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

Introduction

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-méronimie (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).

Les mesures de similarités sémantiques

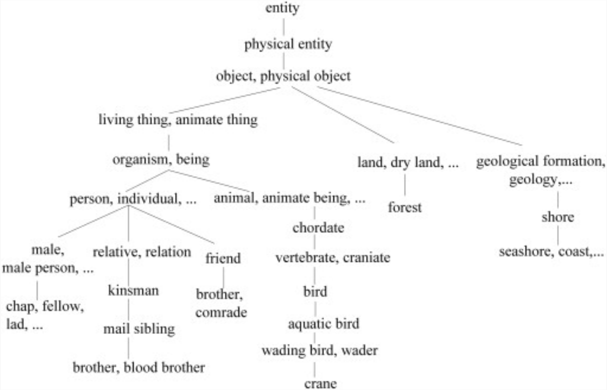
Dans la pratique, il existe plusieurs manières d'obtenir une **mesure de similarité sémantique** entre deux termes. Dans ce travail, ces mesures vont être extraites via deux types d'objet :

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

UNIL | Université de Lausanne

WordNet

WordNet [Fellbaum, 1998] est une ontolgie 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**.



Source : [Gao et al., 2015]

WordNet : les similarités

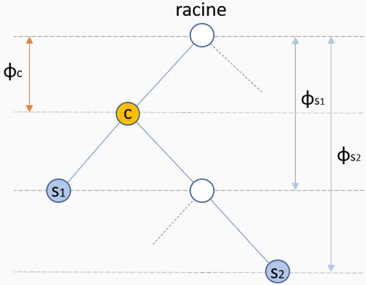
Il existe de nombreuses similarités basées sur WordNet.

La **similarité de Wu-Palmer** [Wu and Palmer, 1994] entre deux synsets s_1 et s_2 :

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

$S(s_1, s_2)$ 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.

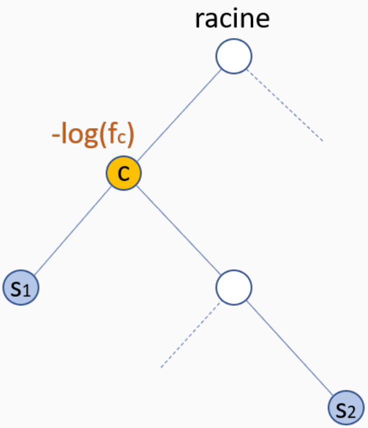


La **similarité de Resnik** [Resnik, 1995] entre deux synsets s_1 et s_2 :

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

$S(s_1, s_2)$ 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)



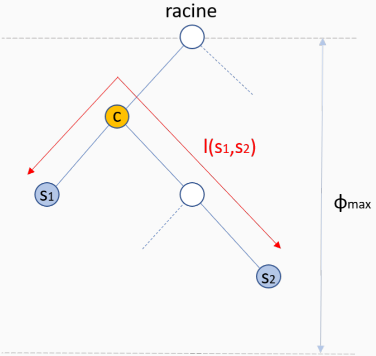
WordNet : les similarités

La **similarité de Leacock-Chodorow** [Leacock and Chodorow, 1998] entre deux synsets s_1 et s_2 :

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

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

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



WordNet : les similarités

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.

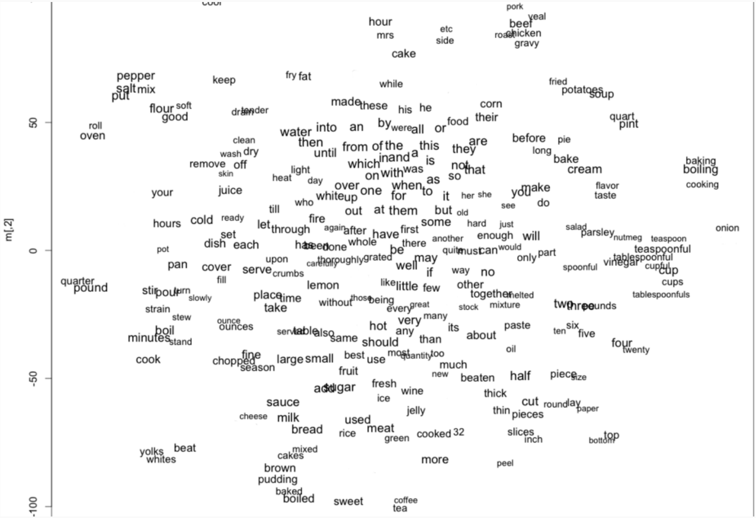
Les plongements lexicaux

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 :

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

UNIL | Université de Lausanne



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

Les plongements lexicaux

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 plongements lexicaux

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**.

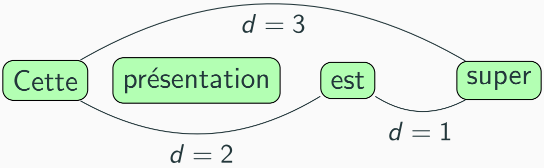
Les similarités seront transformées en dissimilarités avec minus_log.

UNIL | Université de Lausanne

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**.



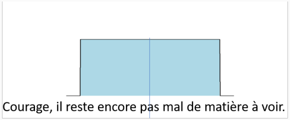
Ici, on prefera définir la spatialité via une **relation de voisinage** c'est-à-dire via une **matrix d'adjacence $\boldsymbol{A}$** de taille $n_{\text{token}} \times n_{\text{token}}$:

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

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 :

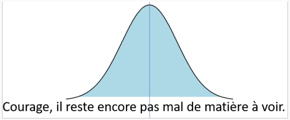
Un **voisinage uniforme** :

$$\boldsymbol{a}_{i:} = \begin{pmatrix} \dots & 0 & 1 & 1 & 1 & 1 & 1 & 0 & \dots \end{pmatrix}$$



Un **voisinage gaussien** :

$$\boldsymbol{a}_{i:} = \begin{pmatrix} \dots & 0.004 & 0.054 & 0.242 & 0.399 & 0.242 & 0.054 & 0.004 & \dots \end{pmatrix}$$



La matrice d'échange

Dans notre formalisme, nous aurons besoin d'une **matrice d'échange $\boldsymbol{E} = (e_{ij})$**. 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 $\boldsymbol{E}$ à partir de $\boldsymbol{A}$**, tout en respectant les poids des tokens f_i.

La spatialité : pour résumé

La **spatialité** des tokens sera définie via une **relation de voisinage** contenue dans une **matrice d'échange** $\boldsymbol{E} = (e_{ij})$ 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 $\pm r$: $\boldsymbol{E}_r^{\text{unif}}$
- Une distribution **normale** de « diffusion » r : $\boldsymbol{E}_r^{\text{norm}}$

Les méthodes

Nous avons défini deux objets sur les tokens :

- Une **matrice de dissimilarité** entre tokens $\boldsymbol{D} = (d_{ij})$, quantifiant à quel point ces derniers sont **sémantiquement éloignés**.
- Une **matrice d'échange** entre tokens $\boldsymbol{E} = (e_{ij})$, 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** [Ceré and Bavaud, 2017], qui permet de **regrouper les tokens proches sémantiquement ET spatialement**.

UNIL | Université de Lausanne

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**.

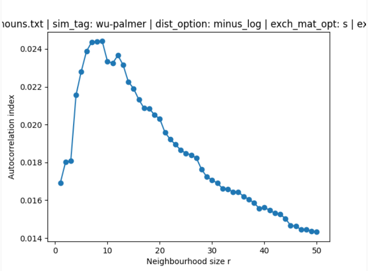
On peut alors définir un **indice d'autocorrelation globale** avec :

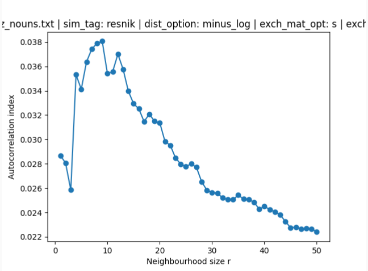
$$\delta := \frac{\Delta - \Delta_{\text{loc}}}{\Delta}$$

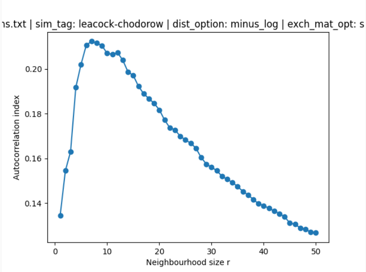
qui peut prendre des valeurs entre -1 et 1.

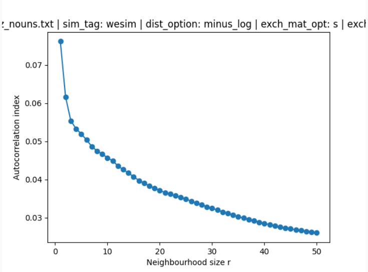
L'autocorrélation globale : résultats

The Wonderful Wizard of Oz (L. Frank Baum) - nouns, $\boldsymbol{E}_r^{\mathrm{unif}}$:



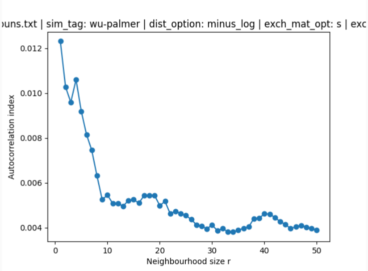


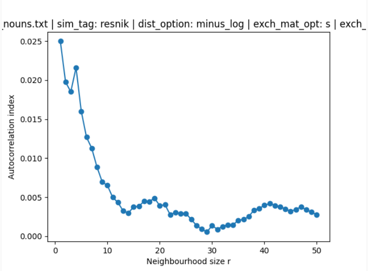


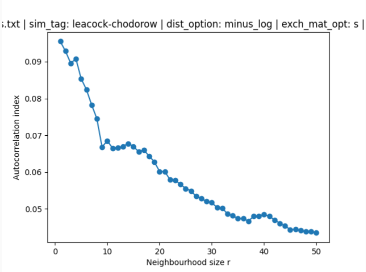


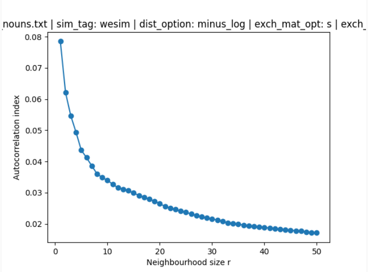
L'autocorrélation globale : résultats

Animal Farm (G. Orwell) - nouns, $\boldsymbol{E}_r^{\text{unif}}$:









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 :

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

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** :

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

UNIL | Université de Lausanne

L'autocorrélation locale : résultats

The Wonderful Wizard of Oz (L. Frank Baum), $\boldsymbol{S}^{\mathrm{we}}$, $\boldsymbol{E}^{\mathrm{unif}}_{100}$:

UNIL | Université de Lausanne

Clustering

Clustering des tokens

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** $\boldsymbol{Z} = (z_{ig})$, 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**.

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

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

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

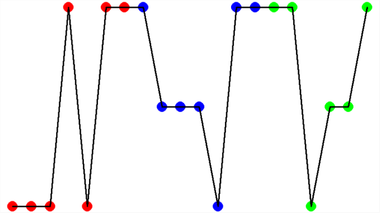
où :

- $\rho_g := \sum_i f_i z_{ig}$ 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

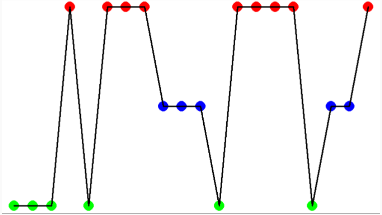
Clustering : l'inertie intra-groupe

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

$\Delta_W[\boldsymbol{Z}]$ élevé



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



UNIL | Université de Lausanne

Clustering : la modularité généralisée

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

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

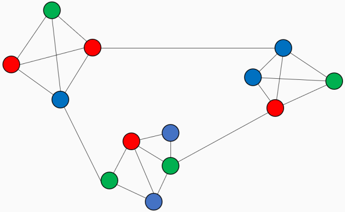
où :

- $\rho_g := \sum_i f_i z_{ig}$ 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 ($\kappa = 1$)

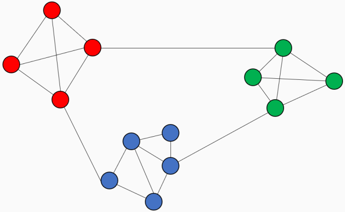
Clustering : l'inertie intra-groupe

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

$C^{\kappa}[\boldsymbol{Z}]$ élevé



$C^{\kappa}[\boldsymbol{Z}]$ faible



UNIL | Université de Lausanne

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 $\boldsymbol{Z}$ qui minimise l'**énergie libre** $F[\boldsymbol{Z}]$, définie comme :

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

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 $\boldsymbol{Z}$ qui minimise cette fonctionnelle.

UNIL | Université de Lausanne

Clustering : résultats préliminaires

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 : $\boldsymbol{E}_r^{\text{unif}}$ ou $\boldsymbol{E}_r^{\text{norm}}$, 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 créer 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.

Clustering : résultats préliminaires

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 :

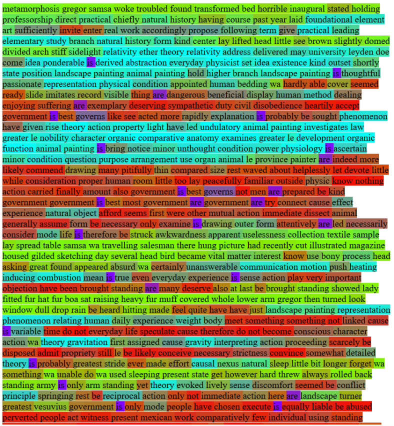
- La similarité sémantique issue du word embedding : $\boldsymbol{S}^{we}$
- La matrice d'échange gaussienne : $\boldsymbol{E}_r^{\text{norm}}$
- Le facteur de diffusion de la matrice d'échange : $r \in \{3, 5, 10, 15\}$.
- Le nombre de groupes sera posé à $p = 4$.
- Le paramètre $\alpha \in \{0.1, 1,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

Clustering : résultats préliminaires

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



Clustering : résultats préliminaires

Gauche : meilleurs résultats. Droite : Mix_sent10

UNIL | Université de Lausanne

Clustering : résultats préliminaires

Appartenance moyenne par type (top10), Mix_sent1

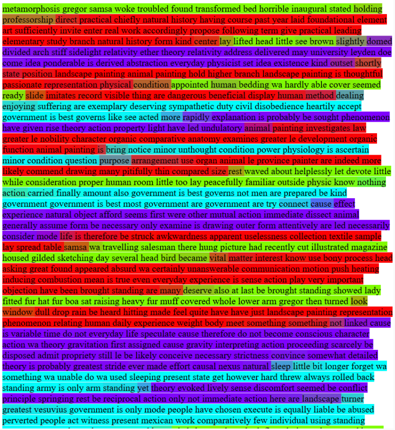
Groupe 1		Groupe 2		Groupe 3		Groupe 4	
ask	0.95	street	0.97	differential	0.99	is	0.96
forgive	0.95	hat	0.97	inertial	0.98	consists	0.90
postpone	0.95	hung	0.96	homogeneous	0.98	are	0.72
consented	0.95	headboard	0.96	variability	0.98	constitutes	0.61
suffer	0.94	slid	0.96	euclidean	0.98	represents	0.48
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



UNIL | Université de Lausanne

Classification semi-supervisée : résultats préliminaires

Gauche : meilleurs résultats 10%. Droite : Mix_sent10

Conclusion

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 adverbes).
- Voir si la notion de *perplexité* peut s'appliquer au clustering.
- Affiner la validation du clustering : textes plus variés, touver le *state-of-the-art*, intervalles de confiance du NMI.

Merci pour votre attention !

(et désolé pour la longueur)

Des questions ?

Bibliographie

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.

Ceré, 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.