

Documentation Iramuteq

version 0.3 alpha 3

Manuel utilisateur

Pierre Ratinaud

Licence GNU FDL

Table des matières

1	Présentation d'iramuteq.....	3
1.1	R.....	3
1.2	Python.....	3
1.3	Lexique 3.....	3
2	Analyses de textes.....	3
2.1	Format des données en entrée.....	3
2.2	Ouverture d'un fichier texte.....	5
2.3	Traitements commun aux analyses.....	5
2.3.1	Nettoyage 1.....	5
2.3.2	Dictionnaire des expressions.....	6
2.3.3	Nettoyage 2.....	6
2.3.4	Lemmatisation.....	6
2.4	Statistiques textuelles.....	6
2.4.1	Description.....	6
2.4.2	Résultats.....	6
2.4.3	Fichiers en sortie.....	6
2.5	Comme Lexico.....	6
2.5.1	Options.....	7
2.5.2	Résultats.....	7
2.6	AFC sur UCI.....	7
2.6.1	Description.....	7
2.6.2	Options.....	7
2.6.3	Résultats.....	7
2.7	Classification.....	7
2.7.1	Méthode ALCESTE.....	7
2.7.1.1	Description.....	7
2.7.1.2	Options.....	9
2.7.1.3	Résultats.....	10
2.7.1.3.1	Options des profils.....	10
2.7.1.4	Fichiers en sortie.....	11
2.7.2	Par matrice des distances.....	12
2.7.2.1	Description.....	12
2.7.2.2	Options.....	12
2.7.2.3	Résultats.....	13
3	Analyses de tableaux de données.....	13
3.1	Format des données en entrée.....	13
3.2	Fréquences.....	13
3.3	Chi 2.....	13
3.4	Classification.....	13
3.5	Par matrice des distances.....	14
3.5.1	Méthode ALCESTE.....	14
3.6	AFCM.....	14
3.7	Graphes.....	14
4	Bibliographie.....	14
5	Annexes.....	14

1 Présentation d'iramuteq

Iramuteq est un logiciel d'analyse de textes et de tableaux de données. Il s'appuie sur le logiciel de statistique R (<http://www.r-project.org>), sur le langage python (<http://www.python.org>) et sur la base de données lexicales Lexique (<http://www.lexique.org>).

ATTENTION

Iramuteq est en cours de développement. Regardez les informations disponibles sur la page <http://repere.no-ip.org/logiciel/iramuteq> pour connaître la fiabilité des différentes analyses.

1.1 R

<http://www.r-project.org>

1.2 Python

<http://www.python.org>

1.3 Lexique 3

<http://www.lexique.org>

2 Analyses de textes

2.1 Format des données en entrée

Les fichiers d'entrée doivent être au format texte brut (.txt) et respecter les règles de formatage des corpus ALCESTE.

Dans ce formatage, l'unité de base est appelée « unité de contexte initiale » (UCI). Une UCI peut représenter un entretien, un article, un livre ou tout autre type de documents. Un corpus peut contenir une ou plusieurs UCI (mais au minimum une).

Les UCI sont introduites par quatre étoiles (****) suivies d'une série de variables étoilées séparées par un espace.



Une uci doit obligatoirement avoir au moins une variable étoilée

Il est possible de placer des variables étoilées à l'intérieur des corpus en les introduisant en début de ligne par un tiret et une étoile (-*). La ligne ne doit contenir que cette variable.

Il est possible d'introduire dans le corps du texte des formes qui seront traitées comme des variables étoilées. Il faut alors que ces formes commencent et se terminent par un _ . :

Exemple

```
texte texte _rire_ texte texte texte
```

Le texte contient, de préférence, les caractères de ponctuations.

Exemple d'un corpus sans thématique :

```
**** *var1_1 *var2_2
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte

**** *var1_2 *var2_3
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
```

Exemple d'un corpus avec thématique :


```
**** *var1_1 *var2_2
-*thematique1
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte

-*thematique2
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte


**** *var1_2 *var2_3
-*thematique1
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte

-*thematique2
```


```
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte texte texte texte
texte texte texte texte texte texte texte
```

 Dans un corpus avec thématique, tous les paragraphes d'une UCI doivent appartenir à une thématique. La construction suivante n'est donc pas possible :

```
**** *var1_1
texte texte texte texte texte
-*thematique1
texte texte texte texte texte texte
```

 Les variables étoilées et les thématiques introduites dans le corpus ne doivent pas contenir d'espaces ou de caractères spéciaux. Elles ne doivent contenir que des caractères parmi a-z, A-Z, 1-9 et des tirets bas (_).

```
*age 18 ans n'est pas un bon codage
*age_18 est un bon codage
*entretien_d'Emilie n'est pas un bon codage
*ent_emilie est un bon codage
```

 Les codages de la forme *variable_modalité doivent être privilégiés pour les variables illustratives. Ils permettent des analyses complémentaires. Exemple : *sex_h pour les hommes et *sex_f pour les femmes permet de repérer la variable sex et les modalités h et f.

2.2 Ouverture d'un fichier texte

Fichier → Ouvrir un corpus texte

Vous devez préciser l'encodage du fichier et la langue du corpus.

2.3 Traitements commun aux analyses

2.3.1 Nettoyage 1

Les corpus texte sont passés en minuscule. Tous les caractères qui ne sont pas dans la liste des caractères retenus sont remplacés par des espaces. Toutes les successions d'espaces ou de sauts de ligne sont remplacés par un espace ou un saut de ligne. Les apostrophes (') sont remplacées par des apostrophes (').

Caractères retenus : a-zA-Z0-9àÀâÄäÁÉÈèÊëËìíîïðÒòÔöÙùÛüÇç'ñ.:;?!?n*'_-

Cette liste devrait devenir paramétrable

2.3.2 Dictionnaire des expressions

Le dictionnaire des expressions contient des expressions ou des mots contenant des tirets (-) ou des apostrophes ('). Il permet de traiter ces expressions comme un tout. Par exemple, le mot aujourd'hui sera traité comme la forme aujourd'hui. L'expression « vis-à-vis » sera transformée en vis_à_vis ». Le dictionnaire des expressions est disponible dans le répertoire d'installation d'iramuteq, dans le sous-répertoire « dictionnaire ».

L'utilisation de ce dictionnaire est optionnel.

2.3.3 Nettoyage 2

Les apostrophes (') et les tirets (-) sont remplacés par des espaces.

2.3.4 Lemmatisation

Les verbes sont réduits à l'infinitif, les noms et les adjectifs sont réduits au masculin singulier.

Exemple :

mangé, mangeons, mangera → manger
professionnelles, professionnelle, professionnels, professionnel → professionnel

La lemmatisation est optionnelle.

2.4 Statistiques textuelles

2.4.1 Description

Analyse de texte → Statistiques textuelles

Cette analyse propose des statistiques simples sur les corpus texte : effectifs de toutes les formes, effectifs des formes actives et supplémentaires, liste des hapax.

2.4.2 Résultats

Les résultats se présentent sous forme de listes. Un clic droit sur une forme permet d'accéder aux formes associées et à un concordancier.

2.4.3 Fichiers en sortie

Répertoire de sortie	NomDuCorpus_Stat_x
Fichiers en sortie :	
total.csv	Toutes les formes et leurs effectifs
formes supplémentaires.csv	Les formes supplémentaires et leurs effectifs
formes actives.csv	Les formes actives et leurs effectifs
hapax.csv	Les hapax

2.5 Comme Lexico

Analyse de texte → Comme lexico

Reproduit une des analyses du logiciel Lexico (<http://www.tal.univ-paris3.fr/lexico/>).

Il s'agit de la description d'un tableau de contingence qui croise formes et groupes d'UCI. Les

groupe d'UCI sont sélectionnées en fonction de variables illustratives. L'objectif est de comparer ces groupes d'UCI.

2.5.1 Options

L'effectif minimum d'une forme sélectionnée peut être paramétré. Par défaut, cette valeur est à 10.

2.5.2 Résultats

Les mêmes résultats sont produits sur les formes et sur les types.

- Onglet Spécificités :
Présente l'exposant du seuil de significativité du chi2 qui mesure la force du lien entre la forme et la variable. Par exemple, si une forme est liée à une variable avec un chi2 dont le seuil de significativité est 0,001, la valeur 3 sera notée car $0,001 = 10^{-3}$.
- Onglet Effectifs :
Les effectifs
- Onglet Effectifs relatifs :
Les effectifs relatifs en 1000ème

2.6 AFC sur UCI

2.6.1 Description

Analyse de texte → AFC sur UCI

Produit une Analyse factorielle des correspondances sur un tableau de contingence qui croise formes actives et UCI.

Cette analyse est immature.

2.6.2 Options

Pas d'options pour l'instant

2.6.3 Résultats

Trois graphiques d'AFC sont proposés : formes actives, formes supplémentaire et variables étoilées.

2.7 Classification

2.7.1 Méthode ALCESTE

2.7.1.1 Description

Analyse de texte → Classification → méthode ALCESTE


Cette analyse propose une classification hiérarchique descendante selon la méthode ALCESTE (Reinert, 1983, 1986, 1991). La classification peut être menée sur les UCI (classification simple sur UCI) ou sur des segments de textes (Unité de Contexte Élémentaire : UCE). Les classifications sur les UCE peuvent être conduites directement sur celles-ci (classification simple sur UCE) ou sur deux tableaux proposant des regroupements de segments de texte (Unité de Contexte : UC) qui diffèrent par le nombre de variables actives (et donc d'UCE) regroupées par ligne (classification double sur UC).

Voir le détail de la classification ALCESTE en annexe.

2.7.1.2 Options


Utiliser le dict. des expressions	<input checked="" type="radio"/> oui <input type="radio"/> non
Lemmatisation	<input checked="" type="radio"/> oui <input type="radio"/> non
Classification	<input checked="" type="radio"/> double sur UC <input type="radio"/> simple sur UCE <input type="radio"/> simple sur UCI
taille uc 1	10
taille uc 2	12
nombre de classes terminales de la phase 1	10
Nombre de d'occurrences par uce (0 = automatique)	0
Nombre minimum d'uce par classe (0 = automatique)	0
Fréquence minimum d'une forme analysée (2 = automatique)	2
Nombre maximum de formes analysées	1500
Configuration des clés d'analyse	Préférences
<input type="button" value="Annuler"/> <input type="button" value="Valeurs par défaut"/> <input type="button" value="Valider"/>	

- Utiliser le dictionnaire des expressions :
voir dictionnaire des expressions
- Lemmatisation :
voir lemmatisation
- Classification :
 - double sur UC : la classification est menée sur deux tableaux qui regroupent sur chaque ligne un certain nombre d'UCE en fonction du nombre de formes actives par ligne des paramètres « taille uc 1 » et « taille uc2 »
 - simple sur UCE : la classification est menée sur les UCE
 - simple sur UCI : la classification est menée sur les UCI
- Nombre de classes terminales de la phase 1 :
Détermine le nombre de classes de la première partie de la classification.
- Nombre d'occurrences par UCE :
Permet de choisir la taille des UCE en fonction du nombre d'occurrences qu'elles regroupent. Par défaut, ce calcul est automatique et la taille des UCE est fonction de la taille du corpus. Plus le corpus est important, plus les UCE seront longues. Dans tous les cas, la ponctuation est prise en compte dans le découpage ; la valeur du nombre d'occurrences est donc « un objectif à atteindre » et pas une valeur stricte.

 Vérifiez la taille des UCE dans vos analyses, elle peut rapidement

devenir trop importante. Des UCE d'une cinquantaine d'occurrences correspondent à environ deux ou trois lignes de texte.

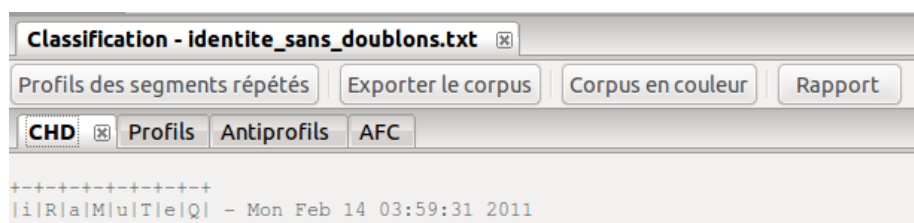
- **Nombre minimum d'UCE par classe :**
Permet de choisir le nombre minimum d'UCE par classe. Par défaut, seules les classes regroupant $1/(10 \times \text{le nombre de classes terminales de la phase 1})$ des UCE pour une classification simple, et $1/(20 \times \text{le nombre de classes terminales de la phase 1})$ des UCE pour une classification double, seront retenues.
- **Nombre maximum de formes analysées :**
Par défaut, les 1500 formes actives les plus fréquentes et les 1500 formes supplémentaires les plus fréquentes seront retenues. Une forme doit avoir au minimum une fréquence de 4 pour être retenue. Si le corpus a moins de 1500 formes, toutes les formes avec une fréquence supérieure à 3 seront retenues.

 Ce paramètre a une forte incidence sur la taille des tableaux analysés et donc sur la quantité de mémoire de l'ordinateur mobilisée. Si votre ordinateur n'a pas assez de mémoire pour analyser un corpus, essayez de baisser ce paramètre. Si votre ordinateur possède « suffisamment » de mémoire pour le corpus et que le corpus possède plus de 1500 formes de fréquence > 3, n'hésitez pas à l'augmenter.

- **Configuration des clés d'analyse :**
Voir clés d'analyse

2.7.1.3 Résultats

Les résultats directement disponibles présentent un résumé de la classification (onglet CHD) les profils des classes (onglet Profils), les antiprofils des classes (onglet Antiprofils) et une analyse factorielle des correspondances menées sur le tableau de contingence croisant formes et classes (onglet AFC).



2.7.1.3.1 Options des profils

A partir d'un clic droit sur une ligne du profil, plusieurs outils complémentaires sont proposés :

Formes associées

Chi2 par classe

Concordancier

Outils du CNRTL

Graph de la classe

Segments répétés

UCE caractéristiques

- **Formes associées :** renvoie les mots associés à la forme sélectionnée et leurs effectifs.
- **Chi2 par classe :** crée un graphique qui présente le chi2 d'association de la forme à chacune des classes. Plusieurs formes peuvent être sélectionnées en même temps.
- **Concordancier :** propose le concordancier de la (ou des) forme(s) sélectionnée(s). Ce concordancier est disponible pour les UCE de la classe, les UCE classées ou toutes les UCE du corpus.

- Outils du CNRTL : interroge la base de données du Centre Nationale de Ressources Textuelles et Lexicales (<http://www.cnrtl.fr/>) à partir de la forme sélectionnée (nécessite d'être connecté à Internet). Permet d'obtenir une définition (Lexicographie), les synonymes (Synonymie), les Antonymes (Antonymie), l'étymologie (Etymologie) et la morphologie (Morphologie) de la forme. Les résultats s'affichent dans le navigateur internet par défaut du système.
- Graph de classe : indépendant de la ligne sélectionnée. Il s'agit d'une analyse de similitude menée sur un tableau absence/présence (0/1) qui croise les unités choisies en ligne (UCI ou UCE) et les formes actives de la classe en colonne. La matrice de similitude est construite sur les colonnes (les formes actives de la classe). Par défaut, l'indice de similitude utilisé est la cooccurrence. Les résultats se présentent sous la forme d'un graphe de similitude réduit à un arbre maximum.
Voir « analyse de similitude » pour plus de détails.
- Segments répétés : indépendant de la ligne sélectionnée. Effectifs et tailles des segments répétés de la classe. Préférez les profils des segments répétés.
- UCE caractéristiques : indépendant de la ligne sélectionnée. Liste les UCE caractéristiques de la classe. Deux scores sont proposés :
 - absolu : les UCE sont classées en fonction de la somme de chi2 de liaison à la classe des formes actives qu'elles contiennent.
 - Relatif : les UCE sont classées en fonction de la moyenne des chi2 de liaison à la classe des formes actives qu'elles contiennent.

Dans le cas d'une classification sur UCI, remplacez UCE par UCI dans la description précédente.

2.7.1.4 Fichiers en sortie

Répertoire de sortie	NomDuCorpus_alceste_x
Fichiers en sortie :	
TableUc1.csv	Le tableau UC1/formes ou UCI/formes ou UCE/formes
TableUc2.csv	Le tableau UC2/formes
listeUCE1.csv	Tableau uce;uc pour les UC1
listeUCE2.csv	Tableau uce;uc pour les UC2
profiles.csv	Profils des classes
antiprofiles.csv	Antiprofiles des classes
info.txt	Résumé de la classification
uce.csv	Les uce par classe
arbre_1.png	Dendrogramme de la première CHD
arbre_2.png	Dendrogramme de la seconde CHD
dendro1.png	Dendrogramme final sur UC1
dendro2.png	Dendrogramme final sur UC2
classe_mod.csv	Tableau de contingence formes actives/classes
RData.RData	Résultats dans R
tablesup.csv	Tableau de contingence formes supplémentaires/classes
tableet.csv	Tableau de contingence variables illustratives/classes
SbyClasseOut.csv	Les uce par classe

chisqtable.csv	Chi2 d'association de chaque formes aux classes
ptable.csv	Seuil de significativité des chi2 d'associations de chaque forme aux classes.
Analyse.ira	Fichier Analyse : permet de ré-ouvrir une analyse.
AFC2DL.png	Graph AFC : Variables actives - coordonnées - facteurs 1 / 2
AFC2DSL.png	Graph AFC : variables supplémentaires - coordonnées - facteurs 1 / 2
AFC2DEL.png	Graph AFC : Variables illustratives - Coordonnées - facteur 1 / 2
AFC2DCL.png	Graph AFC : Classes - Coordonnées - facteur 1 / 2
AFC2DCoul.png	Graph AFC : Variables actives - Corrélation - facteur 1 / 2
AFC2DCoulSup.png	Graph AFC : Variables supplémentaires - Corrélation - facteur 1 / 2
AFC2DCoulEt.png	Graph AFC : Variables illustratives - Corrélations - facteur 1 / 2
AFC2DCoulCl.png	Graph AFC : Classes - Corrélations - facteurs 1 / 2
liste_graph_afc.txt	Liste de s graphiques de l'onglet AFC
liste_graph_chd.txt	Liste de graphiques de l'onglet CHD
afc_row.csv	Résultats de l'AFC ; Coordonnées, corrélation, MASS, contribution des formes : voir le manuel de la librairie ca de R pour plus de détail.
afc_col.csv	Résultats de l'AFC ; Coordonnées, corrélation, MASS, contribution des classes : voir le manuel de la librairie ca de R pour plus de détail.
afc_facteur.csv	Résultats de l'AFC ; Valeurs propres, Pourcentage d'inertie extraite et Pourcentage cumulé des facteurs.
segments_classes.csv	Tableau de contingence segments répétés/classes
prof_segments.csv	Profils des segments répétés
antiprof_segments.csv	Antiprofils des segments répétés
profil_type.csv	Profils des types
antiprof_type.csv	Antiprofils des types
type_cl.csv	Tableau de contingence types/classes
analyse.db	Base de données contenant les résultats

2.7.2 Par matrice des distances

2.7.2.1 Description

Produit une classification à partir d'une matrice de distance construite à partir d'un tableau absence présence qui croise l'unité choisie (UCI ou UCE) et les formes actives. La matrice de distance est construite à partir des lignes de ce tableau (les unités).

2.7.2.2 Options

Lemmatisation	<input checked="" type="radio"/> oui <input type="radio"/> non
Utiliser le dict. des expressions	<input checked="" type="radio"/> oui <input type="radio"/> non
Méthode de construction de la matrice des distances	binary
Analyse	<input checked="" type="radio"/> k-means (pam) <input type="radio"/> fuzzy (fanny)
Classification	<input checked="" type="radio"/> sur UCE <input type="radio"/> sur UCI
Nombre maximum de formes analysées	1500
Nombre de formes par uce (0 = automatique)	0
Nombre de classes	4
Configuration des clés d'analyse	Préférences
<input type="button" value="Annuler"/> <input type="button" value="Valeurs par défaut"/> <input type="button" value="Valider"/>	

- Méthode de construction de la matrice des distances :
Permet de choisir l'indice de distance utilisé dans la matrice des distances. Voir la documentation de la fonction `dist` de R plus de détail sur ces indices. Le fichier traité étant de type absence/présence, seul l'indice « binary » est pertinent. Il s'agit de la distance de Jaccard.
- Analyse :
Deux algorithmes de classification sont proposés : « k-means » par l'intermédiaire de la fonction « pam » la librairie `cluster` de R et « fuzzy clustering » par l'intermédiaire de la fonction « fanny » de la librairie `cluster`. Voir la documentation de la librairie `cluster` pour plus de détails : <http://cran.r-project.org/web/packages/cluster/cluster.pdf>
- Classification :
Permet de choisir les unités en ligne : UCE ou UCI
- Nombre maximum de formes analysées :
Voir Méthode ALCESTE → Options
- Nombre de classes :
Nombre de classes souhaité. Par défaut, 4 classes seront construites.

2.7.2.3 Résultats

Les résultats se présentent comme les résultats de la méthode ALCESTE. Voir méthode ALCESTE → résultats.

3 Analyses de tableaux de données


3.1 Format des données en entrée

Les tableaux de données doivent être du type individus/caractères. Les variables doivent être préférentiellement présentées sous la forme variable_modalité. Dans le cadre des classifications ALCESTE, le tableau d'entrée est transformé en tableau absence/présence (0/1). Il n'est donc généralement pas acceptable que deux colonnes distinctes contiennent des modalités formatées de la même façon. Une étoile peut être introduite devant les modalités qui seront utilisées comme variables illustratives. Cette présentation correspond à un corpus « formaté ».

exemple :

id	var1	var2	...
1	*var1_mod1	var2_mod2	...
2	*var1_mod2	var2_mod1	...
3	*var1_mod3	var2_mod3	...
4	*var1_mod2	var2_mod4	...
5	*var1_mod3	var2_mod6	...
...

Les fichiers acceptés en entrée doivent être au format .xls (Microsoft Excel 97/2003), .csv ou .ods (openoffice, libreoffice, etc...).

 Tous les fichiers transmis à R sont au format .csv avec le ';' comme séparateur de champs. IL EST DONC INDISPENSABLE QUE LE FICHIER EN ENTREE NE CONTIENNE AUCUN ';'.

De façon plus générale, il faut éviter les caractères en dehors des lettres (a-z), des chiffres (0-9) et du tiret bas (_).

3.2 Fréquences

3.3 Chi 2

3.4 Classification

3.5 Par matrice des distances

3.5.1 Méthode ALCESTE

3.6 AFCM

3.7 Graphes

4 Bibliographie

5 Annexes