Information Retrieval

Natural Language Processing

Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

What is IR?

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).

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
Introduction to Information Retrieval

What is IR?


Google
Buscar con Google
Voy a tener suerte
Ofrecido por Google en: English català galego euskara

IR beyond Google

- Finding items (documents, webpages, images) of an **unstructured nature** (usually text) from within large collections.
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval
  - Image Search

Structured vs. Unstructured

```
SELECT ?person WHERE{
   ?person dcterms:subject <http://es.dbpedia.org/resource/Categoría:Científicos_de_España>
}
```

person
http://es.dbpedia.org/resource/José_Antonio_Pavón_y_Jiménez
http://es.dbpedia.org/resource/Carmen_Vela
http://es.dbpedia.org/resource/Fernando_Baquero_Mochales
http://es.dbpedia.org/resource/Juana_Álvarez-Prida_y_Vega
http://es.dbpedia.org/resource/Montserrat_Gomendio
http://es.dbpedia.org/resource/Pedro_Mayoral_Carpintero
http://es.dbpedia.org/resource/Severo_Ochoa
http://es.dbpedia.org/resource/Josep_Trueta
http://es.dbpedia.org/resource/Bernardo_Rodríguez_Largo
http://es.dbpedia.org/resource/David_Vázquez_Martínez
http://es.dbpedia.org/resource/Javier_Gafo
http://es.dbpedia.org/resource/Agustín_Escardino
http://es.dbpedia.org/resource/Elías_Fereres_Castiel
http://es.dbpedia.org/resource/Emilio_Muñoz_Ruiz
http://es.dbpedia.org/resource/José_María_Mato_de_la_Paz
http://es.dbpedia.org/resource/Policarp_Hortolà
http://es.dbpedia.org/resource/Amador_Schüller
http://es.dbpedia.org/resource/Miguel_Guirao_Gea
http://es.dbpedia.org/resource/Antonio_Camacho_Díaz
http://es.dbpedia.org/resource/Montserrat_Soliva_Torrentó
http://es.dbpedia.org/resource/Ricardo_Carmona_(físico)

Structured vs. Unstructured

spanish scientists

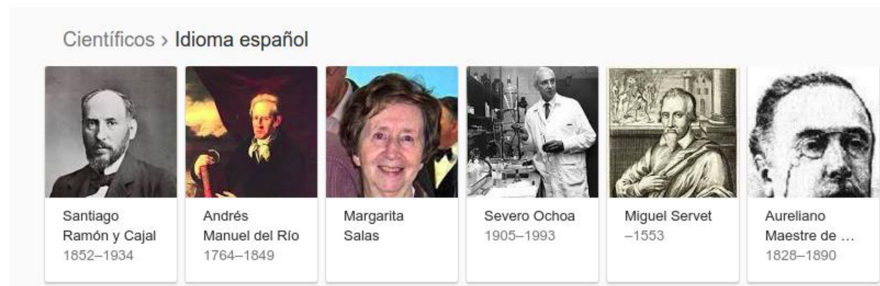
[Category:Spanish scientists - Wikipedia](https://en.wikipedia.org/wiki/Category:Spanish_scientists)
https://en.wikipedia.org/wiki/Category:Spanish_scientists ▾ Traducir esta página
Pages in category **"Spanish scientists"**. The following 41 pages are in this category, out of 41 total. This list may not reflect recent changes (learn more). A. José de Acosta · José María Albareda · María de los Ángeles Alvariño González · Antonio Arnaiz-Villena · Félix de Azara. B. Ferran Sunyer i Balaguer · Xavier Barcons ...

List of Spanish inventors and discoverers - Wikipedia
https://en.wikipedia.org/.../List_of_Spanish_inventors_and_disco... ▾ Traducir esta página
This is a list of **Spanish** inventors and discoverers. Santiago Ramón y Cajal, father of Neuroscience, Nobel prize Laureate. Contents. [hide]. 1 A; 2 B; 3 C; 4 D; 5 E; 6 F; 7 G; 8 H; 9 J; 10 L; 11 M; 12 O; 13 P; 14 R; 15 S; 16 T; 17 U; 18 V; 19 See also; 20 References. A[edit]. José de Acosta (1540–1600), one of the first naturalists ...
C · G · M · R

Famous Scientists from Spain | List of Top Spanish Scientists - Ranker
https://www.ranker.com/list/famous-scientists...spain/reference ▾ Traducir esta página
List of notable or famous scientists from Spain, with bios and photos, including the top scientists born in Spain and even some popular scientists who immigrated to Spain. If you're trying to find out the names of famous **Spanish scientists** then this list is the perfect resource for you. These scientists are among the most ...

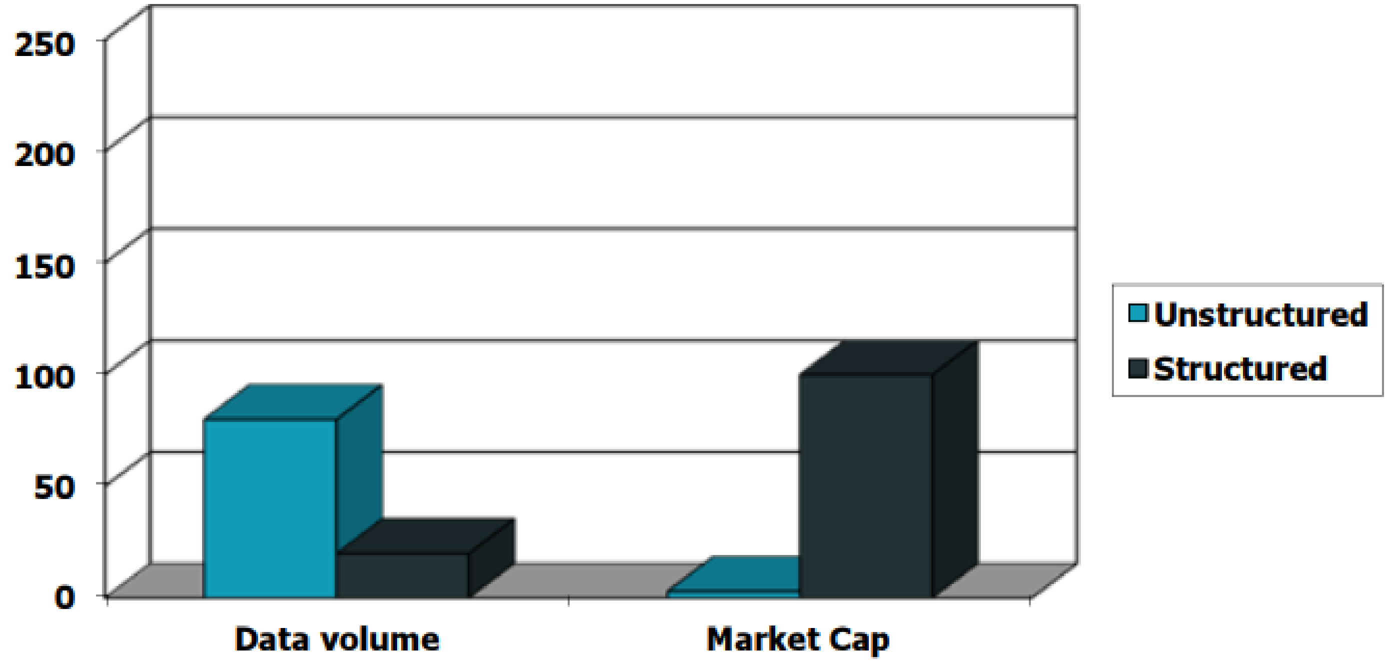
10 Hispanic Scientists You Should Know | HowStuffWorks
https://science.howstuffworks.com › ... › Physicists ▾ Traducir esta página
Over the centuries, many remarkable **scientists** have emerged from **Spanish**-speaking lands, cultures and ancestors. Though grouping such a diverse collection of people under a single rubric -- particularly the politically expedient but dubious term **Hispanic** – isn't ideal, it does make room to explore their wide-ranging array ...

Top H-Index For Scientists in Spain - Guide 2 Research
www.guide2research.com/scientists/ES ▾ Traducir esta página
Top H-Index For **Scientists** in **Spain** : We list only **scientists** having H-Index>=40. If you or other scholars are not listed, we appreciate if you can contact us ... 12,538. 58. 996. 7. Mario Piattini · University of Castile-La Mancha · **Spain**. 15,694. 56. 1034. 8. Carles Sierra · **Spanish** National Research Council · **Spain**. 17,292. 55 ...



Structured vs. Unstructured

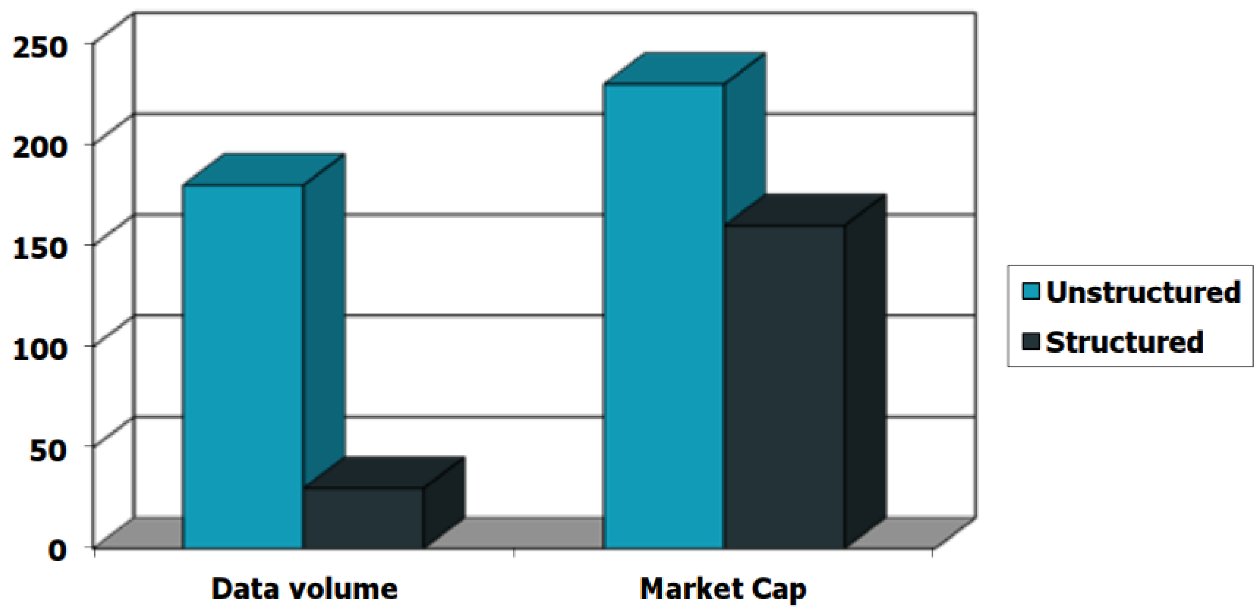
- Unstructured (text) vs. structured (database) data in the mid-nineties.



Source: Stanford NLP

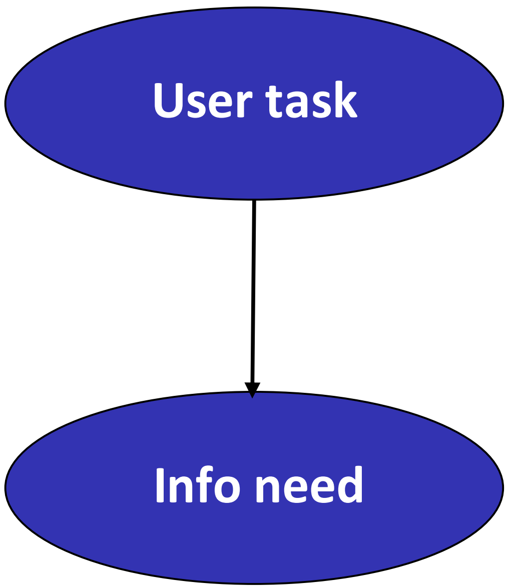
Structured vs. Unstructured

- Unstructured (text) vs. structured (database) data today.



Source: Stanford NLP

Classical Model



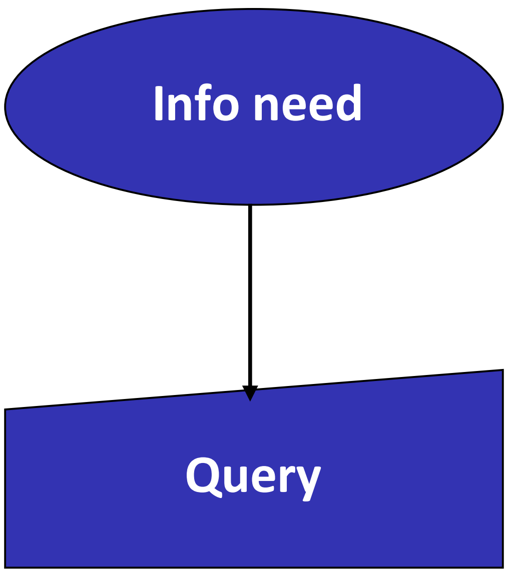

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-289-bcb7ba6c7df2> in <module>()
     28         m.addConstr(s[t] == p[t], name = "charge_{}".format(t))
     29     else:
---> 30         m.addConstr(s[t] == s[t-1] + p[t], name = "charge_{}".format(t))
     31 
     32 # integrate variables and constraints

KeyError: 0
```

Misconception?

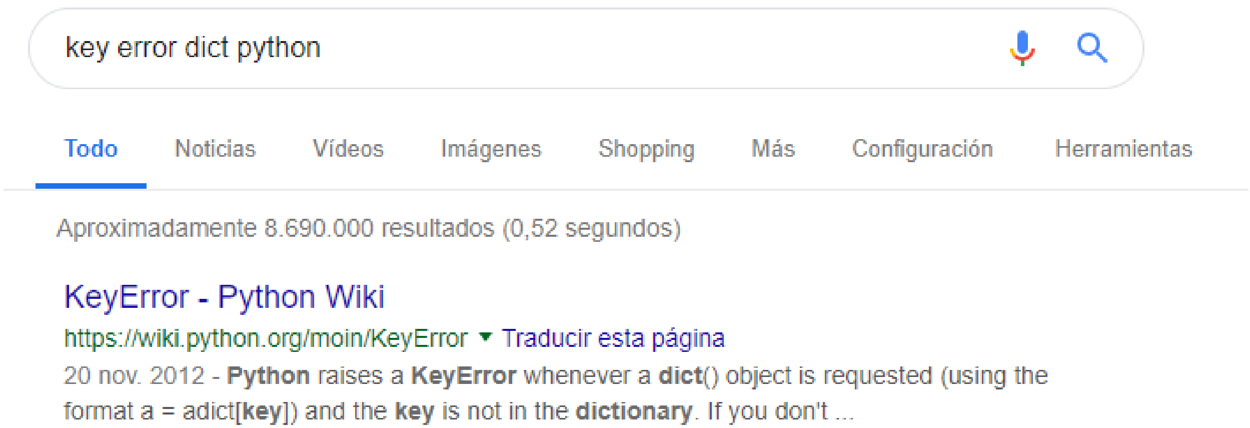
It seems something is wrong with my dict, I guess it's about the key.

Classical Model

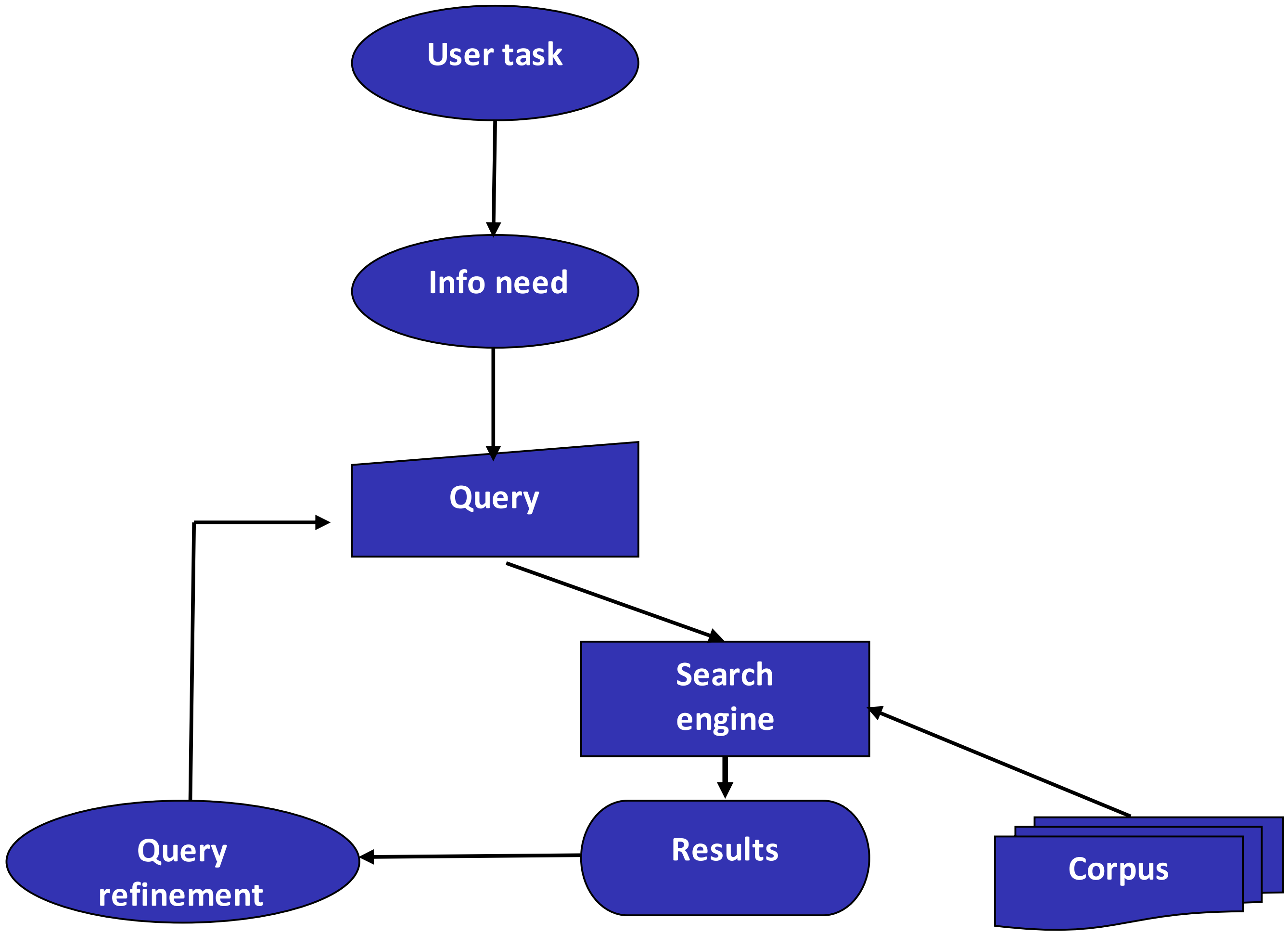


It seems something is wrong with my dict, I guess it's about the key.

Misformulation?



Classical Model



Why is it hard?

- **From user task → Query**
  - **Misconception:** But users don't often know what they want
  - **Misformulation:** Verbalizing information needs
- **Query**
  - **Semantics**
    - *banks in Madrid*
    - *second-hand jaguar*
  - **Context**
    - *funny images about…*
    - *cheap hotels*
  - **Weak signals**
    - Only a few words to express a need

Preliminary Approach

- Which plays of Shakespeare contain the words ***Brutus*** *AND* ***Caesar*** but *NOT* ***Calpurnia***?

```
grep(Shakespeare's plays,(Brutus and Caesar)|strip out(Calpurnia)
```

- Problems:
  - **Slow**: you have to go over all documents
  - ***NOT Calpurnia*** is non-trivial
  - Other operations not feasible:
    - find the word ***Romans*** near ***countrymen***
  - Ranked retrieval (best documents to return)

Term-document incidence matrices

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word,
0 otherwise

Brutus** AND **Caesar** BUT NOT **Calpurnia

Incidence vectors

- We have a 0/1 vector for each term.
- **Answer query**:
  - take the vectors for ***Brutus, Caesar*** and ***Calpurnia*** (complemented) ➔ bitwise *AND*.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

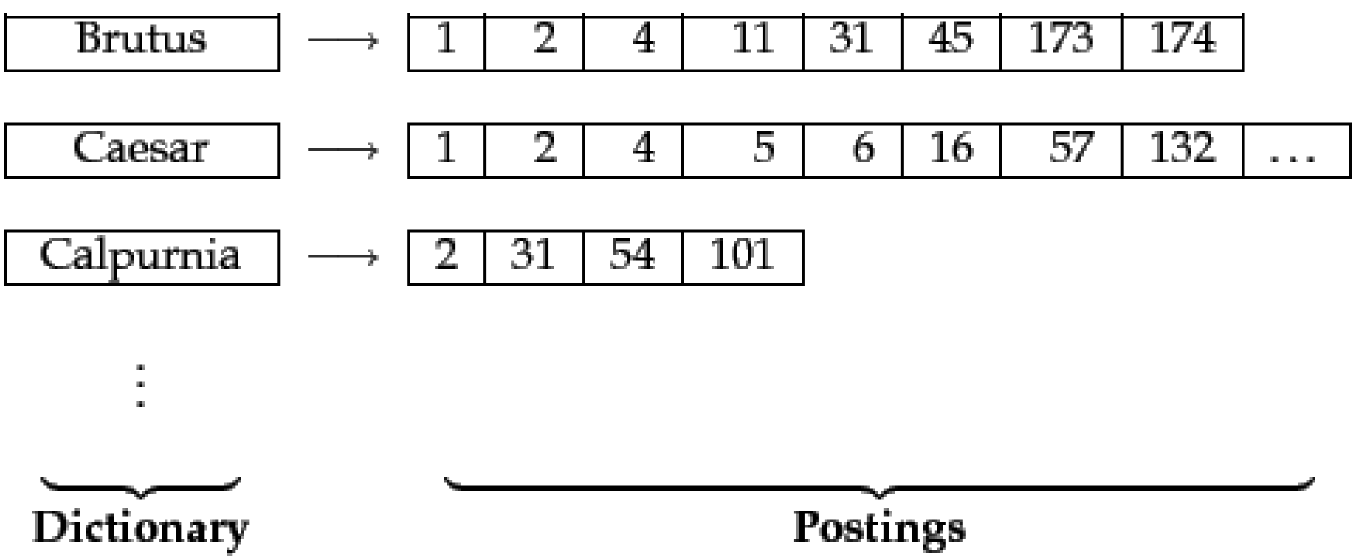
110100 *AND* 110111 *AND* 101111 = **100100**

Bigger collections

- N = 1 million documents, *document* = 1000 words.
- avg. 6 bytes/word including spaces/punctuation
  - **6GB of data** in the documents.
- M = **500K distinct** terms among these.
  - **500K x 1M matrix has half-a-trillion 0's and 1's.**
    - It is extremely sparse
    - We only record the 1 positions.

Inverted index

- For each **term *t***, we must s**tore a list of all documents that contain *t***.



https://nlp.stanford.edu/IR-book/html/htmledition/an-example-information-retrieval-problem-1.html

Indexer steps: Token sequence

- Sequence of (Modified token, Document ID) pairs.

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Indexer steps: Sort

- Sort by terms and docID

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

Indexer steps: Dictionary & Postings

- Multiple term entries in a single document are merged.
- Split into Dictionary and Postings
- Doc. frequency information is added.

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
i	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Query processing: AND

- Consider processing the query: ***Brutus*** *AND* ***Caesar***
  - Locate ***Brutus*** in the Dictionary;
    - Retrieve its postings.
  - Locate ***Caesar*** in the Dictionary;
    - Retrieve its postings.
  - “Merge” the two postings (intersect the document sets):

Brutus ⟶ 1 → 2 → 4 → 11 → 31 → 45 → 173 → 174

Calpurnia ⟶ 2 → 31 → 54 → 101

Intersection ⟹ 2 → 31

https://nlp.stanford.edu/IR-book/html/htmledition/processing-boolean-queries-1.html

Phrase queries

- ***"IE university"* – as a phrase**

  "I went to university at Madrid"

  *"I went to the **IE university** at Madrid"*

- Our **inverted index** is not sufficient

- We need to also store the **position in the document**

A first attempt: Biword indexes

- Index **every consecutive pair of terms** as a **phrase**
- ***““IE university Madrid”*** would generate the bi-words
  - ***IE university***
  - ***University Madrid***
- Each of these **bi-words** is now a **dictionary term**

  <(*term1, term2)* : *docs*>
- Longer phrases can be processed by breaking them down
  - ***“IE university Madrid"***: ***IE university*** *AND* ***university Madrid***

- **Problems?**

Issues for biword indexes

- **False positives**
  - **verify** that the matches of the previous Boolean query:

    *(**IE university** AND **university Madrid**)*

    do **contain the phrase**

- **Index blowup** due to bigger dictionary
  - Huge for bi-words
  - Infeasible for more larger n-grams

Solution 2: Positional indexes

- Store the position(s) of the ***term***

<***term,*** number of docs containing ***term***;

doc1: position1, position2 … ;

doc2: position1, position2 … ;

…

>

Processing a phrase query

- Extract **inverted index entries** for each distinct term: ***to, be, or, not.***
- Merge their ***doc:position*** lists to enumerate all positions with "***to be or not to be***".
  - ***to****:*
    - *2*:1,17,74,222,551; ***4***:8,**16**,**190**,**429**,**433**; *7*:13,23,191; ...
  - ***be****:*
    - *1*:17,19; ***4***:**17**,**191**,291,**430**,**434**; *5*:14,19,101; ...

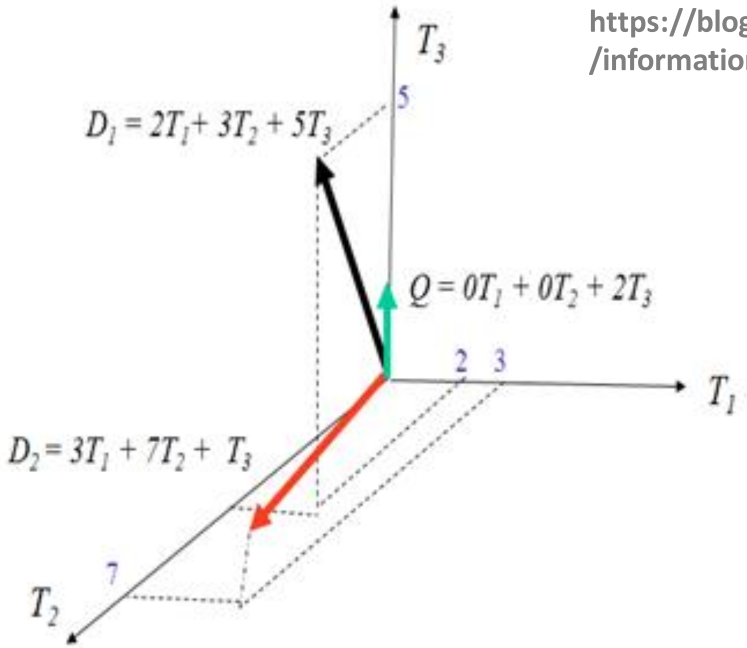
Positional index size

- Positional index ***substantially*** **larger**
  - A positional index is **2–4 as large** as a non-positional index
  - Positional index size **35–50% of volume of original text**

- We **assume** that **because of the power and usefulness** of phrase and proximity queries.

- Bigrams and positional indexes can be **combined**
  - Use bigrams for particular phrases (***"Michael Jackson", "Britney Spears"***)

Solution 3: Vector Space model



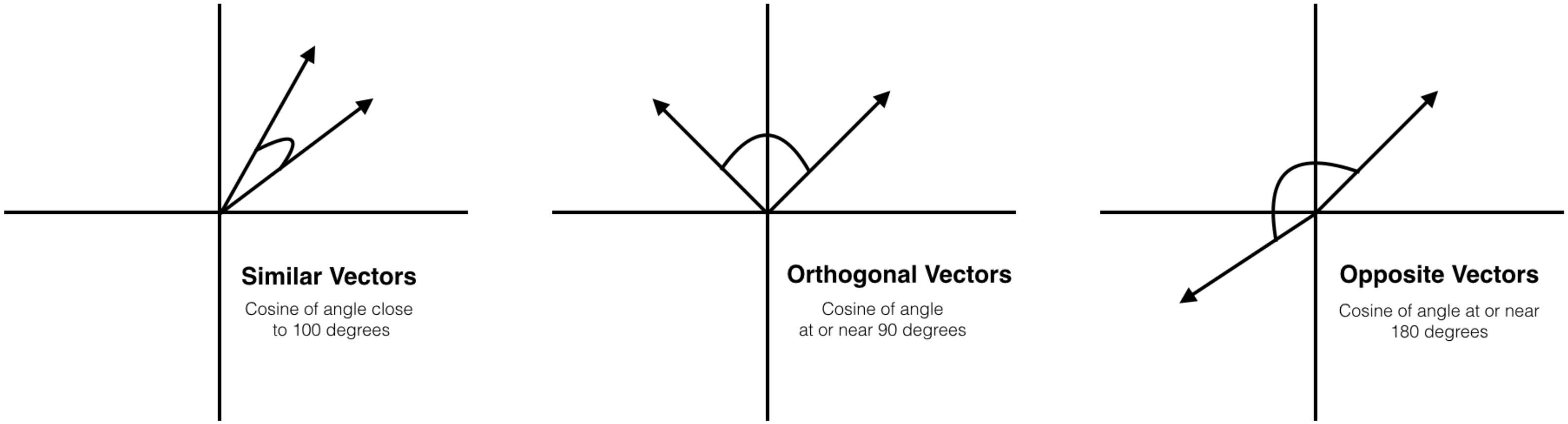
https://blogs.msdn.microsoft.com/spt/2008/03/05/information-retrieval-amp-search-basic-ir-models/

Assumption: Documents close in this space talk about the same things.

Retrieve the **documents close to the query** in this space

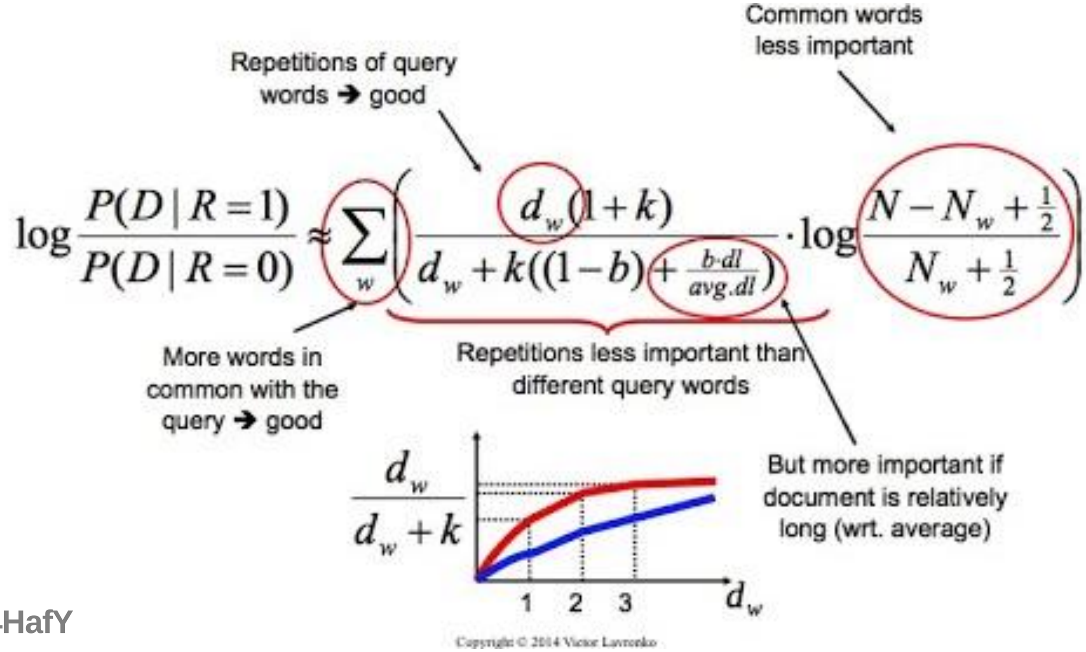
VSM Similarity

- Measure how close two vectors are
  - Cosine Distance



VSM Similarity

- Measure how close two vectors are
  - Cosine Distance
  - Okapi BM25



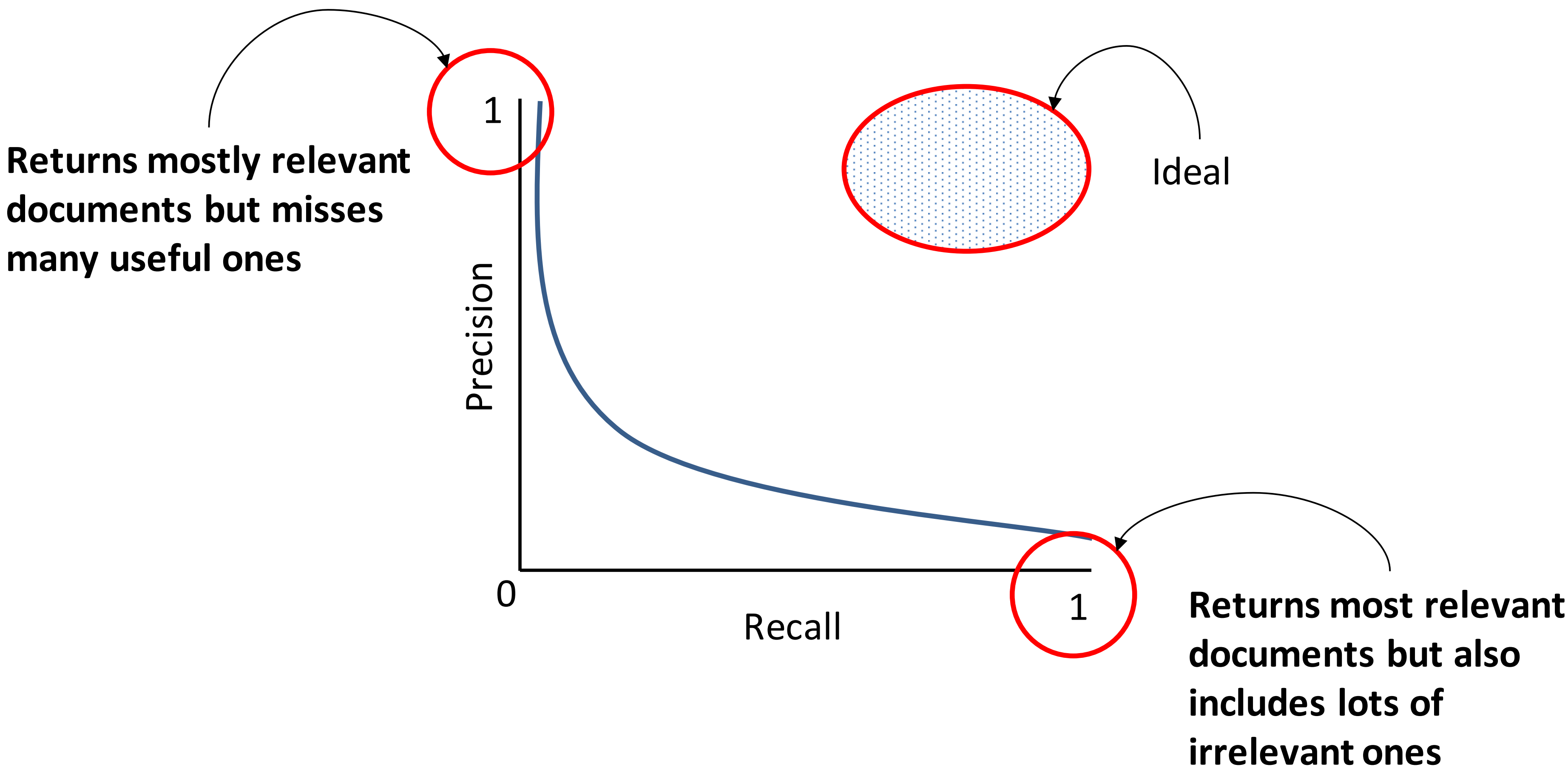
https://www.youtube.com/watch?v=XFIKE34HafY

How good are the retrieved docs?

- **Precision :** Fraction of retrieved docs that are relevant to the user's information need
  - From what you gave, what is what I actually wanted
- **Recall :** Fraction of relevant docs in collection that are retrieved
  - From what I actually wanted, how much have you given me
- **F-measure of both**

- **Why not accuracy?**

Precision-Recall Trade-off



How good are the retrieved docs?

- **Precision/Recall @N**
  - Focus on the N-first results
  - **Why?**
- **MAP (Mean Average Precision):**
  - Average of the **precision** for the **top k documents,** each time a relevant doc is retrieved

$$\text{avg}(P@k \ / \ \text{iff} \ k \ is \ relevant)$$

  - Avoids use of fixed recall levels
- **NDCG (Normalized Cumulative Discounted Gain)**
  - Take into account the ranking of the relevant documents
  - Reward you more for getting rank 1 right than for getting rank 10 right

More practical measures

- **Users finds what they want**
  - **eCommerce**: searching --> buying
  - **Search Engine**: user returns to the engine
  - **Enterprise search:** time spent by employees when looking for information
- **Indexing speed**
- **Search speed**
- **A/B testing**

Improve IR effectiveness

- **Integrate User in the process (Relevant Feedback)**
  - **User annotates query results**

Improve IR effectiveness

- **Integrate User in the process (Relevant Feedback)**
  - **User annotates query results**
  - **User cooperation is rare --> Pseudo RF**
    - **Use the top-ranked documents as user annotation (as if they are relevant)**
    - **~ 10% Improvement**

Improve IR effectiveness

- **Integrate User in the process (Relevant Feedback)**
- **Query expansion**
  - **Add terms related to the query terms**
    - **Thesaurus-based (e.g. Wordnet)**
    - **Corpus-based**
      - Mutual information find terms that co-occur frequently
      - Word2vec
  - **Global vs. Local context**
    - Subset of retrieved documents to find term relationships

Improve IR effectiveness

- **Integrate User in the process (Relevant Feedback)**
- **Query expansion**
- **Importance of the Retrieved Results**
  - **Google's Page Rank**
    - More and more-quality input links = Important the website
    - More important websites are likely to receive more links from other websites

Standard benchmarks

- **TREC** - National Institute of Standards and Technology (NIST) has run a large IR test bed for many years
  - https://trec.nist.gov/
- **Reuters** collections
  - https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection
- **CLEF datasets**
  - European languages and cross-language IR.
  - http://www.clef-initiat:ve.eu/dataset/test-collection