Construction d’indicateurs pour détecter des stratégies d'écriture collaborative dans des documents textes

Atelier RJC-EIAH 2022 : Conception et évaluation de tableaux de bord

Anis M. HADDOUCHE **

En collaboration avec :

- Fahima DJELLIL **
- Jean-Marie GILLIOT **
- Cédric HAM *
- Christian HOFFMANN *
- Nadine MANDRAN *
- Maria Teresa SEGARRA MONTESINON **

** IMT Atlantique ,* Université Grenoble-Alpes

Table des matières

I - Introduction

- Lors de la conception des tableaux de bords, les enseignants demandent souvent des indicateurs pour suivre la collaboration entre les étudiants.
- Nos travaux nous ont conduits à élaborer des indicateurs pour détecter des stratégies d'écriture collaborative sur des documents textes.
- L'élaboration de ces indicateurs est difficile car il est nécessaire de combiner différents outils de traitement des données et de valider ces indicateurs.
- Notre présentation focalisera sur le processus de construction et de validation des ces indicateurs entre statisticien et enseignants.

- Nous nous intéressons à deux stratégies de collaboration:
  - Construction séquentielle sommative :
    - un membre propose un document initial (complet ou pas), les autres ajoutent successivement leurs contribution sans modifier ce qui à été écrit auparavant
  - Construction séquentielle intégrative :
    - un membre propose un document initial (complet ou pas), les autres ajoutent successivement leurs contribution et modifient ce qui à été écrit auparavant
- Nous construisons deux indicateurs pour détecter ces deux stratégies

- Séquentielle : comment mesurer la différence (ou similarité) entre deux séquences de texte ?
- Utilisation de métrique pour les chaines de caractères
- La mesure la plus connue est une mesure « rudimentaire » appelée distance de Levenshtein

- Le coût minimal pour transformer une chaîne A en B en effectuant :
  - Modification
  - Suppression
  - Insertion
- Associer à chaque opération un coût
- La distance (entre 0 et 1) est la somme de ces coûts

- Librairie Difflib de Python
- Recherche les plus grandes séquences similaires
- Accorder des tags pour chaque séquence

En bleu, on a indiqué les parties communes aux deux textes, en rouge les parties supprimées, en vert les parties insérées et en orange les parties modifiées.

Texte initial : « LabNbook est une plateforme gratuite, utilisée par plus de 2 800 élèves chaque année, à l'Université Grenoble-Alpes, Grenoble-INP, dans des collèges, lycées et des CPGE. »

Texte final : «LabNbook est une plateforme open source et gratuite, utilisée par plus de 2 800 étudiants chaque année, à l'Université Grenoble-Alpes, Grenoble-INP, dans des lycées et des CPGE. »

equal	'LabNbook est une plateforme' --> 'LabNbook est une plateforme'
insert	'' --> 'open source et'
equal	'gratuite, utilisée par plus de 2 800' --> 'gratuite, utilisée par plus de 2 800'
replace (0.73)	'élèves' --> 'étudiants'
equal	'chaque année, à l'Université ... dans des' --> 'chaque année, à l'Université ... dans des'
delete	'collèges,' --> ''
equal	'lycées et des CPGE.' --> 'lycées et des CPGE.'

- Algorithme non robuste à la ponctuation qui nécessite de faire un pré-nettoyage de texte
- D'autres traitements possibles :
  - *Lemmatization : Prendre la racine de chaque mot (ex : petites –> petit)*
  - Enlever les *Stop-words : Mots non significatifs dans le texte (ex : la, de, ce…etc)*
- *Nous avons décidé de faire uniquement un nettoyage de ponctuation*

- Comment gérer les formules mathématiques ?
- Il faut d'abord être capable de les détecter :
  - Formule LaTex : utiliser des règles d'occurrences d'expressions régulières (ex : le dollar « $ » pour le début et la fin d'une formule)
  - Plus difficile quand c'est des formules écrites avec des caractères « textes » (ex : f(x) = ax + b)
    - Des approches statistiques sont possibles comme les HMM (Hiden Markov Chain)
- Une fois les formules détectées, on peut utiliser une comparaison caractère par caractère pour mesurer son évolution (à l'instar des logiciels de « versionning » informatiques comme git)
- Pour l'instant, on s'intéresse uniquement aux documents textes avec très peu de formules

- Comment quantifier et agréger cette information pour la rendre exploitables ?
- Construction de matrice de contribution :
  - Pour chaque apprenant on donne un score entre 0 et 1 pour quantifier sa contribution à l'écriture de chaque mot du texte

Entrées :

Etudiant n°1 : « LabNbook est une plateforme utilisée à l'Université Grenoble-Alpes et dans des établissements secondaires »

Etudiant n°2 : « LabNbook est une plateforme open source et gratuite, utilisée par près de 3000 élèves tout au long de l'année, à l'Université Grenoble-Alpes »

Etudiant n°1 : « LabNbook est une plateforme open source et gratuite, utilisée par plus de 2800 étudiants chaque année, à l'Université Grenoble-Alpes mais aussi dans des lycées et des CPGE»

Sorties :

Etudiant n°1 : [1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0.5, 0, 0.53, 0.53, 0.53, 0.53,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Etudiant n°2 : [0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0.5, 1, 0.47, 0.47, 0.47, 0.47, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

- Soit $x_{i,j,l}$ la contribution de l'auteur i au mot l de la phrase j
- Ici, $i \in [1,K]$, $j \in [1,N]$ et $l \in [1,n_j]$ où K est le nombre d'auteurs, N le nombre de phrases et n_j le nombre de mots dans la phrase j

Phrase 1 — Phrase j — Phrase N

$$X = \begin{bmatrix} x_{1,1,1} & \dots & x_{1,1,l} & \dots & x_{1,1,n_1} & \dots\dots & x_{1,j,1} & \dots & x_{1,j,l} & \dots & x_{1,j,n_j} & \dots\dots & x_{1,N,1} & \dots & x_{1,N,l} & \dots & x_{1,N,n_N} \\ \vdots & \ddots & \vdots & \dots & \vdots & \dots\dots & \vdots & \ddots & \vdots & \dots & \vdots & \dots\dots & \vdots & \ddots & \vdots & \dots & \vdots \\ x_{i,1,1} & \dots & x_{i,1,l} & \dots & x_{i,1,n_1} & \dots\dots & x_{i,j,1} & \dots & x_{i,j,l} & \dots & x_{i,j,n_j} & \dots\dots & x_{i,N,1} & \dots & x_{i,N,l} & \dots & x_{i,N,n_N} \\ \vdots & \dots & \vdots & \ddots & \vdots & \dots\dots & \vdots & \dots & \vdots & \ddots & \vdots & \dots\dots & \vdots & \dots & \vdots & \ddots & \vdots \\ x_{K,1,1} & \dots & x_{K,1,l} & \dots & x_{i,1,n_1} & \dots\dots & x_{K,j,1} & \dots & x_{K,j,l} & \dots & x_{i,j,n_j} & \dots\dots & x_{K,N,1} & \dots & x_{K,N,l} & \dots & x_{K,N,n_N} \end{bmatrix}$$

- Nous verrons plus bas comment on découpe un texte en phrases

- L'équilibre de contribution, noté $e(X)$, est un indicateur qui mesure la répartition des contributions au texte final en terme de mots (écrits ou modifiés).
- Exemple :

L'équilibre de contribution mesure la répartition des contributions au texte final

$X =$

Rédacteur 1	1	1	1	0,5	0,5	0,2	1	1	0	0	0
Rédacteur 2	0	0	0	0,5	0,5	0,5	0	0	0	0,8	0,8
Rédacteur 3	0	0	0	0	0	0,3	0	0	1	0,2	0,2

1. Moyenne sur les lignes :
   - R1 = 0,56
   - R2 =0,35
   - R3 = 0,081
2. Dispersion des ces moyennes autours de 1/3
   - $e(X)$ = 0,58

- Matrice de contribution X

$$\begin{bmatrix}
x_{1,1,1} & \cdots & x_{1,1,l} & \cdots & x_{1,1,n_1} & \cdots\cdots & x_{1,j,1} & \cdots & x_{1,j,l} & \cdots & x_{1,j,n_j} & \cdots\cdots & x_{1,N,1} & \cdots & x_{1,N,l} & \cdots & x_{1,N,n_N} \\
\vdots & \ddots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \ddots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
x_{i,1,1} & \cdots & x_{i,1,l} & \cdots & x_{i,1,n_1} & \cdots\cdots & x_{i,j,1} & \cdots & x_{i,j,l} & \cdots & x_{i,j,n_j} & \cdots\cdots & x_{i,N,1} & \cdots & x_{i,N,l} & \cdots & x_{i,N,n_N} \\
\vdots & \cdots & \vdots & \ddots & \vdots & \cdots\cdots & \vdots & \cdots & \vdots & \ddots & \vdots & \cdots\cdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
x_{K,1,1} & \cdots & x_{K,1,l} & \cdots & x_{i,1,n_1} & \cdots\cdots & x_{K,j,1} & \cdots & x_{K,j,l} & \cdots & x_{i,j,n_j} & \cdots\cdots & x_{K,N,1} & \cdots & x_{K,N,l} & \cdots & x_{K,N,n_N}
\end{bmatrix}$$

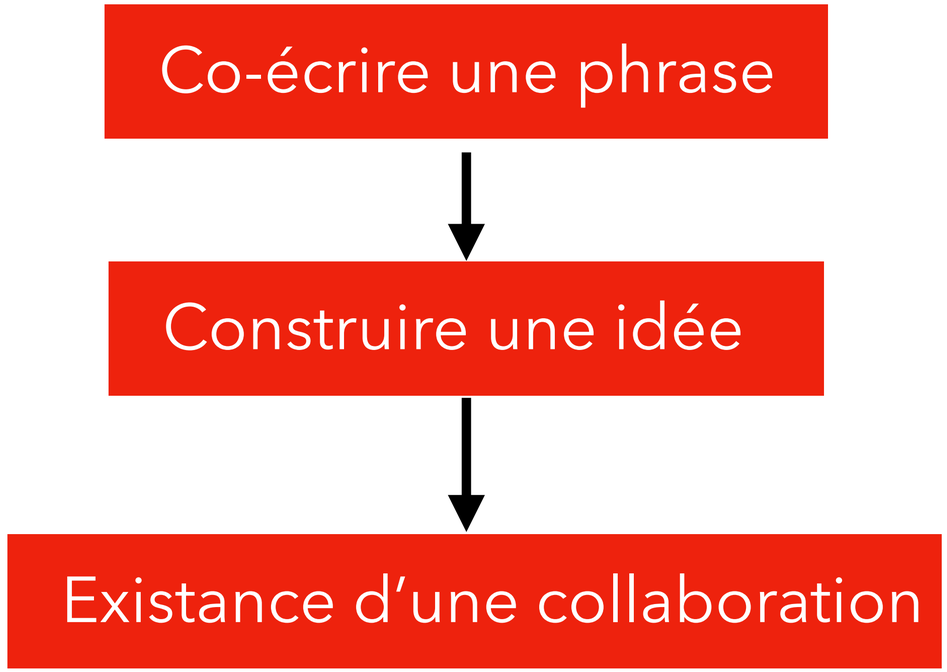
- Équilibre de contribution $e(X)$

$$e(X) = 1 - \frac{K}{K-1} \sum_{i=1}^{K} \left(\bar{x}_{i,\cdot,\cdot} - \frac{1}{K} \right)^2 \quad \text{où} \quad \bar{x}_{i,\cdot,\cdot} = \frac{1}{N} \sum_{j=1}^{N} \sum_{l=1}^{n_j} x_{i,j,l}$$

- Propriétés de $e(X)$:
  - Prend des valeurs dans l'intervalle $[0,1]$
  - $e(X) = 0$: Le texte à été écrit par un seul rédacteur
  - $e(X) = 1$: Le texte à été écrit de façon équilibrée par tous les rédacteurs

- $e(X) \simeq 1$
  - Exemples avec deux rédacteurs :
    - L'équilibre de contribution mesure la répartition des contributions au texte final.
    - L'équilibre de contribution mesure la répartition des contributions au texte final.
    - L'équilibre de contribution mesure la répartition des contributions au texte final.
- $e(X) \simeq 0.50$
  - Exemple avec trois rédacteurs
    - L'équilibre de contribution mesure la répartition des contributions au texte final.

- Rappels :
  - Nous considérons que les phrases sont des unités sémantiques



  - L’équilibre de contribution ne mesure pas ce type de collaboration

- Exemples :

	PHRASE 1			PHRASE 2		
Exemple 1	MOT 1	MOT 2	MOT 3	MOT 1	MOT 2	MOT 3
Exemple 2	MOT 1	MOT 2	MOT 3	MOT 1	MOT 2	MOT 3

- Pour ces deux exemples $e(X) = 1$, alors que, c'est deux stratégies de collaboration sont différentes.
  - Exemple 1 : intégrative
  - Exemple 2 : sommative

1. D'abord segmenter le texte en phrases
   - Segmentation des phrases basée sur des règle (cf *Sadvilkar, N., & Neumann, M. (2020). PySBD: Pragmatic Sentence Boundary Disambiguation. arXiv preprint arXiv:2010.09657.*)
   - Exemple :
     - Un simple point pour terminer une phrase
     - Bonjour à tous. Mon nom est Nicole. ["Bonjour à tous.", "Mon nom est Nicole.]
2. Ensuite calculer un indice de collaboration sur chaque segment
3. Agréger ces indices en faisant la moyenne pondérée

Mat. de contribution

↓

Segm. En phrases

↓

Cont. sur les phrases

↓

Aggérer cont. Phrases

$$\begin{bmatrix} x_{1,1,1} & \cdots & x_{1,1,l} & \cdots & x_{1,1,n_1} & \cdots\cdots & x_{1,j,1} & \cdots & x_{1,j,l} & \cdots & x_{1,j,n_j} & \cdots\cdots & x_{1,N,1} & \cdots & x_{1,N,l} & \cdots & x_{1,N,n_N} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \ddots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ x_{i,1,1} & \cdots & x_{i,1,l} & \cdots & x_{i,1,n_1} & \cdots\cdots & x_{i,j,1} & \cdots & x_{i,j,l} & \cdots & x_{i,j,n_j} & \cdots\cdots & x_{i,N,1} & \cdots & x_{i,N,l} & \cdots & x_{i,N,n_N} \\ \vdots & \cdots & \vdots & \ddots & \vdots & \cdots\cdots & \vdots & \cdots & \vdots & \ddots & \vdots & \cdots\cdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{K,1,1} & \cdots & x_{K,1,l} & \cdots & x_{i,1,n_1} & \cdots\cdots & x_{K,j,1} & \cdots & x_{K,j,l} & \cdots & x_{i,j,n_j} & \cdots\cdots & x_{K,N,1} & \cdots & x_{K,N,l} & \cdots & x_{K,N,n_N} \end{bmatrix}$$

$$\begin{bmatrix} \bar{x}_{1,1,\cdot} & \cdots & \bar{x}_{1,j,\cdot} & \cdots & \bar{x}_{1,N,\cdot} \\ \vdots & \ddots & \vdots & & \vdots \\ \bar{x}_{i,1,\cdot} & \cdots & \bar{x}_{i,j,\cdot} & \cdots & \bar{x}_{i,N,\cdot} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \bar{x}_{K,1,\cdot} & \cdots & \bar{x}_{K,j,\cdot} & \cdots & \bar{x}_{K,N,\cdot} \end{bmatrix}$$

$$v_j = 1 - \frac{K}{K-1} \sum_{i=1}^{K} \left(\bar{x}_{i,j,\cdot} - \frac{1}{K} \right)^2, \qquad \forall j \in [1,N].$$

$$c(X) = \sum_{j=1}^{N} p_j v_j \quad \text{where} \quad p_j = \frac{n_j}{N}.$$

II - Création des indicateurs

Processus de construction
Indice de co-éctiure : Exemple

Mat. de contribution

↓

Segm. En phrases

↓

Cont. sur les phrases

↓

Aggérer cont. Phrases

R. 1	1	1	1	0.5	0.2	0.5	1	1	0	0	0
R. 2	0	0	0	0.5	0.5	0.5	0	0	0.8	0.8	0.8
R. 3	0	0	0	0	0.3	0	0	0	0.2	0.2	0.2

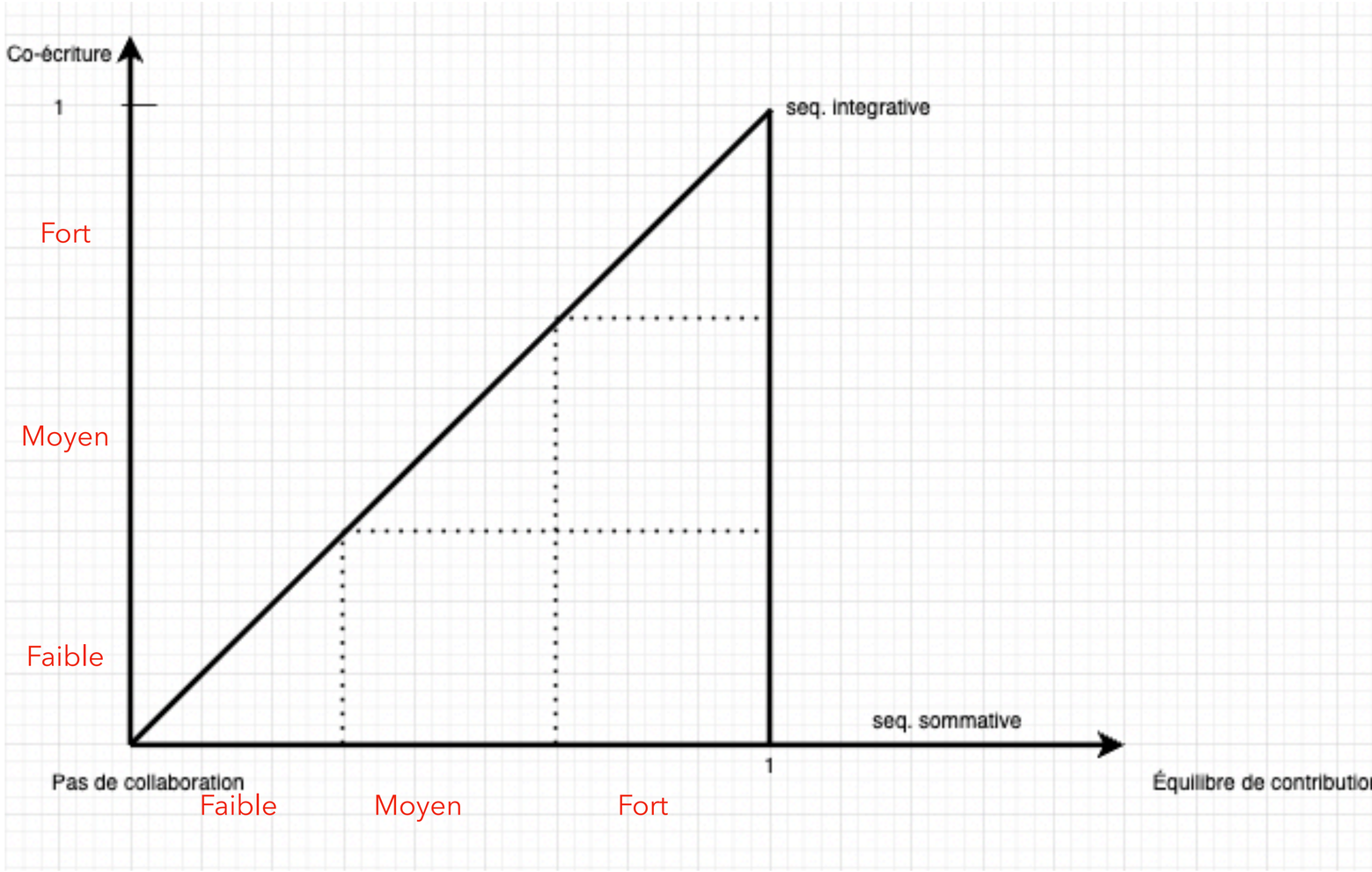
	PHRASE 1	PHRASE 2	PHRASE 3
R. 1	1	0.4	0.4
R. 2	0	0.5	0.48
R. 3	0	0.1	0.12

	PHRASE 1	PHRASE 2	PHRASE 3
V	$v_1 = 0$	$v_1 = 0.63$	$v_1 = 0.67$

$$c(X) = (3/11) \times 0 + (3/11) \times 0{,}63 + (5/11) \times 0{,}67 = 0{,}48$$

IV - Detection des stratégies de collaboration

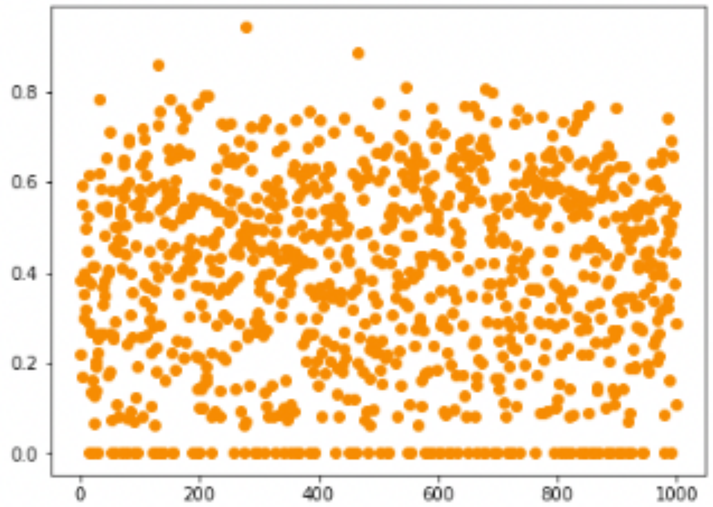
- Rappel
  - Sommative : membre propose un document initial (complet ou pas), les autres ajoutent successivement leurs contributions sans modifier ce qui à été écrit auparavant
  - intégrative : un membre propose un document initial (complet ou pas), les autres ajoutent successivement leurs contributions et modifient ce qui à été écrit auparavant



IV - Detection des stratégies de collaboration

Validation des indicateurs

- Simuler des matrices de contribution aléatoire pour voir la distribution des indicateurs grâce à la loi de Dirichelit (https://fr.wikipedia.org/wiki/Loi_de_Dirichlet)



- Les tester sur un ensemble d'exemple
- Faire une validation empirique à travers un questionnaire :
  - Tirer des exemples d'une base de donnée réelle (LabnBook)
  - Les classer automatiquement dans le triangle de collaboration
  - Comparer cette classification automatique par une classification faite par des humains

V- Perspectives

- Segmentation sémantique
- Extraction des formules
- Meilleurs découpages des valeurs dans l'intervalle [0,1]
- Définir une relation mathématique qui prend en entrée les indicateurs et donne en sortie la stratégie de collaboration