Indexation automatique par termes-clés en domaines de spécialité

Adrien BOUGOUIN

Thèse de doctorat soutenue le
27 Octobre 2015

Jury

Président :	**Marc GELGON**, Professeur des universités, Université de Nantes
Rapporteurs :	**Brigitte GRAU**, Professeur des universités, ENSIIE
	Jacques SAVOY, Professeur des universités, Université de Neuchâtel
Examinatrice :	**Fabienne MOREAU**, Maître de conférences, Université de Rennes
Directrice :	**Béatrice DAILLE**, Professeur des universités, Université de Nantes
Encadrant :	**Florian BOUDIN**, Maître de conférences, Université de Nantes

UNIVERSITÉ DE NANTES

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

TermITH

Accès à l'information des documents numériques

Représentation de l'information numérique en domaines de spécialité

Termes-clés (mots-clés)

- Unités textuelles (mots et expressions)
- Décrivent le contenu principal d'un document
- Utiles pour la Recherche d'Information (RI) :
  - Indexation de document
  - Expansion de requête
  - Résumé automatique

Comment identifier automatiquement les termes-clés d'un document ?

1. Dans le contexte général
2. En domaines de spécialité

$\Rightarrow$ **Indexation (automatique) par termes-clés**

Météo du 19 **août 2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Étude préliminaire de la **céramique non tournée micacée** du bas Languedoc occidental : **typologie**, **chronologie** et aire de **diffusion**

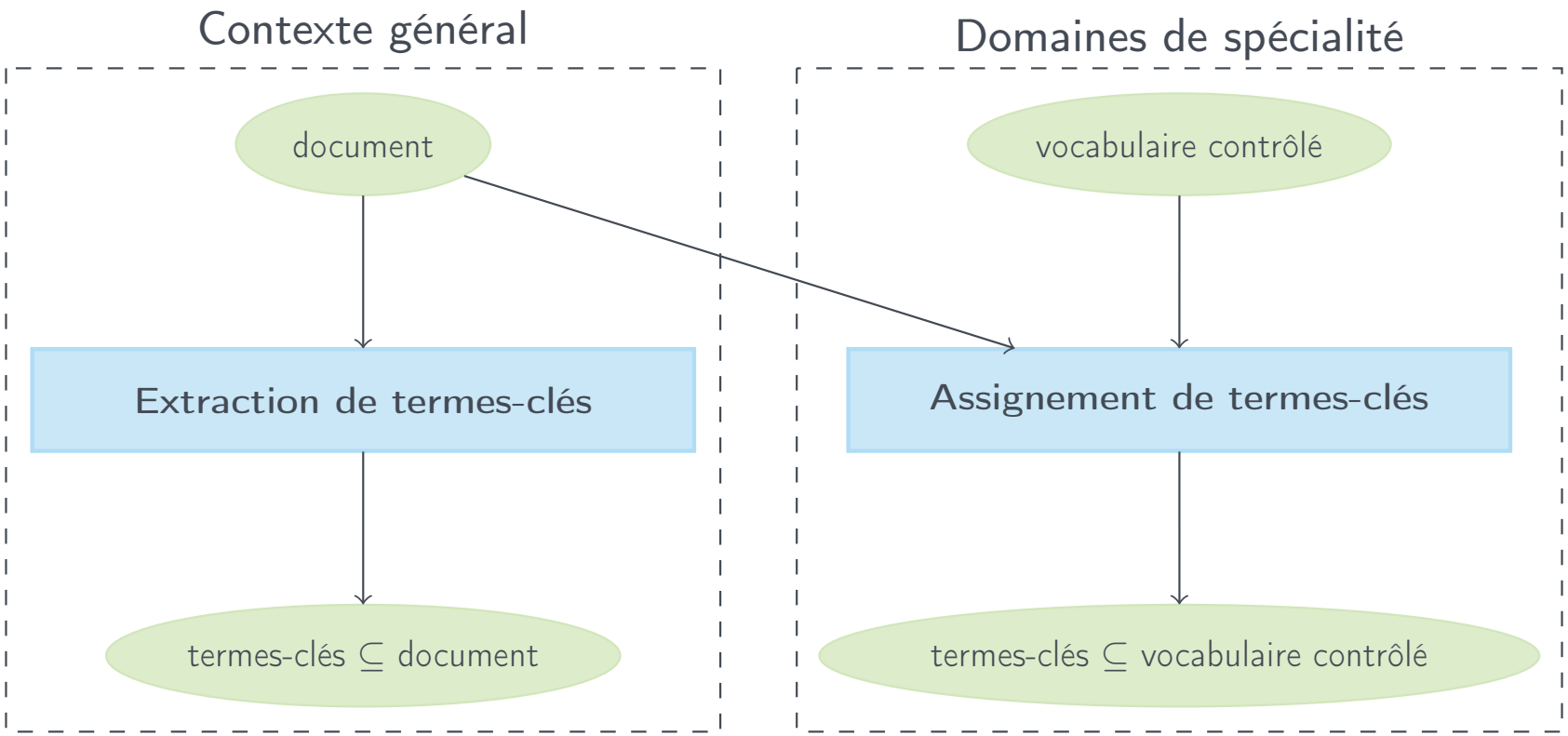
L'étude présente une variété de **céramique non tournée** dont la **typologie** et l'analyse des **décors** permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le **décor** effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de **fouilles anciennes** menées sur les **oppidums du Cayla** à **Mailhac** (**Aude**) et de **Mourrel-Ferrat** à **Olonzac** (**Hérault**). La carte de **répartition** fait état d'**échanges** ou de **commerce** à l'échelon macrorégional rarement mis en évidence pour de la **céramique non tournée**. S'il est difficile de statuer sur l'origine des **décors**, il semble que la **production** s'insère dans une ambiance celtisante. La **chronologie** de cette **production** se situe dans le deuxième **âge du Fer**. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence

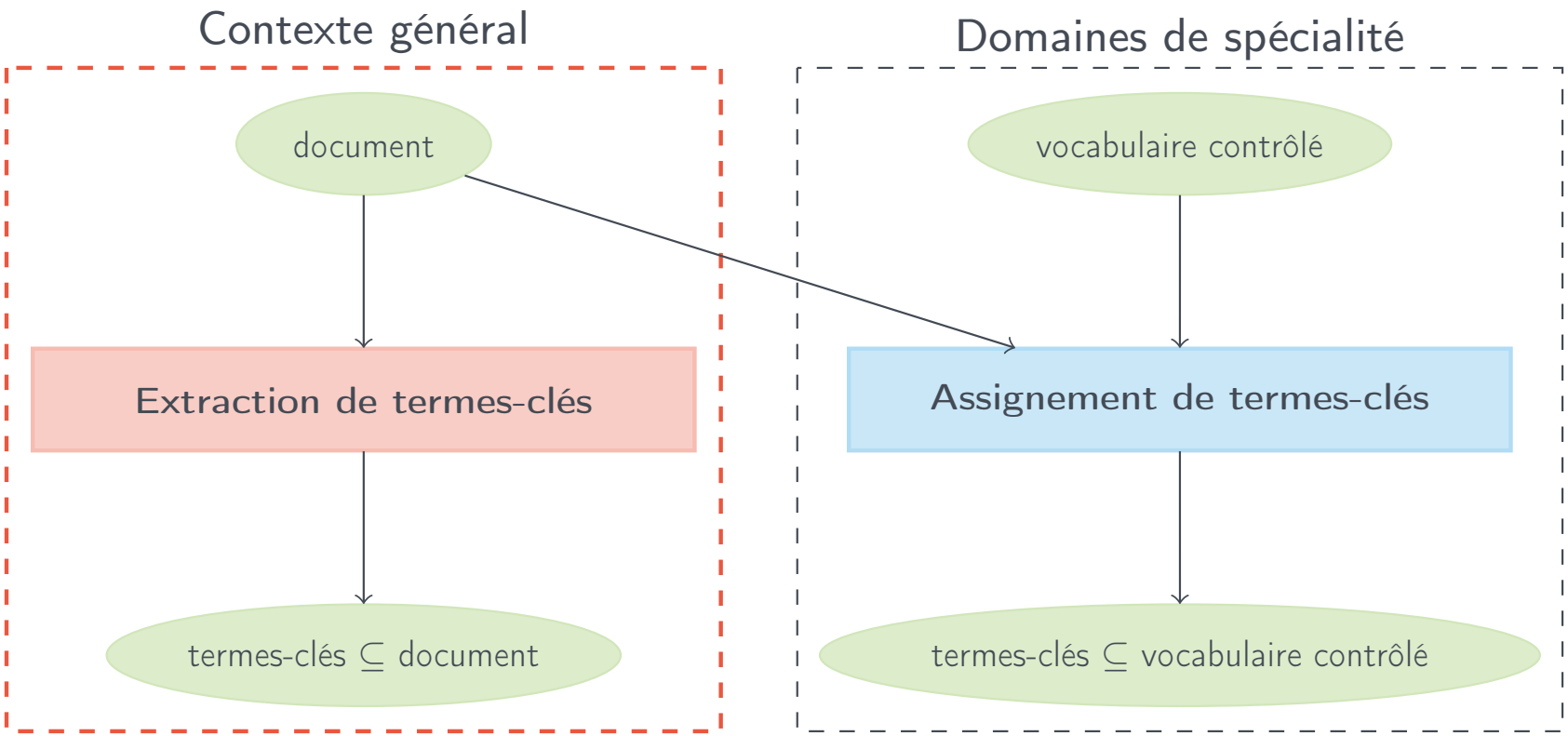
Mailhac ; **Aude** ; **Mourrel-Ferrat** ; **Olonzac** ; **Hérault** ; **céramique** ; **typologie** ; **décor** ; **chronologie** ; **diffusion** ; **production** ; **commerce** ; **répartition** ; **oppidum** ; **analyse** ; **fouille ancienne** ; **le Cayla** ; **micassé** ; **céramique non-tournée** ; **echange** ; **age du Fer** ; La Tène ; Europe ; France ; celtes ; distribution ; cartographie ; habitat ; site fortifié ; identification ; étude du matériel

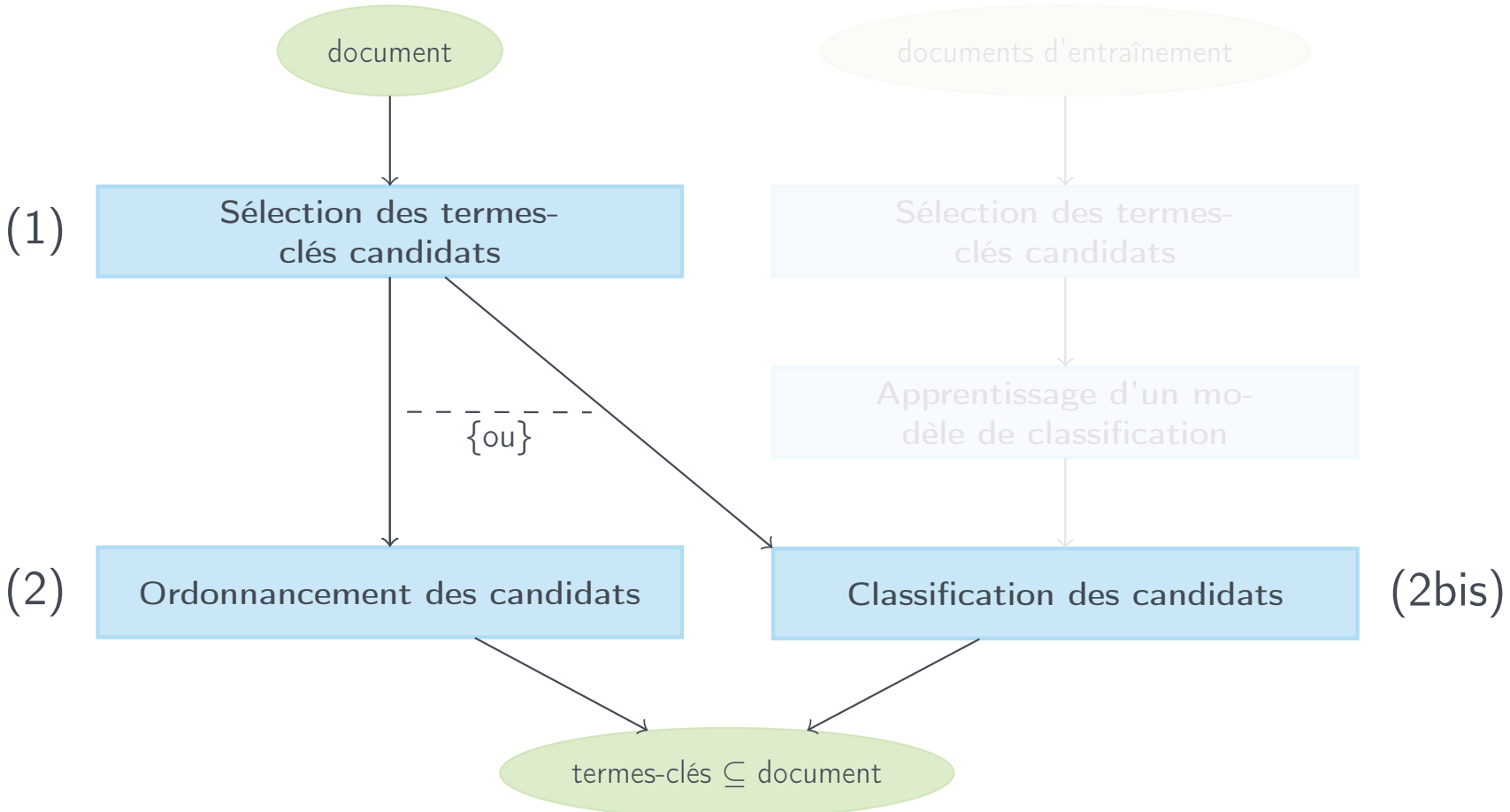
Plan

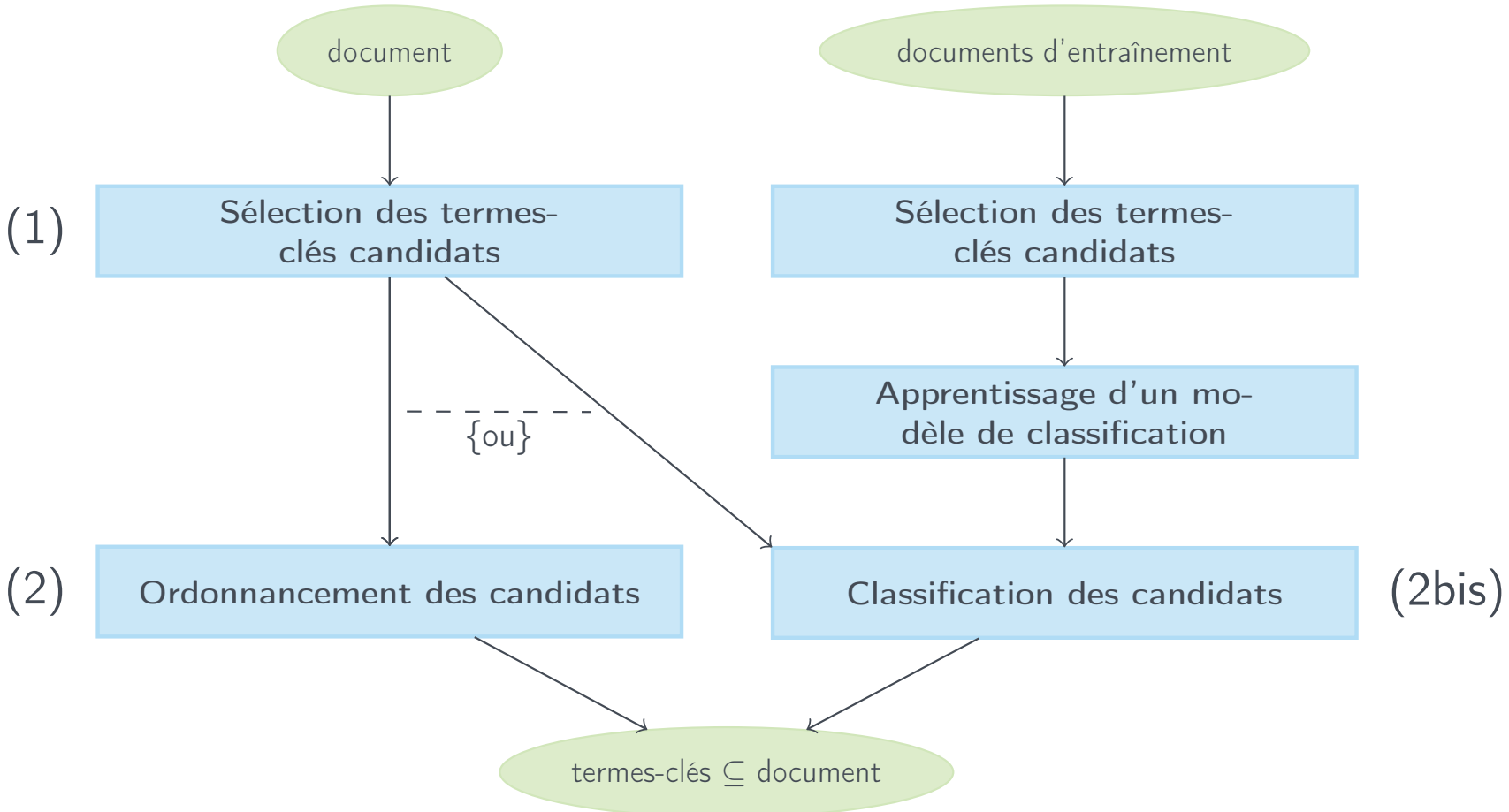
Indexation par termes-clés

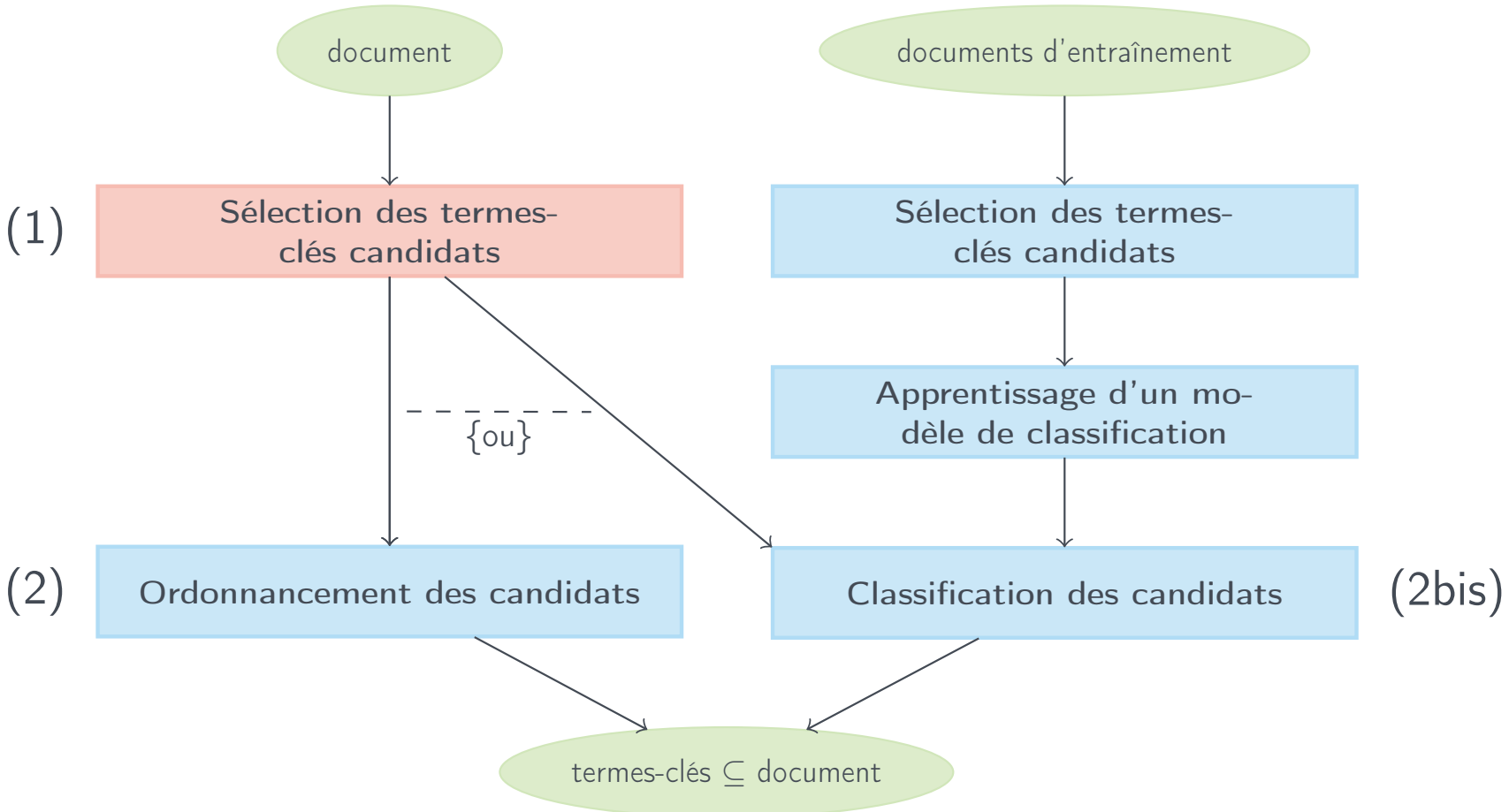


Indexation par termes-clés




document
documents d'entraînement
(1)
Sélection des termes-clés candidats
Sélection des termes-clés candidats
Apprentissage d'un modèle de classification
{ou}
(2)
Ordonnancement des candidats
Classification des candidats
(2bis)
termes-clés ⊆ document


document
documents d'entraînement
(1)
Sélection des termes-clés candidats
Sélection des termes-clés candidats
Apprentissage d'un modèle de classification
{ou}
(2)
Ordonnancement des candidats
Classification des candidats
(2bis)
termes-clés ⊆ document


document
documents d'entraînement
(1)
Sélection des termes-clés candidats
Sélection des termes-clés candidats
Apprentissage d'un modèle de classification
{ou}
(2)
Ordonnancement des candidats
Classification des candidats
(2bis)
termes-clés ⊆ document

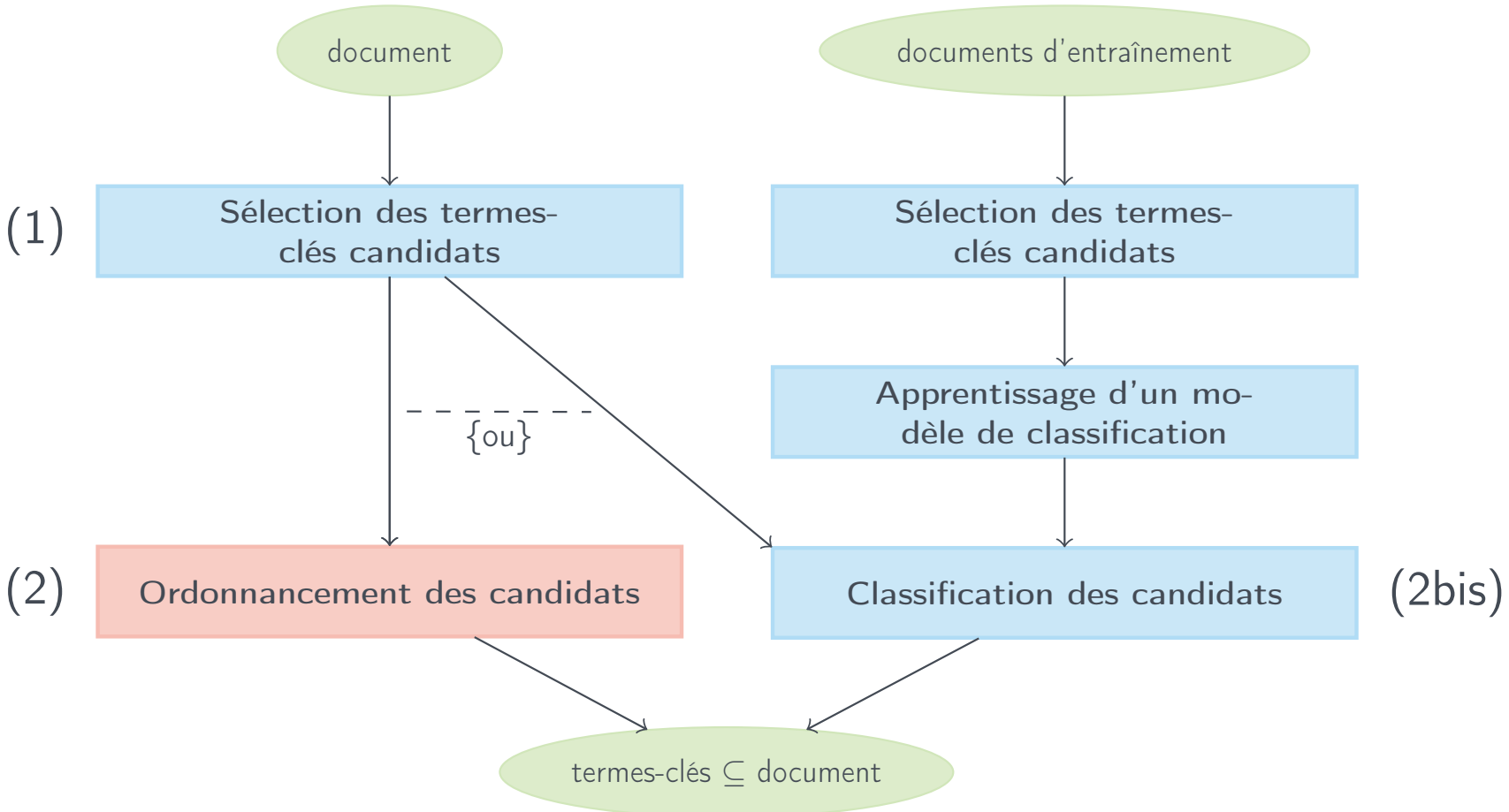
Deux principales méthodes :

- N-grammes
- Séquences grammaticalement définies

Exemple

« À l'exception de la province de Luxembourg, en alerte jaune, l'ensemble de la Belgique est en vigilance orange à la canicule. »

- N-grammes ($\{1..3\}$-grammes) :
  ($N = 1$) exception ; province ; Luxembourg ; alerte ; jaune ; ensemble ; Belgique ; vigilance ; orange ; canicule ;
  ($N = 2$) alerte jaune ; vigilance orange ;
  ($N = 3$) province de Luxembourg
- Séquences grammaticalement définies (`/(NOM | ADJ)+/`) :
  exception ; province ; Luxembourg ; alerte jaune ; ensemble ; Belgique ; vigilance orange ; canicule


document
documents d'entraînement
(1)
Sélection des termes-clés candidats
Sélection des termes-clés candidats
Apprentissage d'un modèle de classification
{ou}
(2)
Ordonnancement des candidats
Classification des candidats
(2bis)
termes-clés ⊆ document

Objectif

Déterminer les termes-clés candidats les plus importants

Ordonnancement des candidats selon diverses approches :

- Statistiques (Salton et al., 1975, TF-IDF)
- Par groupement distributionnel (Matsuo et Ishizuka, 2004)
- **À base de graphe** (Mihalcea et Tarau, 2004, TextRank)

Météo du 19 **août 2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Météo du 19 **août** **2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte** **orange**, tandis que le sud a était placé en **alerte** rouge.

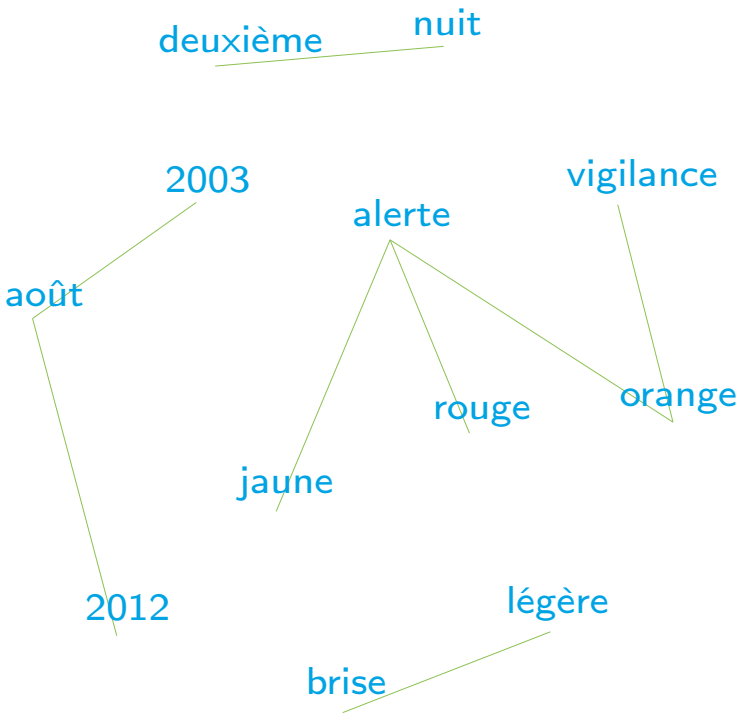
En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus** **chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus** **chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

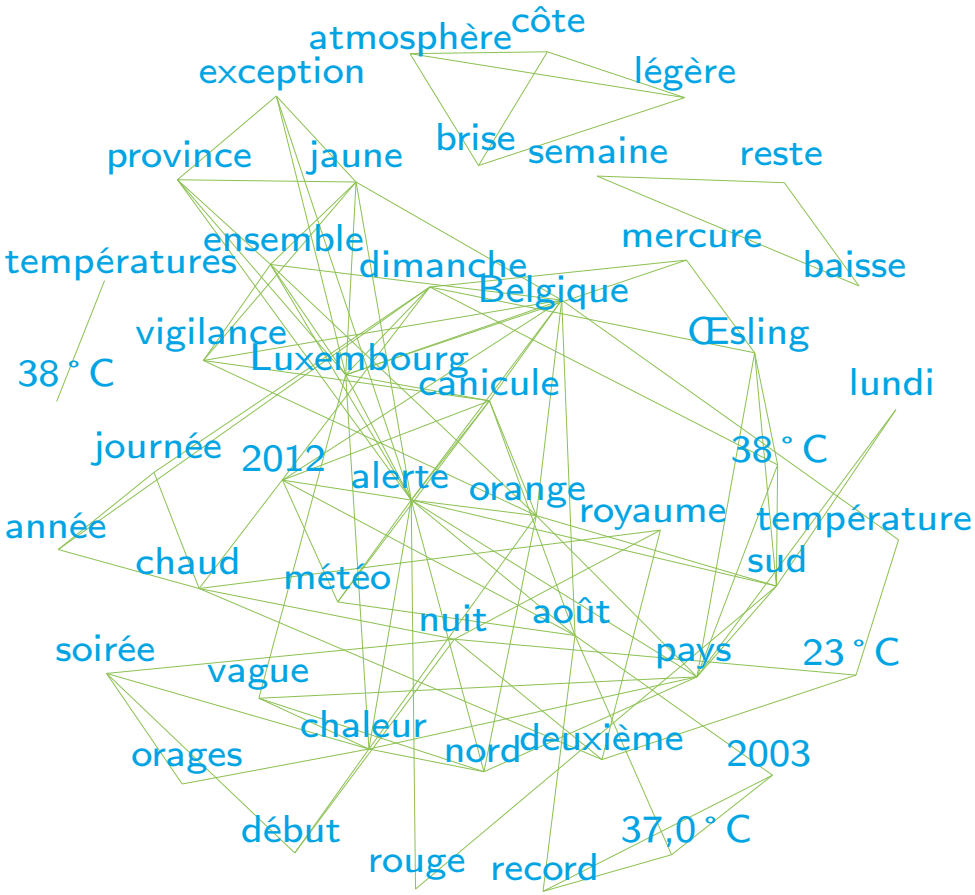
Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

TextRank
(Mihalcea et Tarau, 2004)

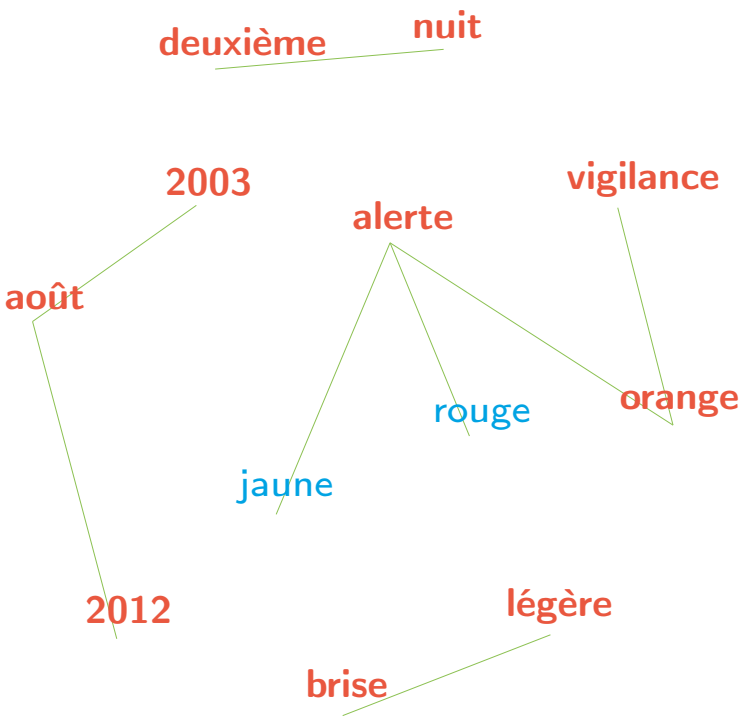


SingleRank
(Wan et Xiao, 2008)

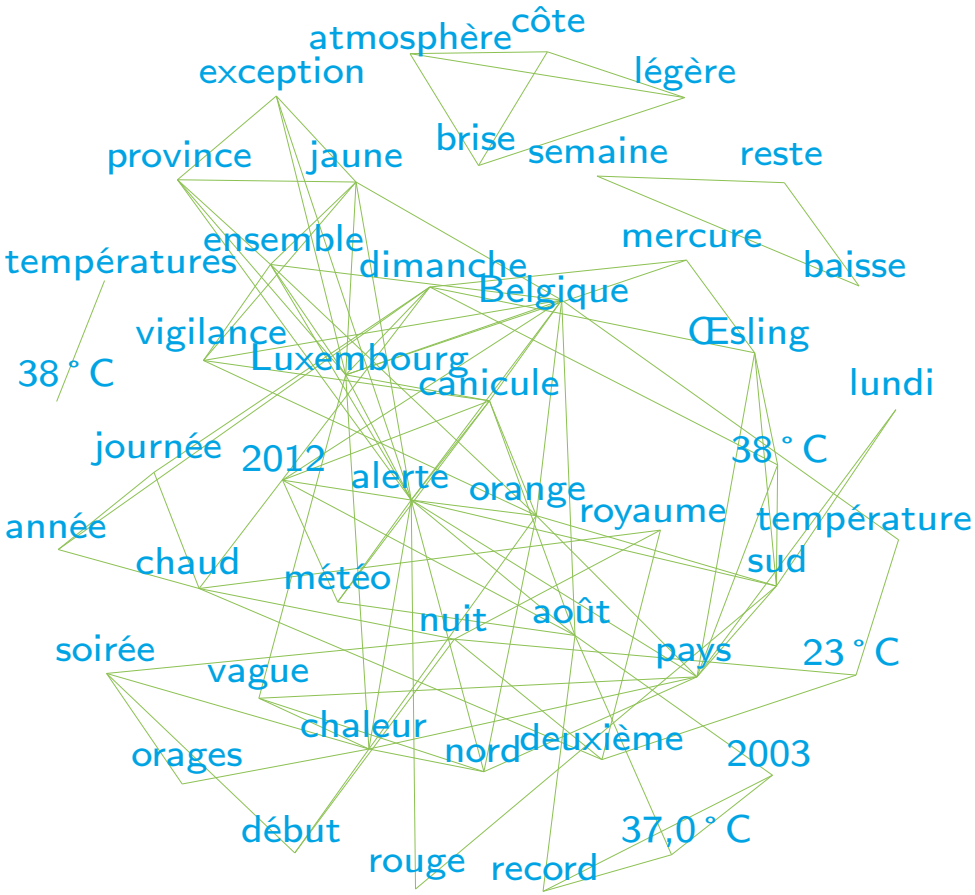


$$\text{Importance}(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times \text{Importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)}$$

TextRank
(Mihalcea et Tarau, 2004)



SingleRank
(Wan et Xiao, 2008)



$$\text{Importance}(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times \text{Importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)}$$

Météo du 19 août 2012 : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Météo du 19 **août 2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

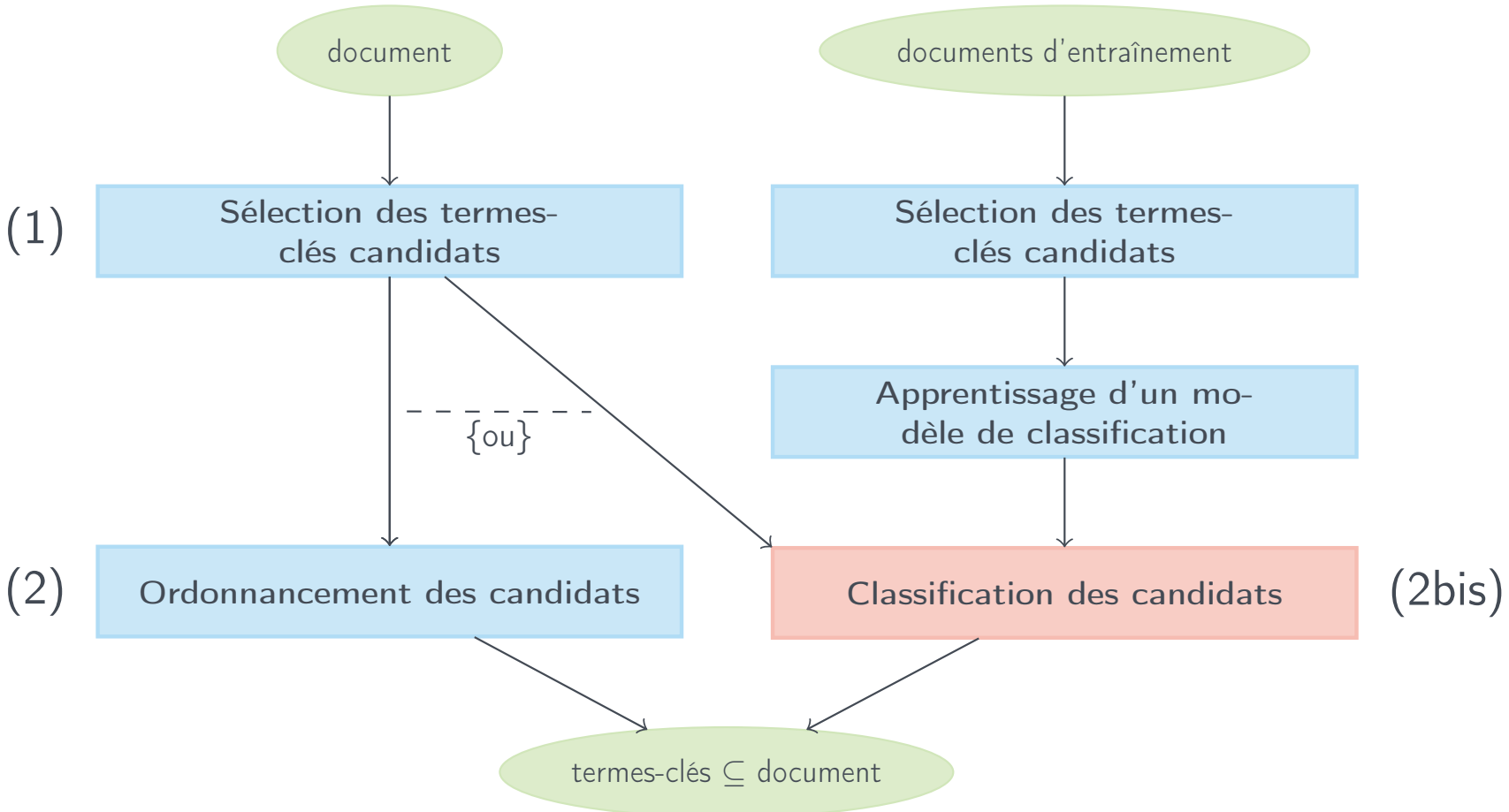
Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Rang	TextRank	SingleRank
01	août 2012	alerte orange
02	août 2003	alerte jaune
03	alerte orange	alerte rouge
04	vigilance orange	alerte
05	deuxième nuit	deuxième nuit
06	légère brise	août 2012
07		août 2003
08		vigilance orange
09		légère brise
10		Luxembourg

Termes-clés

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**


document
documents d'entraînement
(1)
Sélection des termes-clés candidats
Sélection des termes-clés candidats
Apprentissage d'un modèle de classification
{ou}
(2)
Ordonnancement des candidats
Classification des candidats
(2bis)
termes-clés ⊆ document

Objectif

Apprendre à reconnaître les termes-clés parmi les candidats

Classification des candidats selon divers critères (traits) :

- Fréquenciels
  - TF-IDF (Witten et al., 1999, KEA)
- Positionnels
  - Première position (Witten et al., 1999, KEA)
  - Apparition dans une section particulière (Nguyen et Kan, 2007)
- Linguistiques
  - Catégorie grammaticale (Nguyen et Kan, 2007)

Classification à l'aide de techniques d'apprentissage automatique :

- Classification naïve bayesienne (Witten et al., 1999, KEA)
- Réseau de neurones (Sarkar et al., 2010)

Objectif

Apprendre à reconnaître les termes-clés parmi les candidats

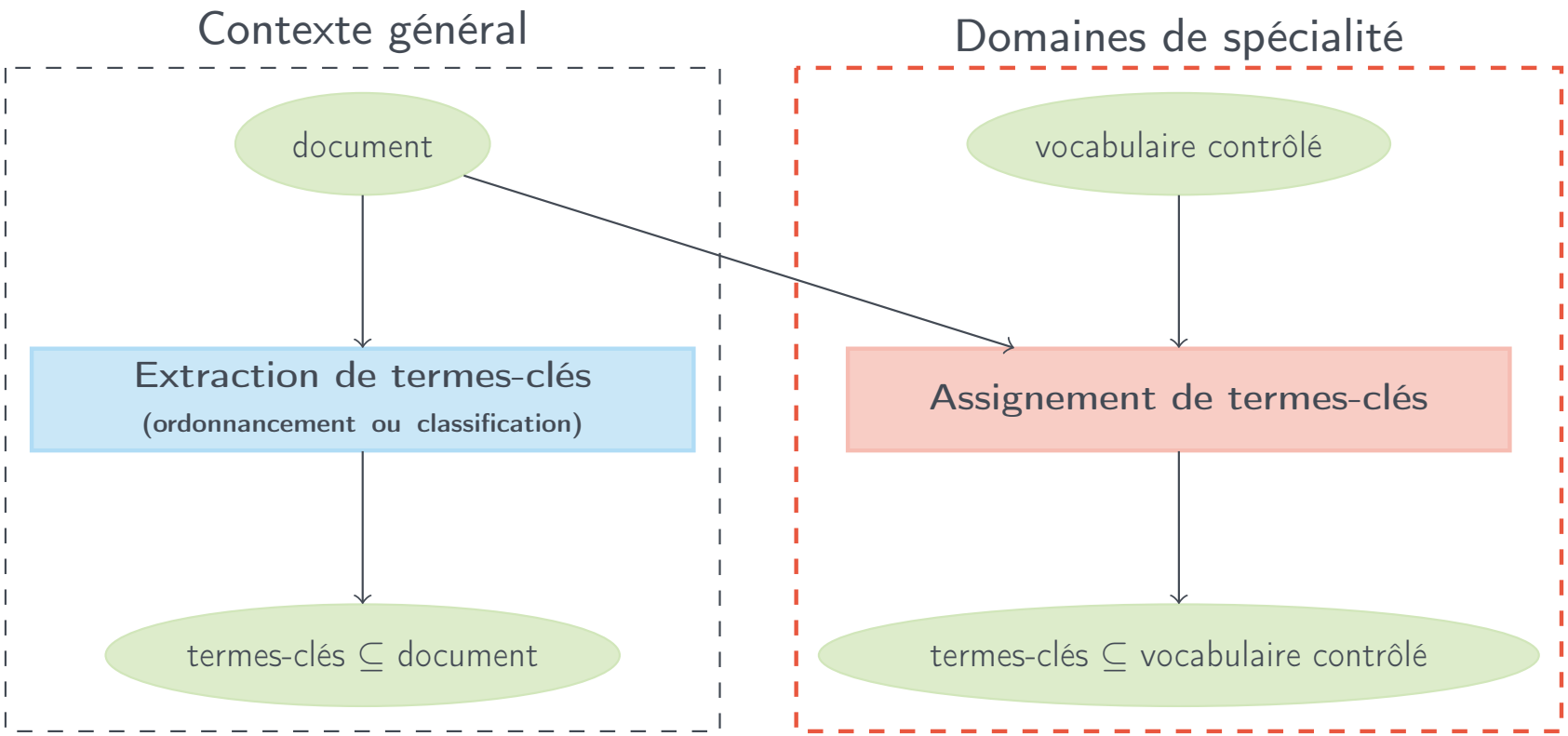
Classification des candidats selon divers critères (traits) :

- Fréquenciels
  - **TF-IDF (Witten et al., 1999, KEA)**
- Positionnels
  - **Première position (Witten et al., 1999, KEA)**
  - Apparition dans une section particulière (Nguyen et Kan, 2007)
- Linguistiques
  - Catégorie grammaticale (Nguyen et Kan, 2007)

Classification à l'aide de techniques d'apprentissage automatique :

- **Classification naïve bayesienne (Witten et al., 1999, KEA)**
- Réseau de neurones (Sarkar et al., 2010)

Indexation par termes-clés



Objectif

Positionner le document vis-à-vis de son domaine

Vocabulaire contrôlé

Terminologie spécifique à un domaine :

- Liste de termes
- Thésaurus

Classification des entrées d'un thésaurus (Medelyan et Witten, 2006, KEA++) :

1. Termes-clés candidats $\subseteq$ thésaurus
2. Trois critères de classification
   - TF-IDF
   - Première position

   } KEA
   - Nombre de relations avec d'autres candidats

Classification multi-étiquette et multi-classe (Partalas et al., 2013) :

- n étiquettes $=$ n termes-clés à assigner
- m classes $=$ termes du vocabulaire contrôlé

Deux catégories réalisées disjointement :

- Extraction de termes-clés
  - Ordonnancement ou classification
  - Applicable dans le contexte général
- Assignement de termes-clés
  - Focalisé sur un vocabulaire contrôlé
  - Adapté aux domaines de spécialité

Limite

Manque d'exhaustivité

- Extraction limitée au contenu du document
- Assignement limitée au vocabulaire du domaine

Étude préliminaire de la **céramique non tournée micacée** du bas Languedoc occidental : **typologie**, **chronologie** et aire de **diffusion**

L'étude présente une variété de **céramique non tournée** dont la **typologie** et l'analyse des **décors** permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le **décor** effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de **fouilles anciennes** menées sur les **oppidums du Cayla** à **Mailhac** (**Aude**) et de **Mourrel-Ferrat** à **Olonzac** (**Hérault**). La carte de **répartition** fait état d'**échanges** ou de **commerce** à l'échelon macrorégional rarement mis en évidence pour de la **céramique non tournée**. S'il est difficile de statuer sur l'origine des **décors**, il semble que la **production** s'insère dans une ambiance celtisante. La **chronologie** de cette **production** se situe dans le deuxième **âge du Fer**. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence

Mailhac ; **Aude** ; **Mourrel-Ferrat** ; **Olonzac** ; **Hérault** ; **céramique** ; **typologie** ; **décor** ; **chronologie** ; **diffusion** ; **production** ; **commerce** ; **répartition** ; **oppidum** ; **analyse** ; **fouille ancienne** ; **le Cayla** ; **micassé** ; **céramique non-tournée** ; **echange** ; **age du Fer** ; La Tène ; Europe ; France ; celtes ; distribution ; cartographie ; habitat ; site fortifié ; identification ; étude du matériel

Plan

Corpus	Documents		Termes-clés de référence		
	Quantité	Mots moy.	Catégorie	Quantité moy.	« À assigner »
DEft *(fr)* (Paroubek et al., 2012)					
→ Appr.	141	7 276,7	Auteur	5,4	18,2 %
→ Test	93	6 839,4	Auteur	5,2	21,1 %
Semeval *(en)* (Kim et al., 2010)					
→ Appr.	144	5 134,6	Auteur / Lecteur	15,4	13,5 %
→ Test	100	5 177,7	Auteur / Lecteur	14,7	22,1 %

Articles scientifiques

Corpus	Documents		Termes-clés de référence		
	Quantité	Mots moy.	Catégorie	Quantité moy.	« À assigner »
Wikinews *(fr)* (Bougouin et al., 2013)	100	308,5	Lecteur	9,6	7,6 %
DUC *(en)* (Wan et Xiao, 2008)	308	900,7	Lecteur	8,1	3,5 %

Dépêches journalistiques

Corpus	Documents		Termes-clés de référence		
	Quantité	Mots moy.	**Catégorie**	Quantité moy.	**« À assigner »**
Linguistique *(fr)*					
→ Appr.	515	160,5	**Professionnel**	8,6	**60,6 %**
→ Test	200	147,0	**Professionnel**	8,9	**62,8 %**
Sciences de l'info. *(fr)*					
→ Appr.	506	105,0	**Professionnel**	7,8	**67,9 %**
→ Test	200	157,0	**Professionnel**	10,2	**66,9 %**
Archéologie *(fr)*					
→ Appr.	518	221,1	**Professionnel**	16,9	**37,0 %**
→ Test	200	213,9	**Professionnel**	15,6	**37,4 %**
Chimie *(fr)*					
→ Appr.	582	105,7	**Professionnel**	12,2	**75,2 %**
→ Test	200	103,9	**Professionnel**	14,6	**78,8 %**

Références bibliographiques

Particularités

- Indexation par termes-clés homogène
- Indexation par termes-clés exhaustive

Plan

Méthode d'extraction de termes-clés à base de graphe

Limites des méthodes à base de graphe

- Ordonnancement des mots
- Redondance des termes-clés extraits
- Contexte (fenêtre de mots) variable selon les méthodes

Propositions

- Ordonnancement des termes-clés candidats
- Groupement en sujets des termes-clés candidats
- Graphe complet + pondération fine des arêtes

Météo du 19 **août 2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Météo du 19 **août 2012** : **alerte** à la **canicule** sur la **Belgique** et le **Luxembourg**

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

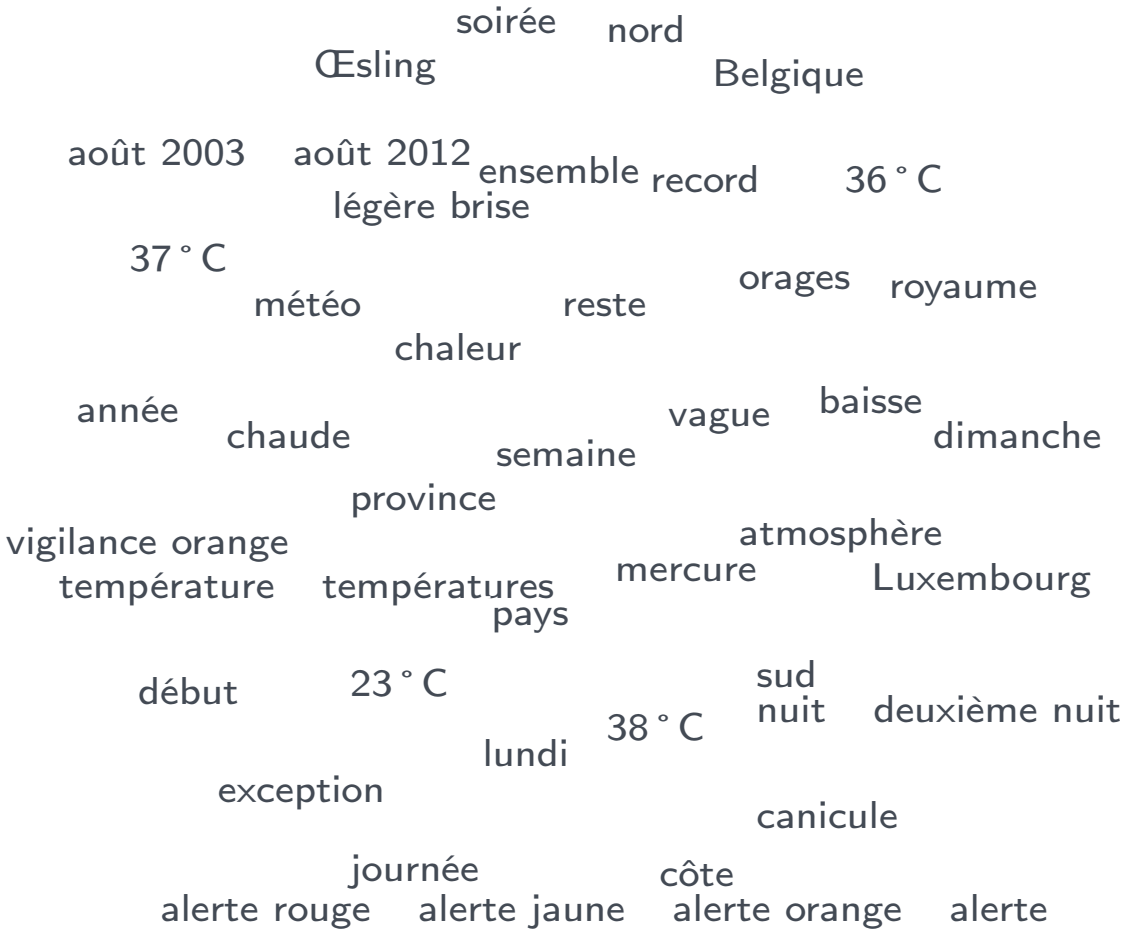
En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

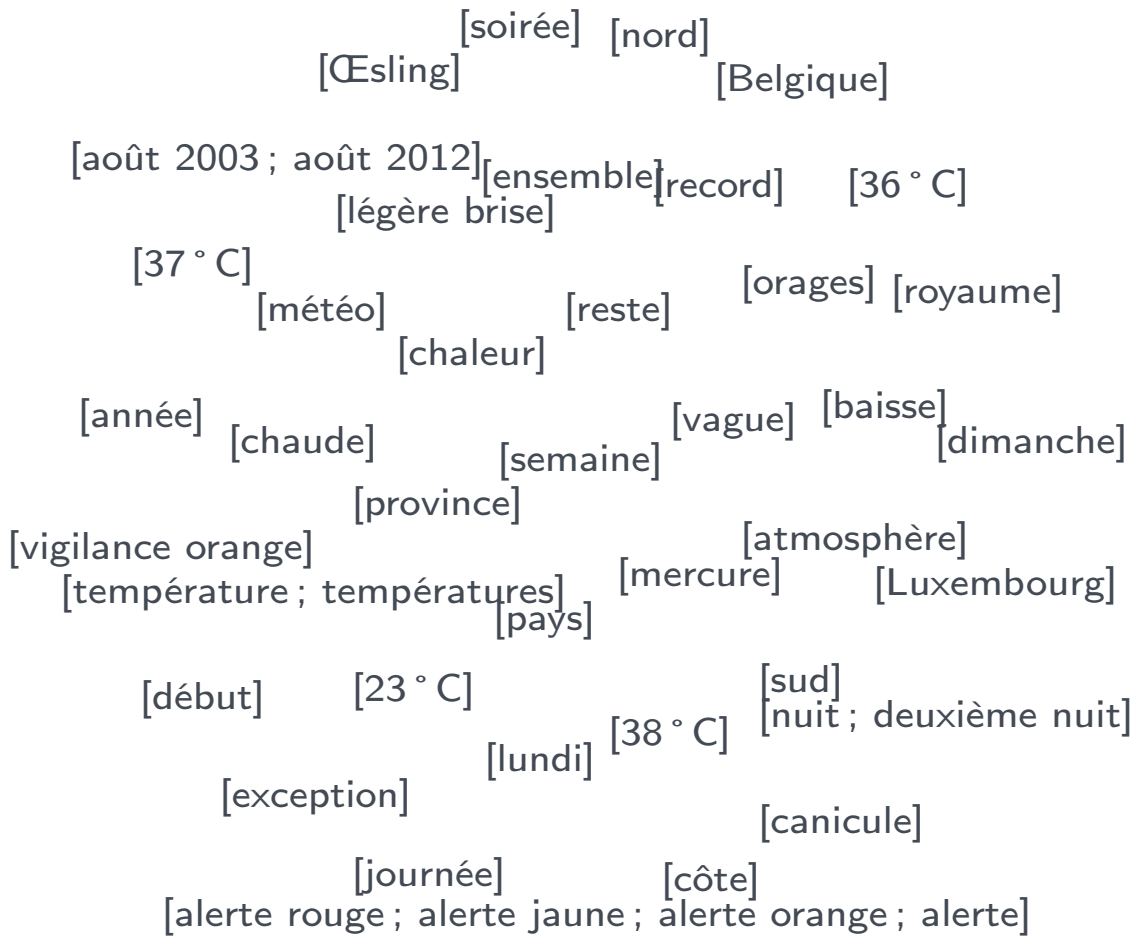
Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

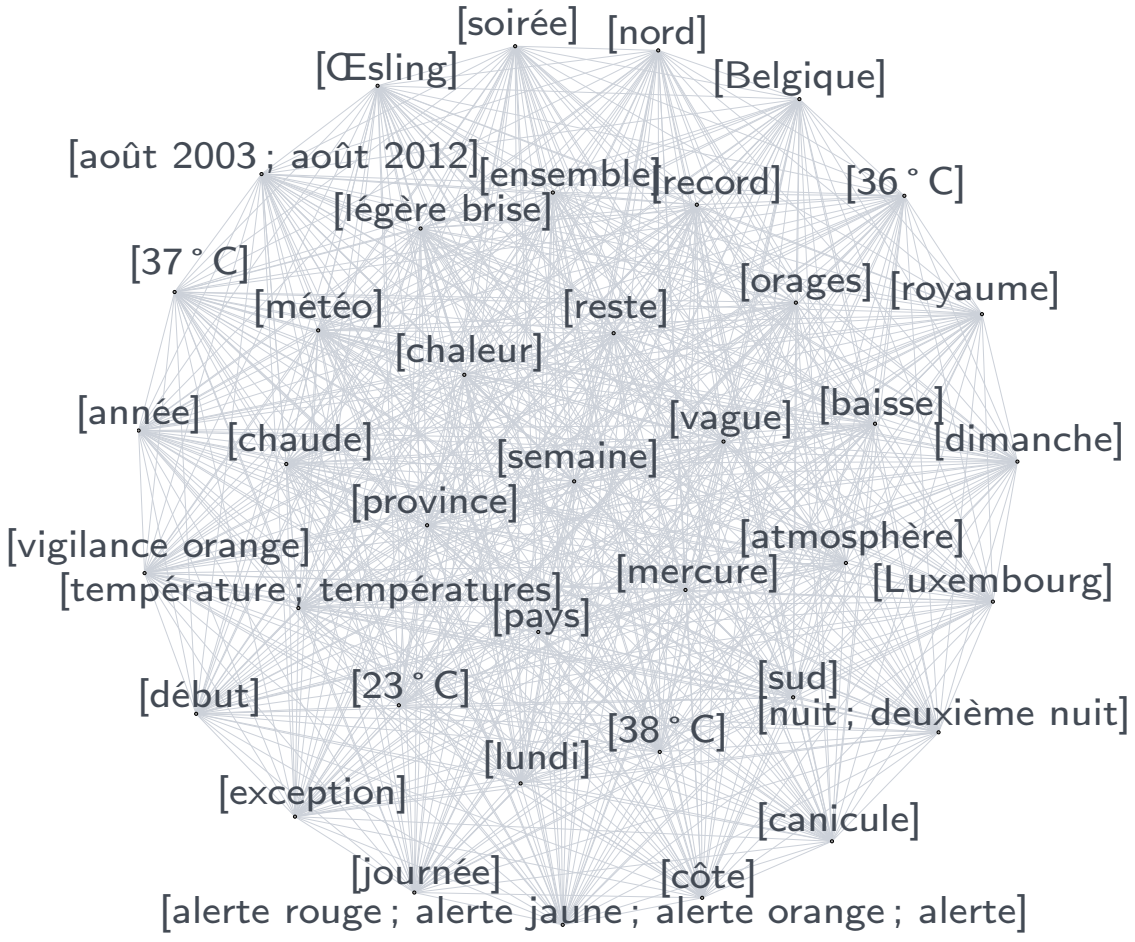
1. Sélection des candidats
   - `/(NOM | ADJ)+/`
2. Groupement des candidats
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
   - $\text{sim}(c_1, c_2) \geq 1/4$
3. Construction du graphe
   - Pondération selon la distance entre les sujets dans le document (en nombre de mots)
4. Ordonnancement des sujets
   - TextRank
5. Sélection des termes-clés
   - Un par sujet
   - Le premier dans le document



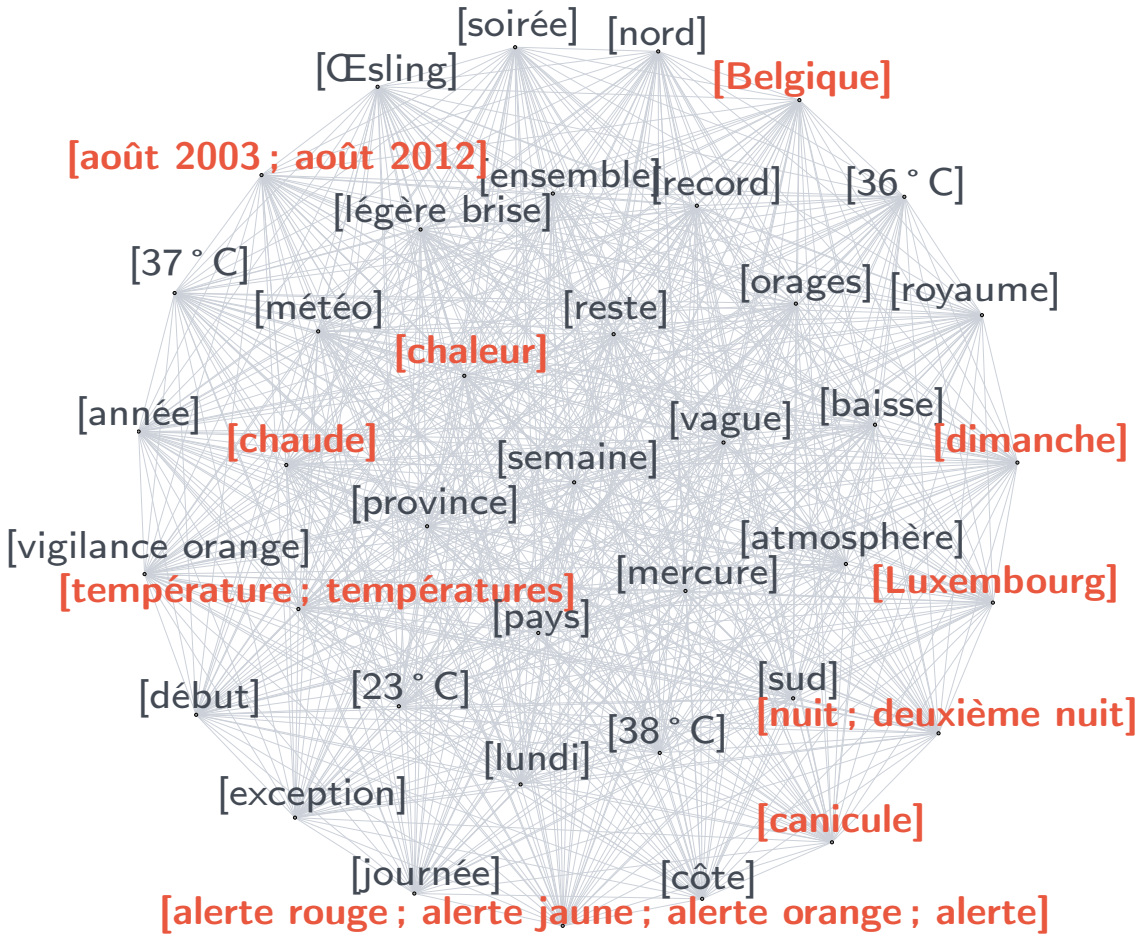
1. Sélection des candidats
   - `/(NOM | ADJ)+/`
2. Groupement des candidats
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
   - $\text{sim}(c_1, c_2) \geq 1/4$
3. Construction du graphe
   - Pondération selon la distance entre les sujets dans le document (en nombre de mots)
4. Ordonnancement des sujets
   - TextRank
5. Sélection des termes-clés
   - Un par sujet
   - Le premier dans le document



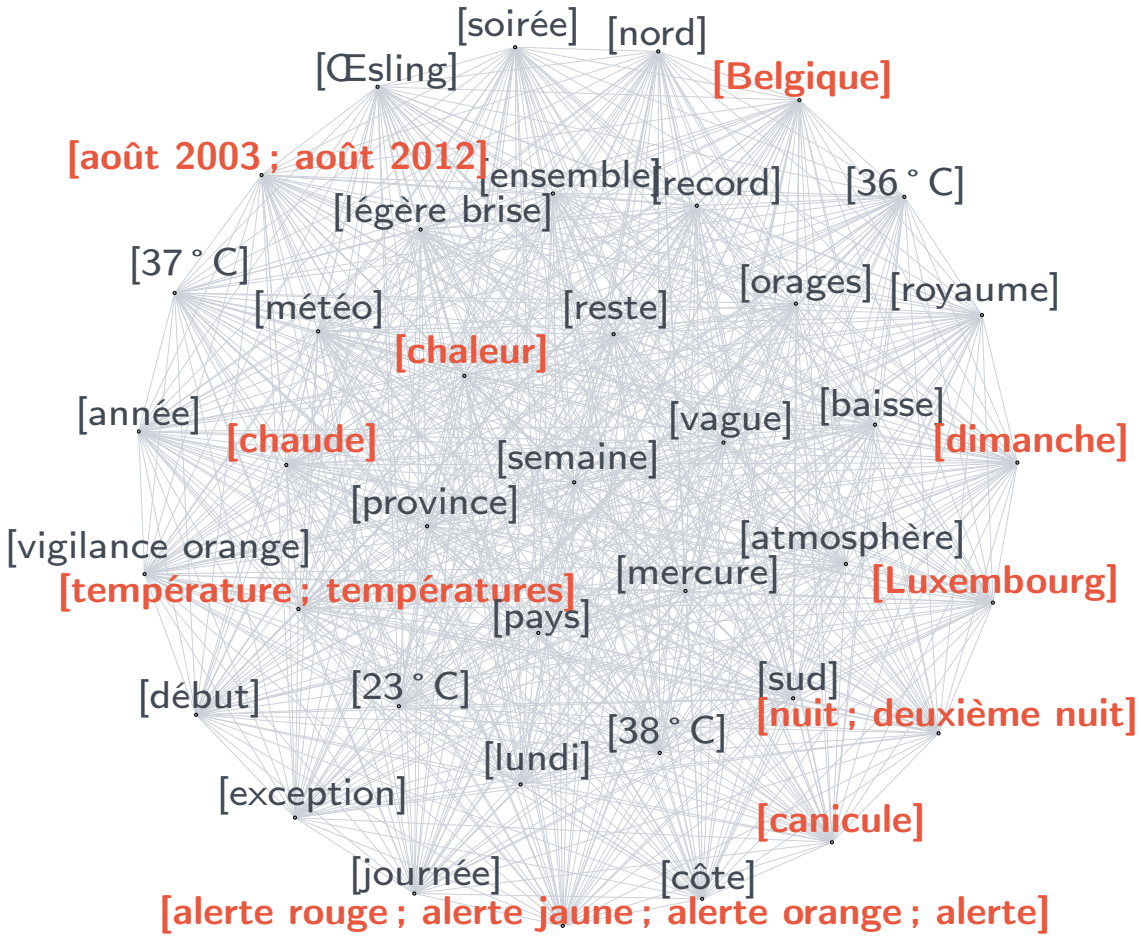
1. Sélection des candidats
   - `/(NOM | ADJ)+/`
2. Groupement des candidats
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
   - $\text{sim}(c_1, c_2) \geq 1/4$
3. Construction du graphe
   - Pondération selon la distance entre les sujets dans le document (en nombre de mots)
4. Ordonnancement des sujets
   - TextRank
5. Sélection des termes-clés
   - Un par sujet
   - Le premier dans le document



1. Sélection des candidats
   - `/(NOM | ADJ)+/`
2. Groupement des candidats
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
   - $\text{sim}(c_1, c_2) \geq 1/4$
3. Construction du graphe
   - Pondération selon la distance entre les sujets dans le document (en nombre de mots)
4. Ordonnancement des sujets
   - TextRank
5. Sélection des termes-clés
   - Un par sujet
   - Le premier dans le document



1. Sélection des candidats
   - `/(NOM | ADJ)+/`
2. Groupement des candidats
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
   - $\text{sim}(c_1, c_2) \geq 1/4$
3. Construction du graphe
   - Pondération selon la distance entre les sujets dans le document (en nombre de mots)
4. Ordonnancement des sujets
   - TextRank
5. Sélection des termes-clés
   - Un par sujet
   - Le premier dans le document



Météo du 19 août 2012 : alerte à la canicule sur la Belgique et le Luxembourg

À l'exception de la province de **Luxembourg**, en **alerte** jaune, l'ensemble de la **Belgique** est en vigilance **orange** à la **canicule**. Le **Luxembourg** n'est pas épargné par la vague du **chaleur** : le nord du pays est en **alerte orange**, tandis que le sud a était placé en **alerte** rouge.

En **Belgique**, la **température** n'est pas descendue en dessous des 23 °C cette nuit, ce qui constitue la deuxième nuit **la plus chaude** jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée **la plus chaude** de l'année. Les **températures** seront comprises entre 33 et 38 °C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de **chaleur** sont a prévoir dans la soirée et en début de nuit.

Au **Luxembourg**, le mercure devrait atteindre 32 °C ce dimanche sur l'Oesling et jusqu'à 36 °C sur le sud du pays, et 31 à 32 °C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9 °C) ne devrait pas être atteint.

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Rang	TextRank	SingleRank	TopicRank
01	août 2012	alerte orange	Luxembourg
02	août 2003	alerte jaune	alerte
03	alerte orange	alerte rouge	nuit
04	vigilance orange	alerte	Belgique
05	deuxième nuit	deuxième nuit	août 2012
06	légère brise	août 2012	chaleur
07		août 2003	température
08		vigilance orange	chaude
09		légère brise	canicule
10		Luxembourg	dimanche

Termes-clés

Termes-clés de référence

Août 2012 ; **canicule** ; **Belgique** ; **Luxembourg** ; **alerte** ; **orange** ; **chaleur** ; **chaude** ; **température** ; **la plus chaude**

Trois méthodes de référence :

- TF-IDF (Salton et al., 1975)
- TextRank (Mihalcea et Tarau, 2004)
- SingleRank (Wan et Xiao, 2008)

Comparaisons à une racine près :

- « températures » accepté pour « température »

Évaluation à 10 termes-clés, en termes de :

- Précision (P) : $\frac{|\text{corrects}|}{|\text{extraits}|}$
- Rappel (R) : $\frac{|\text{corrects}|}{|\text{references}|}$
- F1-mesure (F) : $2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$

Méthode	DEft *(fr)*			Wikinews *(fr)*			SemEval *(en)*			DUC *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	10,3	19,1	13,2	33,9	35,9	34,3	13,2	8,9	10,5	**23,8**	**30,7**	**26,4**
TextRank	4,9	7,1	5,7	9,3	8,3	8,6	7,9	4,5	5,6	4,9	5,4	5,0
SingleRank	4,5	9,0	5,9	19,4	20,7	19,7	4,6	3,2	3,7	22,3	28,4	24,6
TopicRank	**11,7**	**21,7**	**15,1**†	**35,0**	**37,5**	**35,6**†	**14,9**	**10,3**	**12,1**†	18,3	23,8	20,4
Borne haute	14,5	27,0	18,7	41,8	44,1	42,2	30,0	20,7	24,3	30,5	38,7	33,7

Observations

- Faibles résultats globaux
- Meilleure performance globale de TopicRank
- TopicRank significativement meilleur que TextRank et SingleRank †

Perspective d'amélioration de la borne haute

- Obama $\Rightarrow$ Barack Obama $\in$ [Obama ; Barack Obama]
- Romney $\Rightarrow$ Mitt Romney $\in$ [Romney ; Mitt Romney]

Méthode	DEft *(fr)*			Wikinews *(fr)*			SemEval *(en)*			DUC *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	10,3	19,1	13,2	33,9	35,9	34,3	13,2	8,9	10,5	**23,8**	**30,7**	**26,4**
TextRank	4,9	7,1	5,7	9,3	8,3	8,6	7,9	4,5	5,6	4,9	5,4	5,0
SingleRank	4,5	9,0	5,9	19,4	20,7	19,7	4,6	3,2	3,7	22,3	28,4	24,6
TopicRank	**11,7**	**21,7**	**15,1**†	**35,0**	**37,5**	**35,6**†	**14,9**	**10,3**	**12,1**†	18,3	23,8	20,4
Borne haute	**14,5**	**27,0**	**18,7**	**41,8**	**44,1**	**42,2**	**30,0**	**20,7**	**24,3**	**30,5**	**38,7**	**33,7**

Observations

- Faibles résultats globaux
- Meilleure performance globale de TopicRank
- TopicRank significativement meilleur que TextRank et SingleRank †

Perspective d'amélioration de la borne haute

- Obama $\Rightarrow$ Barack Obama $\in$ [Obama ; Barack Obama]
- Romney $\Rightarrow$ Mitt Romney $\in$ [Romney ; Mitt Romney]

Méthode à base de graphe qui ordonne les sujets du document

Avantages

- Applicable dans le contexte général
- Termes-clés extraits peu redondants

Limites

- Groupement naïf des candidats en sujets
- Stratégie sous-optimale pour sélectionner le terme-clé d'un sujet
- Limité au contenu du document

Méthode	Linguistique *(fr)*			Sciences de l'info. *(fr)*			Archéologie *(fr)*			Chimie *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	**13,0**	**15,4**	**13,9**	**13,4**	**14,0**	**13,2**	**28,1**	**19,1**	**22,2**	**14,1**	**11,1**	**11,9**
TextRank	7,1	6,1	6,4	5,8	4,3	4,8	10,2	5,3	6,8	9,4	5,3	6,5
SingleRank	9.0	10,6	9,6	9,5	10,0	9,4	12,7	8,9	10,2	13,0	10,4	11,0
TopicRank	11,2	13,1	11,9	12,1	12,8	12,1	27,5	18,7	21,8	13,8	11,1	11,8
Borne haute	**14,5**	**17,0**	**15,4**	**15,0**	**15,6**	**14,9**	**32,5**	**22,2**	**25,8**	**15,8**	**12,5**	**13,3**

Observations

- TopicRank meilleur que TextRank et SingleRank
- TF-IDF meilleur que TopicRank

Adaptation supervisée de TopicRank aux domaines de spécialité

Objectif

Réaliser conjointement extraction et assignement

Hypothèses

- Les documents d'apprentissage représentent le domaine
  - Termes-clés $\simeq$ vocabulaire du domaine
  - Lien d'association entre les termes-clés d'un même document
- Le domaine complète le document

Propositions

- Représentation du domaine par un graphe
- Unification du graphe du domaine au graphe de sujets
- Ordonnancement conjoint

Phases préparatoires de TopicRank :

1. Sélection des termes-clés candidats du document
2. Groupement des candidats en sujets
3. Création du graphe de sujets

Ajout de la connaissance du domaine :

4. Déduction du vocabulaire contrôlé du domaine
5. Création du graphe des termes-clés du domaine
6. Unification du graphe du domaine au graphe de sujets

Extraction et assignement conjoints :

7. Ordonnancement conjoint des sujets et des termes-clés du domaine
8. Sélection des termes-clés parmi les meilleurs sujets et/ou termes-clés du domaine

Étude préliminaire de la **céramique non tournée micacée** du bas Languedoc occidental : **typologie**, **chronologie** et aire de **diffusion**

L'étude présente une variété de **céramique non tournée** dont la **typologie** et l'analyse des **décors** permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le **décor** effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de **fouilles anciennes** menées sur les **oppidums du Cayla** à **Mailhac** (**Aude**) et de **Mourrel-Ferrat** à **Olonzac** (**Hérault**). La carte de **répartition** fait état d'**échanges** ou de **commerce** à l'échelon macrorégional rarement mis en évidence pour de la **céramique non tournée**. S'il est difficile de statuer sur l'origine des **décors**, il semble que la **production** s'insère dans une ambiance celtisante. La **chronologie** de cette **production** se situe dans le deuxième **âge du Fer**. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence

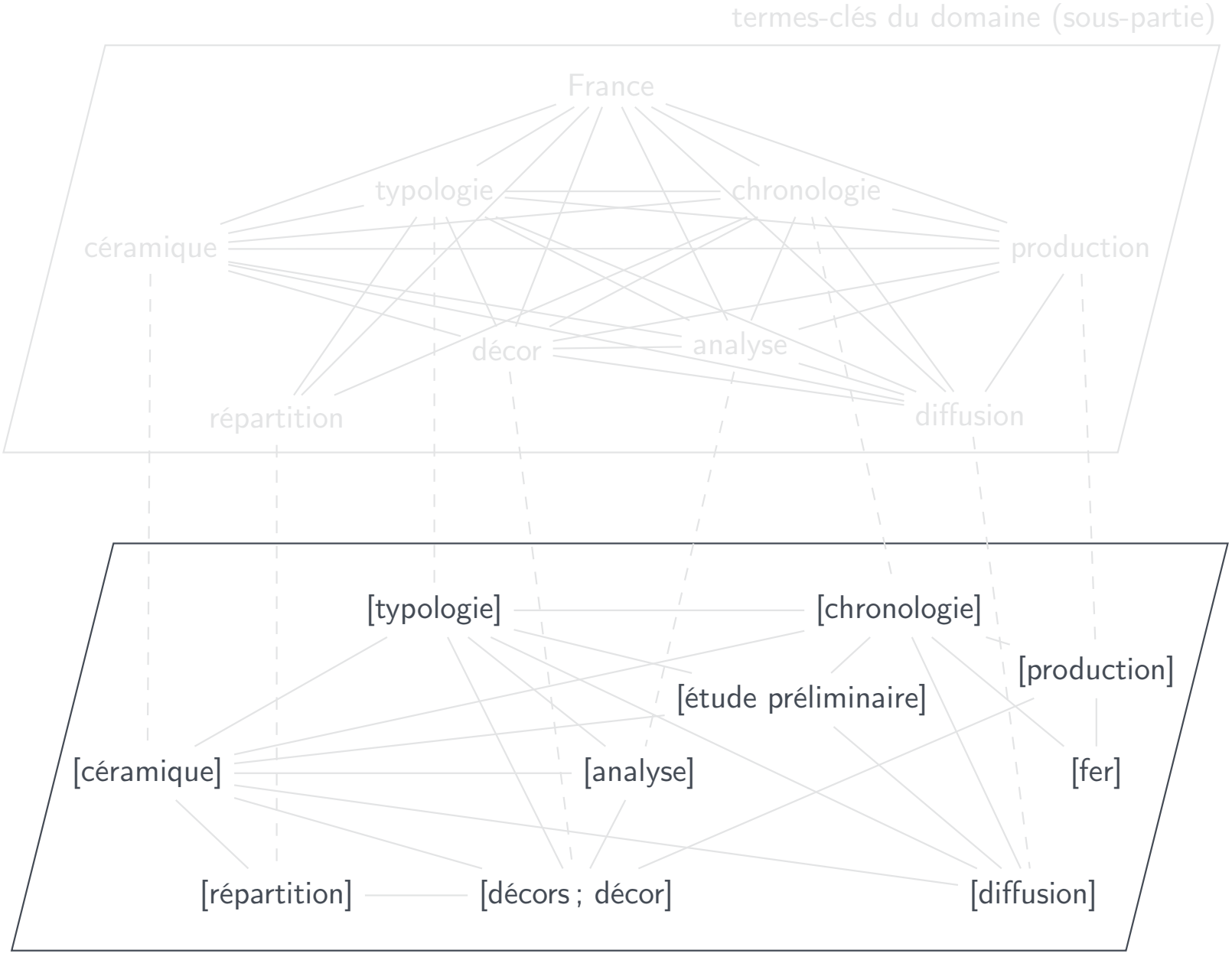
Mailhac ; **Aude** ; **Mourrel-Ferrat** ; **Olonzac** ; **Hérault** ; **céramique** ; **typologie** ; **décor** ; **chronologie** ; **diffusion** ; **production** ; **commerce** ; **répartition** ; **oppidum** ; **analyse** ; **fouille ancienne** ; **le Cayla** ; **micassé** ; **céramique non-tournée** ; **echange** ; **age du Fer** ; La Tène ; Europe ; France ; celtes ; distribution ; cartographie ; habitat ; site fortifié ; identification ; étude du matériel

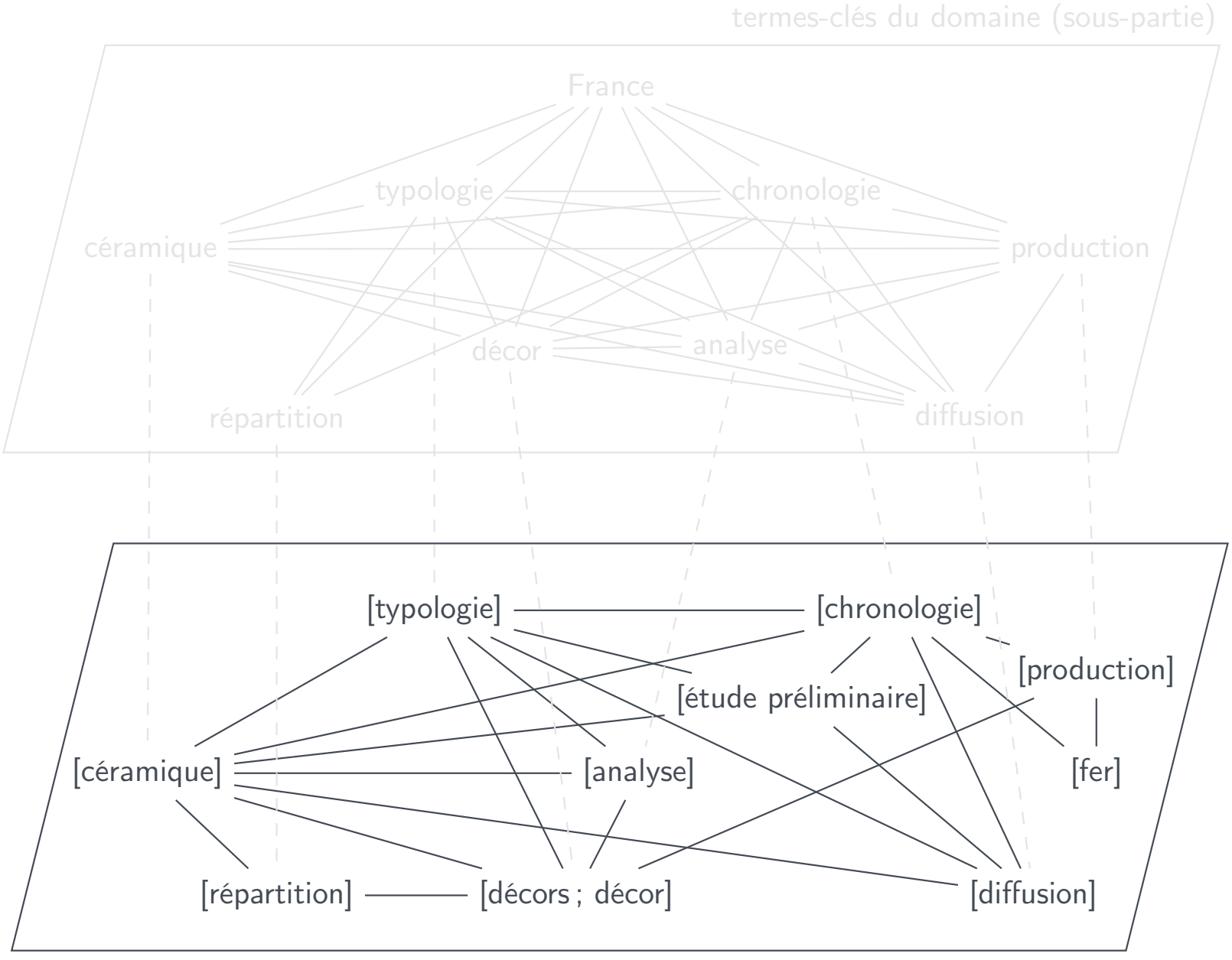
Étude préliminaire de la **céramique non tournée micacée** du bas Languedoc occidental : **typologie**, **chronologie** et aire de **diffusion**

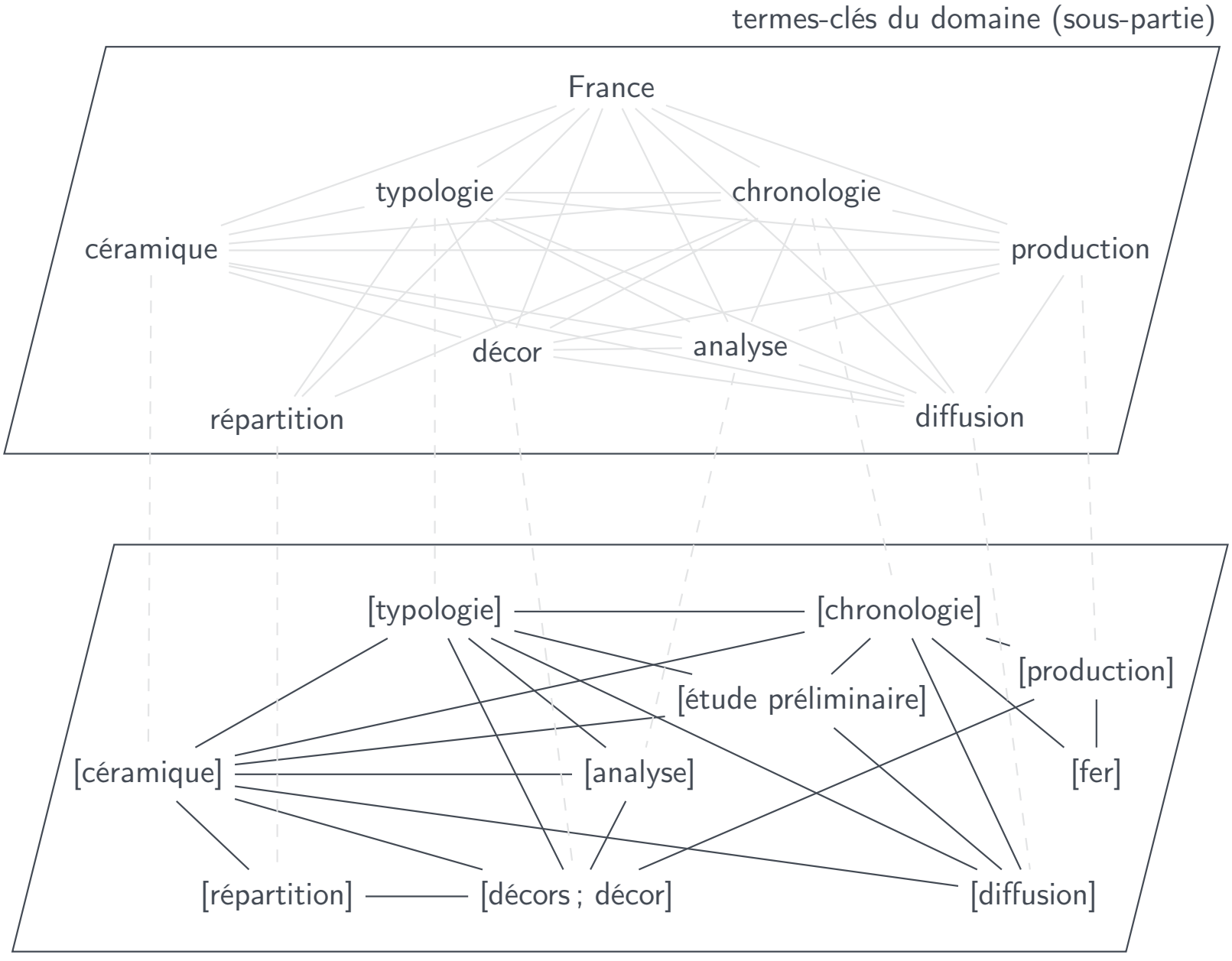
L'étude présente une variété de **céramique non tournée** dont la **typologie** et l'analyse des **décors** permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le **décor** effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de **fouilles anciennes** menées sur les **oppidums du Cayla** à **Mailhac** (**Aude**) et de **Mourrel-Ferrat** à **Olonzac** (**Hérault**). La carte de **répartition** fait état d'**échanges** ou de **commerce** à l'échelon macrorégional rarement mis en évidence pour de la **céramique non tournée**. S'il est difficile de statuer sur l'origine des **décors**, il semble que la **production** s'insère dans une ambiance celtisante. La **chronologie** de cette **production** se situe dans le deuxième **âge du Fer**. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

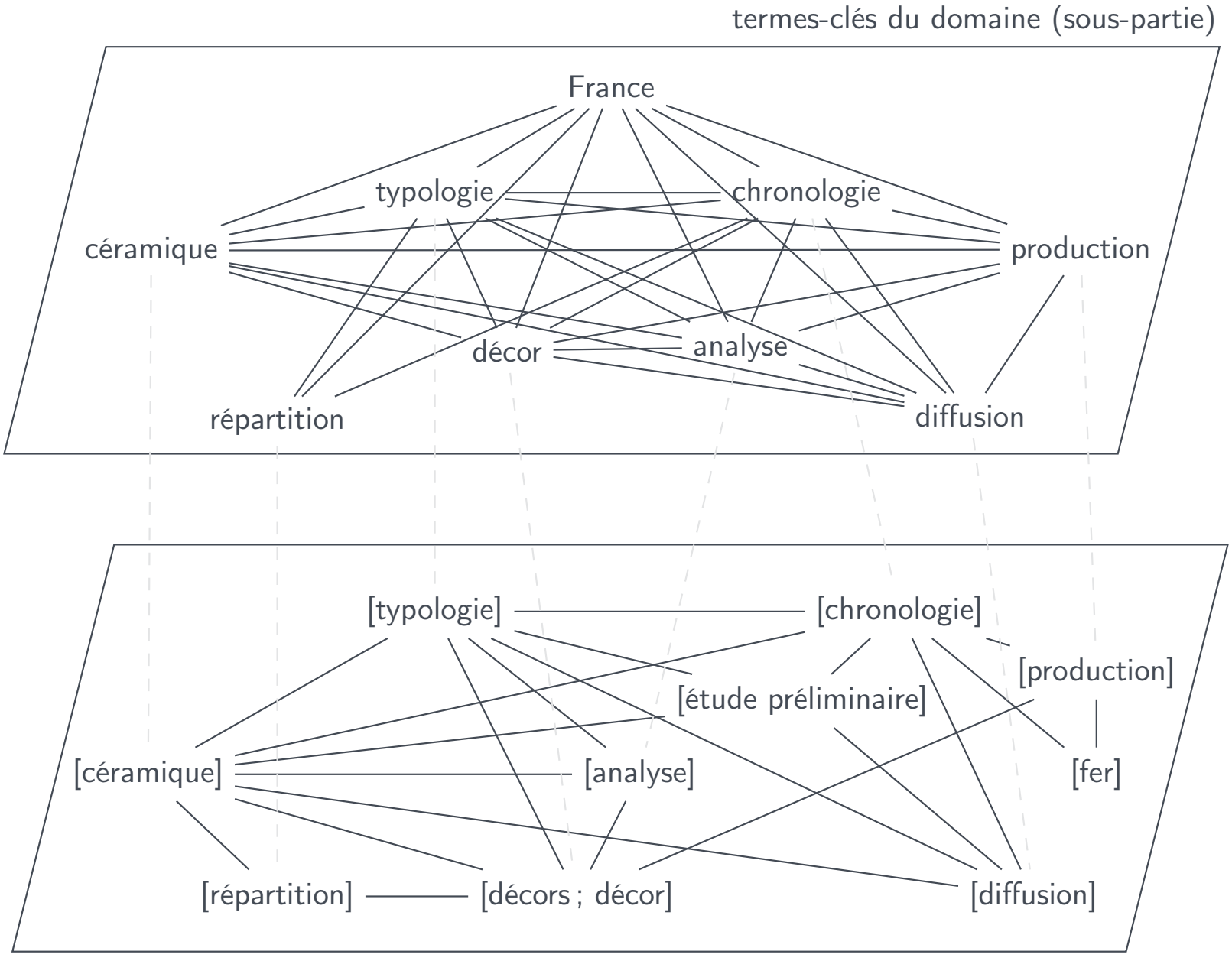
Termes-clés de référence

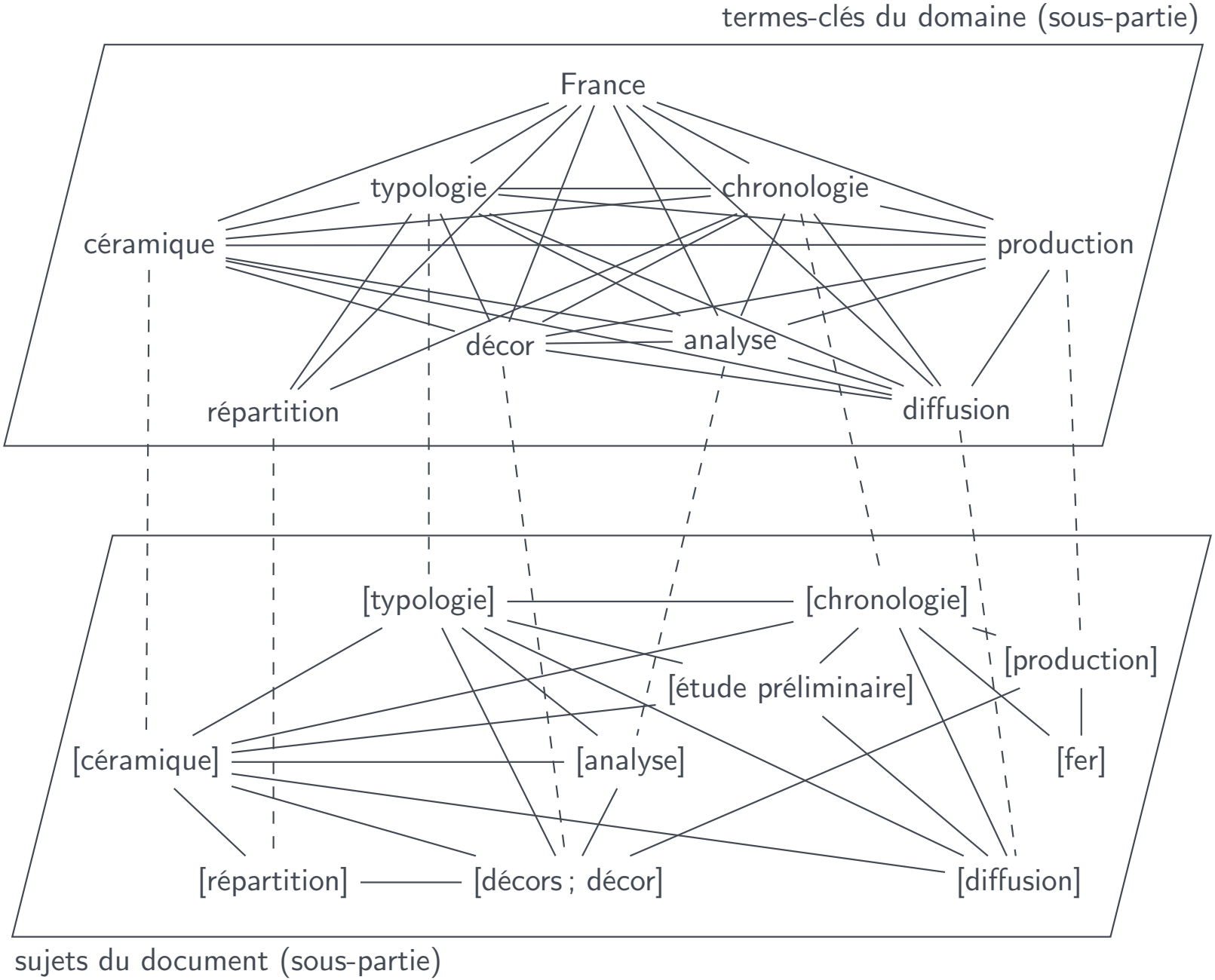
Mailhac ; **Aude** ; **Mourrel-Ferrat** ; **Olonzac** ; **Hérault** ; **céramique** ; **typologie** ; **décor** ; **chronologie** ; **diffusion** ; **production** ; **commerce** ; **répartition** ; **oppidum** ; **analyse** ; **fouille ancienne** ; **le Cayla** ; **micassé** ; **céramique non-tournée** ; **echange** ; **age du Fer** ; La Tène ; Europe ; France ; celtes ; distribution ; cartographie ; habitat ; site fortifié ; identification ; étude du matériel

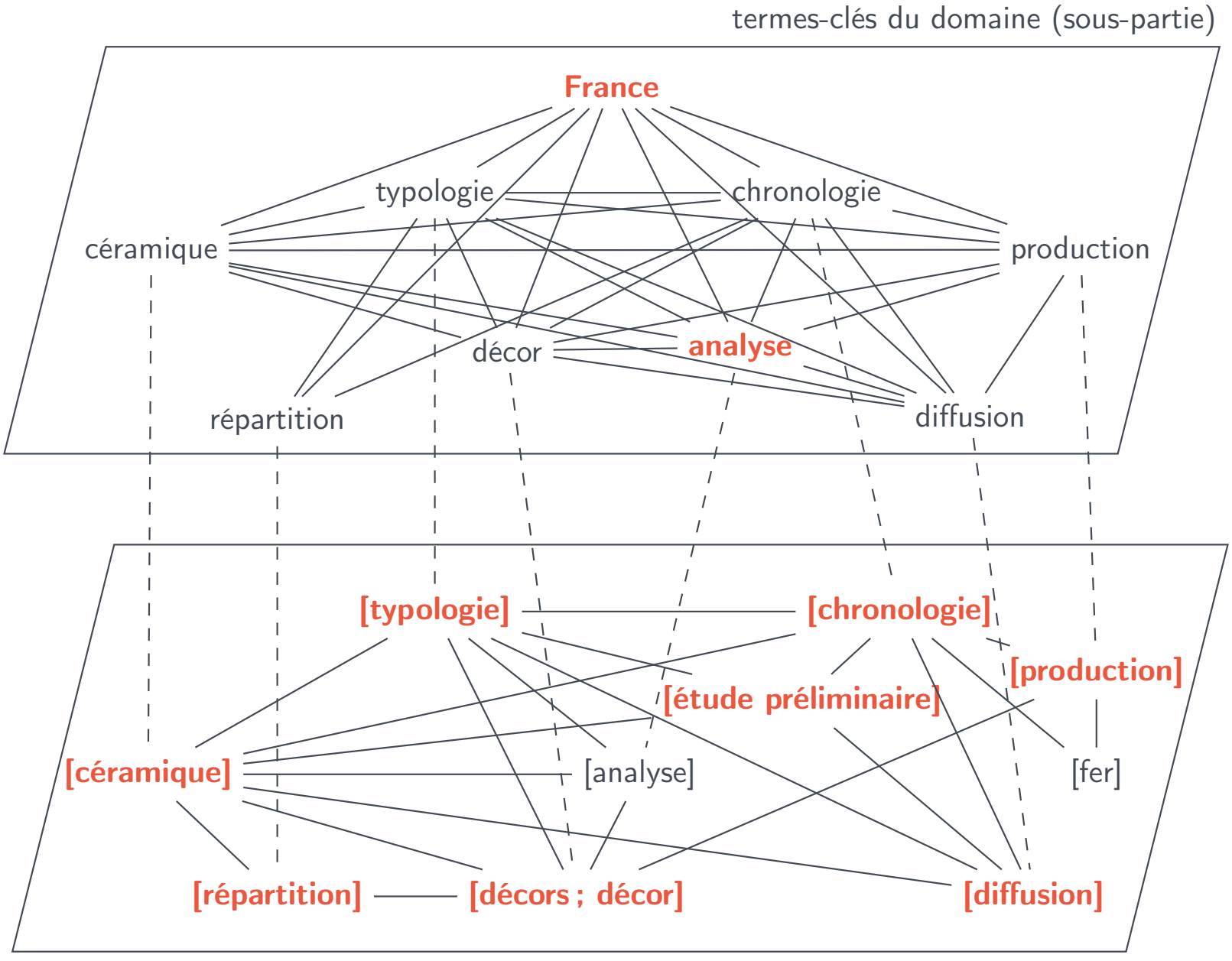

termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)


termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)


termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)


termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)


termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)


termes-clés du domaine (sous-partie)
France
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[production]
[étude préliminaire]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)

Rang	TopicRank	TopicCoRank
01	décors	céramique
02	céramique	décors
03	chronologie	typologie
04	typologie	chronologie
05	production	production
06	fin	étude préliminaire
07	étude préliminaire	diffusion
08	fer	analyse *
09	deuxième âge	France **
10	aire	répartition

Termes-clés

Termes-clés de référence

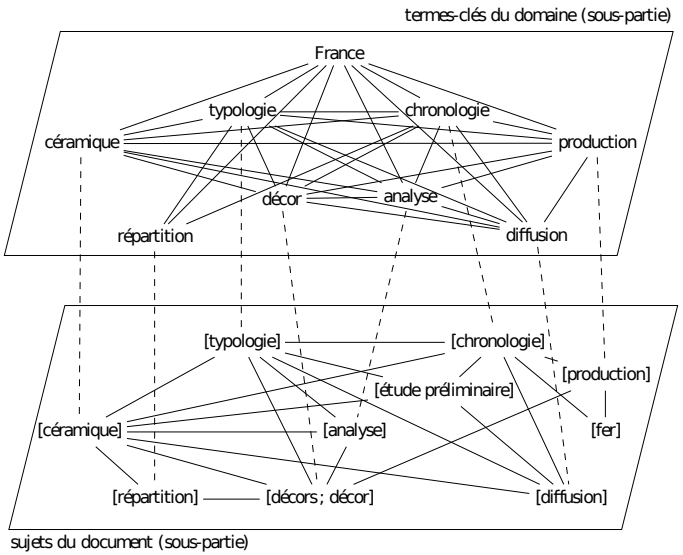
Mailhac ; **Aude** ; **Mourrel-Ferrat** ; **Olonzac** ; **Hérault** ; **céramique** ; **typologie** ; **décor** ; **chronologie** ; **diffusion** ; **production** ; **commerce** ; **répartition** ; **oppidum** ; **analyse** ; **fouille ancienne** ; **le Cayla** ; **micassé** ; **céramique non-tournée** ; **echange** ; **age du Fer** ; La Tène ; Europe ; France ; celtes ; distribution ; cartographie ; habitat ; site fortifié ; identification ; étude du matériel

Trois méthodes de référence :

- TF-IDF (Salton et al., 1975)
- TopicRank (Bougouin et al., 2013)
- KEA++ (Medelyan et Witten, 2006)

Deux variantes :

- $\text{TopicCoRank}_{\text{extr.}}$
- $\text{TopicCoRank}_{\text{assign.}}$



Méthode	Linguistique *(fr)*			Sciences de l'info. *(fr)*			Archéologie *(fr)*			Chimie *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	13,0	15,4	13,9	13,4	14,0	13,2	28,1	19,1	22,2	14,1	11,1	11,9
TopicRank	11,2	13,1	11,9	12,1	12,8	12,1	27,5	18,7	21,8	13,8	11,1	11,8
KEA++	11,6	13,0	12,1	9,5	10,2	9,6	23,5	16,2	18,8	11,4	8,5	9,2
$\text{TopicCoRank}_{\text{extr.}}$	14,3	16,5	15,1	15,4	15,9	15,2†	36,7	24,6	28,8†	15,8	12,1	13,1
$\text{TopicCoRank}_{\text{assign.}}$	**24,5**	**28,3**	**25,8**	**19,7**	**19,8**	**19,2**†	**47,8**	**32,3**	**37,7**†	**20,0**	**14,8**	**16,3**†
TopicCoRank	18,8	21,9	19,9	17,3	17,7	17,0†	38,3	25,7	30,1†	17,2	13,4	14,4†

Observations

- Meilleure performance
- $\text{TopicCoRank}_{\text{extr.}} >$ TopicRank $\Rightarrow$ apport du domaine

Extension supervisée de TopicRank pour assigner des termes-clés

Avantages

- Tire profit de la connaissance du domaine
- Combine extraction et assignement ⇒ meilleure exhaustivité
- Adapté à plusieurs scénarii d'utilisation professionnels :
  - TopicCoRank + validation / enrichissement manuel
  - $\text{TopicCoRank}_{\text{assign.}}$ seul

Limites

- Donne autant d'importance aux deux graphes
- Termes-clés extraits plus redondants que ceux de TopicRank

Plan

Conclusion

Indexation automatique par termes-clés

1. Dans le contexte général
   - Extraction de termes-clés à base de graphe
   - Ordonnancement par importance des sujets du document
   - Amélioration de l'existant
2. En domaines de spécialité
   - Extraction et assignement simultanés
   - Usage de la connaissance du domaine
   - 100 % de couverture théorique

Limites

- TopicCoRank focalisé sur les domaines de spécialité
- Évaluation du point de vue des termes-clés uniquement

Perspectives

TopicCoRank dans le contexte général ?

- Parallèle entre genre et domaine ?
- Quid de l'homogénéité ?
- Quelle influence de chaque graphe sur l'autre ?

Perspectives

Apports de TopicRank :

- Recherche d'information (Jones et Staveley, 1999)
- Lecture pour les dislexiques (Rello et al., 2014)
- Apprentissage des langues secondes (Pressley et al., 1982)

Apports de TopicCoRank :

- Recherche d'information (Jones et Staveley, 1999)
- Veille terminologique
  - Faut-il sélectionner les candidats avec un extracteur terminologique ?

Références

Adrien Bougouin, Florian Boudin et Béatrice Daille : Topicrank : Graph-Based Topic Ranking for Keyphrase Extraction. In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP), pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I13-1062.

Steve Jones et Mark S. Staveley : Phrasier : a System for Interactive Document Retrieval Using Keyphrases. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 160–167, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. URL http://doi.acm.org/10.1145/312624.312671.

Su Nam Kim, Olena Medelyan, Min-Yen Kan et Timothy Baldwin : SemEval-2010 task 5 : Automatic Keyphrase Extraction from Scientific Articles. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), pages 21–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1859664.1859668.

Références

Yutaka Matsuo et Mitsuru Ishizuka : Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools, 13(01):157–169, 2004.

Olena Medelyan et Ian H Witten : Thesaurus Based Automatic Keyphrase Indexing. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, pages 296–297. ACM, 2006.

Rada Mihalcea et Paul Tarau : TextRank : Bringing Order Into Texts. In Dekang Lin et Dekai Wu, éditeurs : Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Thuy Dung Nguyen et Min-Yen Kan : Keyphrase Extraction in Scientific Publications. In Proceedings of the 10th International Conference on Asian Digital Libraries : Looking Back 10 Years and Forging New Frontiers, pages 317–326, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-77093-3, 978-3-540-77093-0. URL `http://dl.acm.org/citation.cfm?id=1780653.1780707`.

Références

Patrick Paroubek, Pierre Zweigenbaum, Dominic Forest et Cyril Grouin : Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012 (Controlled and Free Indexing of Scientific Papers. Presentation and Results of the DEFT2012 Text-Mining Challenge) [in French]. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Workshop DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge), pages 1–13, Grenoble, France, June 2012. ATALA/AFCP. URL `http://www.aclweb.org/anthology/W12-1101`.

Ioannis Partalas, Éric Gaussier et Axel-Cyrille Ngonga Ngomo : Results of the First BioASQ Workshop. In BioASQ@ CLEF, pages 1–8, 2013.

Michael Pressley, Joel R Levin et Harold D Delaney : The Mnemonic Keyword Method. Review of Educational Research, 52(1):61–91, 1982.

Références

Luz Rello, Horacio Saggion et Ricardo Baeza Yates : Keyword Highlighting Improves Comprehension for People with Dyslexia. In Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), pages 30–37, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1204.

Gerard Salton, Andrew Wong et Chungshu Yang : A Vector Space Model for Automatic Indexing. Communication ACM, 18(11):613–620, November 1975. ISSN 0001-0782. URL http://doi.acm.org/10.1145/361219.361220.

Kamal Sarkar, Mita Nasipuri et Suranjan Ghose : A New Approach to Keyphrase Extraction Using Neural Networks. International Journal of Computer Science Issues Publicity Board 2010, 2010.

Xiaojun Wan et Jianguo Xiao : Single Document Keyphrase Extraction Using Neighborhood Knowledge. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620163.1620205.

Références

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin et Craig G. Nevill Manning : KEA : Practical Automatic Keyphrase Extraction. In Proceedings of the 4th ACM Conference on Digital Libraries, pages 254–255, New York, NY, USA, 1999. ACM. ISBN 1-58113-145-3. URL http://doi.acm.org/10.1145/313238.313437.

Approche :

- Analyse de trois collections de documents
  - DEft *(fr)*
  - SemEval *(en)*
  - DUC *(en)*
- Analyse en surface
  - Taille d'un termes-clés ?
  - Classes grammaticales des mots d'un termes-clés ?
- Analyse plus approfondie
  - Quelles classes grammaticales sont intéressantes ?
  - Pouvons-nous affiner la sélection à partir de ces classes ?

Analyse en surface :

	DEft *(fr)*	SemEval *(en)*	DUC *(en)*
Taux (en %) de termes-clés :			
Uni-grammes	60,2	20,2	17,1
Bi-grammes	24,5	53,4	60,8
Tri-grammes	8,8	21,3	17,8
Taux (en %) de termes-clés contenant au moins un(e) :			
Nom commun	93,1	98,7	94,5
Nom propre	6,9	4,3	17,1
Adjectif	65,5	50,2	50,0
Verbe	1,0	4,0	1,0
Adverbe	1,3	0,7	1,6
Préposition	31,2	1,5	0,3
Déterminant	20,4	0,0	0,0

Observations

- Termes-clés rarement composés de plus de trois mots
- Nom omniprésent dans les termes-clés
- **Adjectif très utilisé (modification du nom)**
- Prépositions et déterminants spécifiques au français

Analyse des adjectifs :

Adjectif relationnel

- Dérivé d'un nom
  - « culture » $\rightarrow$ « culturel »
- Établi une relation avec le nom dérivé
  - « héritage culturel » $\Leftrightarrow$ « héritage de la culture »
- Privilégié dans les noms de catégories
  - Catégorie Wikipédia « héritage culturel »

Adjectif composé

- Constitué de plusieurs mots
- Privilégié pour la formation de néologismes

Analyse des adjectifs :

	DEft *(fr)*	**SemEval** *(en)*	**DUC** *(en)*
Adjectifs relationnels (%)	87,1	43,6	53,1
Adjectifs composés (%)	3,3	16,4	10,6
Adjectifs qualificatifs (%)	9,6	40,0	36,3

Taux d'adjectifs dans les termes-clés

	DEft *(fr)*	**SemEval** *(en)*	**DUC** *(en)*
Adjectifs relationnels (%)	61,9	30,7	29,9
Adjectifs composés (%)	0,4	7,9	8,8
Adjectifs qualificatifs (%)	37,7	61,4	61,3

Taux d'adjectifs dans les documents

Observations

- Adjectifs relationnels très utilisés dans les termes-clés
- Adjectifs composés peu utilisés mais peu ambigus
- Ambigüité sur les adjectifs qualificatifs

1. Présélection des candidats
   - `/NOM+ ADJ?/` *(fr)*
   - `/ADJ? NOM+/` *(en)*
2. Filtrage de adjectifs superflus
   - Adjectifs relationnels : OK
   - Adjectifs composés : OK
   - **Adjectifs qualificatifs : OK / KO**

     Si fréquence avec `ADJ` > fréquence sans `ADJ` : OK

Évaluation en deux temps :

1. Intrinsèque
2. Extrinsèque

Trois méthodes de référence :

- {1..3}-grammes
- Syntagmes nominaux minimaux
- `/(NOM | ADJ)+/`

Deux méthode d'extraction de termes-clés :

- Tf-Idf
- KEA

Évaluation intrinsèque :

Méthode	DEFT *(fr)*			SemEval *(en)*			DUC *(en)*		
	Candidats	R_{max}	Q	Candidats	R_{max}	Q	Candidats	R_{max}	Q
N-grammes	2 610,4	**74,1**	0,03	1 652,3	**71,7**	0,04	478,9	**90,4**	0,19
`/(NOM | ADJ)+/`	810,3	61,1	0,08	518,5	62,0	0,12	147,4	88,3	0,60
Syntagmes nominaux	736,5	63,0	**0,09**	478,1	56,3	0,12	141,4	75,6	0,54
LR-NP	**658,2**	60,1	**0,09**	**423,8**	59,0	**0,14**	**135,3**	84,8	**0,63**

Évaluation extrinsèque :

Méthode	**DEft** *(fr)*		**SemEval** *(en)*		**DUC** *(en)*	
	TF-IDF	KEA	TF-IDF	KEA	TF-IDF	KEA
	F	F	F	F	F	F
N-grammes	8,9	**20,0**	7,7	16,2	17,7	14,3
/(NOM | ADJ)+/	13,2	18,4	10,5	17,1	27,3	16,8
Syntagmes nominaux	12,8	18,7	10,6	16,9	24,1	15,7
LR-NP	**13,3**	19,2	**10,9**	**17,7**	**27,4**	**16,9**

- La sélection des termes-clés candidat influe $\pm$ sur l'extraction de termes-clés
- Certaines catégories de mots sont plus utiles dans les termes-clés
- Les adjectifs ne sont pas tous utiles dans les termes-clés

Groupement hiérarchique :

1. Un groupe par candidat
2. Groupement des deux groupes de plus forte similarité moyenne entre tous leurs candidats c_1 et c_2
   - $\text{sim}(c_1, c_2) = \frac{\text{racines}(c_1) \cap \text{racines}(c_2)}{\text{racines}(c_1) \cup \text{racines}(c_2)}$
3. Répéter 2 jusqu'à ce que la plus forte similarité soit $< 1/4$

- Les nœuds sont les sujets
- Tous les nœuds sont connectés entre eux
- Les arêtes entre deux nœuds n_1 et n_2 sont pondérées
  - $\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{dist}(c_i, c_j)$
  - $\text{dist}(c_i, c_j) = \sum_{p_i \in \mathbf{pos}(c_i)} \sum_{p_j \in \mathbf{pos}(c_j)} \frac{1}{|p_i - p_j|}$

Méthode	DEft *(fr)*			Wikinews *(fr)*			SemEval *(en)*			DUC *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	4,5	9,0	5,9	19,4	20,7	19,7	4,6	3,2	3,7	**22,3**	**28,4**	**24,6**
+ complet	4,4	9,0	5,8	20,0	21,4	20,3	5,5	3,8	4,4	22,2	28,1	24,5
+ candidats	10,3	19,2	13,2†	28,5	30,0	28,8†	9,4	6,8	7,8†	10,4	13,5	11,6
+ sujets	11,1	20,4	14,2†	30,7	32,6	31,1†	14,2	9,9	11,6†	18,9	24,2	21,0
TopicRank	**11,7**	**21,7**	**15,1**†	**35,0**	**37,5**	**35,6**†	**14,9**	**10,3**	**12,1**†	18,3	23,8	20,4

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	9.0	10,6	9,6	9,5	10,0	9,4	12,7	8,9	10,2	13,0	10,4	11,0
+ complet	10,0	11,9	10,7	9,9	10,2	9,8	13,5	9,5	11,0	13,0	10,7	11,2
+ candidats	10,8	12,7	11,5	11,1	11,6	11,0	25,7	17,4	20,3	**14,2**	**11,1**	**11,9**
+ sujets	10,6	12,5	11,3	10,9	11,5	10,8	26,5	18,0	20,9	13,5	10,7	11,5
TopicRank	**11,2**	**13,1**	**11,9**	**12,1**	**12,8**	**12,1**	**27,5**	**18,7**	**21,8**	13,8	11,1	11,8

Deux vote pour déterminer l'importance des nœuds :

- Interne, c-à-d depuis les nœuds du même graphe
  - $R_{interne}(n_i) = \sum_{n_j \in A_{\textbf{interne}}(n_i)} \frac{\text{poids}(n_j,n_i) \times S(n_j)}{\sum_{n_k \in A_{\textbf{interne}}(n_j)} \text{poids}(n_j,n_k)}$
- Externe, c-à-d depuis les nœuds de l'autre graphe
  - $R_{externe}(n_i) = \sum_{n_j \in A_{\textbf{externe}}(n_i)} \frac{S(n_j)}{|A_{\textbf{externe}}(n_j)|}$

$$S(n_i) = (1-\lambda)\ R_{externe}(n_i) + \lambda\ R_{interne}(n_i)$$

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TopicRank	**11,2**	**13,1**	**11,9**	**12,1**	**12,8**	**12,1**	**27,5**	**18,7**	**21,8**	13,8	11,1	11,8
TopicRank'	10,5	12,3	11,1	11,3	11,9	11,2	25,0	17,0	19,8	13,1	10,3	11,0