



# A comparison between Latent Semantic Analysis and Correspondence Analysis

Julie Séguéla, Gilbert Saporta

CNAM, Cedric Lab
Multiposting.fr

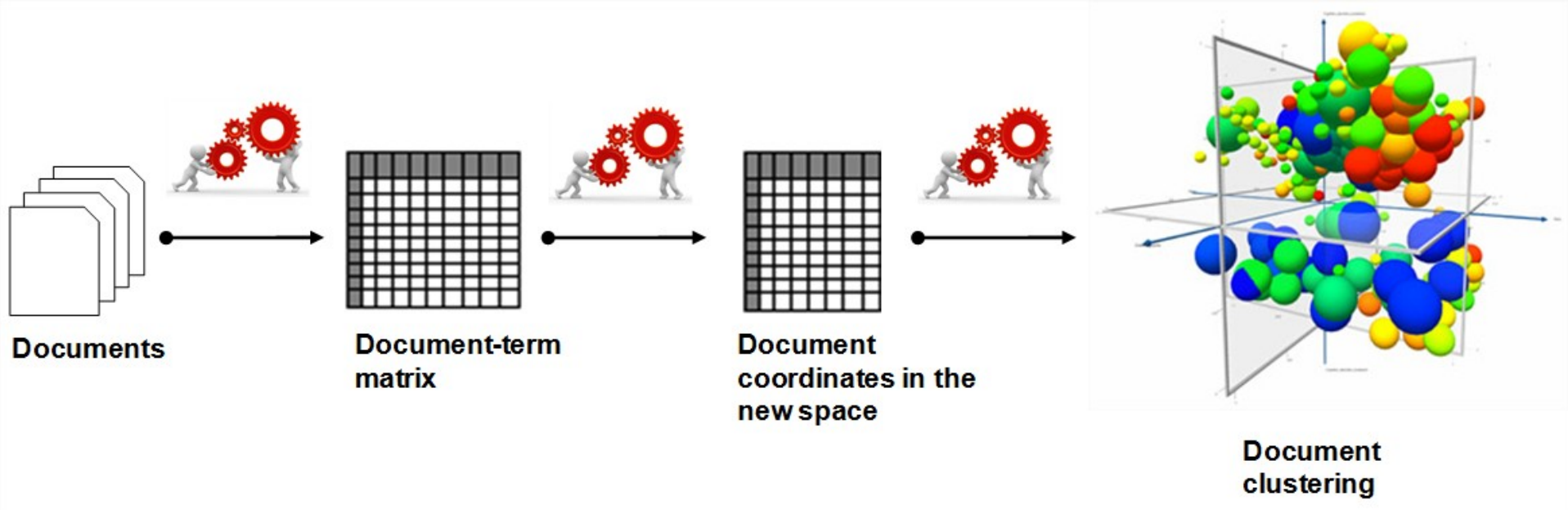
February 9th 2011 - CARME

# Plan

# Plan

# Context

Text representation for categorization task



# Objectives

- Comparison of several text representation techniques through theory and application
- In particular, comparison between a statistical technique : Correspondence Analysis, and an information retrieval (IR) oriented method : Latent Semantic Analysis
- Can we find a representation better than the others to perform document clustering ?
- Discussion about advantages and weaknesses of each method given the type of corpus studied

# Plan

# Uses of LSA

- LSA was patented in 1988 (US Patent 4,839,853) by Deerwester, Dumais, Furnas, Harshman, Landauer, Lochbaum and Streeter.
- Find semantic relations between terms
- Helps to overcome synonymy and polysemy problems
- Dimensionality reduction (from several thousands of features to 40-400 dimensions)

## Applications

- Document clustering and document classification
- Matching queries to documents of similar topic meaning (information retrieval)
- Text summarization
- Essay scoring
- ...

# LSA theory

## How to obtain document coordinates ?

1) Document-Term matrix

$$T = \begin{pmatrix} & \vdots & \\ \dots & f_{ij} & \dots \\ & \vdots & \end{pmatrix}$$

2) Weighting

$$T_W = \begin{pmatrix} & \vdots & \\ \dots & l_{ij}(f_{ij}) \cdot g_j(f_{ij}) & \dots \\ & \vdots & \end{pmatrix}$$

3) SVD

$$T_W = U\Sigma V'$$

4) Document coordinates in the latent semantic space :

$$C = U_k \Sigma_k$$

- We need to find the optimal dimensionality for final representation

# Common weighting functions

## Local weighting

**Term frequency**

$l_{ij}(f_{ij}) = f_{ij}$

**Binary**

$l_{ij}(f_{ij}) = 1$ if term $j$ occurs in document $i$, else 0

**Logarithm**

$l_{ij}(f_{ij}) = log(f_{ij} + 1)$

## Global weighting

**Normalisation**

$g_j(f_{ij}) = \frac{1}{\sqrt{\sum_i f_{ij}^2}}$

**IDF** (Inverse Document Frequency)

$g_j(f_{ij}) = 1 + log(\frac{n}{n_j})$

$n$ : number of documents

$n_j$ : number of documents in which term $j$ occurs

**Entropy**

$g_j(f_{ij}) = 1 - \sum_i \frac{\frac{f_{ij}}{f_{.j}} log(\frac{f_{ij}}{f_{.j}})}{log(n)}$

# LSA vs CA

## Latent Semantic Analysis

1) $T = [f_{ij}]_{i,j}$

2) $T_W = [l_{ij}(f_{ij}) \cdot g_j(f_{ij})]_{i,j}$

3) $T_W = U\Sigma V'$

4) $C = U_k \Sigma_k$

## Correspondence Analysis

1) $T = [f_{ij}]_{i,j}$

2) $T_W = \left[\frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}}\right]_{i,j}$

3) $T_W = U\Sigma V'$

3') $\tilde{U} = diag(\sqrt{\frac{f_{..}}{f_{i.}}})U$

4) $C = \tilde{U}_k \Sigma_k$

- CA : $l_{ij}(f_{ij}) = \frac{f_{ij}}{\sqrt{f_{i.}}}$ and $g_j(f_{ij}) = \frac{1}{\sqrt{f_{.j}}}$

# Plan

# Objectives

- Corpus of job offers
- Find the best representation method to assess "job similarity" between offers in a non-supervised framework
- Comparison of several representation techniques
- Discussion about the optimal number of dimensions to keep
- Comparison between two similarity measures

# Data

- Offers have been manually labeled by recruiters into 8 categories during the posting procedure
- Distribution among job categories :

| **Category** | **Freq.** | **%** | **Category** | **Freq.** | **%** |
|---|---|---|---|---|---|
| Sales/Business Development | 360 | 24 | Marketing/Product | 141 | 10 |
| R&D/Science | 69 | 5 | Production/Operations | 127 | 9 |
| Accounting/Finance | 338 | 23 | Human Resources | 138 | 9 |
| Logistics/Transportation | 118 | 8 | Information Systems | 192 | 13 |
| | | | **Total** | **1483** | **100** |

- We keep only the "title"+"mission description" parts ("firm description" and "profile searched" are excluded)

# Preprocessing of texts

- Lemmatisation and tagging
- Filtering according to grammatical category (we keep nouns, verbs and adjectives)
- Filtering terms occuring in less than 5 offers
- Vector space model ("bag of words")

# Several representations are compared

## Representation method

- LSA, weighting : Term Frequency
- LSA, weighting : TF-IDF
- LSA, weighting : Log Entropy
- CA

## Dissimilarity measure

- Euclidian distance between documents $i$ and $i'$
- 1 - cosine similarity between documents $i$ and $i'$

# Method of clustering

## Clustering steps

- Computing of dissimilarity matrix from document coordinates in the latent semantic space
- Hierarchical Ascendant Classification to obtain a 8 class partition
- Computing of class barycentric coordinates
- K-means clustering initialized from previous barycentric coordinates

# Measures of agreement between two partitions

$P_1$, $P_2$ : two partitions of $n$ objects with the same number of class $k$
$N = [n_{ij}]_{\substack{i=1,..,k \\ j=1,..,k}}$ : corresponding contingency table

## Rand index

$$R = \frac{2\sum_i \sum_j n_{ij}^2 - \sum_i n_{i.}^2 - \sum_j n_{.j}^2 + n^2}{n^2}, \qquad 0 \leq R \leq 1$$

- Rand index is based on the number of pairs of units which belong to the same clusters. It doesn't depend on cluster labeling.

# Measures of agreement between two partitions

- Cohen's Kappa and F-measure values are depending on clusters' labels. To overcome label switching, we are looking for their maximum values over all label allocations.

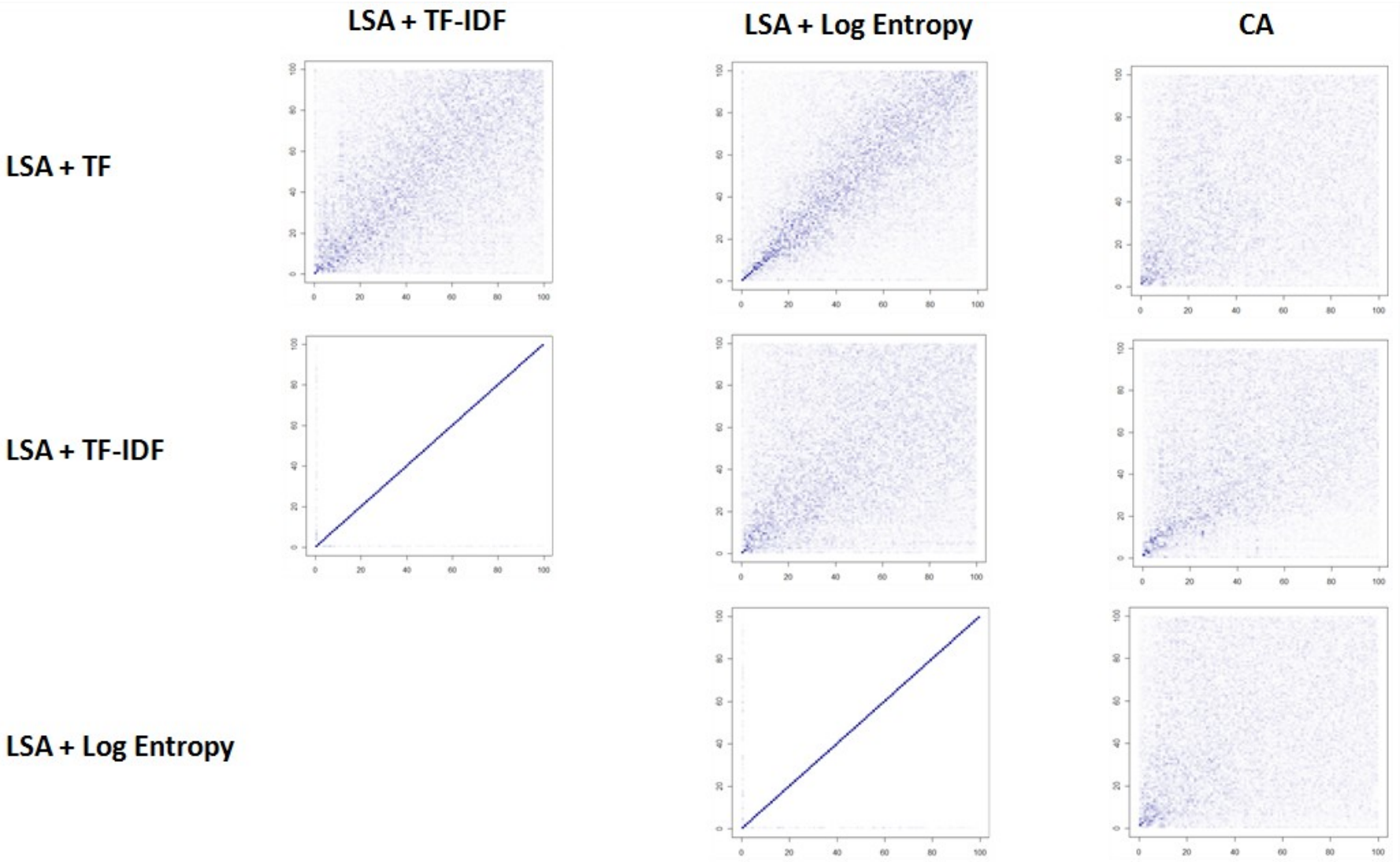
## Cohen's Kappa

$$\kappa^{opt} = max \left\{ \frac{\frac{1}{n}\sum_i n_{ii} - \frac{1}{n^2}\sum_i n_{i.}n_{.i}}{1 - \frac{1}{n^2}\sum_i n_{i.}n_{.i}} \right\}, \qquad -1 \leq \kappa \leq 1$$

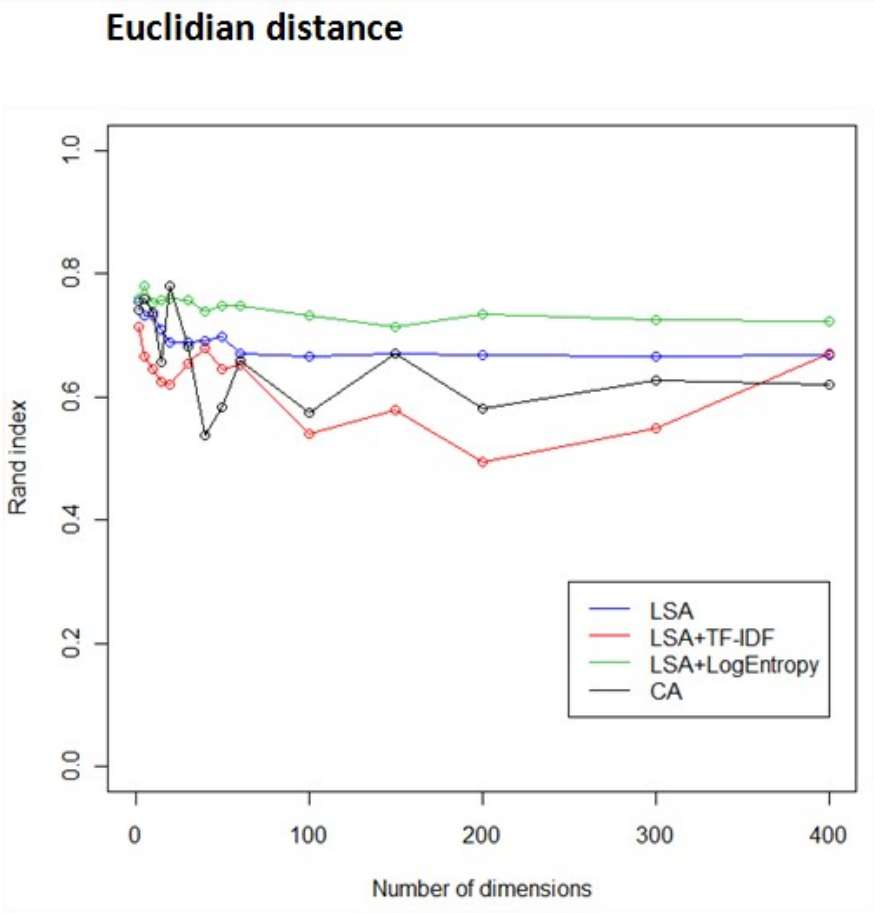
## $F$-measure

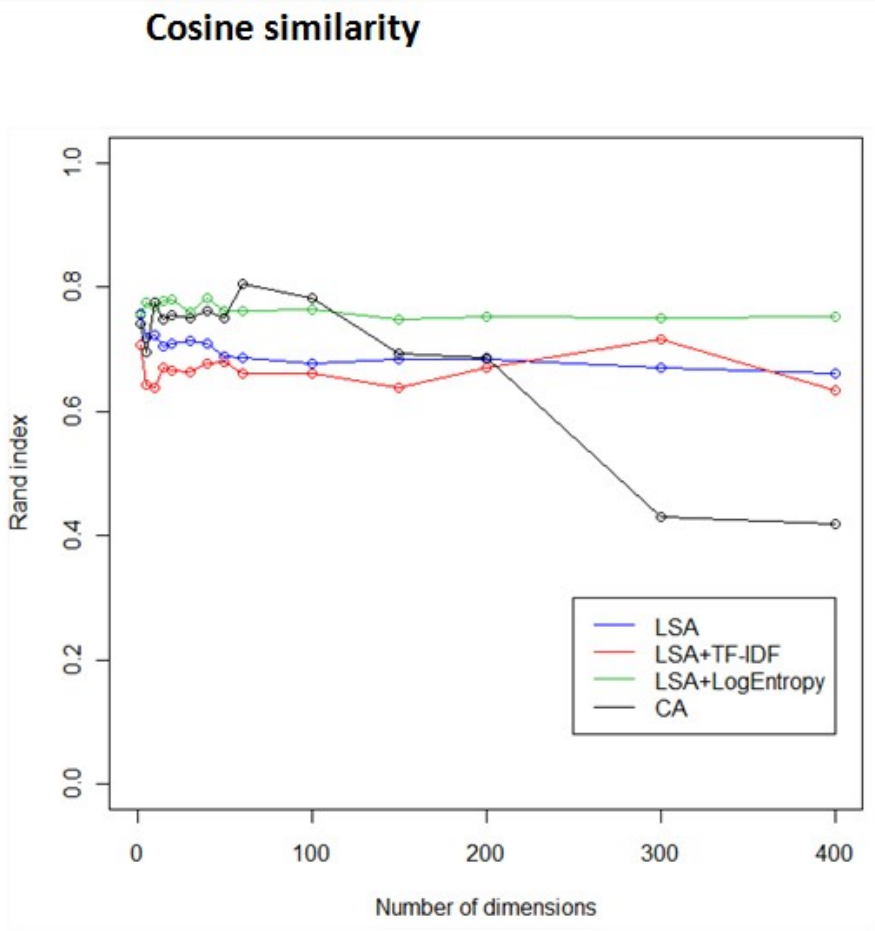
$$F^{opt} = max \left\{ \frac{2 \cdot \frac{1}{k}\sum_i \frac{n_{ii}}{n_{i.}} \cdot \frac{1}{k}\sum_i \frac{n_{ii}}{n_{.i}}}{\frac{1}{k}\sum_i \frac{n_{ii}}{n_{i.}} + \frac{1}{k}\sum_i \frac{n_{ii}}{n_{.i}}} \right\}, \qquad 0 \leq F \leq 1$$

# Correlation between coordinates issued from the different methods

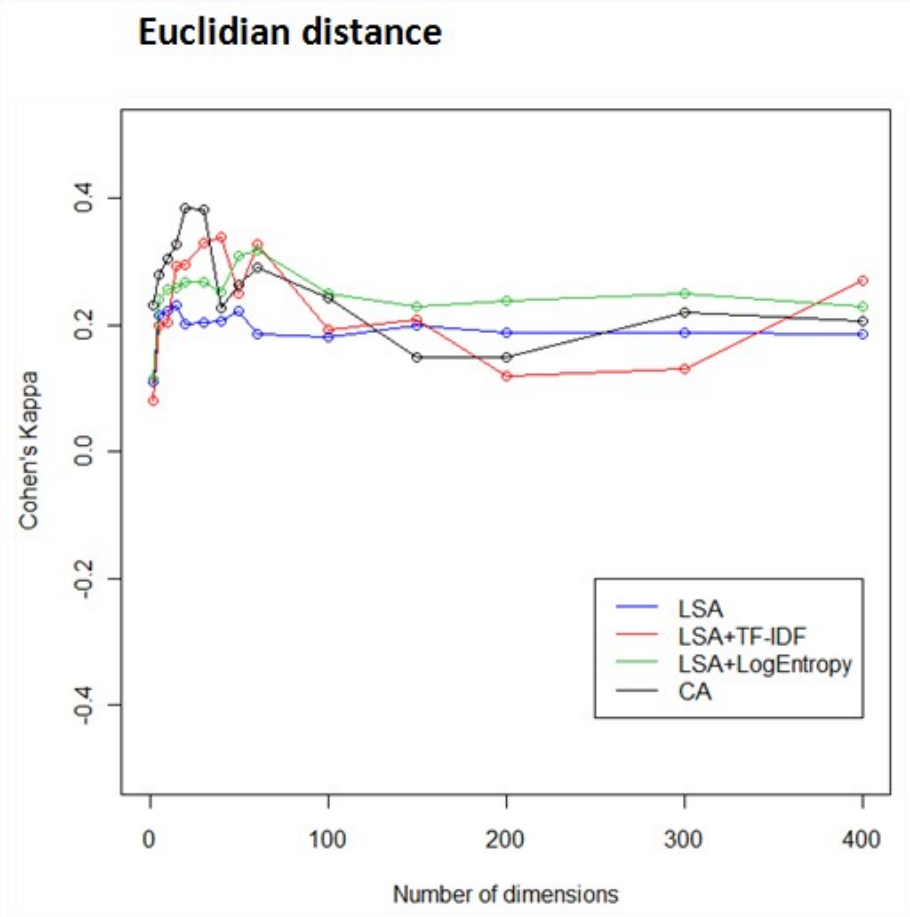


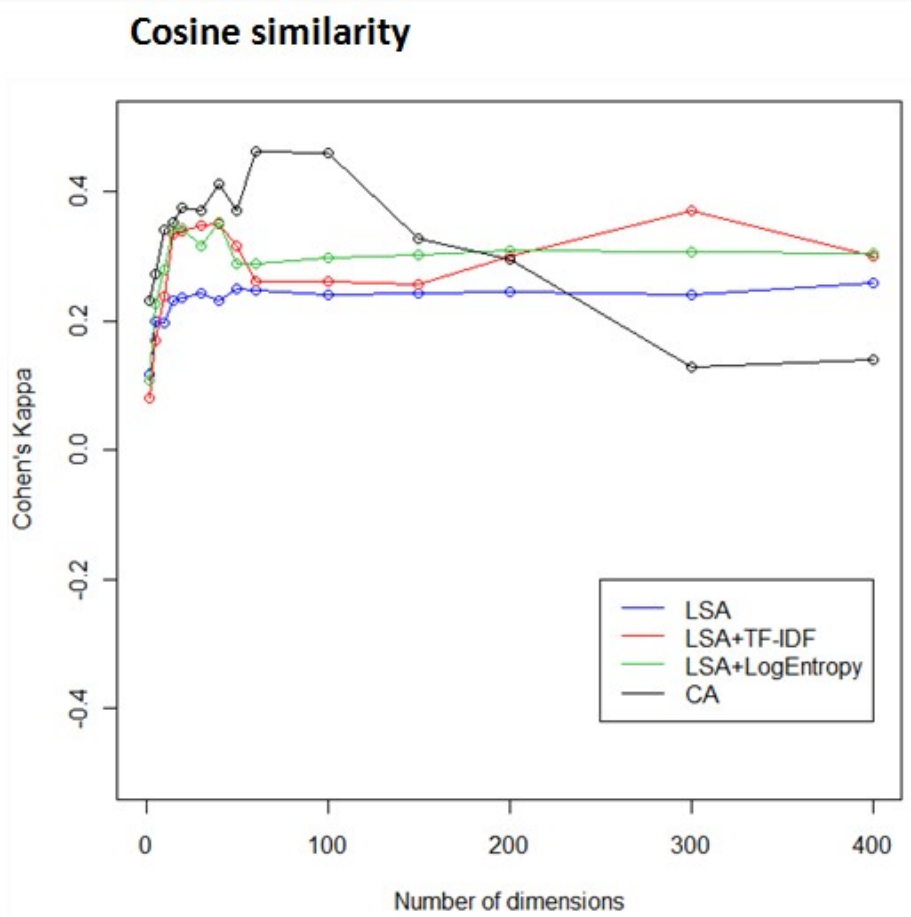
# Clustering quality according to the method and the number of dimensions : Rand index



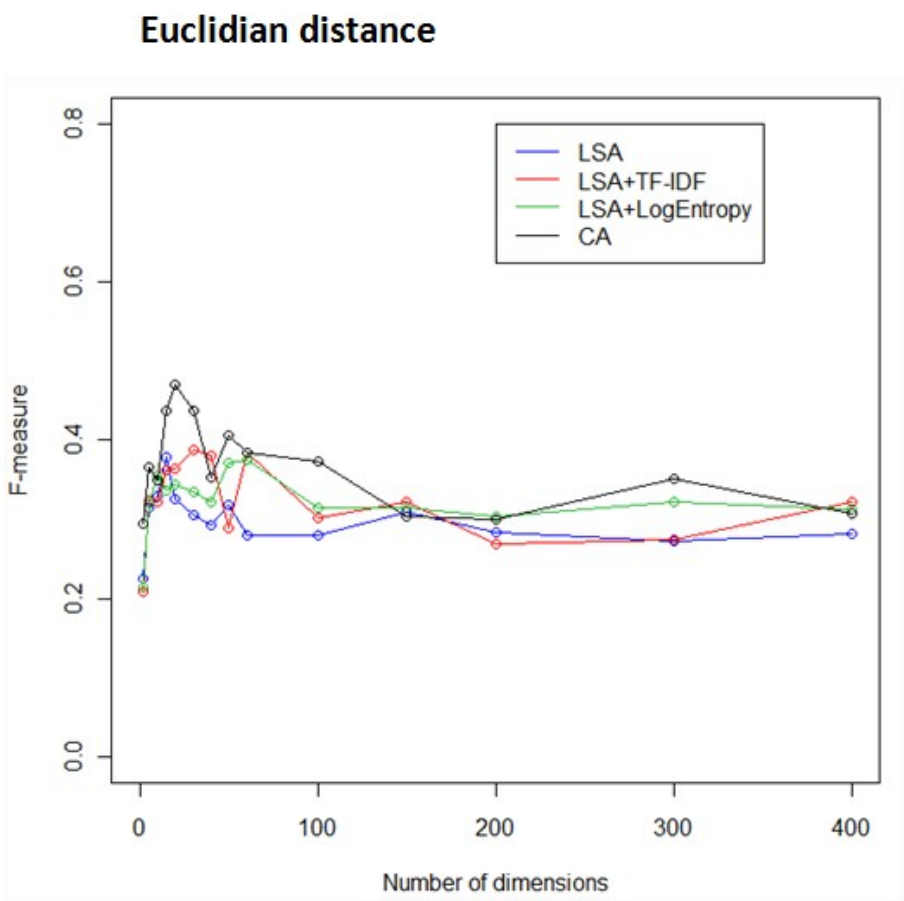



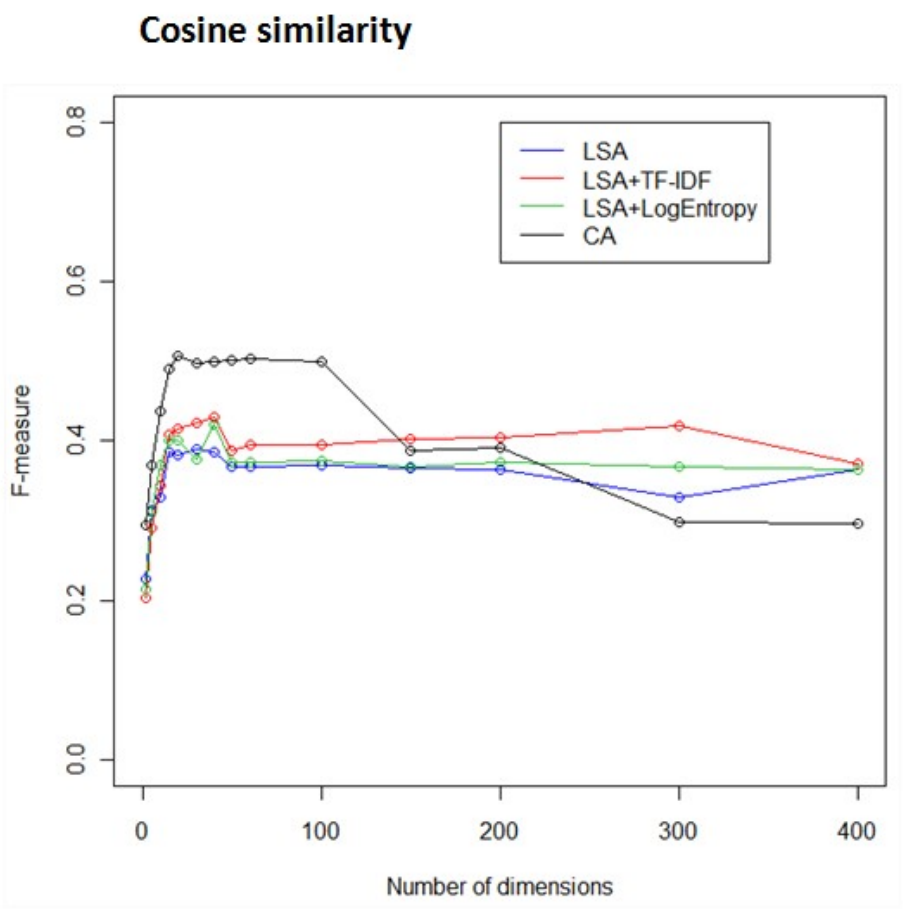
# Clustering quality according to the method and the number of dimensions : Cohen's Kappa



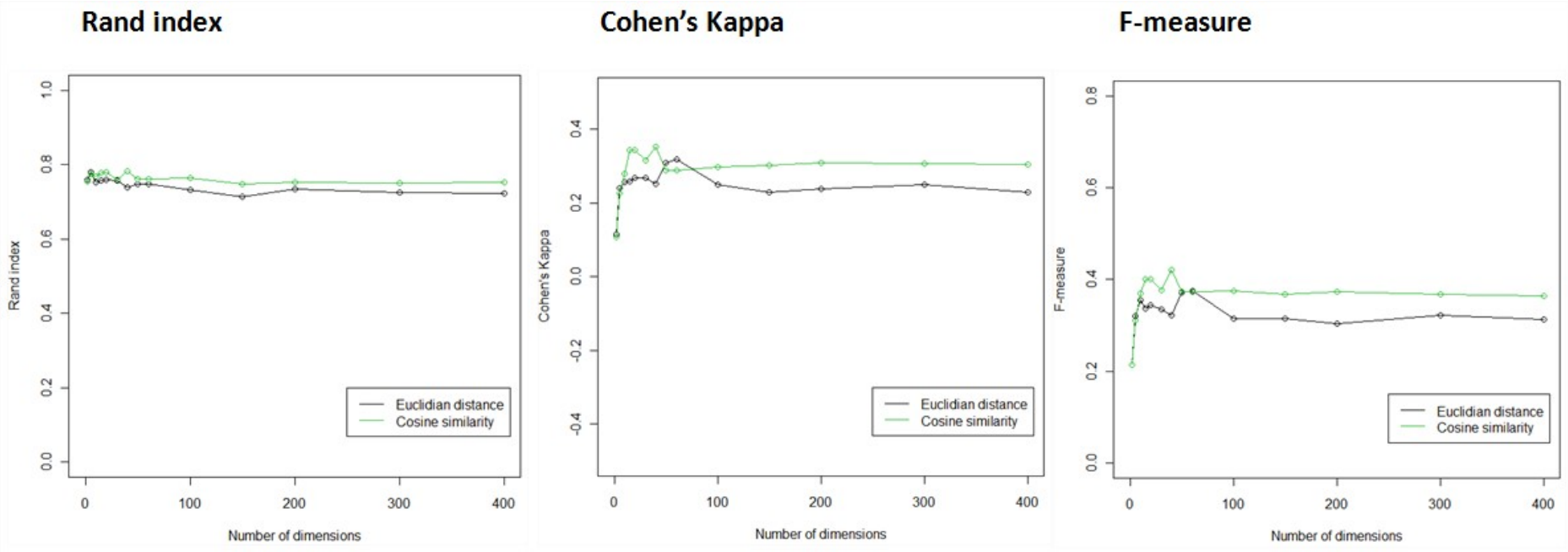



# Clustering quality according to the method and the number of dimensions : F-measure

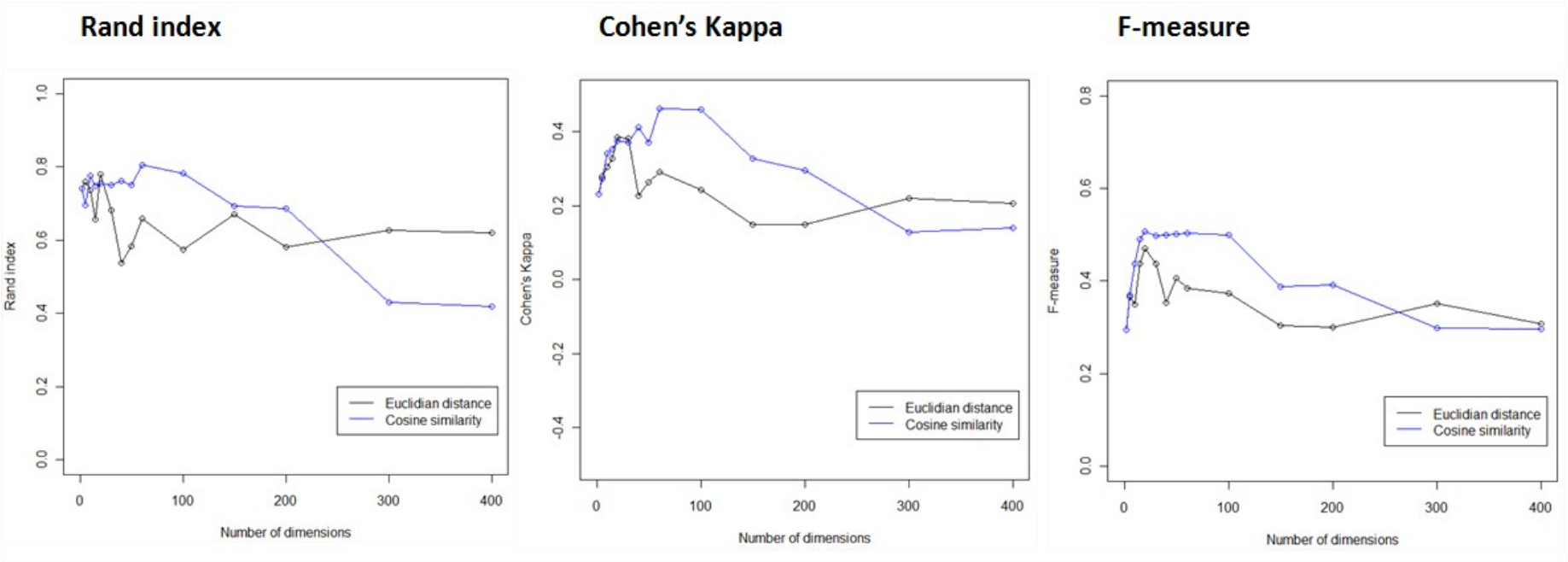


# Clustering quality according to the dissimilarity function : LSA + Log Entropy



# Clustering quality according to the dissimilarity function : CA



# Plan

# Conclusions

- CA seems to be less stable than other methods but provides better results on a few number of dimensions (depending on index)
- As it is said in literature, cosine similarity between vectors seems to be more adapted to textual data than usual dot similarity
- Greater improvement with cosine similarity for CA method (only on the first 200 dimensions)
- Optimal number of dimensions to keep ? It is varying according to the type of text studied and the method used (around 60 dimensions with CA, around 40 dimensions with LSA methods)

# Limitations & future work

## Limitations of the study

- Clusters obtained are compared with categories choosen by recruiters, so it is sometimes a subjective labeling and some errors may appear
- We are working on a very particular type of corpus : shorts texts, variable length, sometimes very similar but not really duplicates

## Future work

- Test other methods of data clustering (the representation to adopt may depend on it)
- Repeat the study with a supervised algorithm for classification (index values are disappointing in unsupervised framework)
- Study the effect of using the different parts of job offers for classification

# Some references


- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.
- Greenacre, M. (2007). *Correspondence Analysis in Practice, Second Edition*. London : Chapman & Hall/CRC.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ : Erlbaum.
- Picca, D., Curdy, B., & Bavaud, F. (2006). *Non-linear correspondence analysis in text retrieval : a kernel view*. In *JADT'06*, pp. 741-747.
- Wild, F. (2007). An LSA package for R. In *LSA-TEL'07*, pp. 11-12.

Thanks !