La matrice terme-document : formalisme, visualisation, clustering

illustrations avec IRaMuTeQ – version 1

François Bavaud

SLI-Lettres et IGD-FGSE

Université de Lausanne

cours-bloc

Statistique textuelle et topic models

novembre 2017

Première partie I

La matrice terme-document : formalisme, visualisation, clustering

La matrice terme-document : un jalon dans les traitements textuels

L'obtention de la matrice terme-document (`term-document matrix`)

- marque la fin des *pré-traitements textuels* :
  choix du corpus ; problèmes d'encodage ; choix des termes retenus ;
  lemmatisation ; choix des parties (textes) et leur catégorisation
- et marque le début des *traitements statistiques*. En particulier :
  - comment représenter la proximité/le contraste entre paires de termes ? entre paires de documents ? → **visualisation**
  - comment classer les termes, ou les documents, au sein de groupes homogènes ? → **classification** (`clustering`)

Exemple (jouet) : brochure 2007 "La Suisse vote en couleurs"

Election du Conseil National du 21 octobre 2007 : 14 partis représentés dans la législature précédente, décrits par de courts textes.
Extrait (fichier `textes_partis_reduit.txt`) :

**** * parti_prd
Le LE est une force politique qui s'engage avec le souci constant de construire des solutions et de préparer la Suisse à affronter les défis de demain. Afin de permettre à notre pays d'évoluer avec un très haut niveau de qualité de vie, le LE parie sur ce qu'il nomme la Suisse de l'intelligence, la Suisse de la croissance, la Suisse de l'équilibre et la Suisse de l'ouverture. Dans son action quotidienne, le LE s'inspire de valeurs libérales telles que la liberté, la responsabilité, la justice ou l'égalité des chances pour chacun. Avec Pascal Couchepin et Hans-Rudolf Merz au Conseil fédéral, le LE peut compter sur deux conseillers fédéraux qui font bouger la Suisse.
1. La Suisse de l'intelligence: la Suisse dispose d'atouts de premier plan dans les domaines de la recherche et de la formation. Le LE souhaite donc se battre afin de développer cette position de pointe reconnue dans le monde entier. Riche de sa diversité linguistique, la Suisse bénéficie d'un patrimoine culturel que le LE entend entretenir et développer.
2. La Suisse de la croissance : la croissance est le fruit du travail de toutes les citoyennes et de tous les citoyens et elle repose sur d'excellentes conditions politiques que le LE entend créer en faveur de l'économie.
3. La Suisse de l'équilibre et de l'ouverture : nous vivons dans un pays qui offre d'excellentes conditions de vie. Notre modèle de protection sociale fonctionne bien. La Suisse est d'autre part un pays ouvert, tolérant, en un mot moderne. Elle entretient des relations étroites et constructives avec l'Europe comme avec le reste du monde.

**** * parti_pss
Le Parti socialiste s'engage pour une Suisse sociale, ouverte et écologique. Avec Micheline Calmy-Rey et Moritz Leuenberger, il dispose au Conseil fédéral de deux représentants crédibles. Il est aussi très engagé dans les exécutifs des grandes villes. Son ambition est de devenir la première force politique du pays après l'élection du 21 octobre 2007. Cela lui permettra de faire sauter le bloc de la droite au Conseil fédéral et d'accroître sa force, tant au gouvernement qu'au Parlement. Pour assurer l'avenir des rentes, garantir des salaires convenables et prélever des impôts justes. Et pour que voient le jour les réformes qui moderniseront la politique en faveur des familles, qui donneront les mêmes chances à tous et à toutes dans le domaine de la formation et qui aideront l'économie suisse à prendre le virage de l'écologie.
1. Pour une Suisse sociale : le LE veut garantir l'avenir de l'AVS, que tout le monde puisse profiter de l'âge flexible de la retraite et ait les mêmes chances de se former. Il est pour l'égalité des sexes et pour que les femmes puissent travailler tout en ayant des enfants.
2. Pour une Suisse ouverte: le LE est pour l'entrée de la Suisse dans l'UE et pour que le pays continue la politique défendue activement par Mme Calmy-Rey: la défense des droits de l'homme, la paix et la coopération au développement.
3. Pour une Suisse écologique : le LE veut une taxe sur le CO2, des transports publics conformes aux voeux des usagers. Il est favorable à l'encouragement des

Matrice terme-document `tdm`

Après remplacement des abrévations telles que `prd`, `udc`, `lega`, etc. par le *mot-outil* `LE`, choix des *mots actifs* (et suppression des mot-outils), lemmatisation des mots actifs, et sélection des seuls mots actifs apparaissant au moins 4 fois, on obtient une matrice terme-document de 80 termes-lignes $\times$ 14 partis-colonne.

Début et fin de la matrice terme-document :

	akz	ds	glp	lega	pcs	pdc	pev	pls	prd	pss	pst	udc	udf	verts	*total*
sécurité	0	0	0	0	2	1	0	1	0	0	1	1	0	0	*6*
liberté	0	1	0	1	0	0	0	2	1	0	0	1	0	0	*6*
bénéficier	0	0	0	1	1	1	0	0	1	0	0	0	0	1	*5*
condition	0	0	1	0	0	2	0	1	2	0	0	0	0	0	*6*
solution	0	0	0	0	0	1	0	0	1	0	0	0	2	1	*5*
droit	2	2	1	0	1	0	0	0	0	1	1	2	0	2	*12*
tessin	0	0	0	3	0	0	0	0	0	0	0	0	1	0	*4*
développement	1	0	1	0	0	0	0	2	0	1	1	0	0	1	*7*
public	1	0	0	2	0	0	0	0	0	1	1	0	0	0	*5*
égalité	2	0	0	0	0	0	0	0	1	1	0	0	0	1	*5*
UE	0	3	0	0	0	0	0	0	0	1	0	0	0	0	*4*
domaine	0	0	0	0	0	0	0	0	1	1	0	0	1	1	*4*
social	0	3	2	1	1	5	1	3	1	2	1	0	1	0	*21*
action	1	0	0	0	0	1	0	0	1	0	1	0	0	0	*4*
formation	1	1	0	0	1	1	0	2	1	1	0	0	0	1	*9*
vie	0	0	0	0	1	2	0	0	2	0	0	0	1	0	*6*

avenir	0	0	1	0	1	0	0	0	0	2	0	0	1	0	*5*
humain	0	0	1	0	1	1	1	0	0	0	0	0	0	0	*4*
chrétien	0	0	0	0	0	1	2	0	0	0	0	0	1	0	*4*
travailler	0	0	0	0	2	0	0	0	0	1	1	0	0	0	*4*
indépendance	0	0	0	3	0	0	0	0	0	0	0	1	0	0	*4*
système	0	0	0	0	0	1	1	2	0	0	0	0	0	0	*4*
présent	0	0	0	0	0	0	0	1	0	0	1	0	1	1	*4*
préserver	0	0	1	1	2	0	0	0	0	0	0	0	0	2	*6*
libéral	0	0	2	0	0	1	0	1	1	0	0	0	1	0	*6*
rester	2	0	0	1	1	0	0	0	1	0	0	1	0	0	*6*
défendre	0	0	0	0	0	3	1	1	0	1	0	1	2	0	*9*
chance	1	0	0	0	0	0	0	0	1	2	0	0	0	3	*7*
fiscal	1	0	0	0	0	0	2	1	0	0	0	0	0	0	*4*
populaire	0	1	0	0	0	0	0	0	0	0	2	1	0	0	*4*
niveau	0	0	0	0	0	1	1	0	1	0	1	0	0	0	*4*
économie	0	0	2	0	0	1	1	0	1	1	0	0	1	0	*7*
total	*41*	*40*	*35*	*35*	*32*	*54*	*31*	*40*	*51*	*52*	*36*	*28*	*38*	*52*	**565**

`tdm` = TABLE DE CONTINGENCE

La matrice terme-document est une matrice $N = (n_{ik})$ dont

- les v lignes correspondent aux termes $i = 1, \ldots, v$
- les p colonnes correspondent aux documents $k = 1, \ldots, p$
- les $v \times p$ cases comptent le nombre de fois n_{ik} que le terme i est apparu dans la colonne k.

De plus :

- la somme de chaque ligne $n_{i\bullet} = \sum_k n_{ik}$ compte le nombre total d'occurrences du terme i
- la somme de chaque colonne $n_{\bullet k} = \sum_i n_{ik}$ mesure la taille du document k (en nombre de termes)
- la somme de toutes les cases $n_{\bullet\bullet} = \sum_{ik} n_{ik}$ mesure la taille totale du corpus.

Quotients d'indépendence

Si la distribution des termes était indépendante des documents, on s'attendrait à observer en moyenne

$$n_{ik}^{\text{theo}} = \frac{n_{i\bullet} n_{\bullet k}}{n_{\bullet\bullet}}$$

occurrences du terme i dans le document k. Les n_{ik}^{theo} sont les *effectifs théoriques* attendus sous l'hypothèse d'indépendance entre termes et documents. Les *quotients d'indépendance*

$$q_{ik} = \frac{n_{ik}^{\text{theo}}}{n_{ik}}$$

quantifient la dépendance entre termes et documents :

- $q_{ik} > 1$ dénote une *attraction* entre i et k
- $q_{ik} < 1$ dénote une *répulsion* entre i et k
- $q_{ik} \cong 1$ dénote une situation de *neutralité* entre i et k.

Le chi-carré et sa significativité

Le coefficient du *chi-carré* est la mesure canonique de dépendance totale entre termes et documents :

$$\texttt{chi2} = \sum_{ik} \frac{(n_{ik} - n_{ik}^{\text{theo}})^2}{n_{ik}^{\text{theo}}} = n_{\bullet\bullet} \sum_{ik} f_i \rho_k (q_{ik} - 1)^2 = n_{\bullet\bullet} \Delta$$

- $f_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$ est le poids relatif (pourcentage) du terme i
- $\rho_k = \frac{n_{\bullet k}}{n_{\bullet\bullet}}$ est le poids relatif du document k
- la quantité Δ (parfois appelée *phi-carré*) est une mesure de dépendance indépendante de la taille $n_{\bullet\bullet}$ du corpus.

Si les termes étaient attribués aléatoirement aux documents, le `chi2` serait "petit", mais pas nul en général (fluctuations autour de l'indépendance). Dès que le `chi2` devient "suffisamment grand" (seuils disponibles dans les tables statistiques), le recours à l'explication par "fluctuations aléatoires" devient intenable : la `valeur p`, qui mesure la plausibilité (probabilité) de l'explication par "fluctuations aléatoires", devient petite (par exemple $\texttt{p} < 0.01$) et le `chi2` est alors déclaré **significatif** (= non dû au hasard).

Interprétation géométrique : distance du chi-carré

La mesure canonique de dissimilarité entre (distributions des) termes i et j est donnée par la *distance du chi-carré*

$$D_{ij}^{\chi} = \sum_k \rho_k (q_{ik} - q_{jk})^2$$

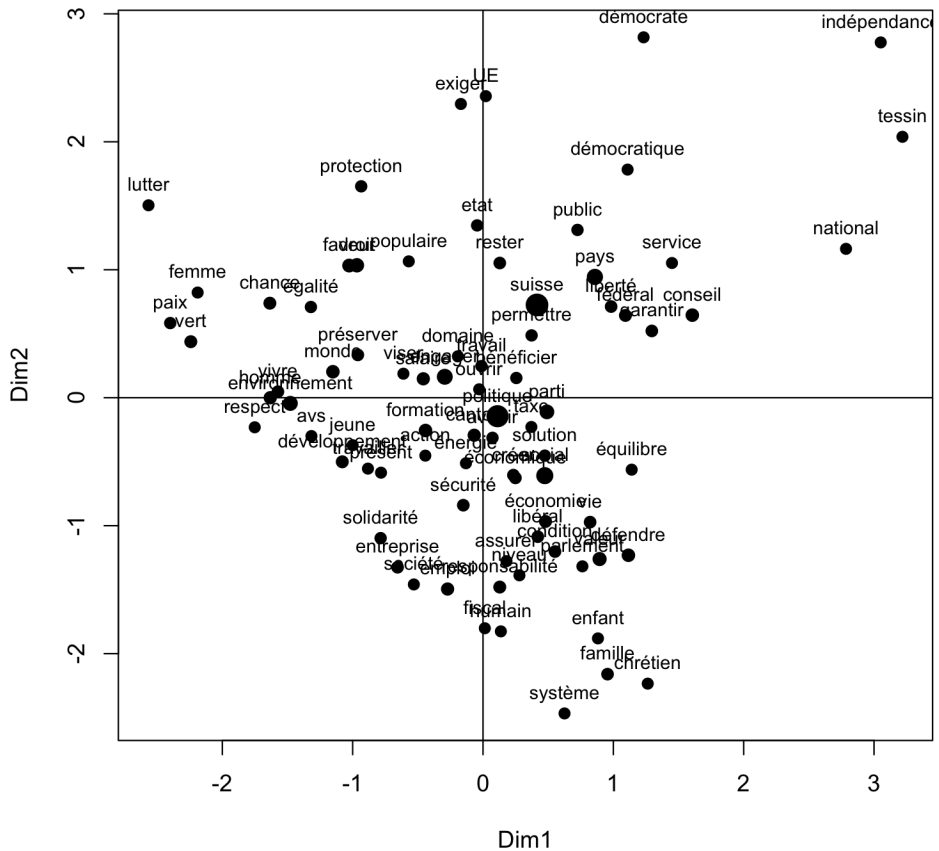
- D'une part, D_{ij}^{χ} représente le carré d'une distance euclidienne entre deux points i et j : on peut donc chercher à *reconstruire* une configuration géométrique de v points-termes de telle sorte que les distances entre paires de points i, j soient précisément D_{ij}^{χ} : c'est l'objet du *multidimensional scaling* (MDS).
  Il faudra pour cela généralement placer les v points dans un espace de grande dimension donnée par $\min(v-1, p-1)$.
- D'autre part, l'*inertie* du nuage des v points (= sa dispersion ou variance totale) n'est autre que le `phi-carré` :

$$\texttt{phi-carré} = \frac{\texttt{chi-carré}}{n_{\bullet\bullet}} = \Delta = \frac{1}{2} \sum_{ij} f_i f_j D_{ij}^{\chi} = \sum_i f_i D_{i0}^{\chi}$$

où D_{i0}^{χ} est le carré de la distance de i à l'origine, laquelle représente le "terme moyen" (fictif).

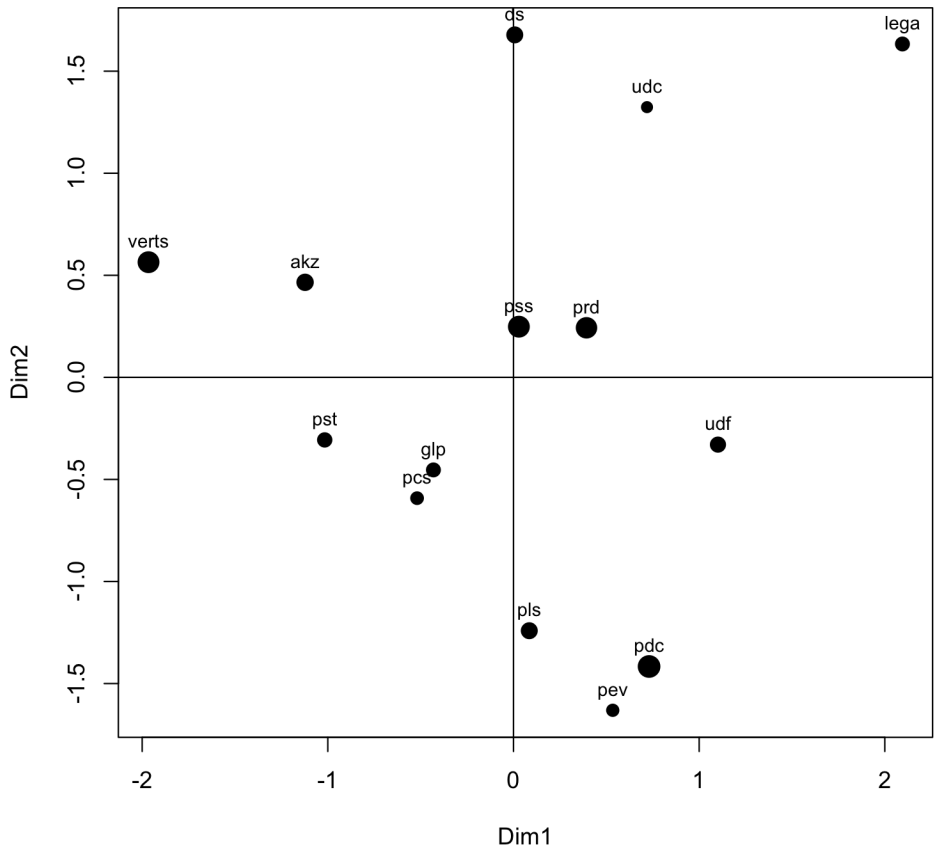
L'analyse factorielle des correspondances (AFC)

L'AFC a précisément pour but de *projeter* le nuage des v points-termes associée à la configuration pondérée (f, D^X), de haute dimensionnalité, dans un espace plus petit, typiquement à 2 dimensions, afin de pouvoir le *visualiser* – tout en veillant à *conserver un maximum de dispersion* Δ .



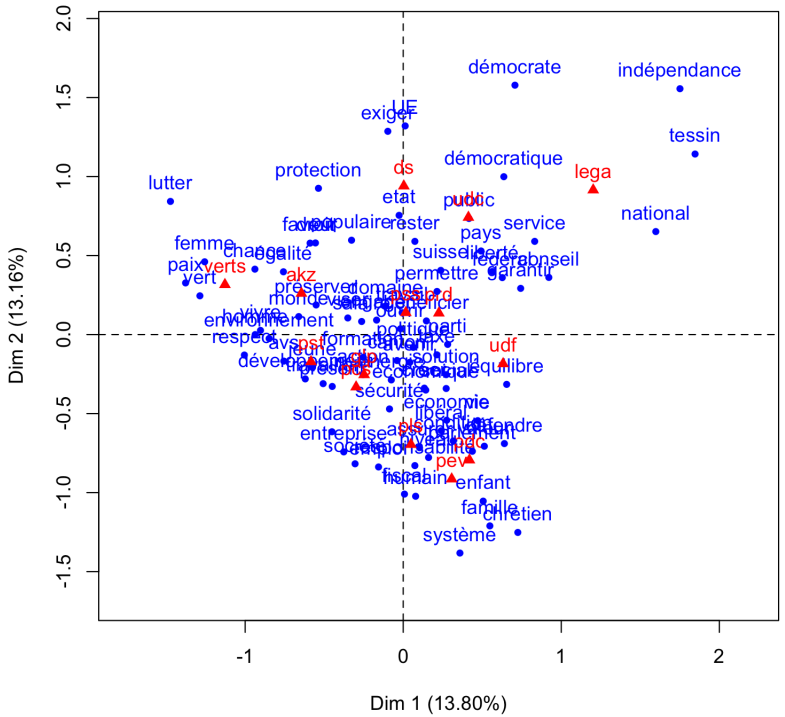
Dualité lignes-colonnes en AFC

Naturellement, on peut échanger le rôle des lignes et des colonnes (dualité) et calculer la distance du chi2 $\tilde{D}_{kl}$ entre documents k et l (ici les partis), ainsi que l'inertie associée $\tilde{\Delta} = \frac{1}{2}\sum_{kl} \rho_k \rho_l \tilde{D}_{kl}$.



Dualité et biplot

Les inerties des lignes et des colonnes coïncident : $\Delta = \tilde{\Delta}$, ainsi que l'espace de basse dimensionalité qui représente au mieux leurs configurations respectives :



Biplot : le *premier facteur* (Dim 1) exprime 13.80% de l'inertie de la configuration ; le *second facteur* (Dim 2) exprime 13.16% de l'inertie de la configuration.

La proximité/éloignement entre termes et documents traduit leur attraction/répulsion

Clustering = Classification non supervisée

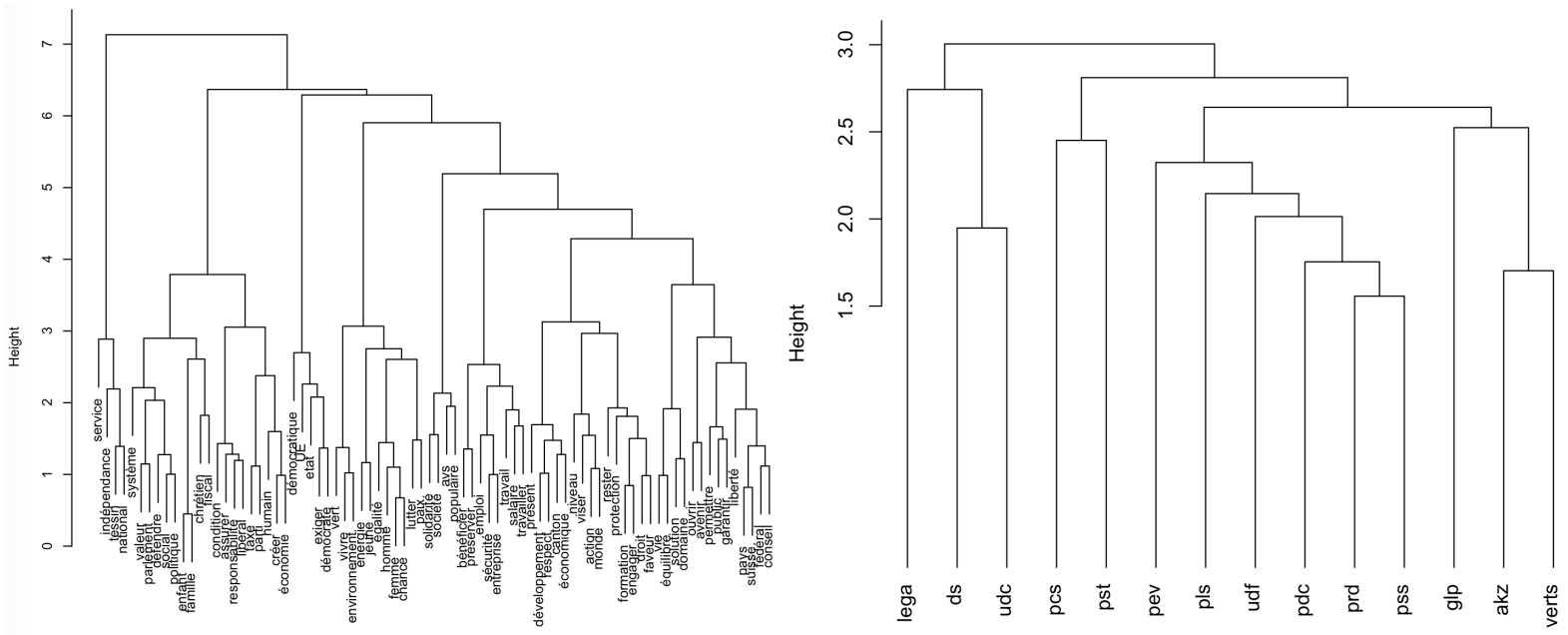
Le *clustering* ou *classification non supervisée* consiste à regrouper des termes (ou, dans l'approche duale, des documents) qui possèdent une *distribution similaire* (dans les documents).

Il s'agit de *créer* une nouvelle variable "groupe" ou "classe" G à K modalités : chaque terme $i = 1, \ldots, v$ appartient à un groupe $g = 1, \ldots, K$. L'appartenance peut être

- *floue* (`soft, fuzzy`) : chaque terme i appartient plus ou moins à une classe g, telle que donnée par la matrice d'appartenance $Z = (z_{ig})$ avec $z_{ig} = \texttt{prob}(g|i)$. Par construction, $z_{ig} \geq$ et $z_{i\bullet} = 1$.
- *dure* (`hard`) : chaque terme i appartient à une classe g et une seule. Par construction, $z_{ig} = 0$ ou $z_{ig} = 1$.

Classification ascendante hiérarchique (CAH)

A partir de la configuration (f, D), regrouper les termes les moins dissimilaires, considérer les agrégats comme de nouveaux termes, et itérer jusqu'à n'obtenir plus qu'un seul groupe.
Le *dendrogramme* résultant peut être coupé à une hauteur "convenable", définissant les K groupes.



Dendrogrammes : termes regroupés par la méthode dite de Ward à partir de leurs dissimilarités du chi2 D^χ (gauche) ; CAH des documents à partir de $\tilde{D}^\chi$ (droite).

Classification descendante hiérarchique (CDH)

A partir de la configuration (f, D), diviser l'ensemble des termes (ou des documents) en deux groupes, de sorte que :

- la dispersion intra-groupe soit minimale
- ou que le contraste inter-groupe (= distance pondérée entre les centres de gravité des groupes) soit maximale.

et itérer la bipartition sur chacun des sous-groupes obtenus.

Ces deux critères sont équivalents lorsque la dissimilarité est euclidienne carrée – comme avec la dissimilarité du chi2 D^{χ} .

On ne dispose pas d'un algorithme efficace (= dont la vitesse d'exécution croîtrait moins vite qu'*exponentiellement avec* v) permettant de bipartitionner une configuration en deux groupes
$\rightarrow$ algorithme approché (cf. "méthode de Reinert" dans IRaMuTeQ)

- couper les deux groupes selon le premier facteur de l'AFC
- éventuellement, procéder à des réallocations individuelles de groupe pour augmenter encore le contraste inter-groupe.

Modèles latents (bi-clustering) : pLSI

On postule l'existence de K thèmes ou `topics` $g = 1, \ldots, K$ tels que, étant donné le thème g, termes et documents sont indépendants :

$$n_{ik} \cong n_{\bullet\bullet} \sum_{g=1}^{K} p(g)p(i|g)p(k|g) \tag{1}$$

L'équation (1) (modèle de `probabilistic latent semantic indexing` ou `pLSI`) propose d'"expliquer" la matrice `tdm` par le recours à une variable latente (inobservée) G à K modalités, qui sont les *groupes* ou *clusters*, ou encore les *thèmes* ou `topics` dans ce contexte textuel.

Etant donné un thème g, de fréquence relative $p(g)$ (avec $\sum_{g=1}^{K} p(g) = 1$),

- un terme i est produit avec probabilité $p(i|g)$. La matrice $A = (a_{gi})$ avec $a_{gi} = p(i|g)$ est la *matrice d'émission des termes*
- un document k est produit avec probabilité $p(k|g)$. La matrice $B = (b_{gk})$ avec $b_{gk} = p(k|g)$ est la *matrice d'émission des documents*
- (1) dit que, pour un thème g donné, les émissions des termes et des documents sont indépendantes.

Modèles latents (bi-clustering) : LDA

Le `Latent Dirichlet Allocation` (LDA) est le modèle le plus fréquemment utilisé en *topic modelling*. Il généralise le modèle `pLSI` en introduisant des paramètres α et β contrôlant les a priori (au sens "bayésien") que l'on peut avoir à propos des données :

- α est le nombre de fois (constant) que l'on aurait observé chaque thème avant d'examiner les données textuelles actuelles. $\alpha \in (0, 1)$ favorise l'émergence de thèmes uniques dans un document donné, $\alpha \gg 1$ tend à produire plusieurs thèmes dans un document donné.
- β est le nombre de fois (constant) que l'on aurait observé chaque mot avant d'examiner les données textuelles actuelles. De faibles valeurs (positives) de β (par exemple $\beta \leq 0.01$) tendent à lier chaque terme à un thème bien spécifique ; des valeurs plus élevées lient chaque terme à davantage de thèmes.

(1) peut aussi se lire comme "choix d'un document k" $\rightarrow$ "choix conditionnel d'un thème" $\rightarrow$ "choix conditionnel d'un terme" :

$$p(i|k) = \frac{n_{ik}}{n_{\bullet k}} \cong \sum_{g=1}^{K} \frac{p(g)p(k|g)}{p(k)} p(i|g) = \sum_{g=1}^{K} p(g|k)p(i|g) = \sum_{g=1}^{K} z^*_{kg} a_{gi}$$

Exemple : avec `topicmodels()`, $K = 3$, $\alpha = 0.1$, β libre

	topic1	topic2	topic3
sécurité	0.0143	0.0049	0.0132
liberté	0.0101	0.0190	0.0000
bénéficier	0.0054	0.0092	0.0132
condition	0.0240	0.0048	0.0000
solution	0.0108	0.0085	0.0066
droit	0.0055	0.0228	0.0408
tessin	0.0000	0.0196	0.0000
développement	0.0167	0.0001	0.0229
public	0.0048	0.0141	0.0075
égalité	0.0000	0.0075	0.0228
UE	0.0000	0.0196	0.0000
domaine	0.0000	0.0141	0.0075
social	0.0608	0.0340	0.0087
action	0.0120	0.0024	0.0066
formation	0.0155	0.0119	0.0220
vie	0.0148	0.0093	0.0066
énergie	0.0074	0.0057	0.0085
entreprise	0.0191	0.0000	0.0132
lutter	0.0000	0.0049	0.0198
salaire	0.0048	0.0128	0.0290
solidarité	0.0179	0.0000	0.0083
enfant	0.0143	0.0049	0.0000
vert	0.0000	0.0000	0.0462
famille	0.0239	0.0049	0.0000

	Topic 1	Topic 2	Topic 3
[1,]	"politique"	"suisse"	"politique"
[2,]	"social"	"pays"	"environnement"
[3,]	"valeur"	"politique"	"vert"
[4,]	"suisse"	"conseil"	"engager"
[5,]	"responsabilité"	"social"	"droit"
[6,]	"défendre"	"engager"	"homme"
[7,]	"parti"	"fédéral"	"faveur"
[8,]	"libéral"	"parti"	"préserver"
[9,]	"économie"	"faveur"	"chance"
[10,]	"condition"	"droit"	"monde"
[11,]	"famille"	"tessin"	"salaire"
[12,]	"créer"	"démocratique"	"emploi"
[13,]	"environnement"	"démocrate"	"femme"
[14,]	"économique"	"national"	"paix"
[15,]	"entreprise"	"indépendance"	"développement"
[16,]	"société"	"UE"	"égalité"
[17,]	"système"	"liberté"	"formation"
[18,]	"emploi"	"garantir"	"vivre"
[19,]	"chrétien"	"etat"	"canton"
[20,]	"engager"	"exiger"	"lutter"
[21,]	"solidarité"	"protection"	"rester"
[22,]	"canton"	"rester"	"travailler"
[23,]	"niveau"	"domaine"	"avenir"
[24,]	"développement"	"public"	"suisse"
[25,]	"pays"	"permettre"	"jeune"

Gauche : premières ligne de la matrice (transposée) d'émission A des termes : la somme de chaque colonne vaut 1. Droite : composition de chaque thème par les 24 mots les plus fréquemment émis, rangés par ordre décroissant.

Exemple (suite)

	topic1	topic2	topic3
sécurité	0.50	0.17	0.33
liberté	0.35	0.65	0.00
bénéficier	0.23	0.37	0.40
condition	0.84	0.16	0.00
solution	0.46	0.34	0.20
droit	0.10	0.39	0.52
tessin	0.00	1.00	0.00
développement	0.50	0.00	0.50
public	0.20	0.57	0.23
égalité	0.00	0.31	0.69
UE	0.00	1.00	0.00
domaine	0.00	0.72	0.28
social	0.61	0.33	0.06
action	0.63	0.12	0.25
formation	0.36	0.27	0.37
vie	0.52	0.31	0.17
énergie	0.39	0.29	0.32
entreprise	0.67	0.00	0.33
lutter	0.00	0.25	0.75
salaire	0.13	0.32	0.55
solidarité	0.75	0.00	0.25
enfant	0.75	0.25	0.00
vert	0.00	0.00	1.00
famille	0.83	0.17	0.00

	topic1	topic2	topic3
akz	0.00	0.00	1.00
ds	0.00	1.00	0.00
glp	0.57	0.00	0.43
lega	0.00	1.00	0.00
pcs	0.00	0.00	1.00
pdc	1.00	0.00	0.00
pev	1.00	0.00	0.00
pls	1.00	0.00	0.00
prd	0.21	0.79	0.00
pss	0.00	0.78	0.22
pst	1.00	0.00	0.00
udc	0.00	1.00	0.00
udf	0.49	0.51	0.00
verts	0.00	0.00	1.00

Gauche : matrice d'appartenance $Z = (z_{ig})$ des termes aux thèmes. Droite : matrice d'appartenance $Z^* = (z^*_{kg})$ des documents aux thèmes. La somme de chaque ligne vaut 1. L'appartenance peut être rendue dure ("forcée") en attribuant chaque terme (ou document) au thème auquel il appartient le plus.

Deuxième partie II

Illustrations avec IRaMuTeQ

Documentation de IRaMuTeQ

IRaMuTeQ (conçu par Pierre Ratinaud, dès 2009) est un logiciel libre, ouvert, multiplateforme, convivial et permettant de nombreuses analyses lexicométriques.

Son mode d'emploi et le déroulement de ses analyses peuvent être obtenues dans plusieurs sites ou documents utiles, dont :

- Documentation générale du site http://www.iramuteq.org/ (Pierre Ratinaud)
- Initiation à la lexicométrie: approche pédagogique à partir de l'étude d'un corpus avec le logiciel Iramuteq (D. Pélissier)
- IRaMuteQ 0.7 alpha 2 Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Elodie Baril et Bénédicte Garnier)

Un exposé unifié des différentes méthodes quantitatives utilisées dans IRaMuTeQ n'est apparemment pas disponible (voir cependant "Statistique Textuelle", Lebart, L. & Salem, A. (1994)) ; assez souvent, la reconstitution de ces méthodes quantitatives implique une bonne dose de culture générale, et de rétro-ingénierie.

Format général

"Les fichiers d'entrée doivent être au format texte brut (.txt) et respecter les règles de formatage suivantes : dans ce formatage, l'unité de base est appelée « texte ». Un texte peut représenter un entretien, un article, un livre ou tout autre type de documents. Un corpus peut contenir un ou plusieurs texte (mais au minimum un). Les textes sont introduits par quatre étoiles (****) suivies d'une série de variables étoilées séparées par un espace. (...) La ligne ne doit contenir que cette variable."
(P.Ratinaud, formatage-des-corpus-texte)

Au minimum quatre étoiles marquant le début d'un (du) texte. Exemple :

```
**** * parti_prd
Le LE est une force politique qui s'engage avec le souci constant de construire des solutions et de préparer la Suisse à affronter les défis de demain. Afin de permettre à notre pays d'évoluer avec un très haut niveau de qualité de vie, le LE parie sur ce qu'il nomme la Suisse de l'intelligence, la Suisse de la croissance, la Suisse de l'équilibre et la Suisse de l'ouverture. Dans son action quotidienne, le LE s'inspire de valeurs libérales telles que la liberté, la responsabilité, la justice ou l'égalité des chances pour chacun. Avec Pascal Couchepin et Hans-Rudolf Merz au Conseil fédéral, le LE peut compter sur deux conseillers fédéraux qui font bouger la Suisse.
1. La Suisse de l'intelligence: la Suisse dispose d'atouts de premier plan dans les domaines de la recherche et de la formation. Le LE souhaite donc se battre afin de développer cette position de pointe reconnue dans le monde entier. Riche de sa diversité linguistique, la Suisse bénéficie d'un patrimoine culturel que le LE entend entretenir et développer.
2. La Suisse de la croissance : la croissance est le fruit du travail de toutes les citoyennes et de tous les citoyens et elle repose sur d'excellentes conditions politiques que le LE entend créer en faveur de l'économie.
3. La Suisse de l'équilibre et de l'ouverture : nous vivons dans un pays qui offre d'excellentes conditions de vie. Notre modèle de protection sociale fonctionne bien. La Suisse est d'autre part un pays ouvert, tolérant, en un mot moderne. Elle entretient des relations étroites et constructives avec l'Europe comme avec le reste du monde.

**** * parti_pss
Le Parti socialiste s'engage pour une Suisse sociale, ouverte et écologique. Avec Micheline Calmy-Rey et Moritz Leuenberger, il dispose au Conseil fédéral de deux représentants crédibles. Il est aussi très engagé dans les exécutifs des grandes villes. Son ambition est de devenir la première force politique du pays après l'élection du 21 octobre 2007. Cela lui permettra de faire sauter le bloc de la droite au Conseil fédéral et d'accroître sa force, tant au gouvernement qu'au Parlement. Pour assurer l'avenir des rentes, garantir des salaires convenables et prélever des impôts justes. Et pour que voient le jour les réformes qui moderniseront la politique en faveur des familles, qui donneront les mêmes chances à tous et à toutes dans le domaine de la formation et qui aideront l'économie suisse à prendre le virage de l'écologie.
1. Pour une Suisse sociale : le LE veut garantir l'avenir de l'AVS, que tout le monde puisse profiter de l'âge flexible de la retraite et ait les mêmes chances de se former. Il est pour l'égalité des sexes et pour que les femmes puissent travailler tout en ayant des enfants.
2. Pour une Suisse ouverte: le LE est pour l'entrée de la Suisse dans l'UE et pour que le pays continue la politique défendue activement par Mme Calmy-Rey: la défense des droits de l'homme, la paix et la coopération au développement.
3. Pour une Suisse écologique : le LE veut une taxe sur le CO2, des transports publics conformes aux voeux des usagers. Il est favorable à l'encouragement des
```

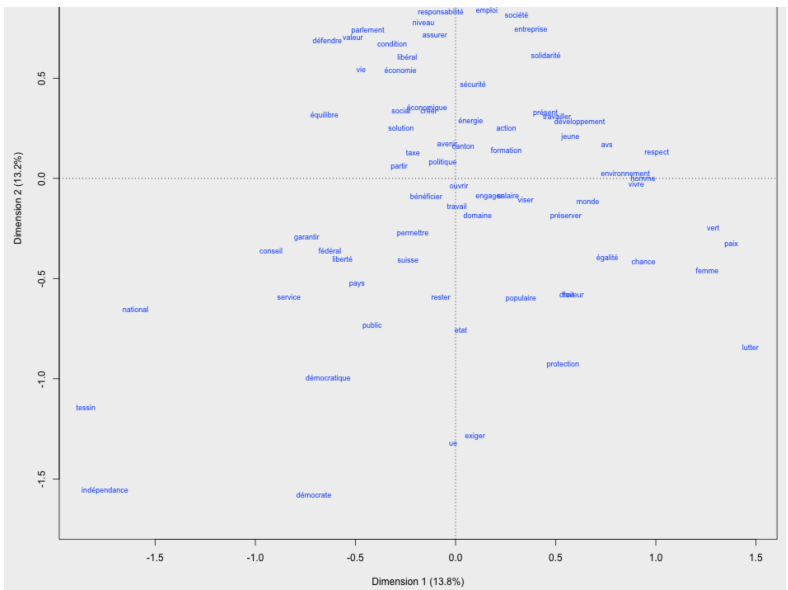
Segmentation

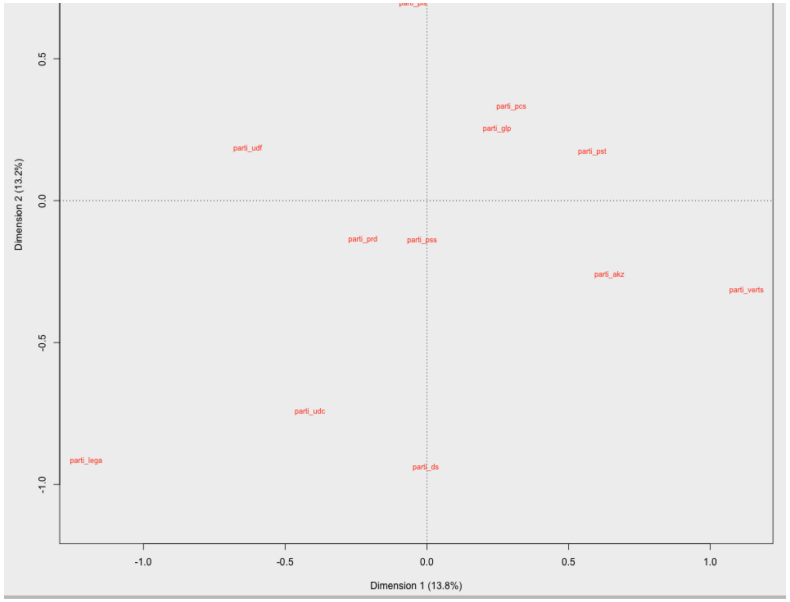
Au début, le texte est **segmenté** en suite de mots ou occurrences (= `tokens`) de longueur (à peu près) constante, en essayant de respecter les séparateurs (ponctuation et paragraphes). Par défaut, la longueur d'un **segment** est de 40 occurrences.

Dans le menu "Statistique", les mots sont *lemmatisés* (au moyen d'un dictionnaire par défaut) et regroupés en **formes** (= `types`), elles-mêmes considérées comme **actives** (=porteuses de sens) ou **supplémentaires** (=mots-outils ne participant pas aux analyses elles-mêmes, mais qui peuvent être représentés (ou non) à la fin ; paramétrisation par défaut)

Spécificités et AFC

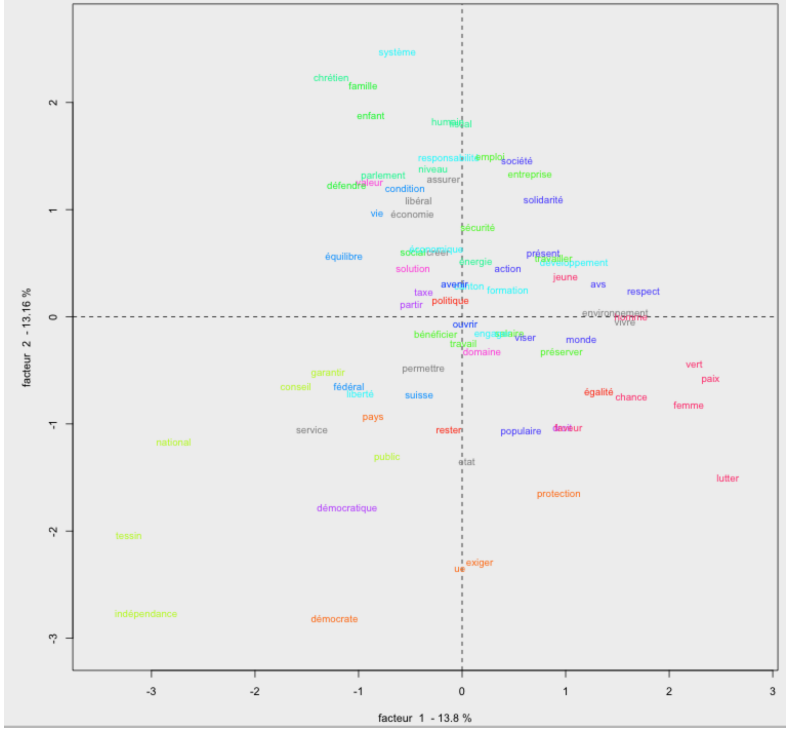
Typiquement, Iramuteq réalise une AFC sur la matrice `tdm` donnant les effectifs croisés entre les formes actives (d'effectif plus grand qu'un seuil fixé) et les modalités d'une variable catégorielle "étoilée". Cette table de contingence est appelée *tableau lexical agrégé*.





Spécificités et AFC (suite)

Iramuteq réalise également une AFC "colorée" des termes, où la couleur représente la modalité k la plus associée au terme i, au sens de leur *spécificité* (= mesure de la sur-représentation de la forme i dans la modalité k, selon deux méthodes possibles "hypergéométrique" et "chi2") :



Formes	Formes banales	Types	Fréquences des formes	Fréquences des types	Fréquences rel

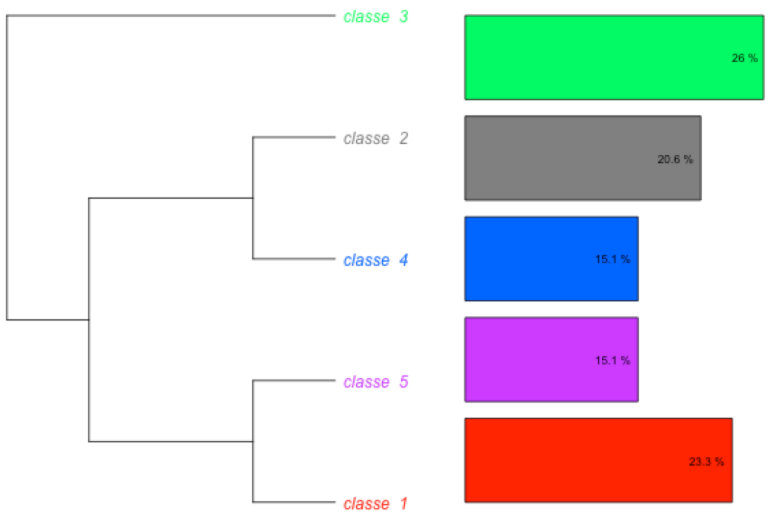
formes	parti_akz	parti_ds	parti_glp	parti_lega	parti_pcs
ue	-0.1312	2.9009	-0.1114	-0.1114	-0.1016
démocrate	-0.1312	1.5719	-0.1114	0.6454	-0.1016
exiger	0.5837	1.5719	-0.1114	-0.1114	-0.1016
pays	-0.2082	1.4951	-0.5078	0.5114	-0.463
protection	0.5019	1.3699	-0.1394	-0.1394	-0.1271
etat	-0.1642	1.3699	1.4795	-0.1394	-0.1271
suisse	-1.358	1.2684	-1.1525	0.3387	-1.0507
salaire	0.3422	0.9816	-0.2236	-0.2236	2.1175
social	-0.7	0.7447	0.4219	-0.2066	-0.1781
droit	0.6696	0.6862	0.2681	-0.3367	0.2952
populaire	-0.1312	0.5933	-0.1114	-0.1114	-0.1016
viser	0.5837	0.5933	-0.1114	-0.1114	-0.1016
démocratique	-0.1312	0.5933	-0.1114	-0.1114	-0.1016
lutter	-0.1312	0.5933	-0.1114	-0.1114	-0.1016
travail	-0.1642	0.5111	-0.1394	-0.1394	0.5957
créer	-0.1972	0.4466	1.3205	-0.1674	-0.1526
liberté	-0.1972	0.4466	-0.1674	0.495	-0.1526
formation	0.3055	0.3133	-0.2518	-0.2518	0.3866
faveur	0.7966	0.2814	-0.28	-0.28	-0.2553

Classification descendante hiérarchique (CDH)

Ce sont les *segments de texte* (et non les formes) qui sont classés, selon les bipartitions successives "à la Reinert" associées à la distance du chi2 D_{ab}^{χ} entre les paires de segments a et b, relativement à leur profil lexical sur les formes actives. Certains segments ne sont pas classés (car pas assez loin du centre de gravité ? critère à préciser...) et lorsque le nombre de classes devient trop petit les bipartitions ne sont plus poursuivies.

Nombre de textes: 14
Nombre de segments de texte: 101
Nombre de formes: 1109
Nombre d'occurrences: 3539
Nombre de lemmes: 930
Nombre de formes actives: 770
Nombre de formes supplémentaires: 160
Nombre de formes actives avec une fréquence >= 3: 133
Moyenne de formes par segment: 35.039604
Nombre de classes: 5
73 segments classés sur 101 (72.28%)

##########
temps : 0h 0m 5s
##########

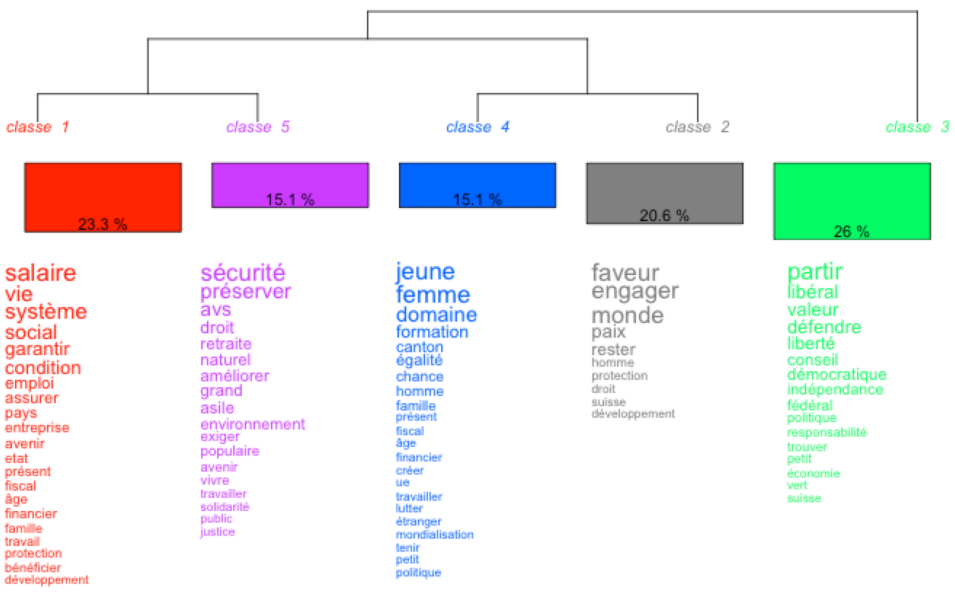


CDH (suite)

Pour chaque forme i et chaque groupe g, on peut calculer χ^2_{ig} qui est le coefficient du chi2 correspondant à la table de contingence

	groupe g	autres groupes $\bar{g}$
forme i	nombre de seg. de g contenant i	nombre de seg. de $\bar{g}$ contenant i
autres formes $\bar{i}$	nombre de seg. de g contenant $\bar{i}$	nombre de seg. de $\bar{g}$ contenant $\bar{i}$

Chaque groupe g est alors caractérisé par les formes i (et types grammaticaux, et modalités) ordonnées par valeurs décroissantes de χ^2_{ig} :



1 Classe 1 17/73 23.29% | 2 Classe 2 15/73 20.55% | 3 Classe 3 19/73 26.03% | 4 Classe 4 11/73 15.07% | 5 Classe 5 11/73 15.07%

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	6	7	85.71	16.89	nom	salaire	< 0,0001
1	4	4	100.0	13.94	nom	vie	0.00018
2	4	4	100.0	13.94	nom	système	0.00018
3	9	16	56.25	12.46	adj	social	0.00041
4	4	5	80.0	9.66	ver	garantir	0.00187
5	4	5	80.0	9.66	nom	condition	0.00187
6	4	6	66.67	6.89	nom	emploi	0.00868
7	3	4	75.0	6.33	ver	assurer	0.01183
8	6	13	46.15	4.63	nom	pays	0.03143
9	3	5	60.0	4.05	nom	entreprise	0.04418
10	3	5	60.0	4.05	nom	avenir	0.04418
11	2	3	66.67	3.3	nr	etat	NS (0.06947)
12	2	3	66.67	3.3	adj	présent	NS (0.06947)
13	2	3	66.67	3.3	adj	fiscal	NS (0.06947)
14	2	3	66.67	3.3	nom	âge	NS (0.06947)
15	2	3	66.67	3.3	adj	financier	NS (0.06947)
16	3	3	100.0	10.31	adj_sup	juste	0.00132
17	7	13	53.85	8.27	num	3	0.00403
18	3	4	75.0	6.33	adv_sup	bien	0.01183
19	5	11	45.45	3.56	con	qu	NS (0.05910)
20	2	3	66.67	3.3	pre	sans	NS (0.06947)
21	2	3	66.67	3.3	adv_sup	ainsi	NS (0.06947)
22	2	3	66.67	3.3	pro_per	y	NS (0.06947)
23	9	26	34.62	2.9	pre	pour	NS (0.08854)
24	5	13	38.46	2.04	ver_sup	devoir	NS (0.15335)
25	3	4	75.0	6.33		parti_pls	0.01183
26	3	6	50.0	2.61		parti_ds	NS (0.10611)

CDH (suite)

Les cases des tables produites par l'analyse en CDH sont actives, et permettent de générer des concordanciers, sous-corpus, regroupements de formes, graphiques de profils des termes dans les documents, "antiprofils" etc. Le dossier produit par la CDH contient également un *corpus en couleur*, attribuant à chaque segment une couleur correspondant à son groupe (en noir : segments non attribués)

**** * parti_ds

chez les démocrates suisses le le patriotisme n est pas un vain mot notre idée de la suisse est celle d une nation libre viable et indépendante notre politique vise à créer un espace vital sain

stable et social pour nous autres suisses nous demandons l arrêt de l immigration en provenance des pays non européens et le renvoi des étrangers qui refusent de s adapter nous nous opposons aussi bien à la dangereuse islamisation qu à l américanisation rampante du pays

nous rejetons catégoriquement l adhésion à l ue la mondialisation à outrance doit cesser avec le chômage la pression sur les salaires et le démantèlement de l etat social qu elle entraîne

nous nous engageons en faveur de la sauvegarde de l environnement de la nature et d une protection rigoureuse des animaux 1 la suisse doit devenir moins attrayante pour les illégaux et les profiteurs

nous exigeons que la frontière soit mieux gardée et que la justice réprime avec plus de rigueur les abus du droit d asile la criminalité et la violence 2

les démocrates suisses rejettent l entrée de notre pays dans une ue centraliste bureaucratique et non démocratique notre liberté et les droits populaires de notre démocratie directe ne doivent en aucun cas être sacrifiés sur l autel de l adhésion à l ue

3 nous exigeons la protection absolue des salariés et des apprentis du pays sur le marché du travail ainsi que des étudiants suisses dans le secteur de la formation et nous luttons contre les baisses de salaire et le démantèlement de l etat social

---

**** * parti_glp

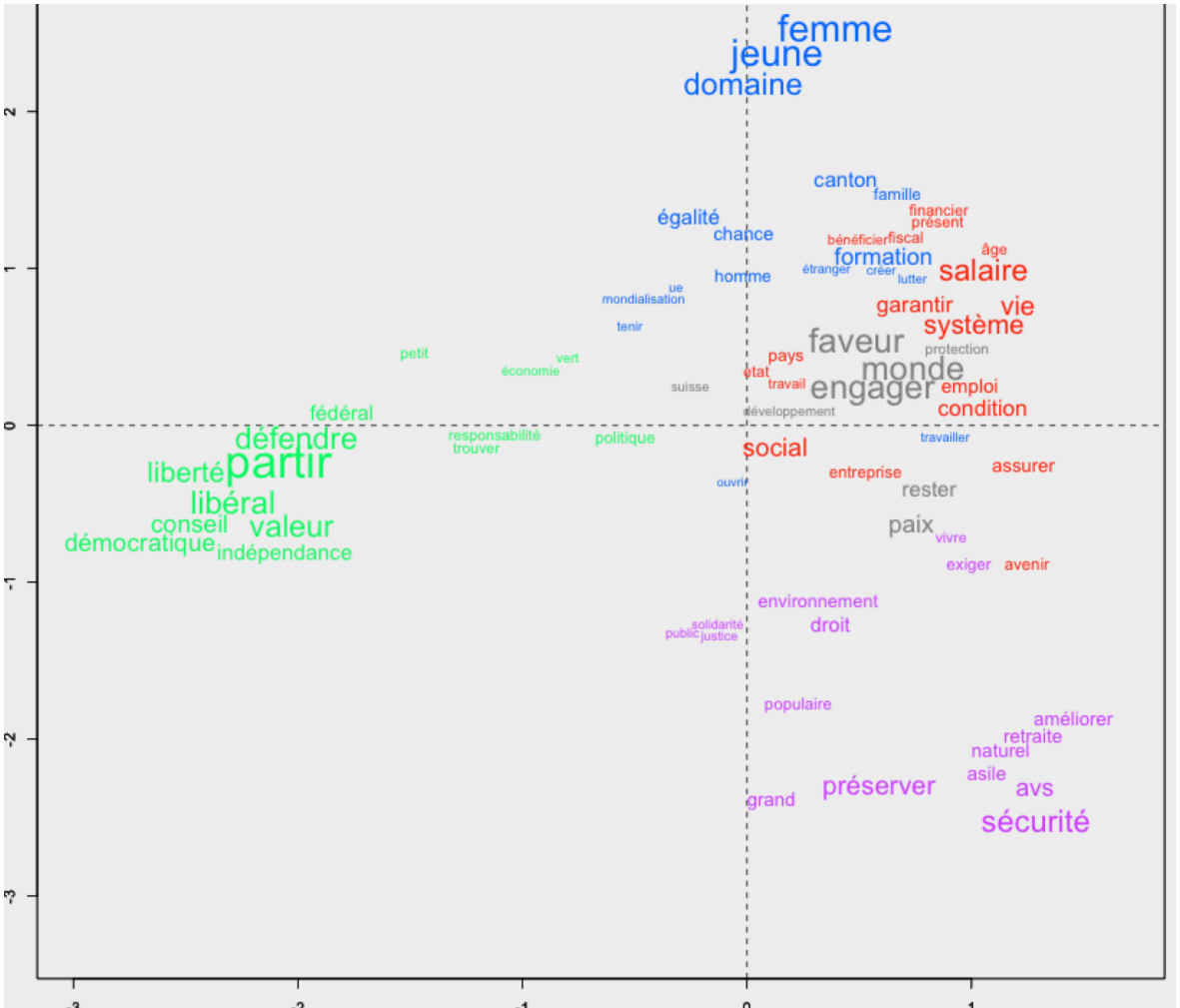
la politique de notre parti à la fois libéral et vert consiste à équilibrer le triangle formé par l environnement les services sociaux et l économie dans la perspective d un développement durable

comme l avenir ne peut être assuré que par un environnement préservé et par des finances saines nous n avons pas le droit de vivre aux dépens des générations futures

l etat doit créer les conditions qui permettront à nos descendants de vivre sans avoir à assumer des charges que notre époque leur aura léguées pour atteindre ce but nous misons essentiellement sur les instruments de l économie de marché et les taxes incitatives qui créent un effet modérateur

CDH (suite)

Enfin, une seconde ACP peut être produite, cette fois basée sur la `tdm` "termes" × "groupes", de taille $v \times K$, dont la case ig est codée "1" si le terme i apparaît dans (au moins) un segment classé en g, et "0" sinon. Cette analyse, très particulière ("itération" d'une information déjà exploitée), est propre à Iramuteq :



Analyses de similitude

Les "analyses de similitude" sont basées sur une matrice de similarité entre formes, qui définit le poids des arrêtes (entre deux formes-noeuds) d'un réseau non orienté. Les indices de similarité sont ceux du package `R proxy()` (à savoir : **cooccurrence**, `Braun-Blanquet`, `Chi-squared`, `correlation`, `cosine`, `Cramer`, `Dice`, `eDice`, `eJaccard`, `Fager`, `Faith`, `Gower`, `Hamman`, `Jaccard`, `Kulczynski1`, `Kulczynski2`, `Michael`, `Mountford`, `Mozley`, `Ochiai`, `Pearson`, `Phi`, `Phi-squared`, `Russel`, `simple matching`, `Simpson`, `Stiles`, `Tanimoto`, `Tschuprow`, `Yule`, `Yule2`), basés sur les "co-présences/co-absences" ; plus précisément :

- a_{ij} = nombre de segments contenant les formes i et j
- b_{ij} = nombre de segments contenant la forme i mais pas j
- c_{ij} = nombre de segments contenant la forme j mais pas i
- d_{ij} = nombre de segments ne contenant ni i ni j.

Iramuteq produit une visualisation en *arbre maximal*, constitué du sous-réseau connexe contenant le moins d'arrêtes possible (=arbre) dont la somme des similarités des arrêtes est la plus grande possible (=maximal).

Plusieurs fonctionalités du package `R igraph()` sont disponibles, en particulier ses algorithmes de *positionnement des noeuds-formes* (`layout`) et de *détection de communautés* (=groupes)

Gaphe de similarité : illustration