Probabilistic Topic Models
- TBFY -

Carlos Badenes-Olmedo
Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

✉ cbadenes@fi.upm.es

@carbadol

oeg-upm.net

github.com/cbadenes

What is it?

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

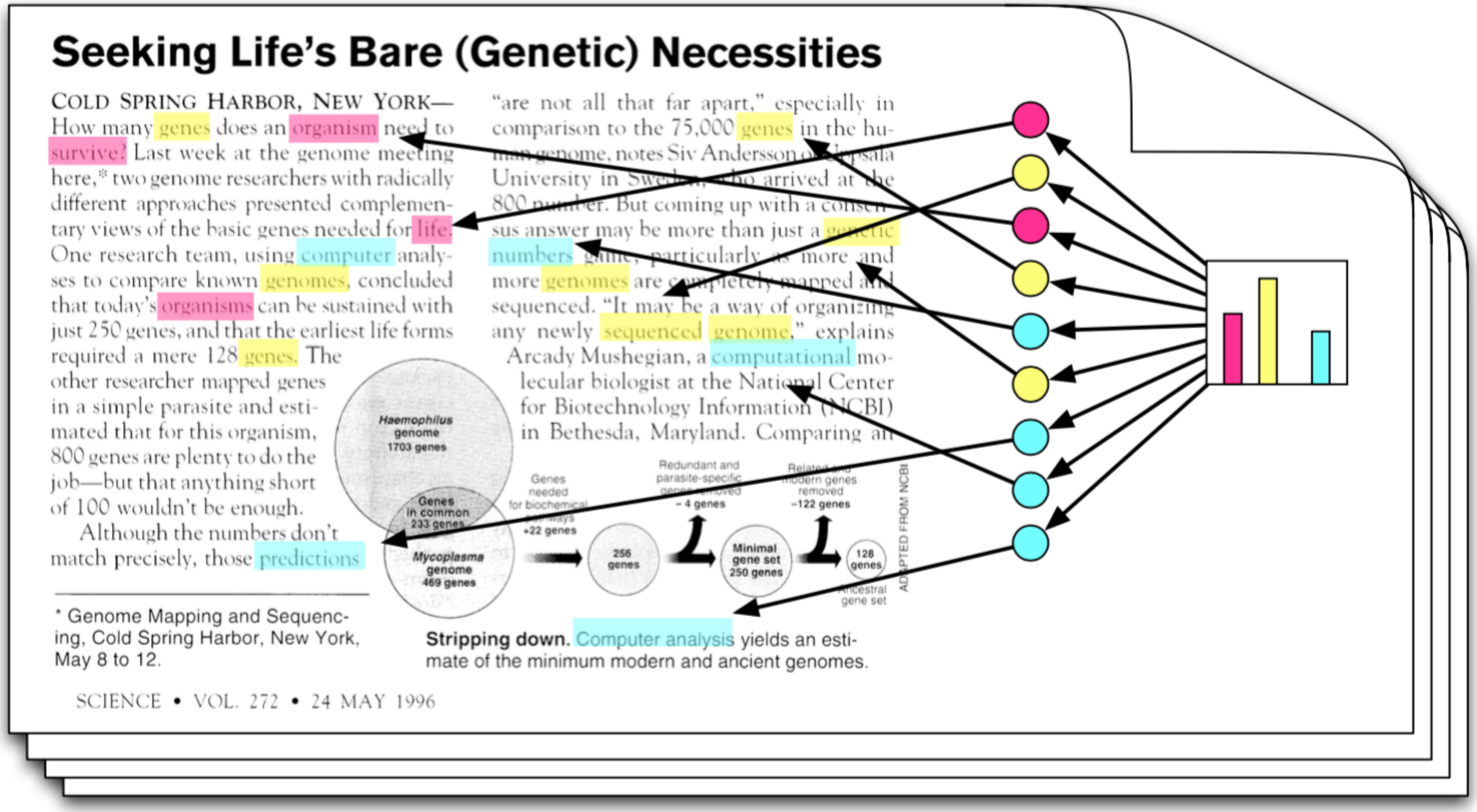
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

[*] Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

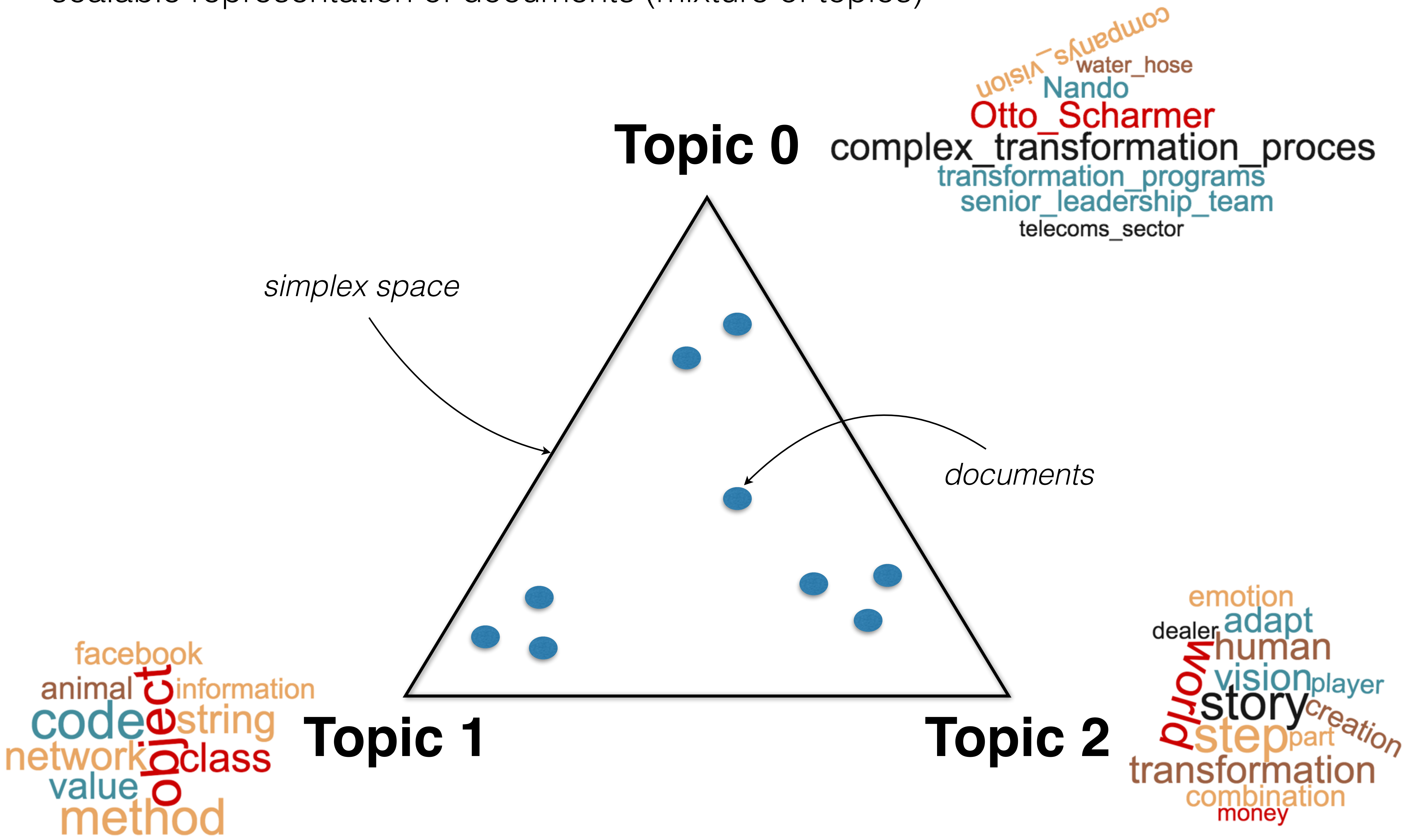
SCIENCE • VOL. 272 • 24 MAY 1996

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Blei, D. M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4–5), 993–1022.

What does it offer?

- corpus summarised by topics (ranked list of words)
- scalable representation of documents (mixture of topics)



How could it be used?

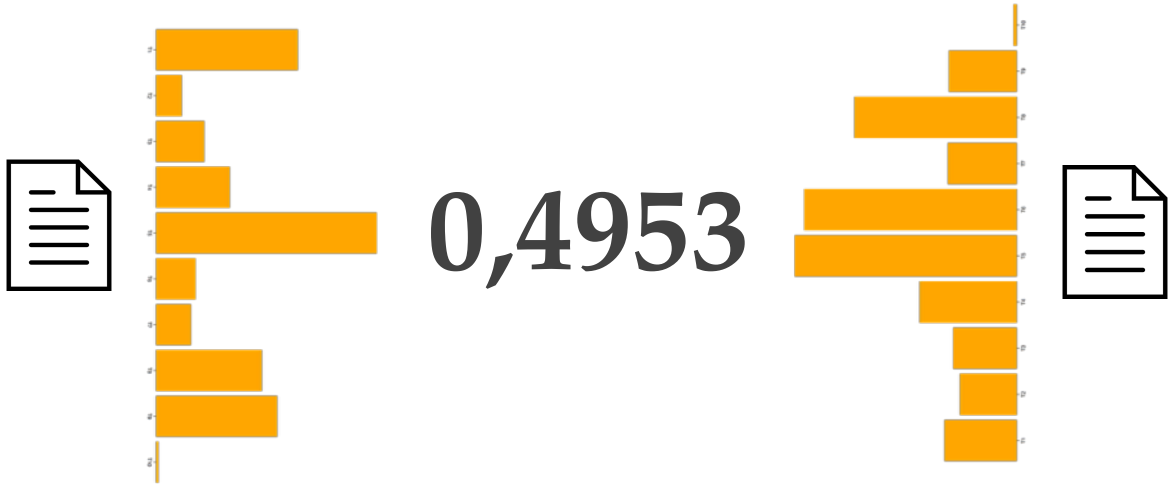
- ***scalable document repository***

  - *topics as annotations to **organize**, **summarise** and **search** documents*
  - ***explore** it in a way that you can index of ideas contained in them*
  - ***browse** it in a way that you can find documents are about the same kinds of ideas*

How could it be used?

- *scalable document repository*
- ***semantic similarity***

- *documents as vectors of topic distributions*
- *similarity metric based on* ***Jensen-Shannon Divergence*** *(JSD)*
- *documents and texts*

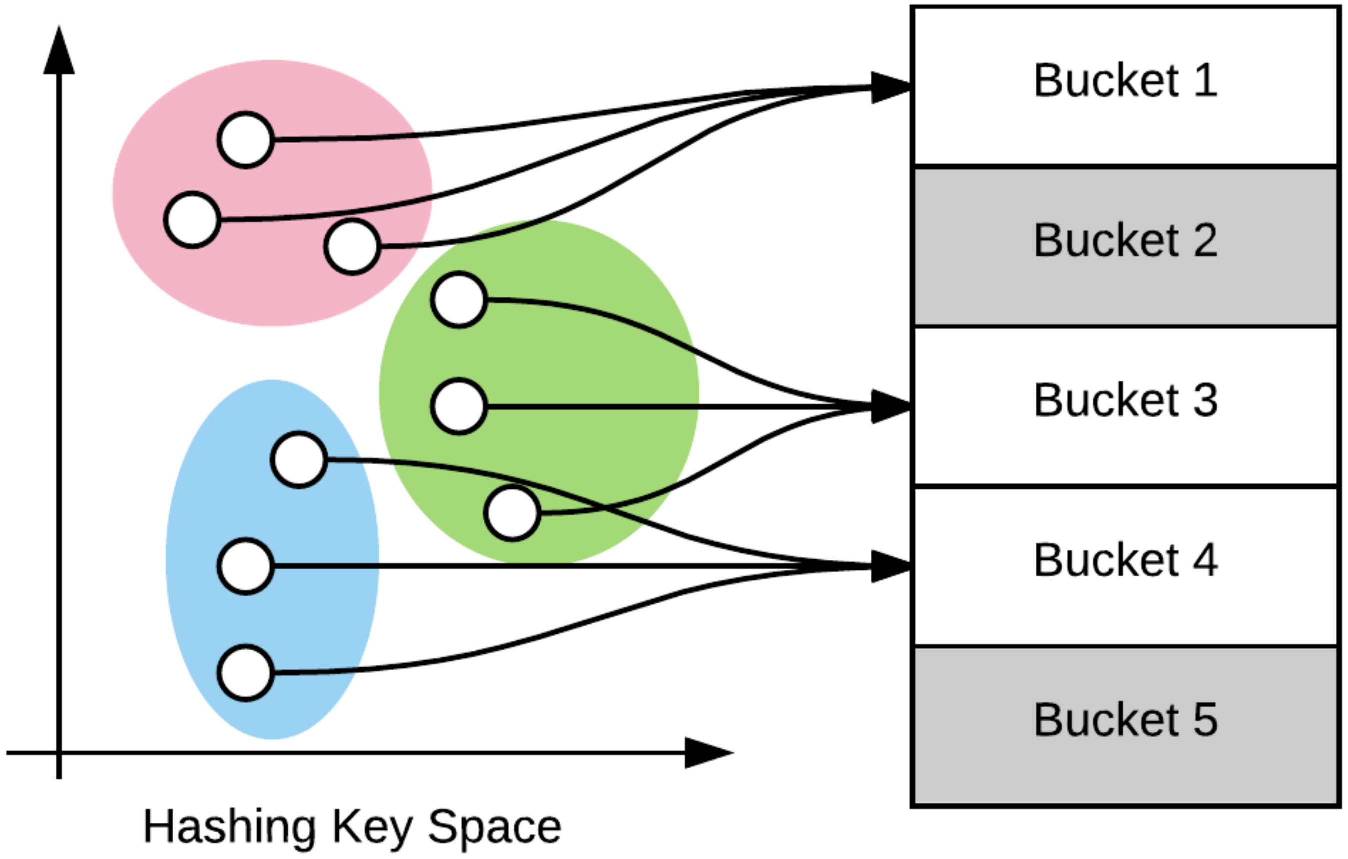


0,4953

How could it be used?

- *scalable document repository*
- *semantic similarity*
- ***alerts***

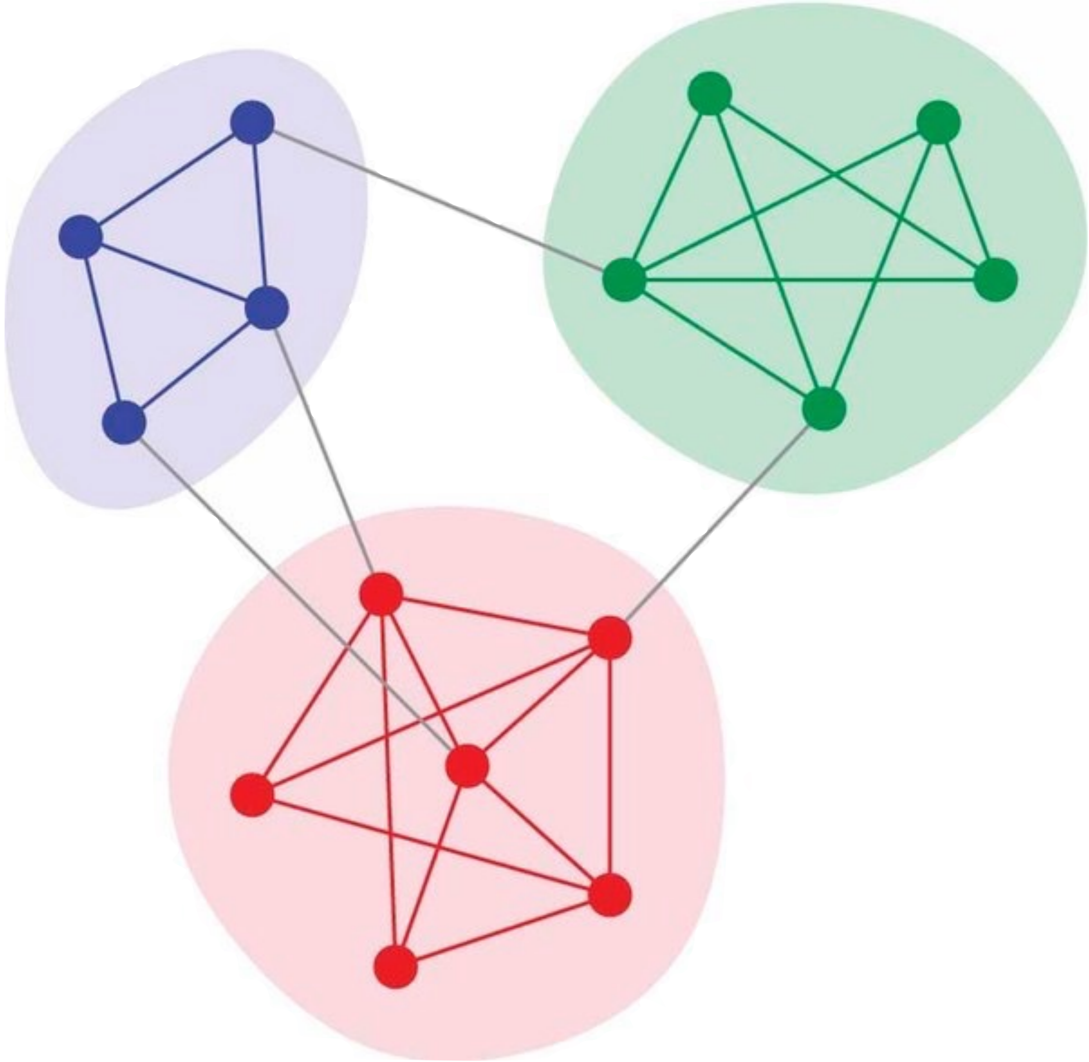
  - *hashing based on topic distribution*
  - *non-static approach*
  - *categorical similarities*
  - *duplicate detection*



How could it be used?

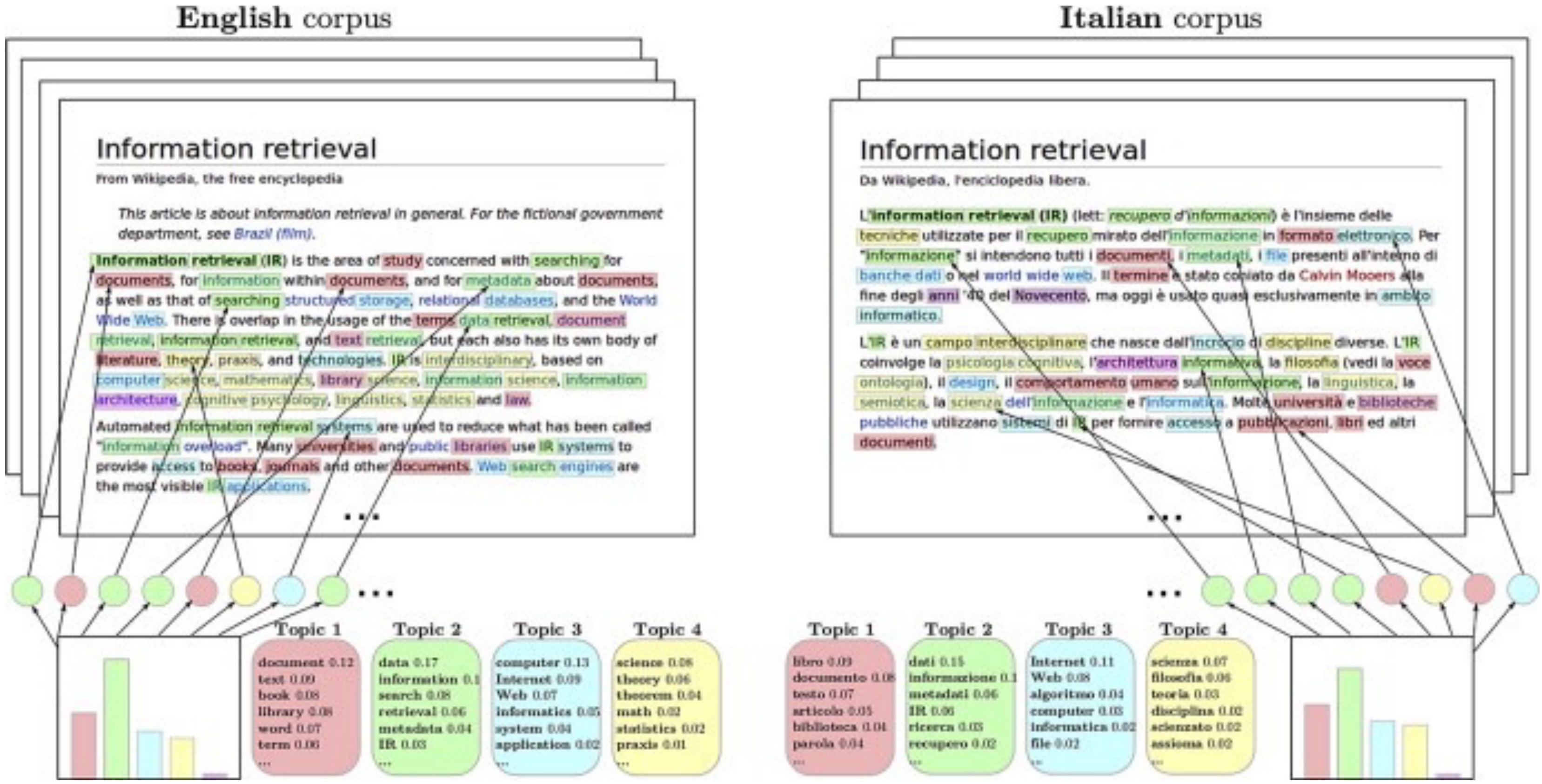
- *scalable document repository*
- *semantic similarity*
- *alerts*
- ***semantic exploration***

  - *topics linked to entities from external Knowledge Graph*
  - *multi-valued classification*
  - *hierarchical-classification*
  - *profiling*



How could it be used?

- *scalable document repository*
- *semantic similarity*
- *alerts*
- *semantic exploration*
- ***multi-lingual recommendation***
  - *topics from different languages (models)*
  - *automatic alignment*



Information retrieval

From Wikipedia, the free encyclopedia

This article is about information retrieval in general. For the fictional government department, see Brazil (film).

Information retrieval (**IR**) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and law.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

...

Information retrieval

Da Wikipedia, l'enciclopedia libera.

L'**information retrieval (IR)** (lett: recupero d'informazioni) è l'insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico. Per "informazione" si intendono tutti i documenti, i metadati, i file presenti all'interno di banche dati o nel world wide web. Il termine è stato coniato da Calvin Mooers alla fine degli anni '40 del Novecento, ma oggi è usato quasi esclusivamente in ambito informatico.

L'IR è un campo interdisciplinare che nasce dall'incrocio di discipline diverse. L'IR coinvolge la psicologia cognitiva, l'architettura informativa, la filosofia (vedi la voce ontologia), il design, il comportamento umano sull'informazione, la linguistica, la semiotica, la scienza dell'informazione e l'informatica. Molte università e biblioteche pubbliche utilizzano sistemi di IR per fornire accesso a pubblicazioni, libri ed altri documenti.

...

Vulic, Ivan, Wim De Smet, Jie Tang and Marie-Francine Moens. "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications." *Inf. Process. Manage.* 51 (2015): 111-147.