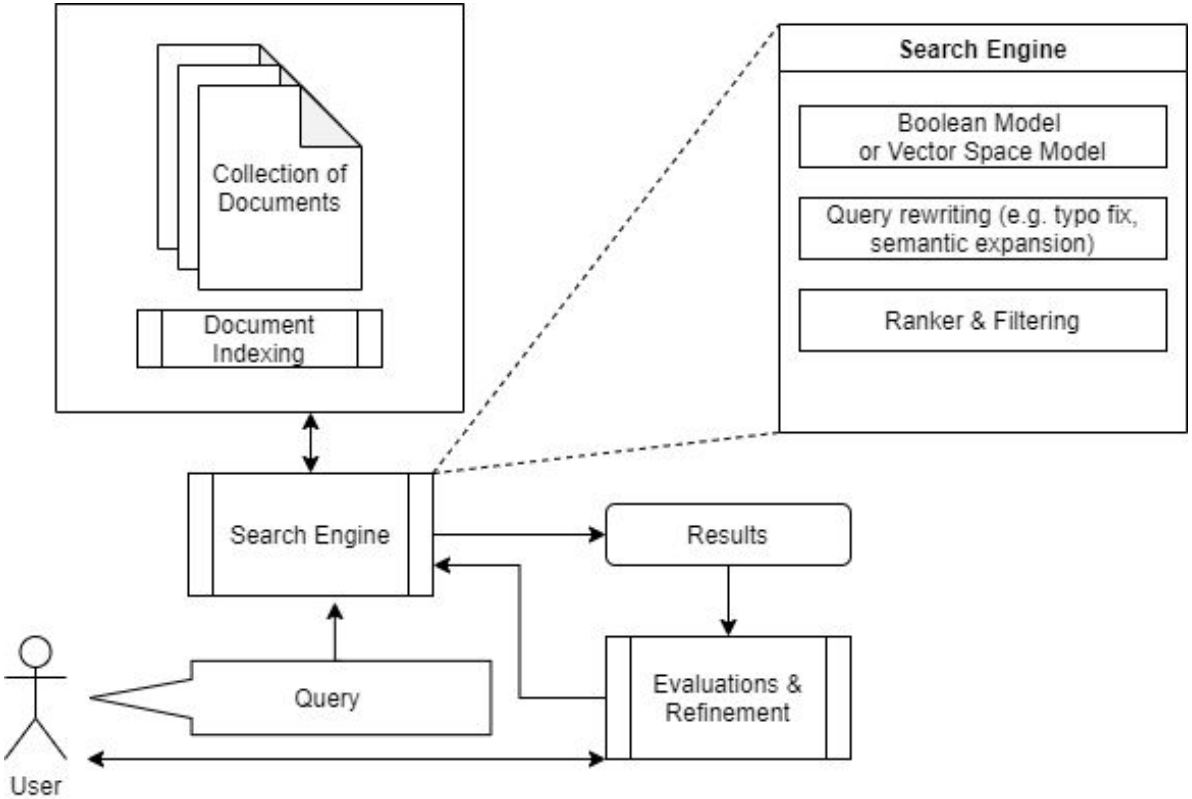
North American jaguar · Jaguar Cars · South American jaguar · El Jefe

https://www.wwf.org.uk › learn › fascinating-facts › jag...

Top 10 facts about Jaguars | WWF

3. They're on the chunky side. The **jaguar** is the third biggest **cat** in the world - after the tiger and the lion - and is ...

Overview of an IR System



Boolean Model

- Document either matches a query or does not match, no inbetween
- Can add boolean operators between keywords

“Information **AND** retrieval”
“Information **AND NOT** retrieval”
“Information **OR** retrieval”

“Information **AND** retrieval”

Doc1

Information retrieval is a branch of computer science that ...

Doc2

The **retrieval** of the shipwreck was difficult ... **Information** about its location was not precise ...

Doc3

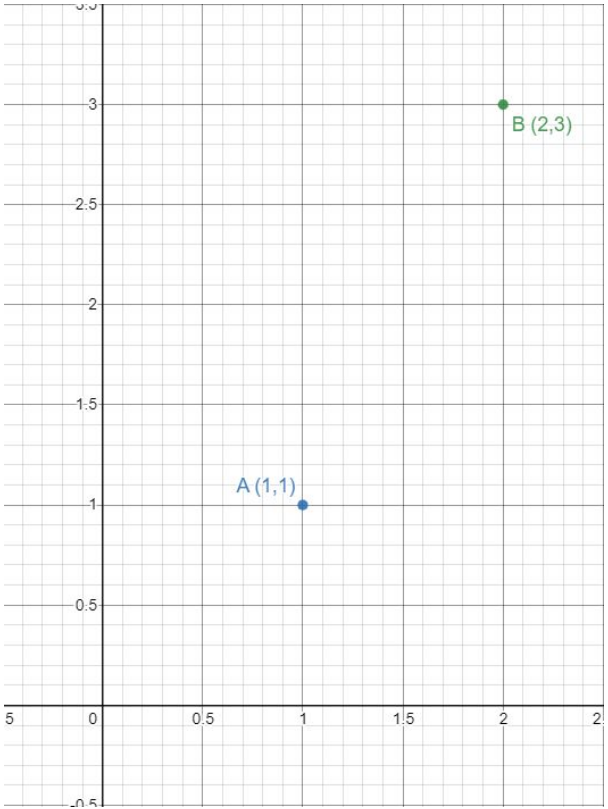
Rendang is one of the most popular dishes in Indonesia. It has a savoury, sweet taste, with mild spiciness ...

Ranker

Relevant documents still needs be ordered!

Vector Space Model

- Let A and B be two points in a 2D space, what would be the distance?



Euclidean Distance

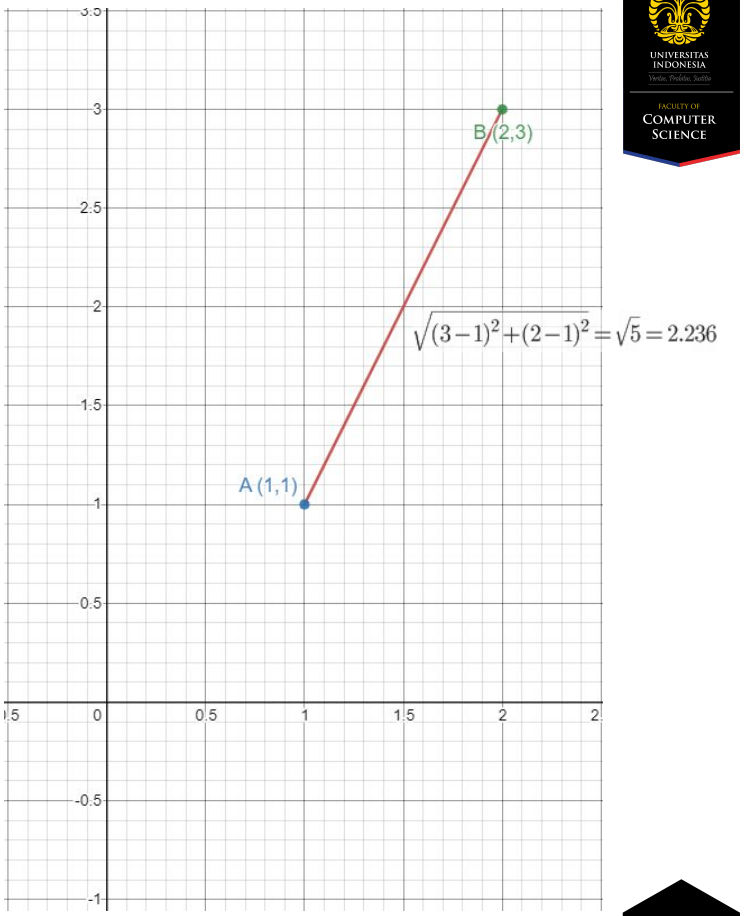
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

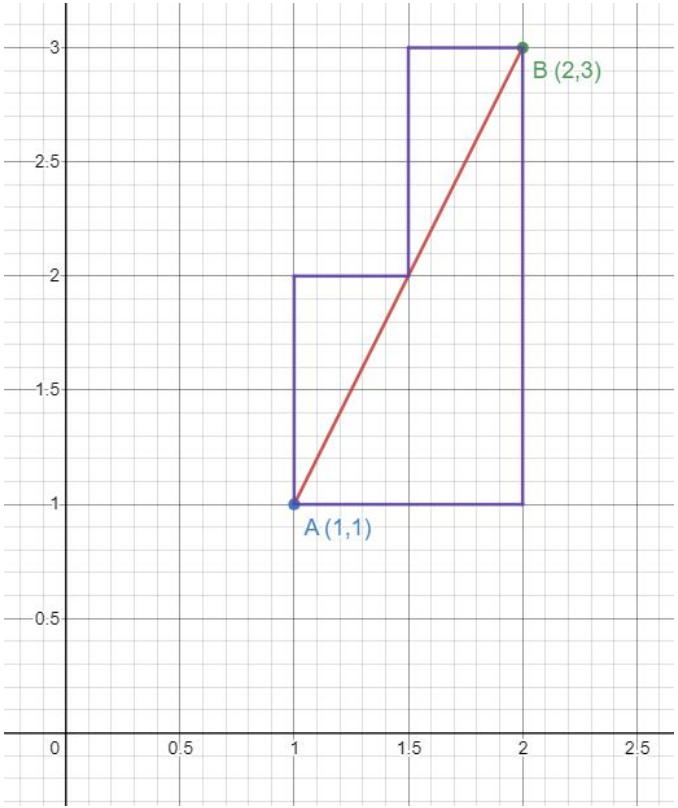
n = n-space

- Note that this is extendable to any big dimension *n*



Manhattan Distance

$$d(p,q) = \sum_{i=1}^{n} |q_i - p_i|$$

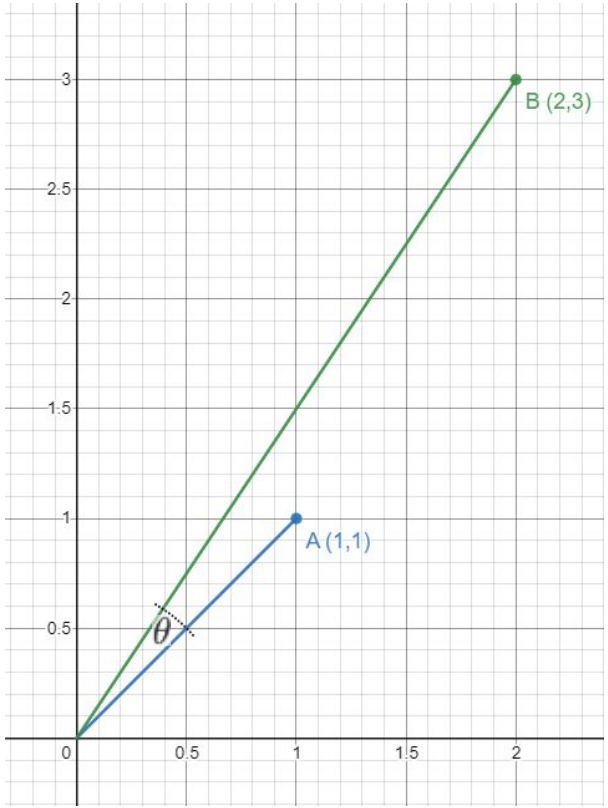


- **Purple** = manhattan distance
- **Red** = euclidean distance
- Note that this is also extendable to any big dimension *n*

Cosine Similarity

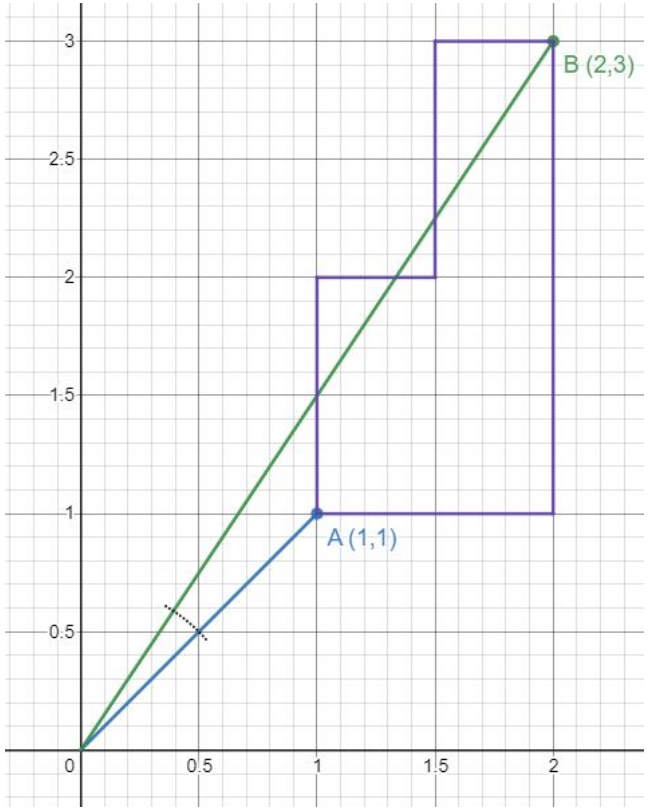
$$d(p,q) = cos(\theta) = \frac{p \cdot q}{||p||\,||q||}$$

$$= \frac{\sum_{i=1}^{n} p_i\, q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\ \sqrt{\sum_{i=1}^{n} q_i^2}}$$



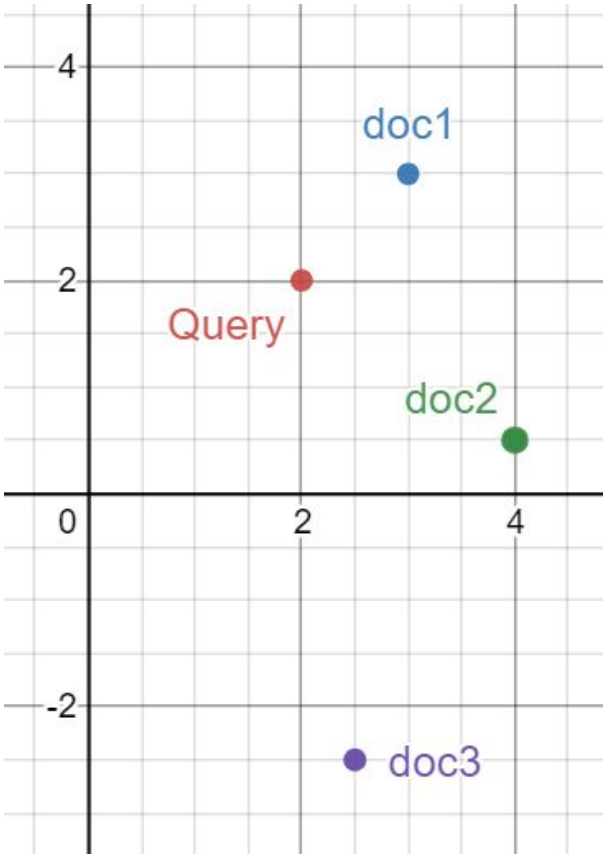
Vector Space Model

- Note again: this is extensible to any large *n*
- In n = 5, a point C can be **[1,2,6,3,-2]**
- In n = 1000, a point D = **[6,100,2.5,...]**
- This representation of points as list of numbers is referred to as **vector**
- In any dimension n, we can compute the distance between any two points P1 and P2.



Vector Space Model

- Suppose that we have a **mechanism** to map queries and documents into points in a dimension of *n*.
- Then, selecting and ranking relevant documents can be done by simply computing the distances between the query and each document.
- **The closer the document to a query is, the higher it ranks!**
- In this course, we will study several of such **mechanisms**.



Ranker

- In the context of **search engine**, there are at least two factors in ranking: document **relevance** and document **importance**.
- In IR, document relevance is the focus.
- In commercial search engine though, algorithms such as **PageRank** are implemented to measure document importance or popularity in order to influence the final ranking.

Nasi padang yang enak

Doc4

"yang" appears 10 times, "nasi" "padang" and "enak" 0 times

Doc5

"nasi" appears 6 times, "padang" and "enak" 1 time

Doc6

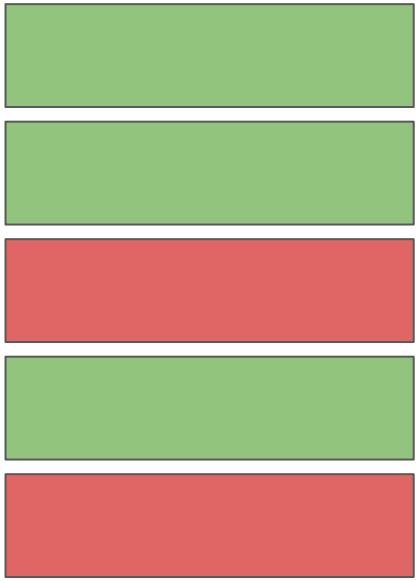
"nasi" "padang" and "enak" all appear 2 times

If we count pure word occurrences, Doc 4 would rank higher, *is that correct*?

We will learn the fact that not every word carries the same **weight**, which can be leveraged for ranking

Evaluation

Top 5 results



- How good are the resulted documents?
- **Precision:** among the returned documents, how are relevant?
- **Recall**: among the documents that should be relevant, how many are returned?
- Q: Suppose there are 4 relevant documents in our collection, what is the precision and recall of the figure on the left?
- We will learn these evaluation metrics in more detail

Feedback, refinement, and query rewriting

- User may give **feedback** on certain pages, e.g. by clicking.
- This feedback can be used to affect the ranking in the future (though, because this is metadata, it is outside the scope of this course)

- **Refinement** can be done automatically, e.g. by detecting that a query resulted in very few documents, due to **typos** or **unusual keyword**, which can be **rewritten**.
- E.g. “nasi dimasakkan” can be rewritten into “nasi masak”, which potentially yield more results.

Indexing

- No indexes - no scalable IR system!
- IR systems employs **inverted indexing**

```
"Nasi" → {doc5, doc6, doc13, …}
"Retrieval" → {doc1, doc2, doc9, …}
```

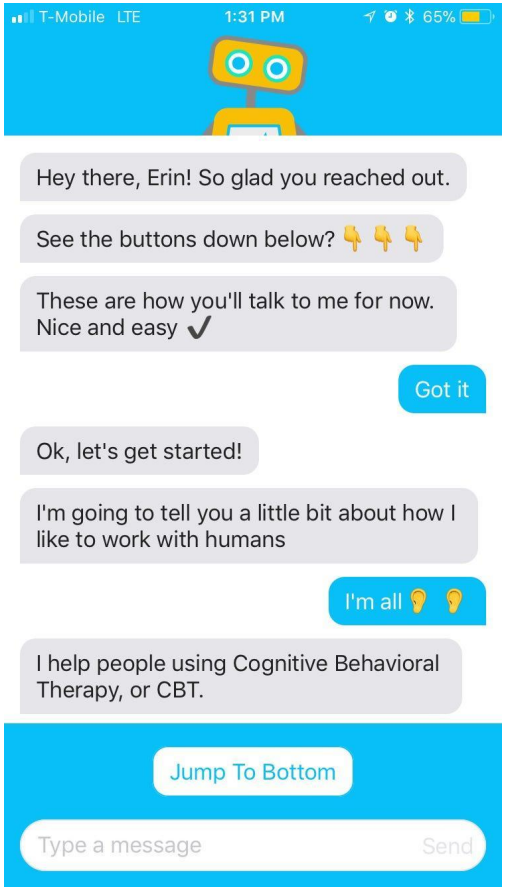
- This works well with the **boolean model,** making use of **boolean operators**.
- To achieve this, the documents need to be **tokenized**, **preprocessed** and sometimes **annotated**

Information Retrieval In other applications

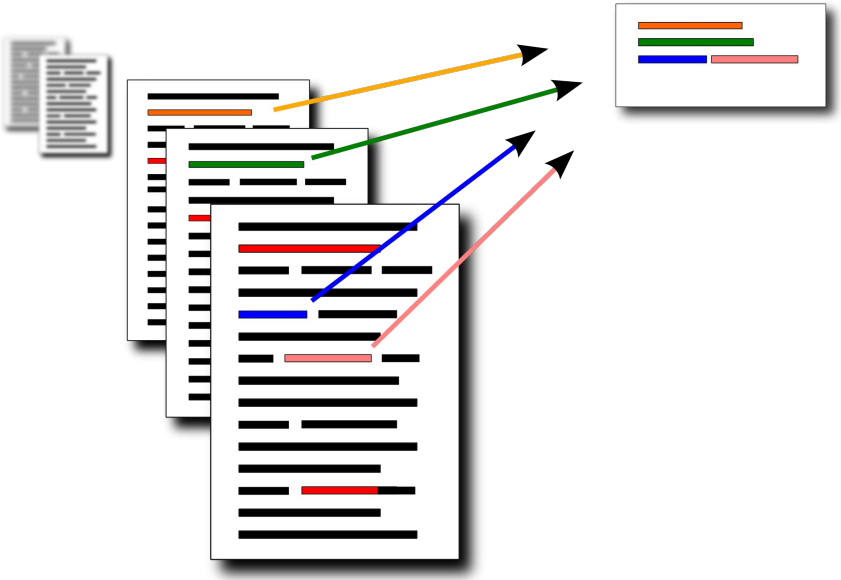
Chatbots


TSEL 11.04 99%
reynirfauzan 9m
Jemma
Today
Jemma
MERRY XMAS!
Let the spirit of love fill your heart and home this Christmas!
10.00
Selamat Natal, Sahabat Jemma!
Apa yang kamu lakukan untuk merayakan hari natalmu tahun ini? Yuk, share ke Jemma
10.00
Read 11.02
emang boleh ngucapin selamat natal jem?
Keyakinanku sih, kita
Send message
TSEL 11.05 99%
reynirfauzan 7m
Jemma
Read 11.02
emang boleh ngucapin selamat natal jem?
Keyakinanku sih, kita diciptakan untuk berbuat baik sesama ciptaanNya
11.02
Read 11.03
ga takut diciduk ormas?
Kenapa emangnya?
11.03
Read 11.03
ya karena kamu ngucapin selamat natal
Maksudnya?
11.03
#SAVAGE
Read 11.03
iya kan kalo soal agama gitu orang indonesia sensitip
Aku percaya adanya Tuhan
11.03
Dia selalu membalas kebaikan dengan kebaikan juga :D
11.03
Send message

Expert systems

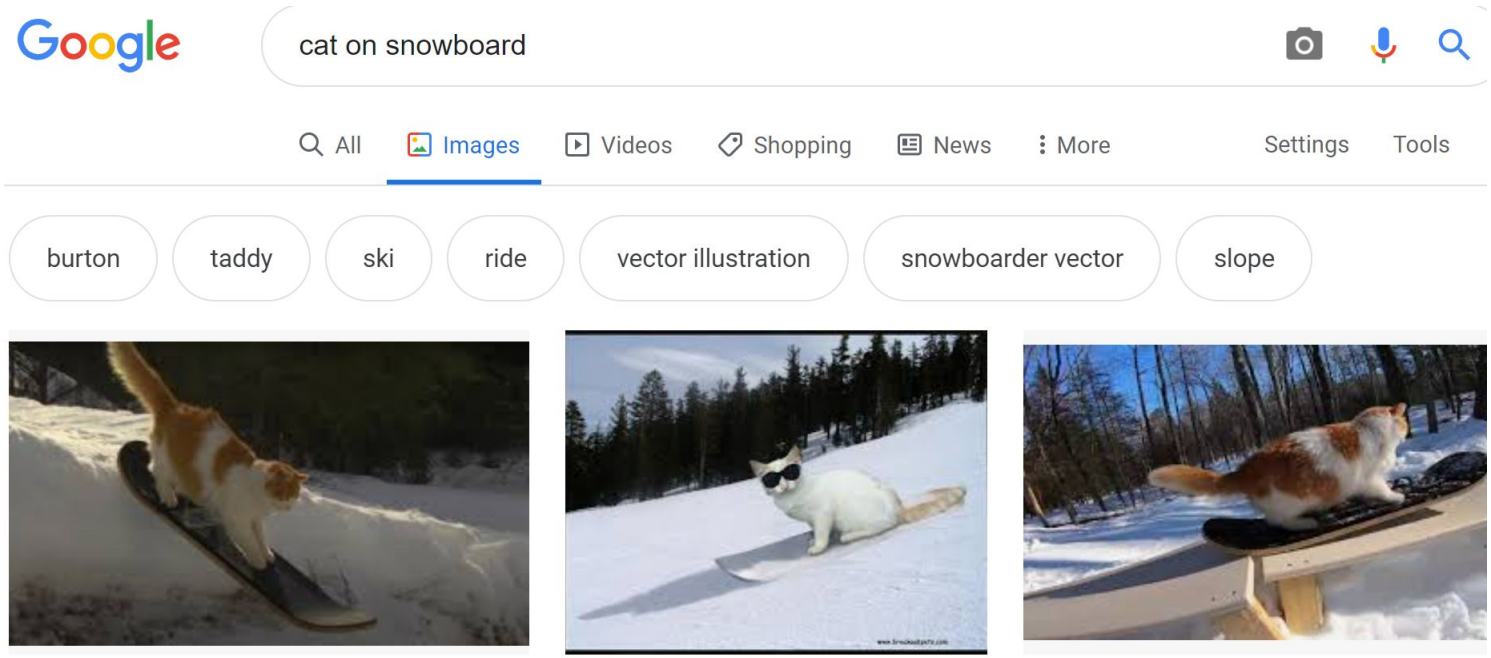


Summarization



Complex filtering: Sentiment analysis

Multimodal IR



Recap: what we will learn

- We will learn how to **tokenize, preprocess, and annotate text data**, as this is an important prerequisite for both query processing and document processing for indexing.
- We will learn how to build inverted index and how to make them efficient and compact.
- We will learn about boolean models and vector space models, and how to improve them with query rewriting (such as typo detection)
- We will learn about how to rank documents
- We will learn how to evaluate IR models, and
- We will learn several technologies that are relevant to IR

Supported by the Kampus Merdeka grant of
Ministry of Education, Culture, Research, and Technology
of Republic of Indonesia