Information retrieval

Evaluation of retrieval systems and learn to rank

Clovis Galiez

Laboratoire Jean Kuntzmann, Statistiques pour les sciences du Vivant et de l'Homme

November 14, 2018

Objectives of the course

- Acquire a culture in information retrieval
- Master the basics concepts allowing to understand:
  - what is at stake in novel IR methods
  - what are the technical limits

This will allow you to have the basics tools to analyze current limitations or lacks, and imagine novel solutions.

Today's outline

- Guided correction of the hands-on
- Evaluation of IR systems
- Learning to search
- Wrap-up

Evaluation of IR systems

How to evaluate the performances of an IR system?

Evaluation of IR systems

How to evaluate the performances of an IR system?

Need a gold standard  and indicators  .

Evaluation of IR systems

How to evaluate the performances of an IR system?

Need a gold standard  and indicators  .

What is a gold standard?

Evaluation of IR systems

How to evaluate the performances of an IR system?

Need a gold standard  and indicators  .

What is a gold standard?

IR: Answering a query by extracting **relevant information** from **a collection of documents**.

Evaluation of IR systems

How to evaluate the performances of an IR system?

Need a gold standard  and indicators  .

What is a gold standard?

IR: Answering a query by extracting **relevant information** from **a collection of documents**.

Can be seen as a partial function $g : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ associating to a couple query-document its quantification of *correctness*, *relevance* or *truth*.
What can be an indicator?

Evaluation of IR systems

How to evaluate the performances of an IR system?

Need a gold standard  and indicators  .

What is a gold standard?

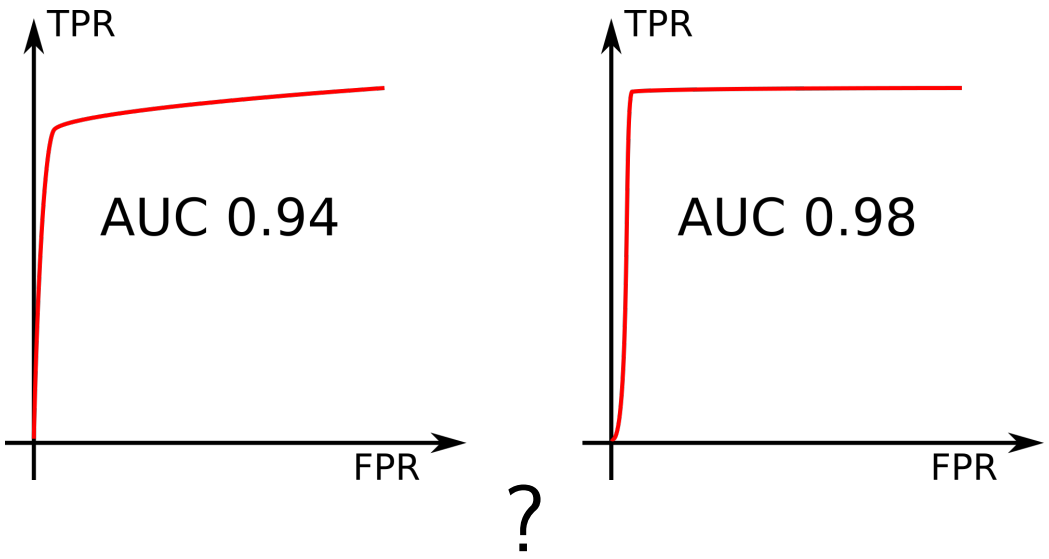
IR: Answering a query by extracting **relevant information** from **a collection of documents**.

Can be seen as a partial function $g : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ associating to a couple query-document its quantification of *correctness*, *relevance* or *truth*.
What can be an indicator?
We distinguish the special case of $\text{dom}(g) = \{0, 1\}$.

Indicators for binary gold standards[1]

Standard indicators: FP, FP, FN, TN.

Use the right summary indicator (e.g. AUC-ROC/AUC-ROC5).



[1] To be used for assessing correctness for instance.

Indicators for rank gold standards[2]

Ranking functions $f_1, f_2 : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$. How can we say that f_1 is better than f_2?

[2] To be used for evaluation of relevance ranking for instance.

Indicators for rank gold standards[2]

Ranking functions $f_1, f_2 : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$. How can we say that f_1 is better than f_2?
Given a gold standard $g : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$, one can measure the cumulative gain:

$$\mathsf{CG}_n(q, f) = \sum_{k:\mathsf{ord}_n(q,f)} g(q, k)$$

where $\mathsf{ord}_n(q, f)$ are the n first elements of $\mathcal{D}$ when sorting by $f(q, \_)$.

[2] To be used for evaluation of relevance ranking for instance.

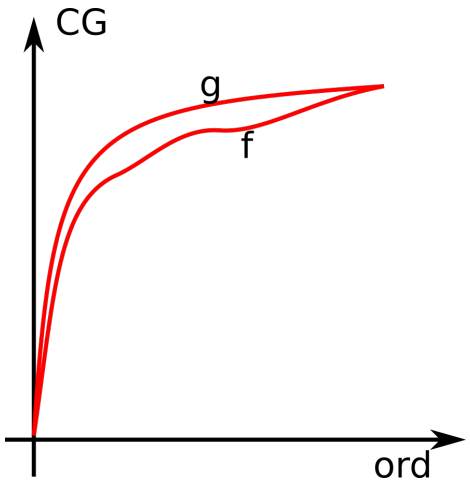
Indicators for rank gold standards[2]

Ranking functions $f_1, f_2 : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$. How can we say that f_1 is better than f_2?
Given a gold standard $g : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$, one can measure the cumulative gain:

$$\mathsf{CG}_n(q, f) = \sum_{k:\mathsf{ord}_n(q,f)} g(q, k)$$

where $\mathsf{ord}_n(q, f)$ are the n first elements of $\mathcal{D}$ when sorting by $f(q, \_)$.
The normalized cumulative gain is:

$$\mathsf{NCG}_n(q, f) = \frac{\mathsf{CG}_n(q,f)}{\mathsf{CG}_n(q,g)}$$



[2] To be used for evaluation of relevance ranking for instance.

Indicators for ranking, stressing first results

In the same spirit of the difference AUC/AUC5, one can stress more the first results, by weighting the relevance by a *discount*[3] function:

$$\mathsf{DCG}(q, f) = \sum_{k:\mathsf{ord}_{N_q}(q,f)} \frac{g(q,k)}{\log(i+1)}$$

where N_Q is $\mathsf{card}\{d|(q, d) \in \mathsf{dom}(g)\}$ and i is the index in the summation.

The NDCG is defined as:

$$\mathsf{NDCG}(q, f) = \frac{\mathsf{DCG}_{N_Q}(q,f)}{\mathsf{DCG}_{N_Q}(q,g)}$$

[3] One can choose different discount functions, but $\frac{1}{\log i}$ has nice theoretical foundations [Wang et al. JMLR 13] and is good in practice.

Comparing ranking strategies

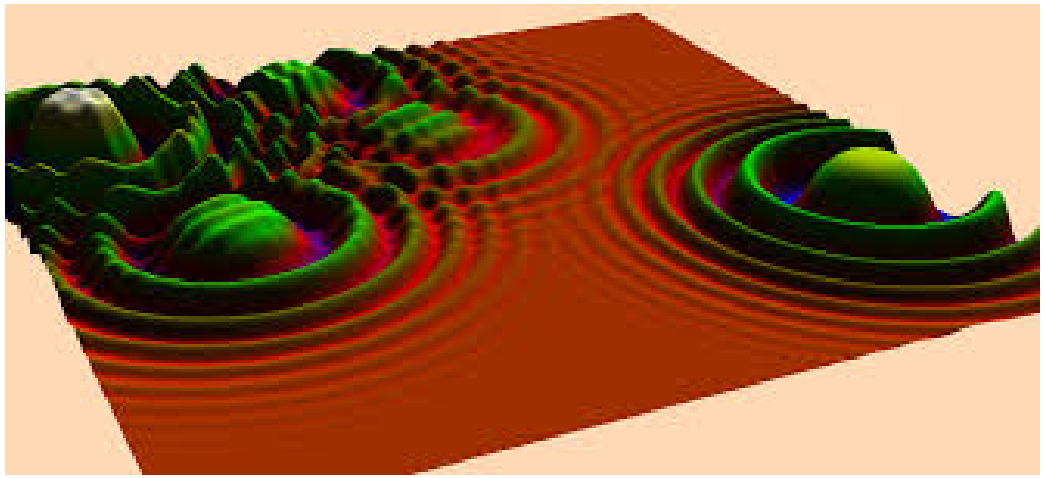
Ranking functions $f_1, f_2 : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}_+^*$. How can we say that f_1 is better than f_2?
Can use the expected NDCG(q, f_i) summary statistic:

$$\rho_i = \frac{1}{Q} \sum_q \text{NDCG}(q, f_i)$$

As usual, averaging can hide bad performance when the gold standard is dominated by high performance on many similar queries.

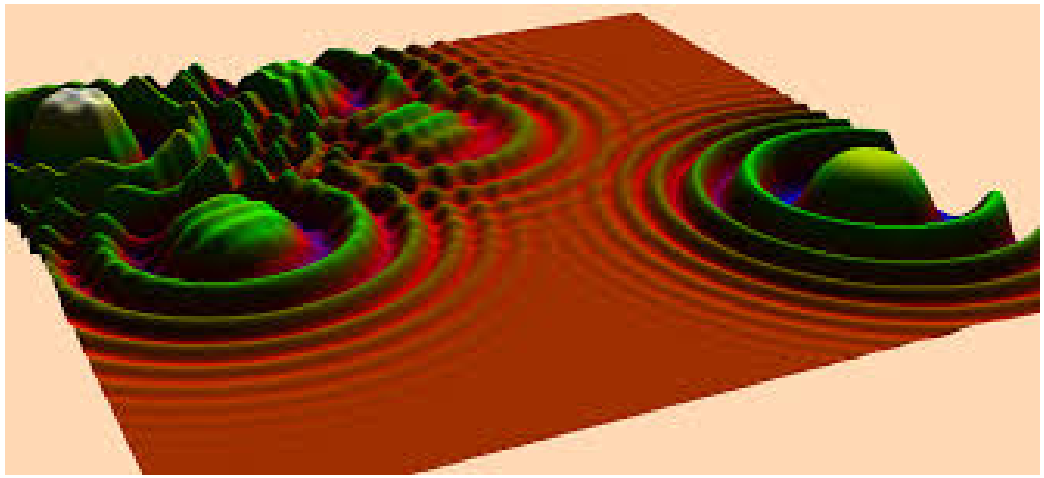
Learning the parameters

Having a gold standard not only allows evaluation, but also optimization of the parameters.



Learning the parameters

Having a gold standard not only allows evaluation, but also optimization of the parameters.



What parameters are we talking about?

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)
- Latent semantics (e.g. size of the space)

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)
- Latent semantics (e.g. size of the space)

At the ranking level:

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)
- Latent semantics (e.g. size of the space)

At the ranking level:

- Source vector choice

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)
- Latent semantics (e.g. size of the space)

At the ranking level:

- Source vector choice
- Prior distribution on links in a web page (e.g. important links)

Some parameters to tune

At the semantics level:

- Tokenization (parameters in *phrase as tokens* cf. patent in 1st tutorial session)
- Stemming
  - not enough: mother - maternal
  - too much: police - policy
- Important text scoring
- The scoring system (cosine similarity, Euclidena distance, etc.)
- Latent semantics (e.g. size of the space)

At the ranking level:

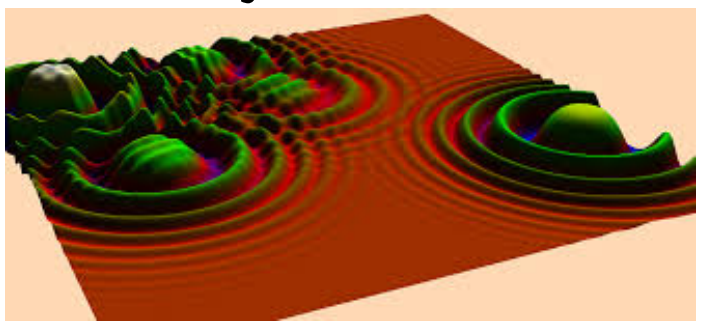
- Source vector choice
- Prior distribution on links in a web page (e.g. important links)

But also...

...the **trade-off** between the semantic scores (e.g. tf-idf vector model, text importance score) and the authority ranking score.

Optimize the objective function by tuning the parameters

$$\text{obj} : \mathbb{R}^p \to \mathbb{R}$$



This is an optimization problem:

$$\arg\max_{\mathbb{R}^p} \text{obj}$$

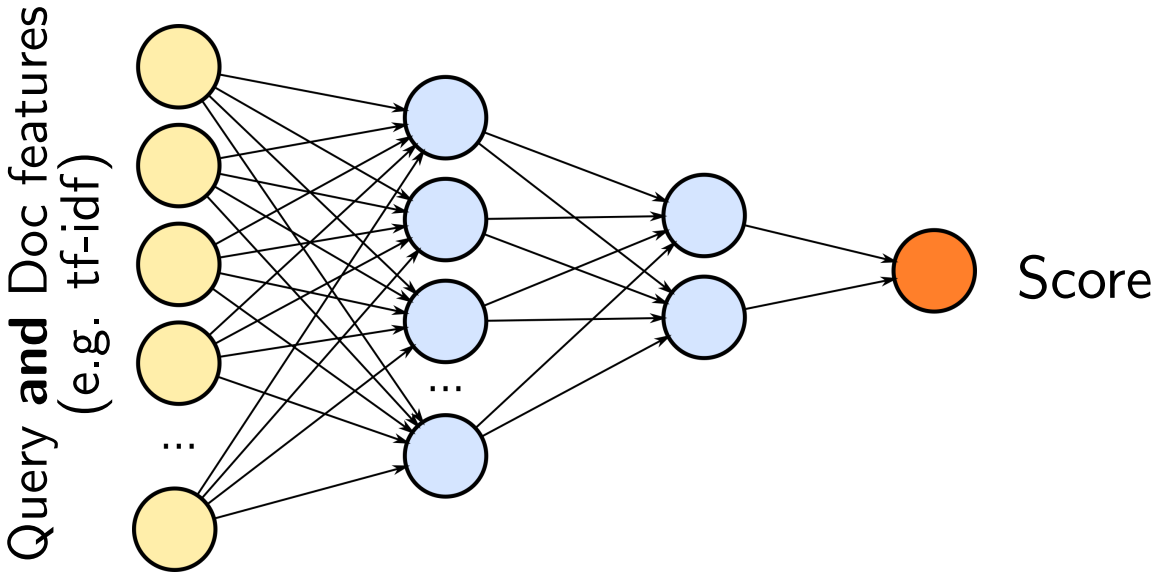
One can therefore use optimization strategies to maximize performance indicators by tuning p parameters.

Why not going further?

Why learning only few parameters?

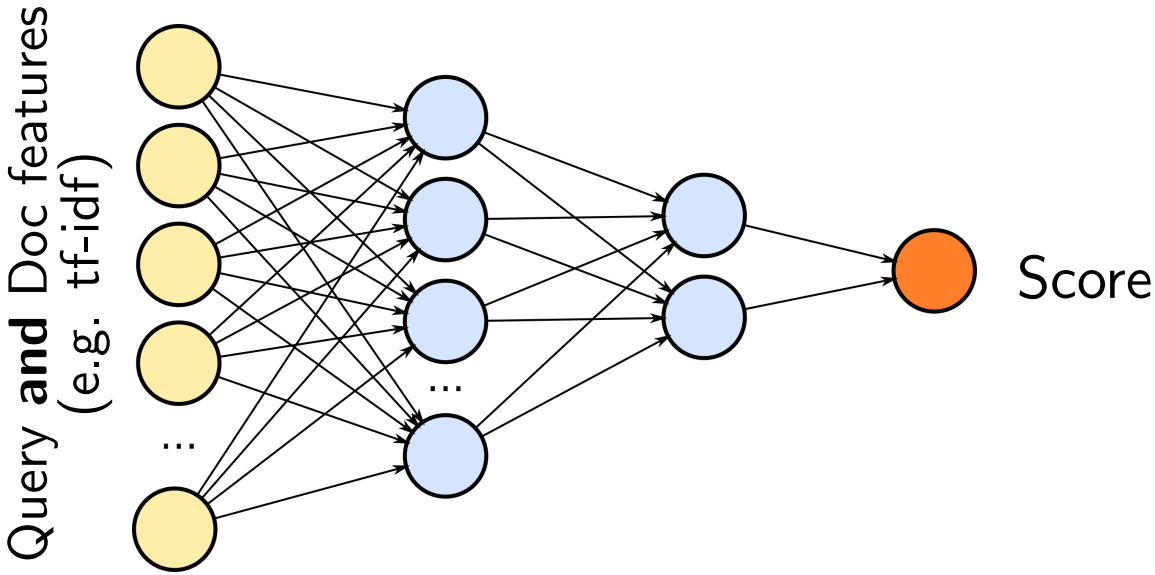
Why not going further?

Why learning only few parameters? Why not learning directly:



Why not going further?

Why learning only few parameters? Why not learning directly:

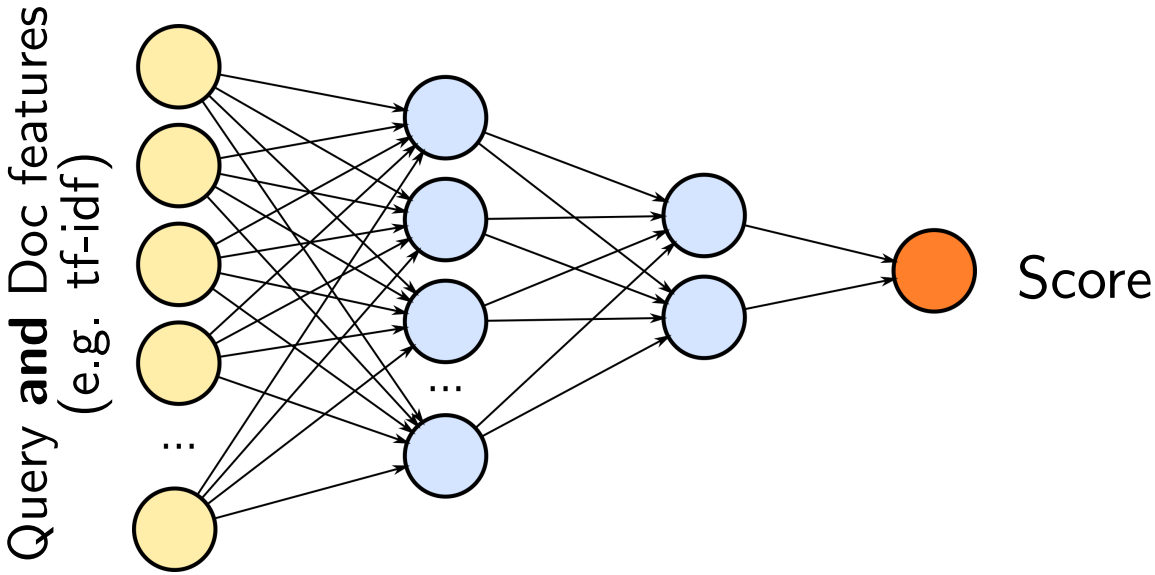


Issue

Curse of dimensionality, training set limitation, overtraining.

Why not going further?

Why learning only few parameters? Why not learning directly:



Issue

Curse of dimensionality, training set limitation, overtraining.

Way out

Decrease the number of features or/and **expand the training set**.

Increasing the training set

If you own a popular search engine, imagine a simple way of increasing your training set.

Increasing the training set

If you own a popular search engine, imagine a simple way of increasing your training set.

Click-through strategy: $g : (q, d) \mapsto$ nb of clicks on d

Reducing the dimensionality

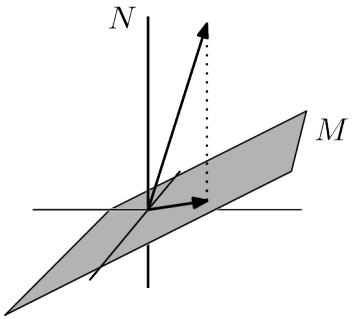
Remember

What dimensionality reduction technique have we seen before?

Reducing the dimensionality

Remember

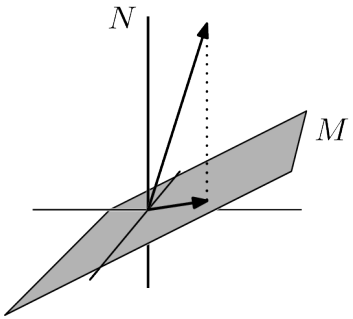
What dimensionality reduction technique have we seen before?



Reducing the dimensionality

Remember

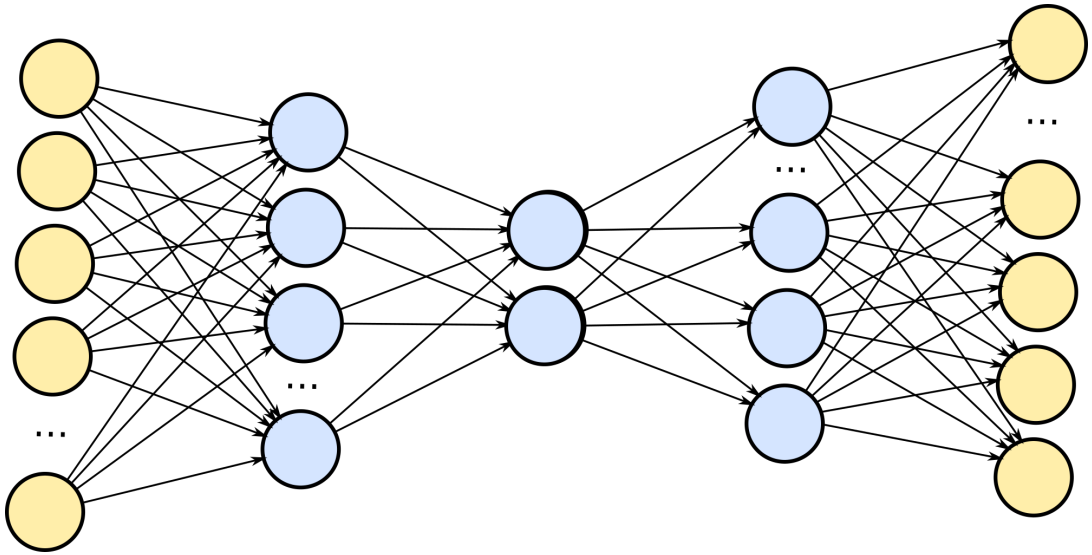
What dimensionality reduction technique have we seen before?



Note that for reducing the dimension we do not lack of data! The Internet is big enough :)

$\rightarrow$ More powerful techniques such as autoencoders.

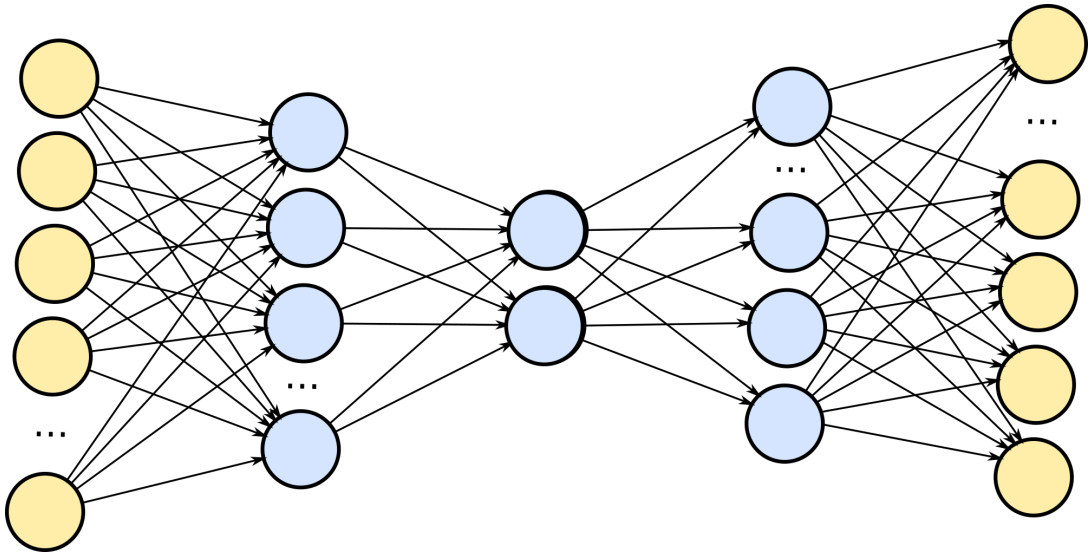
Autoencoders



If you were an autoencoder...

If you were an autoencoder with **few** intermediate neurons, how would you encode a document?

Autoencoders



If you were an autoencoder...

If you were an autoencoder with **few** intermediate neurons, how would you encode a document?

If one wants to minimize the loss between the input and the output, a neuron of the intermediate layer represents a topic, or a concept.

Wrap-up I

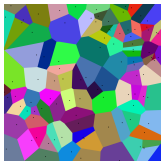
WIKIPEDIA
The Free Encyclopedia

Main page
Contents

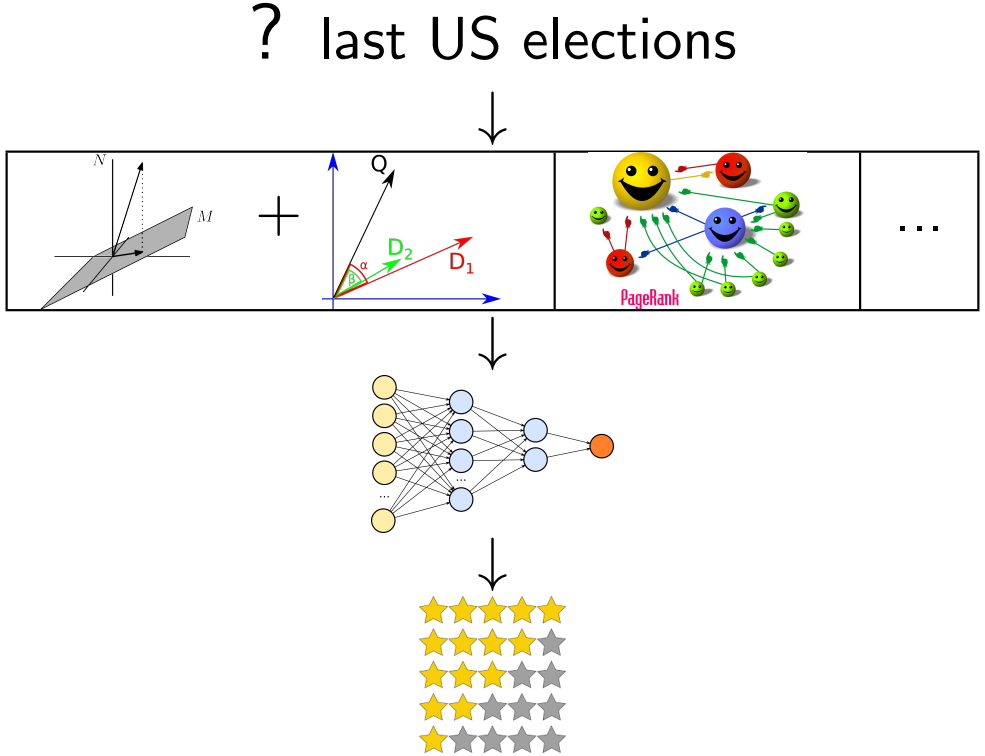
Information retrieval

From Wikipedia, the free encyclopedia

Information retrieval (**IR**) is the activity of obtaining information system resources relevant to an information need from a collection of information resources. Searches can be based on full-text



Wrap-up II



Hope you enjoyed.

Find the material on `http://clovisg.github.io`

Extras