

Similarités sémantiques et spatialité : méthodes d'autocorrelation spatiale appliquées au texte

Guillaume Guex guillaume.guex@unil.ch

Mattia Egloff mattia.egloff@unil.ch

François Bavaud francois.bavaud@unil.ch

SLI - UNIL

Décembre 2020

Les similarités sémantiques

La spatialité

Autocorrélation Globale

Autocorrélation Locale

Clustering

Dans ce travail, les unités étudiées sont les **tokens** d'un corpus.

Deux aspects sont mis en relation :

- Les **similarités sémantiques** entre tokens, obtenues via des sources externes.
- La **spatialité** des tokens, c'est-à-dire leur position les uns par rapport aux autres.

De ces deux ingrédients, nous pourrons obtenir :

- Un indice d'**autocorrélation globale** des tokens d'un corpus.
- Un indice d'**autocorrélation locale** pour chacun des tokens.
- Un **clustering** des tokens.

GitHub : https://github.com/sliunil/SemSim_AutoCor

Les similarités sémantiques

Les similarités sémantiques

En anglais, il existe une distinction entre *semantic similarity* et *semantic relatedness* :

- La **semantic similarity** désigne le fait que deux mots possèdent une relation de type hyponymie-hyperonymie (p.ex : chien et animal)
- La **semantic relatedness** est plus large, désignant des relations de n'importe quel type, comme l'holonymie-méronymie (p.ex : chien et truffe), l'antonymie (p.ex : mort et vivant) ou même une relation de type « fait partie du même univers » (p.ex : voiture et route).

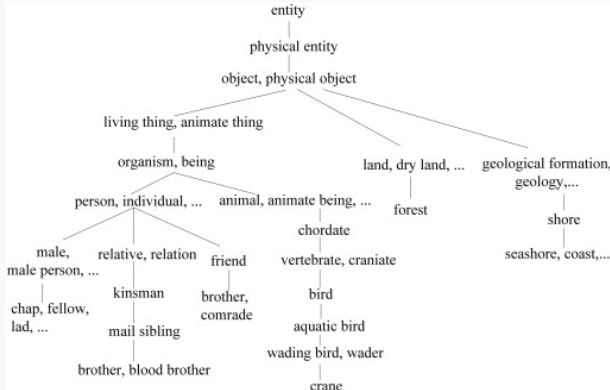
Dans ce travail, lorsque nous parlons de **similarités sémantiques**, il s'agit donc plutôt de la deuxième définition :

Deux mots sont proches sémantiquement s'il existe une forte relation, de n'importe quel type, entre (l'un de) leur(s) signifié(s).

Dans la pratique, il existe plusieurs manières d'obtenir une **mesure de similarité sémantique** entre deux termes. Dans ce travail, ces mesures vont être extraites via deux types d'objet :

- En utilisant l'ontologie **WordNet**.
- En utilisant un **plongement lexical (Word Embedding)** pré-entraîné sur un très grand corpus.

WordNet [Fellbaum, 1998] est une ontologie entre différents **synsets** (ensembles de mots représentant un concept). Plusieurs relations (antonymie, holonymie-meronymie, etc.) sont représentées dans cette ontologie, mais la plus complète est celle d'**hyponymie-hyperonymie**.



Il existe de nombreuses similarités basées sur WordNet.

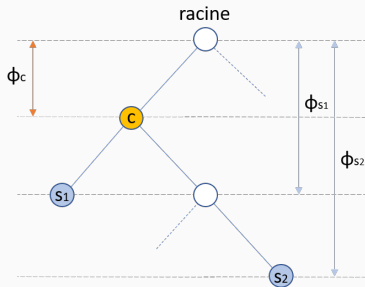
La **similarité de Wu-Palmer**

[Wu and Palmer, 1994] entre deux synsets s_1 et s_2 :

$$s_{s_1 s_2}^{wup} = \max_{c \in S(s_1, s_2)} \left(\frac{2\phi_c}{\phi_{s_1} + \phi_{s_2}} \right)$$

$S(s_1, s_2)$ est l'**ensemble des synsets hyperonymes** à s_1 et s_2 .

ϕ_s est la **profondeur** du synset s , c'est-à-dire la longueur du chemin le reliant à la racine.



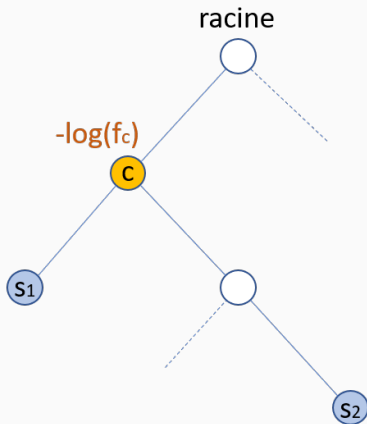
La **similarité de Resnik**

[Resnik, 1995] entre deux synsets
 s_1 et s_2 :

$$s_{s_1 s_2}^{res} = \max_{c \in S(s_1, s_2)} (-\log(f_c))$$

$S(s_1, s_2)$ est l'**ensemble des synsets hyperonymes** à s_1 et s_2 .

f_c est la **probabilité d'utilisation** du synset c (estimée depuis un corpus donné)



La **similarité de Leacock-Chodorow**

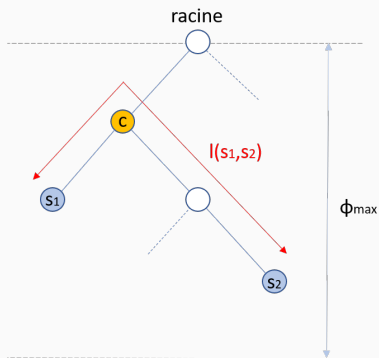
[Leacock and Chodorow, 1998]

entre deux synsets s_1 et s_2 :

$$s_{s_1 s_2}^{lec} = -\log \left(\frac{l(s_1, s_2)}{2\phi_{\max}} \right)$$

$l(s_1, s_2)$ est la **longueur du chemin** reliant s_1 à s_2 .

ϕ_{\max} est la **profondeur maximale** de WordNet.



Il existe cependant des **difficultés** à utiliser les similarités WordNet afin de calculer une **similarité entre deux tokens** d'un corpus donné :

- **Sans désambiguïsation**, il n'est pas possible de savoir à quels synsets ces tokens appartiennent.
- WordNet contient **plusieurs racines**, pour les noms, verbes, adjectifs et adverbes. Il est difficile de calculer une similarité en passant d'un arbre à un autre.

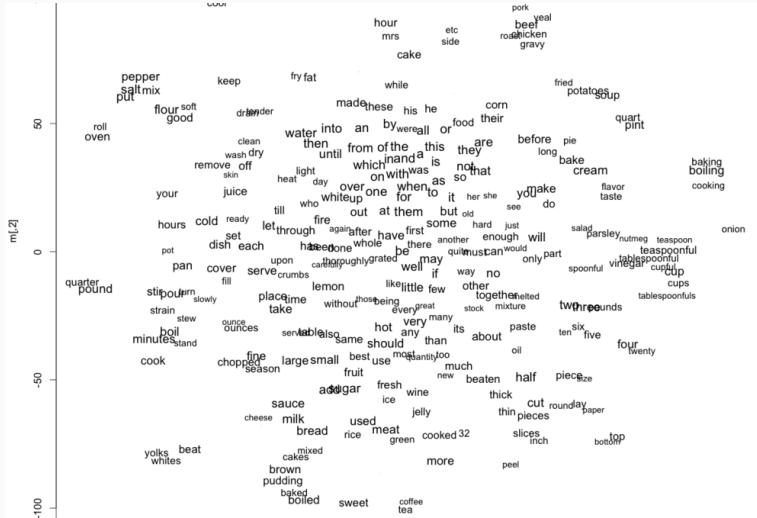
Les plongements lexicaux

Les méthodes de **plongement lexical** sont nombreuses : Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], FastText [Bojanowski et al., 2017], etc. mais elles partagent toutes le même principe :

- On utilise un grand corpus pour **transformer chaque type en un vecteur** de dimension r .
- Ces vecteurs sont calculés depuis le corpus pour que **les types possédant un contexte similaire** (i.e. une fenêtre de taille $\pm t$ tokens) donnent **des vecteurs proches dans l'espace**.
- La similarité sémantique entre deux types peut ensuite être obtenue grâce à la **similarité du cosinus** entre leurs deux vecteurs :

$$s_{kl}^{\text{we}} = \cos(\mathbf{v}_k, \mathbf{v}_l) = \frac{\mathbf{v}_k^\top \mathbf{v}_l}{\|\mathbf{v}_k\| \|\mathbf{v}_l\|}$$

Les plongements lexicaux



Source : <https://www.adityathakker.com/introduction-to-word2vec-how-it-works/>

Ici, nous allons utiliser :

- La méthode **Word2Vec**
- appliquée sur **Wikipedia** (anglais)
- avec une fenêtre de **± 5 tokens**
- qui donneront des vecteurs à **300 dimensions**.

Ces vecteurs pré-entraînés (dans plusieurs langues) se trouvent sur <https://wikipedia2vec.github.io/wikipedia2vec/>

Les similarités obtenues via des plongements lexicaux présentent également des **désavantages** :

- Il y a un **unique vecteur par type**, représentant donc une moyenne de tous les sens possibles.
- Ces vecteurs représentent les **sens trouvés dans le corpus** utilisé pour l'entraînement et peuvent être différents dans le texte étudié.
- Plus gênant pour nos méthodes, ces méthodes utilisent la **spatialité** pour construire les **similarités sémantiques**.

Transformation d'une similarité en dissimilarité

Dans la suite, nous avons généralement besoin d'une **dissimilarité plutôt qu'une similarité**. Il existe plusieurs transformations possibles, par exemple :

$$d_{ij}^{\text{minus_log}} = \begin{cases} -\log(s_{ij}) & \text{si } s_{kl} > 0, \forall k, l \\ -\log(s_{ij} - \min_{kl} s_{kl} + \epsilon) & \text{sinon} \end{cases}$$
$$d_{ij}^{\text{max_minus}} = \max_{kl} s_{kl} - s_{ij}$$

Dans la pratique, la transformation **minus_log** semble donner des résultats plus convaincants, nous n'allons donc utiliser que cette dernière.

Les similarités sémantiques : pour résumé

L'étude portera sur 3 similarités sémantiques issues de WordNet pour calculer la similarité entre deux tokens i et j :

- La **similarité de Wu-Palmer** : s_{ij}^{wup}
- La **similarité de Resnik** : s_{ij}^{res}
- La **similarité de Leacock-Chodorow** : s_{ij}^{lec}

Pour ces similarités, le **sens le plus fréquemment utilisé** définira les synsets de i et j et elles ne seront appliquées qu'à des **tokens d'une même catégorie (nom, verbe, adjectif, adverbe)**. De plus, on aura :

- La **similarité issue de Word2Vec sur Wikipedia** : s_{ij}^{we}

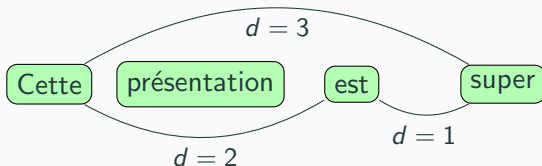
où cette similarité sera obtenue **en fonction du type de i et du type de j** .

Les similarités seront transformées en dissimilarités avec `minus_log`.

La spatialité

La spatialité

La **spatialité** des tokens est plus facile à définir, il s'agit **leur position dans le texte, vu comme un espace unidimensionnel**.



Ici, on préfère définir la spatialité via une **relation de voisinage** c'est-à-dire via une **matrix d'adjacence A** de taille $n_{\text{token}} \times n_{\text{token}}$:

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

Il est cependant possible de définir ce voisinage avec **une fenêtre de taille r** , permettant de représenter plus fidèlement notre manière de lire. Il y a cependant deux choix possibles :

Un **voisinage uniforme** :

$$a_{i:} = (\dots \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ \dots)$$



Un **voisinage gaussien** :

$$a_{i:} = (\dots \ 0.004 \ 0.054 \ 0.242 \ 0.399 \ 0.242 \ 0.054 \ 0.004 \ \dots)$$



La matrice d'échange

Dans notre formalisme, nous aurons besoin d'une **matrice d'échange** $E = (e_{ij})$. Les composantes e_{ij} **représentent les liens de voisinage entre les tokens i et j** et cette dernière doit vérifier :

- $e_{ij} \geq 0$
- $e_{ij} = e_{ji}$
- $e_{i\bullet} = f_i$

où f_i **sont les poids des tokens** (généralement uniforme).

Des méthodes particulières (utilisant le Laplacien d'un graphe et un algorithme de Metropolis-Hastings) permettent d'**obtenir E à partir de A** , tout en respectant les poids des tokens f_i .

La **spatialité** des tokens sera définie via une **relation de voisinage** contenue dans une **matrice d'échange** $E = (e_{ij})$ symétrique, où e_{ij} représente le lien entre les tokens i et j , et $e_{i\bullet} = f_i$ le poids du token i .

Un entier $r \geq 1$ va définir la largeur de voisinage que l'on considère.

Il y a deux choix sur la distribution des liens dans le voisinage :

- Une distribution **uniforme** sur une fenêtre de $\pm r$: E_r^{unif}
- Une distribution **normale** de « diffusion » r : E_r^{norm}

Nous avons défini deux objets sur les tokens :

- Une **matrice de dissimilarité** entre tokens $D = (d_{ij})$, quantifiant à quel point ces derniers sont **sémantiquement éloignés**.
- Une **matrice d'échange** entre tokens $E = (e_{ij})$, définissant leurs **relations spatiales**.

Ces deux ingrédients vont nous permettre d'obtenir :

- Une mesure d'**autocorrelation globale** [Bavaud et al., 2015], quantifiant en moyenne sur un corpus donné à quel point les tokens voisins sont similaires **sémantiquement**.
- Une mesure d'**autocorrelation locale**, quantifiant **pour chaque token** d'un corpus donné à quel point celui-ci est similaire **sémantiquement** à son **voisinage**.
- Un **clustering des tokens** [Céré and Bavaud, 2017], qui permet de **regrouper les tokens proches** **sémantiquement ET spatialement**.

Autocorrélation Globale

L'autocorrélation globale

Sur un corpus donné, on peut définir l'**inertie globale** des tokens (de poids f_i) avec :

$$\Delta := \frac{1}{2} \sum_{ij} f_i f_j d_{ij}$$

cette dernière quantifie l'**éloignement sémantique moyen entre tous les tokens**.

De manière similaire, on peut définir l'**inertie locale** avec :

$$\Delta_{\text{loc}} := \frac{1}{2} \sum_{ij} e_{ij} d_{ij}$$

cette fois-ci, il s'agit de l'**éloignement sémantique moyen entre tokens voisins**.

On peut alors définir un **indice d'autocorrélation globale** avec :

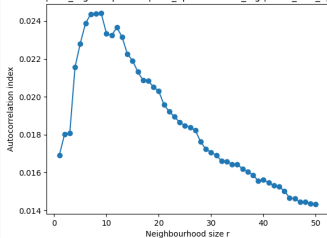
$$\delta := \frac{\Delta - \Delta_{\text{loc}}}{\Delta}$$

qui peut prendre des valeurs entre -1 et 1 .

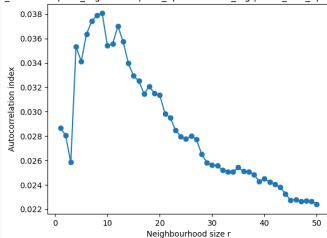
L'autocorrélation globale : résultats

The Wonderful Wizard of Oz (L. Frank Baum) - nouns, E_r^{unif} :

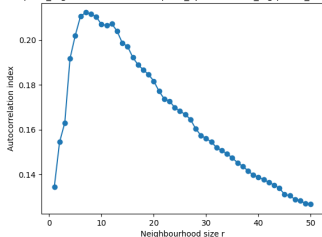
nouns.txt | sim_tag: wu-palmer | dist_option: minus_log | exch_mat_opt: s | exch



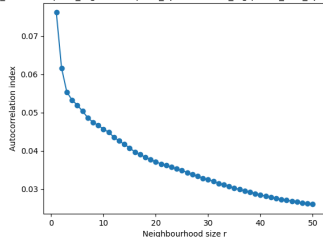
nouns.txt | sim_tag: resnik | dist_option: minus_log | exch_mat_opt: s | exch



nouns.txt | sim_tag: leacock-chodorow | dist_option: minus_log | exch_mat_opt: s | exch



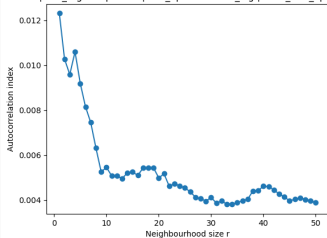
nouns.txt | sim_tag: wesim | dist_option: minus_log | exch_mat_opt: s | exch



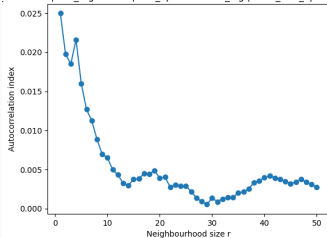
L'autocorrélation globale : résultats

Animal Farm (G. Orwell) - nouns, E_r^{unif} :

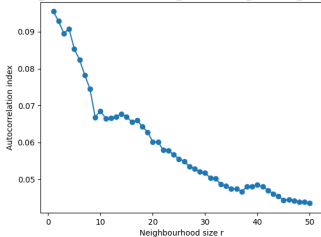
nouns.txt | sim_tag: wu-palmer | dist_option: minus_log | exch_mat_opt: s | exch_mat_opt: s



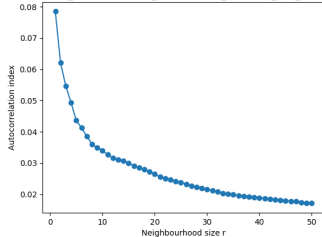
nouns.txt | sim_tag: resnik | dist_option: minus_log | exch_mat_opt: s | exch_mat_opt: s



nouns.txt | sim_tag: leacock-chodorow | dist_option: minus_log | exch_mat_opt: s | exch_mat_opt: s



nouns.txt | sim_tag: wesim | dist_option: minus_log | exch_mat_opt: s | exch_mat_opt: s



Autocorrélation Locale

L'autocorrélation locale

On peut également définir un **indice d'autocorrélation locale** pour chaque token i avec :

$$\delta_i = \frac{\sum_k w_{ik} b_{ki}}{\Delta}$$

où :

- $W = (w_{ij})$ est une **matrice de transition** de la chaîne de Markov issue de E , i.e. $W := \text{Diag}(f)^{-1}E$.
- $B = (b_{ij})$ est la **matrice des produits scalaires**, issue de D , i.e. $B := -\frac{1}{2}HDH^\top$ avec $H := I_n - \mathbf{1}_n f^\top$.

Cet indice mesure donc **la similarité sémantique moyenne** entre i et ses voisins.

On peut montrer que l'**autocorrélation globale** est la **moyenne (pondérée)** des autocorrélations locales :

$$\delta = \sum_i f_i \delta_i$$

L'autocorrélation locale : résultats

The Wonderful Wizard of Oz (L. Frank Baum), S^{we} , E_{100}^{unif} :

chapter 1 dorothy lived midst great kansa prairie uncle henry wa farmer aunt em wa farmer wife house wa small lumber build had 12 carried wagon many mile were wall floor roof made room room contained rusty looking cooking stove cupboard dish table chair bed uncle henry aunt em had big bed corner dorothy little best corner wa garret at all cellar small hole bug ground called cyclone bella family no case great whirlwind arose mighty enough crush building path wa reached trap door middle floor ladder led down small dark hole dorothy stood doorway looked around see nothing great gray prairie side not tree house broke broad sweep flat country reached edge sky direction sun had baked plowed land gray mass little crack running even grass wa not green sun had burned top long blade were same gray color be seen everywhere once house had been painted sun blistered paint rain washed away now house wa a dull gray everything else aunt em came there live wa young pretty wife sun wind had changed too had taken sparkle eye left sober gray had taken red cheek lip were gray also wa thin gaunt never smiled now dorothy wa orphan first came aunt em had been so startled child laughter scream press hand heart dorothy merry voice reached ear still looked little girl wonder find anything laugh uncle henry never laughed worked hard morning night did not know joy wa wa gray also long beard rough boot looked stern solemn rarely spoke wa to to made dorothy laugh saved growing a gray other surroundings to to wa not gray wa little black dog long silky hair small black eye twinkled merrily side funny wee nose too played day long dorothy played loved dearly to day however were not playing uncle henry sat door step looked anxiously sky wa even grayer usual dorothy stood door arm looked sky too aunt em wa washing dish far north heard low wail wind uncle henry dorothy see long grass bowed wave coming storm there now came sharp whistling air south turned eye way saw ripple grass coming direction also suddenly uncle henry stood cyclone coming called wife go look stock then ran shed cow horse were kept aunt em dropped work came door glance told danger close hand quick dorothy screamed run cellar took jumped dorothy arm had bed girl started get aunt em badly frightened threw trap door floor climbed ladder small dark hole dorothy caught too last started follow aunt wa half way room came great shriek wind house shook so hard lost footing sat suddenly floor strange thing then happened house whirled around time rose slowly air dorothy felt were going balloon north south wind met house stood made exact center cyclone middle cyclone air is generally still great pressure wind side house raised higher higher wa very top cyclone there remained wa carried mile away a easily easy feather wa very dark wind howled horribly dorothy found wa riding quite easily first few whirl around other time house tipped badly felt were being rocked gently baby cradle toto did not like ran room now here now there barking loudly dorothy sat quite still floor waited see happen once toto got too open trap door fell first little girl thought had lost soon saw ear sticking hole strong pressure air wa keeping not fall creep hole caught too ear dragged room again afterward closing trap door more accident happen hour hour passed slowly dorothy got fright felt quite lonely wind shrieked so loudly all nearly became deaf first had wondered be dashed piece house fell again hour passed nothing terrible happened stopped worrying resolved wait calmly see future lying soft bed crawled swaying floor bed lay toto followed lay spate swaying house wailing wind dorothy soon closed eye fell fast asleep chapter wa awakened shock so sudden severe dorothy had not been lying soft bed have been hurt wa jar made catch breath wonder had happened toto put cold little nose face whined dimally dorothy sat noticed house wa not moving wa dark bright sunshine came in window flooding little room sprang bed toto heel ran opened door little girl gave cry amazement looked eye growing bigger bigger wonderful sight saw cyclone had set house very gently cyclone midst country marvelous beauty were lovely patch green sward all about stately tree bearing rich luxuriant luscious fruit bank gorgeous flower were hand bird rare brilliant plumage sang fluttered tree bush little way off wa small brook rushing sparkling along green bank murmuring voice very grateful little girl had lived so long dry gray prairie stood looking eagerly strange beautiful sight noticed coming group queerest people had ever seen were not a big grown folk had always been used were very small fact seemed about a tall dorothy wa well grown child age were so far look so many year older were men woman were oddly dressed wore round hat rose small point foot head little bell brim sweetly moved hat men were blue little woman hat wa white wore white gown hung plait shoulder were sprinkled little star glinted sun diamond men were dressed blue same shade hat wore well polished boot deep roll blue top men dorothy thought were about a old uncle henry had beard little woman wa doubtless much older face wa covered wrinkle hair wa nearly white walked rather stiffly people drew house dorothy wa standing doorway paused whispered afraid come farther little old woman walked dorothy made low bow said sweet voice are welcome most noble sorceress land munchkins are so grateful having killed wicked witch asked setting people free bondage dorothy listened speech wonder little woman possibly mean calling sorceress saying had killed wicked witch east dorothy wa innocent harmless little girl had been carried cyclone many mile house had never killed anything life little woman evidently expected answer so dorothy said hesitation are very kind be mistake have not killed anything house did anyway replied little old woman laugh is same thing see continued pointing corner house are too still sticking block wood dorothy looked gave little cry fright there indeed just corner great beam house rested foot were sticking shod silver shoe pointed toe dear cried dorothy clasping hand together dismay house have fallen ever do is nothing be done said little woman calmly wa asked dorothy wa wicked witch east said answered little woman ha held munchkins bondage many year masking slave night day now are all set free are grateful favour are munchkins enquired dorothy are people live land east wicked witch ruled are munchkin asked dorothy am friend live land north saw witch east wa dead munchkins sent swift messenger came once am witch north cried dorothy are real witch indeed answered little woman am good witch people love am not a powerful wicked witch wa ruled here have set people free thought witch were wicked said girl wa half frightened facing real witch is great mistake were only witch land oz live north south are good witch know is true am be mistaken dwell east west were indeed wicked witch now have killed is wicked witch land oz one life west said dorothy moment thought aunt em ha told witch were all dead year year ago is aunt em inquired little old woman is aunt life kansa came witch north seemed think time head bowed eye ground then looked said do not know kansa is have never heard country mentioned before tell is civilized country replied dorothy then account civilized country believe are witch left wizard sorceress magician see land oz ha never been civilized are cut rest world therefore still have witch wizards are wizard asked dorothy oz is great wizard answered witch sinking voice whisper is more powerful rest together life city emerald dorothy wa going ask question just then munchkins had been standing silently by gave loud shout pointed corner house wicked witch had been lying is asked little old woman looked began laugh

Clustering

Il est également possible d'utiliser ces deux ingrédients pour obtenir un **clustering flou des tokens** d'un corpus.

Posons p le nombre de groupes. Le résultat d'un clustering flou sur les n tokens peut être donnée par une **matrice d'appartenance** $\mathbf{Z} = (z_{ig})$, de taille $(n \times p)$, qui vérifie :

- $z_{ig} \geq 0$
- $z_{i\bullet} = 1$

La quantité z_{ig} nous donnera donc l'**appartenance (en pourcent)** du token i au groupe g .

Clustering : l'inertie intra-groupe

Plusieurs quantités peuvent être calculée à partir d'une matrice d'appartenance \mathbf{Z} .

L'**inertie intra-groupe** $\Delta_W[\mathbf{Z}]$, définie par

$$\Delta_W[\mathbf{Z}] := \sum_g \rho_g \Delta_g$$

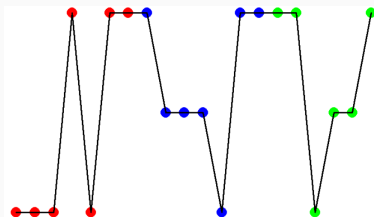
où :

- $\rho_g := \sum_i f_i z_{ig}$ est le **poids du groupe** g .
- $\Delta_g := \sum_{ij} \frac{f_i z_{ig}}{\rho_g} \frac{f_j z_{jg}}{\rho_g} d_{ij}$ est l'**inertie du groupe** g

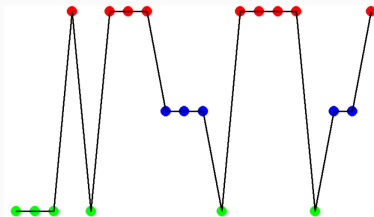
Clustering : l'inertie intra-groupe

Cette quantité mesure donc **la variation sémantique des tokens à l'intérieur des groupes** :

$\Delta_W[\mathbf{Z}]$ élevé



$\Delta_W[\mathbf{Z}]$ faible



La **modularité généralisée** $C^\kappa[\mathbf{Z}]$, définie par

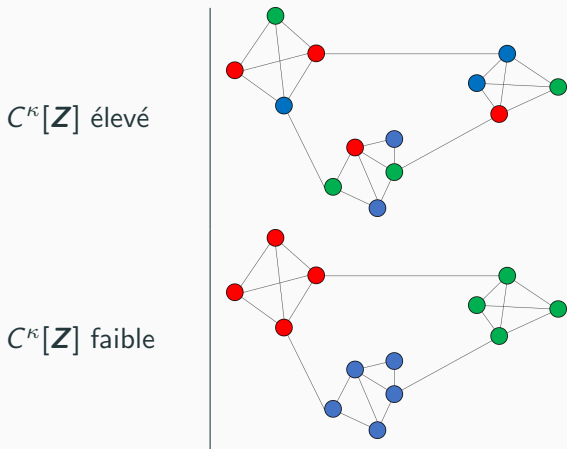
$$C^\kappa[\mathbf{Z}] := \sum_g \frac{\rho_g^2 - e(g, g)}{\rho_g^\kappa}$$

où :

- $\rho_g := \sum_i f_i z_{ig}$ est le **poids du groupe** g .
- $e(g, g) := \sum_{ij} e_{ij} z_{ig} z_{jg}$ est la **moyenne pondérée du nombres de liens (de voisinage) dans le groupe** g .
- $\kappa \in [0, 1]$ un **paramètre libre**, permettant de passer d'un objectif de modularité ($\kappa = 0$) à un objectif de N-cut ($\kappa = 1$)

Clustering : l'inertie intra-groupe

Cette quantité mesure donc **la quantité des liens (de voisinage) entre les groupes** :



Finalement, on peut aussi définir l'**information mutuelle** $K[\mathbf{Z}]$ par :

$$K[\mathbf{Z}] = \sum_{ig} f_i z_{ig} \frac{z_{ig}}{\rho_g}$$

Cette quantité mesure la **dépendance entre les groupes et les tokens**, i.e. :

- $K[\mathbf{Z}]$ sera **élevé** si chaque token appartient à un groupe bien défini (solution dure).
- $K[\mathbf{Z}]$ sera **faible** si chaque token appartient en partie à chaque groupe (solution floue).

Clustering : la fonctionnelle

Le but sera de chercher un clustering \mathbf{Z} qui minimise l'énergie libre $F[\mathbf{Z}]$, définie comme :

$$F[\mathbf{Z}] = \beta \Delta_w[\mathbf{Z}] + \frac{\alpha}{2} C^\kappa[\mathbf{Z}] + K[\mathbf{Z}]$$

où $\alpha > 0$, $\beta > 0$ et $\kappa \in [0, 1]$ sont des paramètres libres.

- Augmenter α relativement à β favorisera l'émergence de groupes spatialement continus, i.e. de longues suites de tokens.
- Augmenter β relativement à α favorisera l'émergence de groupes homogènes, sémantiquement parlant.
- Baisser α ET β favorisera l'émergence de groupes flous.
- κ fait passer l'objectif spatial d'un critère de modularité ($=0$) à un critère du N-cut ($=1$).

Un algorithme itératif permet de trouver la solution \mathbf{Z} qui minimise cette fonctionnelle.

Il est difficile de se rendre compte si les groupes de tokens trouvés correspondent à quelque chose, surtout qu'il y a **beaucoup d'hyperparamètres** :

- Choix de la similarité sémantique.
- Choix du voisinage : E_r^{unif} ou E_r^{norm} , définir r .
- Choix du nombre de groupes p .
- Choix des hyperparamètres de la fonctionnelle : α , β et κ .

Pour essayer de voir si ce clustering est **capable de regrouper des passages sémantiquement proches**, nous avons créé une **technique d'évaluation originale**.

Clustering : résultats préliminaires

Quatre livres, parlant de sujets à priori bien différents, ont été sélectionnés sur le projet Gutenberg. Il s'agit de :

- *Sidelights on relativity*, livre de vulgarisation de physique de Albert Einstein,
- *Metamorphosis*, roman de Franz Kafka.
- *On the Duty of Civil Disobedience*, essai politique de Henry David Thoreau.
- *Lectures On Landscape*, recueil de cours académiques sur la peinture de paysage de John Ruskin.

A partir de ces livres, **4 corpus ont été créés en mélangeant les différents livres** :

- **Mix_word1**, où chaque **token** est tiré aléatoirement d'un des livres.
- **Mix_word3**, où chaque séquence de **3 token** est tirée aléatoirement d'un des livres.
- **Mix_sent1**, où chaque **phrase** est tirée aléatoirement d'un des livres.
- **Mix_sent10**, où chaque séquence de **10 phrases** est tirée aléatoirement d'un des livres.

Chaque phrase doit comporter 5 tokens minimum et les corpus sont construits afin d'avoir des nombres relativement égaux de tokens issus de chaque livre.

On va effectuer un **recherche sur grille** pour les hyperparamètres et mesurer à chaque fois l'adéquation entre les groupes obtenus et les vrais groupes grâce à l'indice d'**information mutuelle normalisé (NMI)**. Les paramètres posés/explorés sont :

- La similarité sémantique issue du word embedding : S^{we}
- La matrice d'échange gaussienne : E_r^{norm}
- Le facteur de diffusion de la matrice d'échange :
 $r \in \{3, 5, 10, 15\}$.
- Le nombre de groupes sera posé à $p = 4$.
- Le paramètre $\alpha \in \{0.1, 1, 2, 5, 10, 50, 100\}$.
- Le paramètre $\beta \in \{0.1, 1, 5, 10, 50, 100, 300\}$.
- Le paramètre $\kappa \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$.

Dataset	Best parameters	NMI	NMI - LDA
Mix_word1	$r = 3, \alpha = 2, \beta = 100, \kappa = 0$	0.05	0.0026
Mix_word3	$r = 3, \alpha = 2, \beta = 10, \kappa = 1/3$	0.06	0.0029
Mix_sent1	$r = 15, \alpha = 2, \beta = 10, \kappa = 1/3$	0.16	0.0025
Mix_sent10	$r = 15, \alpha = 2, \beta = 10, \kappa = 1/3$	0.35	0.0025

Clustering : résultats préliminaires

Gauche : meilleurs résultats. Droite : Mix_sent1

metamorphosis gregor samza woke troubled found transformed bed horrible inaugural stated holding
professorship direct practical chiefly natural history having course past year laid foundational element
art sufficiently invite enter real work accordingly propose following term give practical leading
elementary arch branch natural history form kind center lay lifted head little see brown slightly domed
divided arch sniff sidelight relativity ether theory relativity address delivered may university leyden des
come idea ponderable derived abstraction everyday physicist set idea existence kind outset shortly
state position landscape painting animal painting hold higher branch landscape painting thoughtful
passionate representation physical condition appointed human bedding wa hardly able cover seemed
ready slide imitates record visible thing are dangerous beneficial display human method dealing
enjoying suffering are exemplary deserving sympathetic duty civil disobedience heartily accept
government is best governs like see acted more rapidly explanation is probably be sought phenomenon
have given rise theory action property light have led undulatory animal painting investigates law
greater le nobility character organic comparative anatomy examines greater le development organic
function animal painting is bring notice minor unthought condition power physiology is ascertain
minor condition question purpose arrangement use organ animal le province painter are indeed more
likely commend drawing many pitifully then compared size rest waved about helplessly let devote little
while consideration proper human room little too lay peacefully familiar outside physic know nothing
action carried finally amount also government is best governs not men are prepared be kind
government government is best most government are government are try connect cause effect
experience natural object afford seems first were other mutual action immediate dissect animal
generally assume form be necessary only examine is drawing outer form attentively are led necessarily
consider mode life is therefore be struck awkwardness apparent uselessness collection textile sample
lay spread table samza wa travelling salesman there hung picture had recently cut illustrated magazine
housed gilded sketching day several head bird became vital matter interest know use bony process head
taking great found appeared abroad wa certainly unanswerable communication notion push heating
inducing combustion mean is true even everyday experience is sense action play very important
objection have been brought standing are many deserve also at last be brought standing showed lady
fitted fur hat fur box sat raising heavy fur muff covered whole lower arm gregor then turned look
window dull drop rain be heard hitting made feel quite have have just landscape painting representation
phenomenon relating human daily experience weight body meet something something not linked cause
is variable time do not everyday life speculate cause therefore do not become conscious character
action wa theory gravitation first assigned cause gravity interpreting action proceeding scarcely be
disposed admit propriety still le be likely conceive necessary strictness convince somewhat detailed
theory is probably greatest stride ever made effort causal nexus natural sleep little bit longer forget wa
something wa unable do wa used sleeping present state get however hard threw always rolled back
standing army is only arm standing yet theory evoked lively sense discomfort seemed be conflict
principle springing rest be reciprocal action only not immediate action here are landscape turner
greatest vesuvius government is only mode people have chosen execute is equally liable be abused
perverted people act witness present mexican work comparatively few individual using standing

metamorphosis gregor samza woke troubled found transformed bed horrible inaugural stated holding
professorship direct practical chiefly natural history having course past year laid foundational element
art sufficiently invite enter real work accordingly propose following term give practical leading
elementary arch branch natural history form kind center lay lifted head little see brown slightly domed
divided arch sniff sidelight relativity ether theory relativity address delivered may university leyden des
come idea ponderable derived abstraction everyday physicist set idea existence kind outset shortly
state position landscape painting animal painting hold higher branch landscape painting thoughtful
passionate representation physical condition appointed human bedding wa hardly able cover seemed
ready slide imitates record visible thing are dangerous beneficial display human method dealing
enjoying suffering are exemplary deserving sympathetic duty civil disobedience heartily accept
government is best governs like see acted more rapidly explanation is probably be sought phenomenon
have given rise theory action property light have led undulatory animal painting investigates law
greater le nobility character organic comparative anatomy examines greater le development organic
function animal painting is bring notice minor unthought condition power physiology is ascertain
minor condition question purpose arrangement use organ animal le province painter are indeed more
likely commend drawing many pitifully then compared size rest waved about helplessly let devote little
while consideration proper human room little too lay peacefully familiar outside physic know nothing
action carried finally amount also government is best governs not men are prepared be kind
government government is best most government are government are try connect cause effect
experience natural object afford seems first were other mutual action immediate dissect animal
generally assume form be necessary only examine is drawing outer form attentively are led necessarily
consider mode life is therefore be struck awkwardness apparent uselessness collection textile sample
lay spread table samza wa travelling salesman there hung picture had recently cut illustrated magazine
housed gilded sketching day several head bird became vital matter interest know use bony process head
taking great found appeared abroad wa certainly unanswerable communication notion push heating
inducing combustion mean is true even everyday experience is sense action play very important
objection have been brought standing are many deserve also at last be brought standing showed lady
fitted fur hat fur box sat raising heavy fur muff covered whole lower arm gregor then turned look
window dull drop rain be heard hitting made feel quite have have just landscape painting representation
phenomenon relating human daily experience weight body meet something something not linked cause
is variable time do not everyday life speculate cause therefore do not become conscious character
action wa theory gravitation first assigned cause gravity interpreting action proceeding scarcely be
disposed admit propriety still le be likely conceive necessary strictness convince somewhat detailed
theory is probably greatest stride ever made effort causal nexus natural sleep little bit longer forget wa
something wa unable do wa used sleeping present state get however hard threw always rolled back
standing army is only arm standing yet theory evoked lively sense discomfort seemed be conflict
principle springing rest be reciprocal action only not immediate action here are landscape turner
greatest vesuvius government is only mode people have chosen execute is equally liable be abused
perverted people act witness present mexican work comparatively few individual using standing

Clustering : résultats préliminaires

Gauche : meilleurs résultats. Droite : Mix_sent10

metamorphosis gregor samia woke troubled found transformed bed horrible lay lifted head little see
brown slightly domed divided arch stiff bedding wa hardly able cover seemed ready slide many
pitifully thin compared size rest waved about helplessly proper human room little too lay peacefully
familiar collection textile sample lay spread table samia wa travelling salesman there hung picture had
recently cut illustrated magazine housed gilded showed lady fitted fur hat fur boa sat raising heavy fur
muff covered whole lower arm gregor then turned look window dull drop rain be heard hitting made
feel quite sleep little bit longer forget wa something wa unable do wa used sleeping present state get
inaugural stated holding professorship direct practical chiefly natural history having course past year
laid foundational element art sufficiently invite enter real work accordingly propose following term
give practical leading elementary study branch natural history form kind center outset shortly state
position landscape painting animal painting hold higher branch landscape painting is thoughtful
passionate representation physical condition appointed human imitates record visible thing are
dangerous beneficial display human method dealing enjoying suffering are exemplary deserving
sympathetic animal painting investigates law greater le nobility character organic comparative anatomy
examines greater le development organic function animal painting is bring notice minor unthought
condition power physiology is ascertain minor condition question propose arrangement use organ
animal le province painter are indeed more likely commend drawing dissect animal generally assume
form be necessary only examine is drawing outer form attentively are led necessarily consider mode
life is therefore be struck awkwardness apparent uselessness sketching day several head bird became
vital matter interest know use bony process head asking great found appeared absurd wa certainly
unanswerable have have just landscape painting representation phenomenon relating human duty civil
disobedience heartily accept government is best governs like see acted more rapidly carried finally
amount also government is best governs not men are prepared be kind government government is best
most government are government are objection have been brought standing are many deserve also at
last be brought standing standing army is only arm standing government is only mode people have
chosen execute is equally liable be abused perverted people act witness present mexican work
comparatively few individual using standing government people not have consented american is recent
endeavoring transmit unimpaired instant losing ha not vitality force single living single man bend is
sort wooden gun people is not le necessary people have complicated machinery hear satisfy idea
government government show thus successfully men be imposed even impose own government never
furthered alacrity got character inherent american people ha done ha been have done somewhat
government had not sometimes got government is men fan success letting one ha been is most
governed are most let alone trade were not made never manage bounce obstacle legislator are
continually putting were judge men wholly effect action not partly deserve be classed punished person
put obstruction speak practically call ask not once better let man make known kind government
command be step obtaining after practical reason power is once hand majority are long period rule is
not are most likely be seems fairest are physically government majority rule case not be based even a

metamorphosis gregor samia woke troubled found transformed bed horrible lay lifted head little see
brown slightly domed divided arch stiff bedding wa hardly able cover seemed ready slide many
pitifully thin compared size rest waved about helplessly proper human room little too lay peacefully
familiar collection textile sample lay spread table samia wa travelling salesman there hung picture had
recently cut illustrated magazine housed gilded showed lady fitted fur hat fur boa sat raising heavy fur
muff covered whole lower arm gregor then turned look window dull drop rain be heard hitting made
feel quite sleep little bit longer forget wa something wa unable do wa used sleeping present state get
inaugural stated holding professorship direct practical chiefly natural history having course past year
laid foundational element art sufficiently invite enter real work accordingly propose following term
give practical leading elementary study branch natural history form kind center outset shortly state
position landscape painting animal painting hold higher branch landscape painting is thoughtful
passionate representation physical condition appointed human imitates record visible thing are
dangerous beneficial display human method dealing enjoying suffering are exemplary deserving
sympathetic animal painting investigates law greater le nobility character organic comparative anatomy
examines greater le development organic function animal painting is bring notice minor unthought
condition power physiology is ascertain minor condition question propose arrangement use organ
animal le province painter are indeed more likely commend drawing dissect animal generally assume
form be necessary only examine is drawing outer form attentively are led necessarily consider mode
life is therefore be struck awkwardness apparent uselessness sketching day several head bird became
vital matter interest know use bony process head asking great found appeared absurd wa certainly
unanswerable have have just landscape painting representation phenomenon relating human duty civil
disobedience heartily accept government is best governs like see acted more rapidly carried finally
amount also government is best governs not men are prepared be kind government government is best
most government are government are objection have been brought standing are many deserve also at
last be brought standing standing army is only arm standing government is only mode people have
chosen execute is equally liable be abused perverted people act witness present mexican work
comparatively few individual using standing government people not have consented american is recent
endeavoring transmit unimpaired instant losing ha not vitality force single living single man bend is
sort wooden gun people is not le necessary people have complicated machinery hear satisfy idea
government government show thus successfully men be imposed even impose own government never
furthered alacrity got character inherent american people ha done ha been have done somewhat
government had not sometimes got government is men fan success letting one ha been is most
governed are most let alone trade were not made never manage bounce obstacle legislator are
continually putting were judge men wholly effect action not partly deserve be classed punished person
put obstruction speak practically call ask not once better let man make known kind government
command be step obtaining after practical reason power is once hand majority are long period rule is
not are most likely be seems fairest are physically government majority rule case not be based even a
far men understand delight relativity ether theory relativity address delivered may university leyden
doe come idea ponderable is derived abstraction everyday physicist set idea existence kind explanation

Appartenance moyenne par type (top10), Mix_sent1

Groupe 1		Groupe 2		Groupe 3		Groupe 4	
ask	0.95	street	0.97	differential	0.99	is	0.96
forgive	0.95	hat	0.97	inertial	0.98	consists	0.90
postpone	0.95	hung	0.96	homogeneous	0.98	are	0.72
consented	0.95	headboard	0.96	variability	0.98	constitutes	0.61
suffer	0.94	slid	0.96	euclidean	0.98	represents	0.48
decide	0.94	boa	0.96	comparative	0.98	conforms	0.39
believed	0.94	buried	0.96	scalar	0.98	governs	0.30
accuse	0.94	muff	0.96	continuum	0.98	exists	0.21
expect	0.94	gilded	0.96	computation	0.97	enjoys	0.14
actually	0.94	sat	0.96	finite	0.97	commonly	0.14

Classification semi-supervisée : résultats préliminaires

Dataset	NMI - 5%	NMI - 10%	NMI - 20%
Mix_word1	0.06 $r = 10, \alpha = 5$ $\beta = 100, \kappa = 0$	0.06 $r = 15, \alpha = 0.1$ $\beta = 300, \kappa = 1$	0.06 $r = 3, \alpha = 10$ $\beta = 300, \kappa = 1$
Mix_word3	0.09 $r = 3, \alpha = 2$ $\beta = 10, \kappa = 1/3$	0.10 $r = 3, \alpha = 5$ $\beta = 50, \kappa = 2/3$	0.14 $r = 3, \alpha = 5$ $\beta = 50, \kappa = 1/3$
Mix_sent1	0.29 $r = 3, \alpha = 5$ $\beta = 10, \kappa = 0$	0.40 $r = 5, \alpha = 5$ $\beta = 10, \kappa = 0$	0.54 $r = 3, \alpha = 5$ $\beta = 10, \kappa = 0$
Mix_sent10	0.77 $r = 15, \alpha = 5$ $\beta = 10, \kappa = 0$	0.90 $r = 10, \alpha = 2$ $\beta = 10, \kappa = 2/3$	0.95 $r = 15, \alpha = 2$ $\beta = 5, \kappa = 1$

Classification semi-supervisée : résultats préliminaires

Gauche : meilleurs résultats 10%. Droite : Mix_sent1

metamorphosis gregor samoa woke troubled found transformed bed horrible inaugural stated holding
professionist direct practical chiefly natural history having course past year laid foundational element
an sufficiently invite enter real work accordingly propose following item give practical leading
elementary study branch natural history from kind center lay lifted head little see brown slightly domes
divided arch stiff sidewalk relativity ether theory relativity address delivered many university leyden dose
come idea ponderable is derived abstraction everyday physicist set idea existence kind outset shortly
state position landscape painting animal painting hold higher branch landscape painting is thoughtful
passionate representation physical condition appointed human bedding wa hardly able cover seemed
ready slide imitates record visible thing are dangerous beneficial display human method dealing
enjoying suffering are exemplary deserving sympathetic duty civil disobedience heartily accept
government is best governs like see acted more rapidly explanation is probably be sought phenomenon
have given rise theory action property light have led undulatory animal painting investigates lay
greater le nobility character organic comparative anatomy examines greater le development organic
function animal painting is bring notice minor unthought condition power physiology is ascertain
minor condition question purpose arrangement use organ animal le province painter are indeed more
likely commend drawing many pitifully thin compared size rest waved about helplessly let devote little
while consideration proper human room little too lay peacefully familiar outside physic know nothing
action carried finally amount also government is best governs not men are prepared be kind
government government is best most government are government are try connect cause effect
experience natural object afford seems first were other mutual action immediate dissect animal
generally assume form be necessary only examine is drawing outer form attentively are led necessarily
consider mode life is therefore be struck awkwardness apparent uselessness collection textile sample
lay spread table samoa wa travelling salesman there hung picture had recently cut illustrated magazine
housed gilded sketching day several head bird became vital matter interest know use bony process heart
asking great found appeared ahead wa certainly unanswerable communication motion push hearing
inducing combustion means is true even everyday experience is cause action play very important
objection have been brought standing are many deserve also at last be brought standing showed lady
fitted fur hat fur boa sat raising heavy fur muff covered whole lower arm gregor then turned look
window dull drop rain be heard hitting made feel quite have have just landscape painting representation
phenomenon relating human daily experience weight body meet something something not linked cause
is variable time do not everyday life speculate cause therefore do not become conscious character
action wa theory gravitation first assigned cause gravity interpreting action proceeding scarcely be
disposed admit propriety still le be likely conceive necessary strictness convince somewhat detailed
theory is probably greatest stride ever made effect causal nexus natural sleep little bit longer forget wa
something wa unable do wa used sleeping present state get however hard threw always rolled back
standing army is only arm standing yet theory evoked lively sense discomfort seemed be conflict
principle springing rest be reciprocal action only not immediate action here are landscape turner
greatest verusius government is only mode people have chosen execute is equally liable be abused
perverted people act witness present mexican work comparatively few individual using standing

metamorphosis gregor samoa woke troubled found transformed bed horrible inaugural stated holding
professionist direct practical chiefly natural history having course past year laid foundational element
an sufficiently invite enter real work accordingly propose following form give practical leading
elementary study branch natural history from kind center lay lifted head little see brown slightly domes
divided arch stiff sidewalk relativity ether theory relativity address delivered many university leyden dose
come idea ponderable is derived abstraction everyday physicist set idea existence kind outset shortly
state position landscape painting animal painting hold higher branch landscape painting is thoughtful
passionate representation physical condition appointed human bedding wa hardly able cover seemed
ready slide imitates record visible thing are dangerous beneficial display human method dealing
enjoying suffering are exemplary deserving sympathetic duty civil disobedience heartily accept
government is best governs like see acted more rapidly explanation is probably be sought phenomenon
have given rise theory action property light have led undulatory animal painting investigates lay
greater le nobility character organic comparative anatomy examines greater le development organic
function animal painting is bring notice minor unthought condition power physiology is ascertain
minor condition question purpose arrangement use organ animal le province painter are indeed more
likely commend drawing many pitifully thin compared size rest waved about helplessly let devote little
while consideration proper human room little too lay peacefully familiar outside physic know nothing
action carried finally amount also government is best governs not men are prepared be kind
government government is best most government are government are try connect cause effect
experience natural object afford seems first were other mutual action immediate dissect animal
generally assume form be necessary only examine is drawing outer form attentively are led necessarily
consider mode life is therefore be struck awkwardness apparent uselessness collection textile sample
lay spread table samoa wa travelling salesman there hung picture had recently cut illustrated magazine
housed gilded sketching day several head bird became vital matter interest know use bony process heart
asking great found appeared ahead wa certainly unanswerable communication motion push hearing
inducing combustion means is true even everyday experience is cause action play very important
objection have been brought standing are many deserve also at last be brought standing showed lady
fitted fur hat fur boa sat raising heavy fur muff covered whole lower arm gregor then turned look
window dull drop rain be heard hitting made feel quite have have just landscape painting representation
phenomenon relating human daily experience weight body meet something something not linked cause
is variable time do not everyday life speculate cause therefore do not become conscious character
action wa theory gravitation first assigned cause gravity interpreting action proceeding scarcely be
disposed admit propriety still le be likely conceive necessary strictness convince somewhat detailed
theory is probably greatest stride ever made effect causal nexus natural sleep little bit longer forget wa
something wa unable do wa used sleeping present state get however hard threw always rolled back
standing army is only arm standing yet theory evoked lively sense discomfort seemed be conflict
principle springing rest be reciprocal action only not immediate action here are landscape turner
greatest verusius government is only mode people have chosen execute is equally liable be abused
perverted people act witness present mexican work comparatively few individual using standing

Classification semi-supervisée : résultats préliminaires

Gauche : meilleurs résultats 10%. Droite : Mix_sent10

metamorphosis gregor samsa woke troubled found transformed bed horrible lay lifted head little see brown slightly domed divided arch stiff bedding was hardly able cover seemed ready slide many pitifully thin compared size rest waved about helplessly proper human room little too lay peacefully familiar collection textile sample lay spread table samsa was travelling salesman there hung picture had recently cut illustrated magazine housed gilded showed lady fitted fur hat fur box sat raising heavy fur muff covered whole lower arm gregor then turned look window dull drop rain be heard hitting madd feel quite sleep little bit longer forget was something was unable do was used sleeping present state get inaugural stated holding professorship direct practical chiefly natural history having course past year laid foundational element art sufficiently invite enter real work accordingly propose following term give practical leading elementary study branch natural history form kind center outset shortly state position landscape painting animal painting hold higher branch landscape painting is thoughtful passionate representation physical condition appointed human imitates record visible thing are dangerous beneficial display human method dealing enjoying suffering are exemplary deserving sympathetic animal painting investigates law greater le nobility character organic comparative anatomy examines greater le development organic function animal painting is bring notice minor unthought condition power physiology is ascertain minor condition question purpose arrangement use organ animal le province painter are indeed more likely commend drawing dissect animal generally assume form be necessary only examine is drawing outer form attentively are led necessarily consider mode life is therefore be struck awkwardness apparent uselessness sketching day several head bird became vital matter interest know use bony process head asking great found appeared absurd was certainly unanswerable have have not landscape painting representation phenomenon relating human duty civil disobedience heartily accept government is best governs like see acted more rapidly carried finally amount also government is best governs not men are prepared be kind government government is best most government are objection have been brought standing are many deserve also at last be brought standing standing army is only arm standing government is only mode people have chosen execute is equally liable be abused perverted people act witness present mexican work comparatively few individual using standing government people not have consented american is recent endeavoring transmit unimpaired instant losing ha not vitality force single living single man bend is sort wooden gun people is not le necessary people have complicated machinery hear satisfy idea government government show thus successfully men be imposed even impose own government never furthered alacrity got character inherent american people ha done ha been have done somewhat government had not sometimes got government is men fan succeed letting one ha been is most governed are most let alone trade were not made never manage bounce obstacle legislator are continually putting were judge men wholly effect action not partly deserve be classed punished person put obstruction speak practically call ask not once better let man make known kind government command be step obtaining after practical reason power is once hand majority are long period rule is not are most likely be seems fairest are physically government majority rule case not be based even a far men understand sidewalk relativity other theory relativity address delivered may university leyden doe come idea ponderable is derived abstraction everyday physicist set idea existence kind explanation

metamorphosis gregor samsa woke troubled found transformed bed horrible lay lifted head little see brown slightly domed divided arch stiff bedding was hardly able cover seemed ready slide many pitifully thin compared size rest waved about helplessly proper human room little too lay peacefully familiar collection textile sample lay spread table samsa was travelling salesman there hung picture had recently cut illustrated magazine housed gilded showed lady fitted fur hat fur box sat raising heavy fur muff covered whole lower arm gregor then turned look window dull drop rain be heard hitting madd feel quite sleep little bit longer forget was something was unable do was used sleeping present state get inaugural stated holding professorship direct practical chiefly natural history having course past year laid foundational element art sufficiently invite enter real work accordingly propose following term give practical leading elementary study branch natural history form kind center outset shortly state position landscape painting animal painting hold higher branch landscape painting is thoughtful passionate representation physical condition appointed human imitates record visible thing are dangerous beneficial display human method dealing enjoying suffering are exemplary deserving sympathetic animal painting investigates law greater le nobility character organic comparative anatomy examines greater le development organic function animal painting is bring notice minor unthought condition power physiology is ascertain minor condition question purpose arrangement use organ animal le province painter are indeed more likely commend drawing dissect animal generally assume form be necessary only examine is drawing outer form attentively are led necessarily consider mode life is therefore be struck awkwardness apparent uselessness sketching day several head bird became vital matter interest know use bony process head asking great found appeared absurd was certainly unanswerable have have not landscape painting representation phenomenon relating human duty civil disobedience heartily accept government is best governs like see acted more rapidly carried finally amount also government is best governs not men are prepared be kind government government is best most government are objection have been brought standing are many deserve also at last be brought standing standing army is only arm standing government is only mode people have chosen execute is equally liable be abused perverted people act witness present mexican work comparatively few individual using standing government people not have consented american is recent endeavoring transmit unimpaired instant losing ha not vitality force single living single man bend is sort wooden gun people is not le necessary people have complicated machinery hear satisfy idea government government show thus successfully men be imposed even impose own government never furthered alacrity got character inherent american people ha done ha been have done somewhat government had not sometimes got government is men fan succeed letting one ha been is most governed are most let alone trade were not made never manage bounce obstacle legislator are continually putting were judge men wholly effect action not partly deserve be classed punished person put obstruction speak practically call ask not once better let man make known kind government command be step obtaining after practical reason power is once hand majority are long period rule is not are most likely be seems fairest are physically government majority rule case not be based even a far men understand sidewalk relativity other theory relativity address delivered may university leyden doe come idea ponderable is derived abstraction everyday physicist set idea existence kind explanation

En conclusion :

- L'indice d'**autocorrélation globale** peut servir d'**indicateur global** sur un corpus donné.
- L'indice d'**autocorrélation locale** permet de repérer la **redondance sémantique** des tokens.
- Le **clustering** fonctionne bien pour découper le texte en **parties grossières (p.ex. paragraphes) parlant d'un thème similaire**.

Pistes d'amélioration :

- Arriver à construire des similarités WordNet entre tokens de différentes catégories (noms, verbes, adjectifs et adverbess).
- Voir si la notion de *perplexité* peut s'appliquer au clustering.
- Affiner la validation du clustering : textes plus variés, trouver le *state-of-the-art*, intervalles de confiance du NMI.

Merci pour votre attention !
(et désolé pour la longueur)

Des questions ?



Bavaud, F., Cocco, C., and Xanthos, A. (2015).
Textual navigation and autocorrelation.
Sequences in Language and Text, 69 :35–56.



Bojanowski, P., Grave, E., Joulin, A., and
Mikolov, T. (2017).
**Enriching word vectors with subword
information.**

*Transactions of the Association for
Computational Linguistics*, 5 :135–146.



Céré, R. and Bavaud, F. (2017).
**Soft image segmentation : on the clustering of
irregular, weighted, multivariate marked
networks.**

In *International Conference on Geographical
Information Systems Theory, Applications and
Management*, pages 85–109. Springer.



Fellbaum, C. (1998).
WordNet : an electronic lexical database.
MIT Press, Cambridge, Mass.



Gao, J.-B., Zhang, B.-W., and Chen, X.-H.
(2015).
**A WordNet-based semantic similarity
measurement combining edge-counting and
information content theory.**

*Engineering Applications of Artificial
Intelligence*, 39 :80–88.



Leacock, C. and Chodorow, M. (1998).
**Combining local context and wordnet similarity
for word sense identification.**
WordNet : An electronic lexical database,
49(2) :265–283.



Mikolov, T., Chen, K., Corrado, G., and Dean,
J. (2013).
**Efficient estimation of word representations in
vector space.**
arXiv preprint arXiv :1301.3781.



Pennington, J., Socher, R., and Manning, C. D.
(2014).
Glove : Global vectors for word representation.
In *Proceedings of the 2014 conference on
empirical methods in natural language
processing (EMNLP)*, pages 1532–1543.



Resnik, P. (1995).
**Using information content to evaluate
semantic similarity in a taxonomy.**
arXiv preprint cmp-lg/9511007.



Wu, Z. and Palmer, M. (1994).
Verb semantics and lexical selection.
arXiv preprint cmp-lg/9406033.