Segmentation exploratoire de texte

Dominique Gay (1), Romain Guigourès (2), Marc Boullé (3) & Fabrice Clérot (3)

(1) Université de La Réunion
dominique.gay@univ-reunion.fr

(2) Zalando, Germany
romain.guigoures@zalando.de

(3) Orange Labs, Lannion
prenm.nom@orange.com

Segmentation de texte

Text/Topic Segmentation

Topic segmentation: *automatically dividing single long recordings or transcripts into shorter, topically coherent segments*

(Matthew Purver, Topic Segmentation 2011)

Text/Topic Segmentation : Applications

- tâche préliminaire au résumé de texte et la recherche d'information
- identification des différents *moments* dans un texte
- résumé de flux de tweets
- ...

Segmentation de texte

Contributions majeures (TAL/NLP)

- Hearst (1997) travail pionnier basé sur la cohésion lexicale
- Utiyama et Isahara (2001)
- **MCSeg** : Malioutov et Barzilay (2006), **BayesSeg** : Eisenstein et Barzilay (2008), **HierBayes** : Eisenstein (2009), **APS** : Kazantseva et Szpakowicz (2011), **TSM** : Du et al. (2013)
- **LDA** et ses dérivés pour le (dynamic) topic modeling : David Blei et son équipe

Certaines limites

- Dédiées aux textes courts (quelques centaines/milliers de mots)
- Méthodes à multiples paramètres (paramétrage non-trivial)
- Exploitation/Exploration du résumé peu ou pas traité

Contributions

Instanciation des modèles en grilles pour la segmentation exploratoire de texte

Plan

1 Modèles en grilles : principes

2 Expérimentations / Applications

3 Conclusion & Perspectives

Plan

1 Modèles en grilles : principes

2 Expérimentations / Applications

3 Conclusion & Perspectives

Modèles en grilles : Une méthode de coclustering

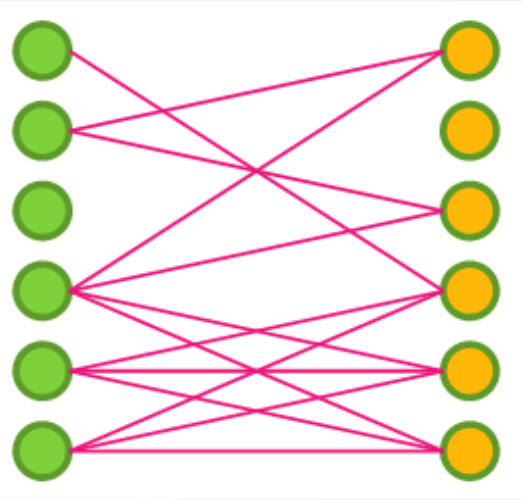
Données

Deux dimensions (entités) reliées par des observations

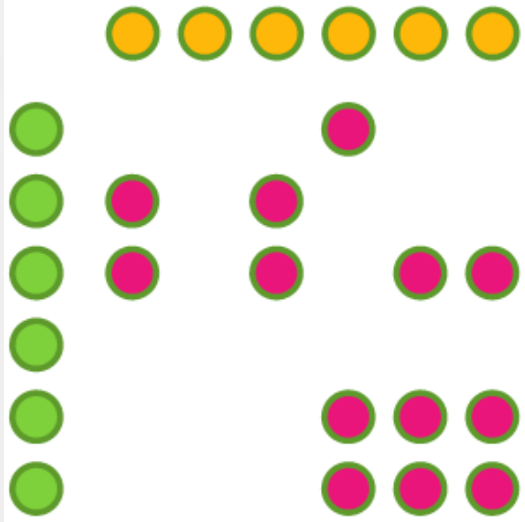
- Des documents, des mots
- Des cookies, des pages web
- Des clients, des produits
- Données de co-occurrence, dyadiques, tables de contingence

e.g., une observation : un document donné utilise un mot donné

Représentation par graphe bi-parti dont les arêtes sont les observations



Représentation matricielle



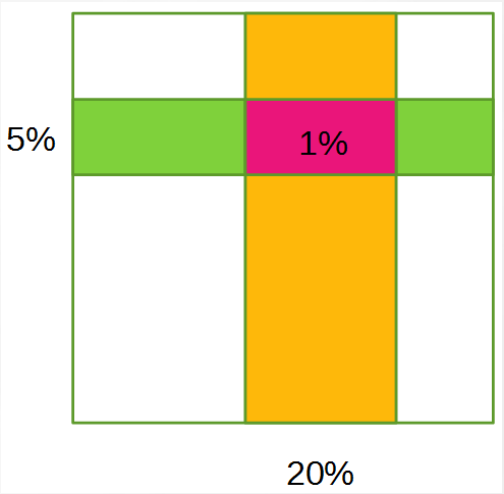
Modèles en grilles : Une méthode de coclustering

Coclustering

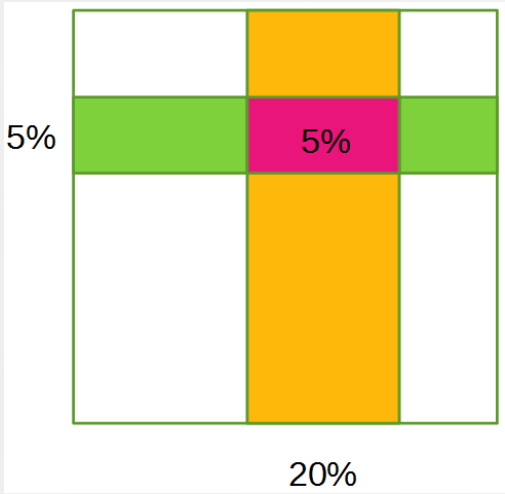
Aussi appelé, bi-clustering, bi-partitioning, double clustering, two-way clustering, block modeling, . . .

But : Résumer les données

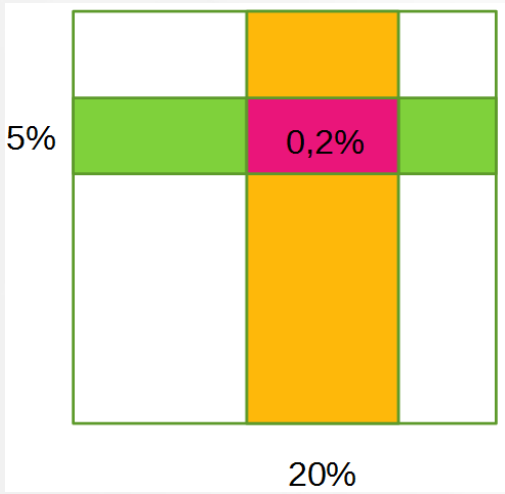
- Partitionner les lignes et colonnes d'une matrice
- $\rightarrow$ partition d'une matrice en coclusters
- Maximiser l'homogénéité des coclusters
- Mettre en évidence des coclusters informatifs (contraste entre la distribution des coclusters et la distribution espérée sous l'hypothèse d'indépendance (connaissant les marginales)



cocluster non-informatif



cocluster informatif



cocluster informatif

Coclustering : une littérature fournie, mais ...

Certaines limites

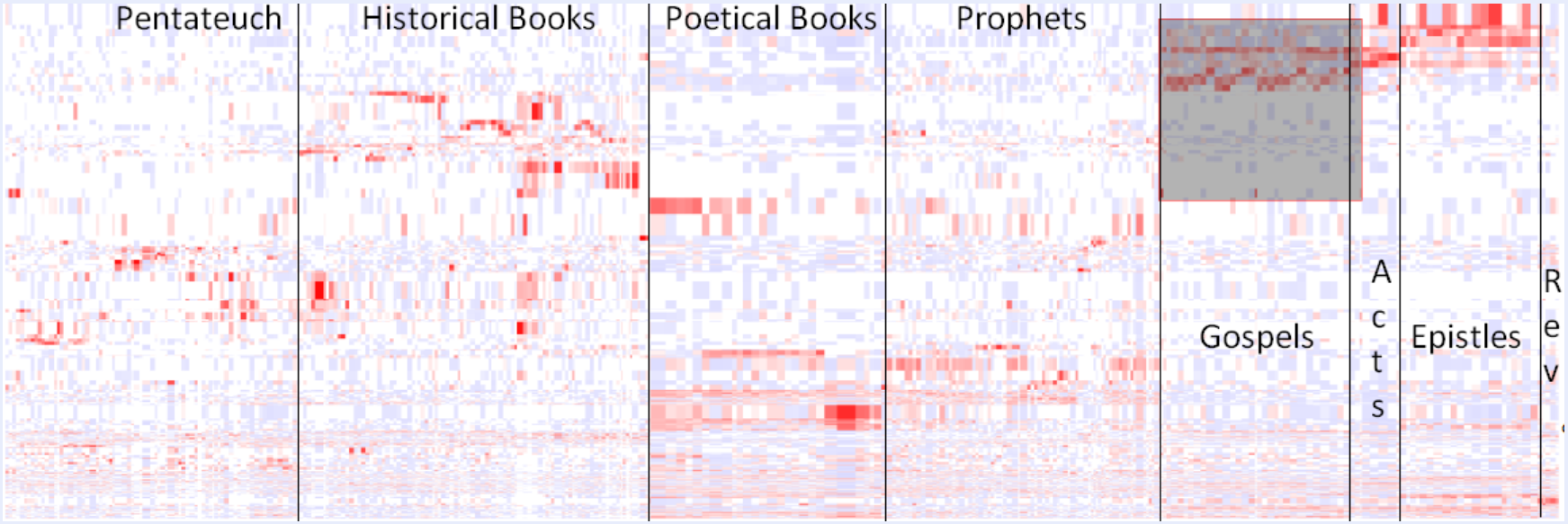
- Paramétrage (nombre de clusters/intervalles)
- $\rightarrow$ Risque under / over fitting
- Résistance non-prouvée au bruit / données aléatoires
- Limité aux dimensions catégorielles

Contributions des modèles en grille

- Pas de paramétrage utilisateur
- Méthode Bayésienne régularisée
- Résistance au bruit / données aléatoires
- Adapté aux dimensions numériques

Modèles en grilles pour la segmentation de texte

Applications à la segmentation de texte

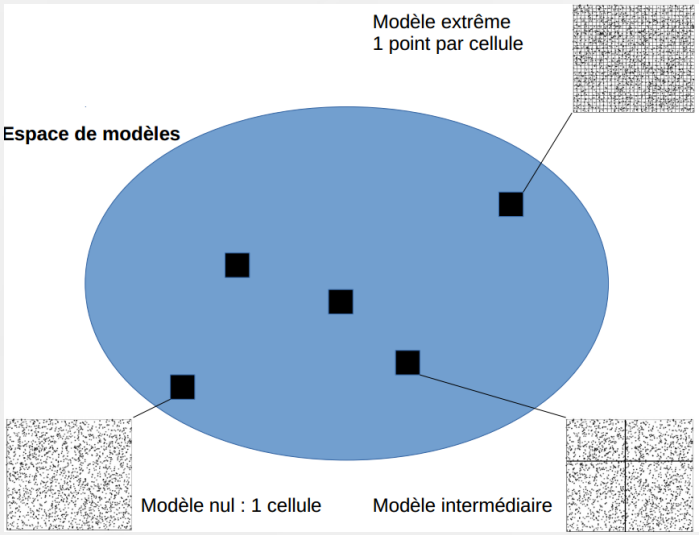


Données : “La Bible” (mot $\times$ position). Segmentation du vocabulaire en cluster de mots et segmentation de la position en intervalles (parties de la bible)

Modèles en grilles : Principes

Méthode

- Définir $\mathcal{M}$ l'espace des modèles en grilles
- Définir un prior sur $\mathcal{M}$
  - exploite la hiérarchie des paramètres
  - uniforme à chaque étage de la hiérarchie
  - prior dépendant des données
- Obtenir une expression analytique exacte du critère d'évaluation
- Exploiter des heuristiques d'optimisation



Modèles en grilles : approche Bayésienne

A la recherche du meilleur modèle de grille

- Meilleure grille : le modèle le plus probable connaissant les données
- Approche Bayésienne "Maximum A Posteriori" (MAP)

Vers le meilleur modèle : Maximisation

$$P(M \mid D) = \frac{P(M)P(D \mid M)}{P(D)}$$

Minimization

$$cost(M) = -\log(\underbrace{P(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{P(M)}_{\text{prior}} \times \underbrace{P(D \mid M)}_{\text{vraisemblance}})$$

Modèles en grilles : approche Bayésienne

$cost$: interprétation

- cas extrême : modèle le plus fin
  forte vraisembance, mais très faible probabilité a priori
- cas extrême : modèle nul
  forte probabilité a priori mais très faible vraisemblance
- cas intermédiaire : modèle le plus probable
  compromis entre précision et généralité du modèle

$cost$: interprétation Minimum Description Length (MDL)

- logarithme négatif de probabilités $\simeq$ longueur de codage (compression)
  $cost(M) = L(M) + L(D \mid M)$
- $-\log P(M)$: longueur de codage des paramètres du modèle
- $-\log P(D \mid M)$: longueur de codage des données connaissant le modèle

Exploration / Exploitation des grilles

MAP vs Pragmatisme

- Exemple : Données de la Bible ($\simeq$ 790K mots (dont $\simeq$ 12900 uniques) répartis en 31102 versets)
- Grille optimale avec 329 segments et 252 clusters de mots
- Difficile à analyser.

Les caractéristiques de l'approche (pas de paramètre, grille optimale la plus probable, MDL) sont intéressantes et garantissent la robustesse mais pas que le résultat soit analysable par un utilisateur.

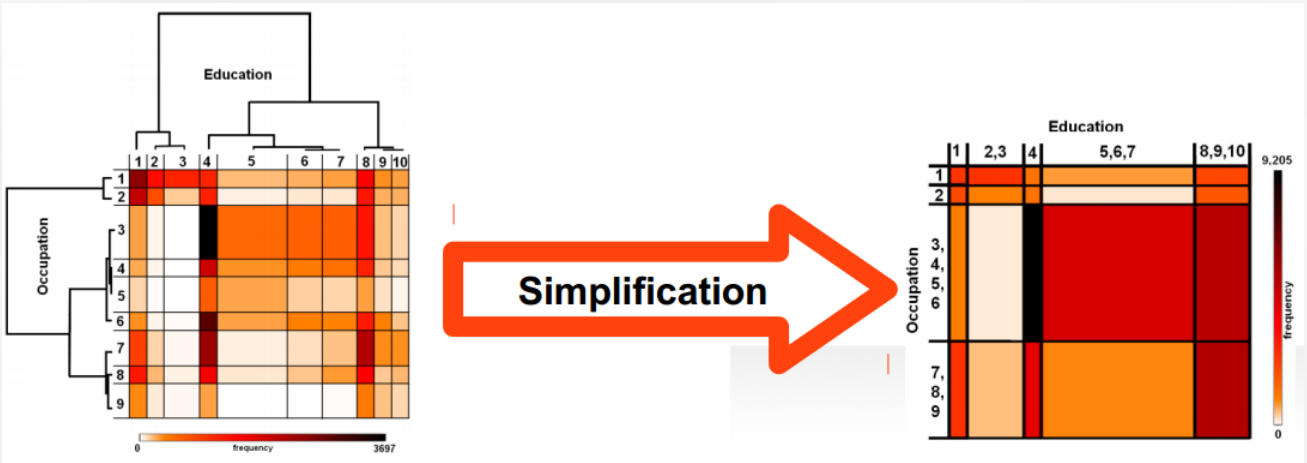
Outils d'analyse

- Outil de Post-traitement pour explorer la grille à diverses granularités
- Réduire le nombre de clusters $\Rightarrow$ Augmenter le nombre d'éléments par cluster
- Critère pour ordonner par importance les éléments d'un cluster

Exploration à diverses granularités

Simplification d'une grille

- De la grille optimale M^*, au modèle nul $M_\emptyset$
- Par hiérarchie ascendante
- Fusion de clusters/intervalles pour chaque dimension en minimisant Δ jusqu'à atteindre la granularité désirée
  $\Delta(c_{.1}, c_{.2}) = cost(M_{c_{.1} \cup c_{.2}}) - cost(M)$
- Contrôle du nombre de clusters/cellules ou de l'*information ratio*
  $IR(M') = (cost(M') - cost(M_\emptyset))/(cost(M^*) - cost(M_\emptyset))$



Typicité des éléments d'un cluster

But

Identifier les éléments représentatifs d'un cluster composé de nombreux éléments

Typicité d'une valeur v d'un cluster c : Intuition

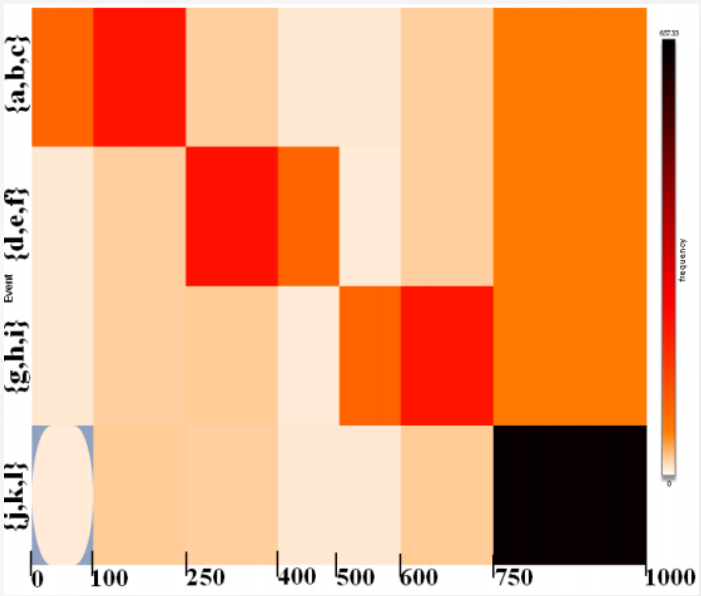
La **typicité** évalue l'**impact moyen** en terme de $cost$ sur la **qualité de la grille** de retirer une valeur v de son cluster c et de la placer dans un autre cluster $c_j \neq c$.
Alors, une valeur v sera dite **typique** d'un cluster c si elle **"proche"** de c et **"différente en moyenne"** des autres clusters $c_j \neq c$.

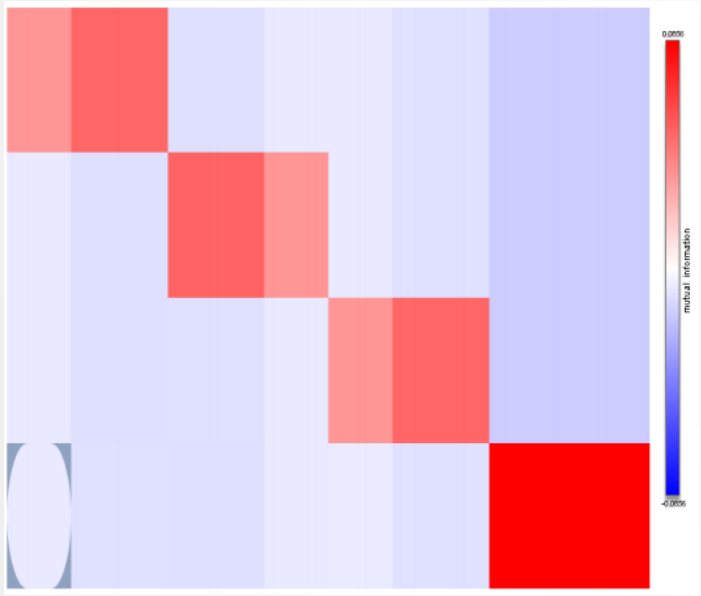
Permet d'identifier les mots les plus typiques d'un cluster de mots

Visualisation des co-clusters pertinents

Contribution à l'information mutuelle (CMI)

- Visualisation de la fréquence (heat map) : fréquence absolue
- Visualisation de la CMI : contribution locale de chaque co-cluster à l'information mutuelle globale ; caractérise l'excès/déficit d'interactions entre un intervalle et un cluster (donc dans un co-cluster par rapport aux marginales)





Plan

1 Modèles en grilles : principes

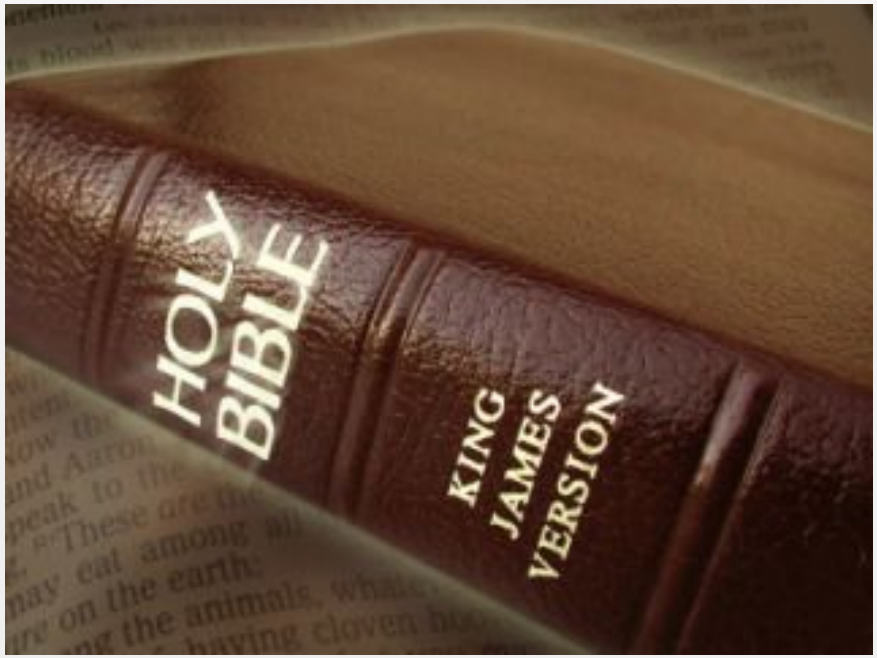
2 Expérimentations / Applications

3 Conclusion & Perspectives

Applications / Expérimentations

Données : La Bible

- livre le plus vendu et le plus étudié
- 66 livres composants
- Vocabulaire W : 12918 mots uniques
- Nombre mots N : 789628 mots
- Nombre de versets T : 31102 versets (pré-segmentation, donc autant de positions)
- Pré-traitement : stop words and stemming (Porter)



La Bible : vue d'ensemble (en 67 minutes)

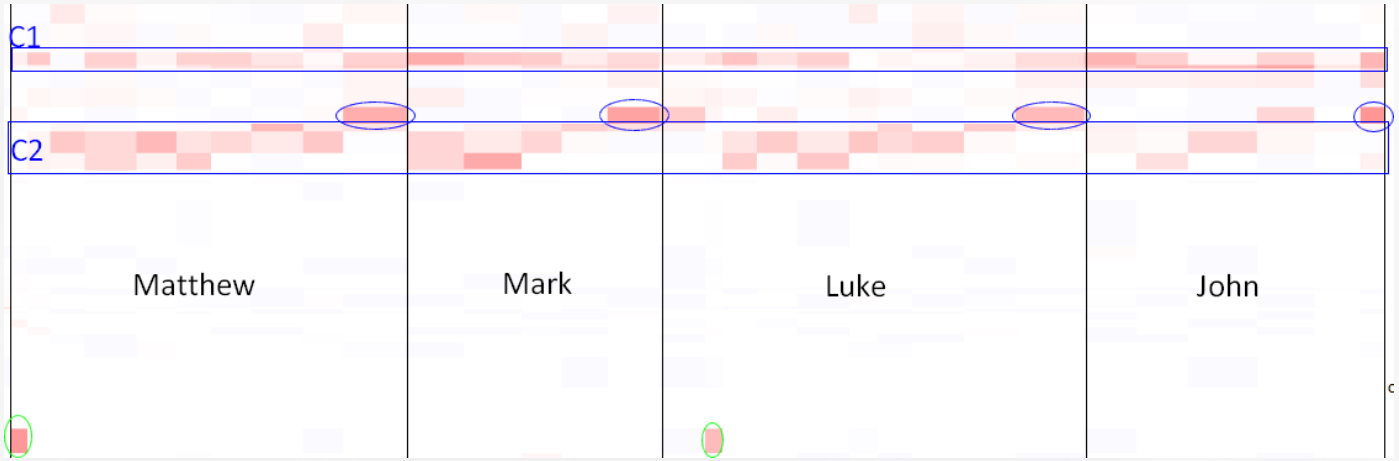
Grille optimale : 252 clusters de mots / 329 intervalles

O : *begat* (enegendré), mot le plus typique du cluster.
Généalogie des personnages clés : Adam, Noah (Génèse) ; Saul (Chroniques) ; Jesus (Evangiles)

O : *angel, repent, satan, throne*, mots typiques de l'Apocalypse de Saint-Jean

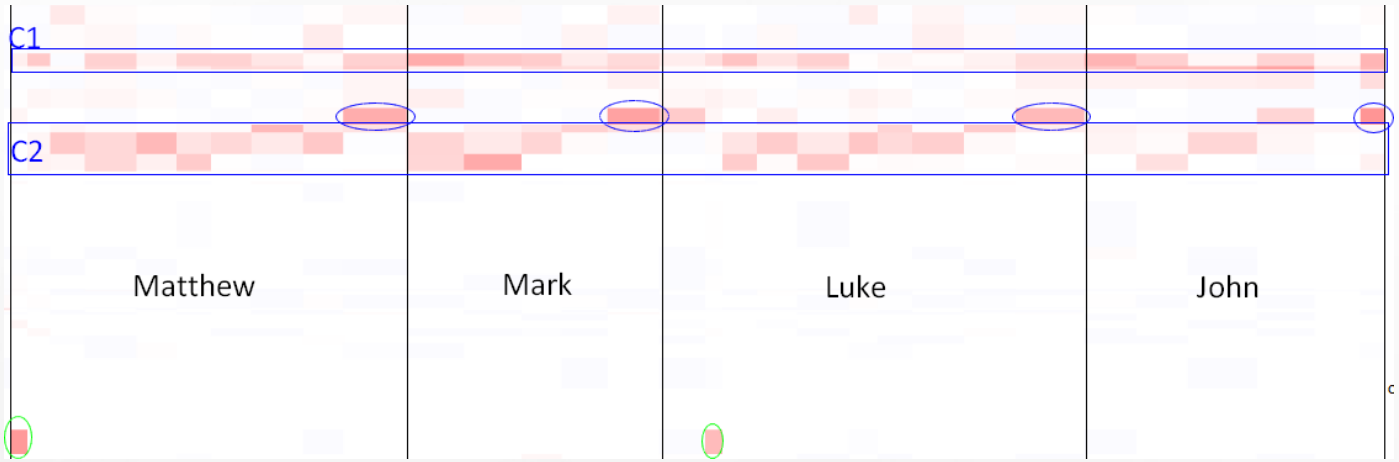
La Bible : Les évangiles (1)

- Les évangiles (Gospels) sont parfaitement séparés des autres livres de la Bible.
- Les différents évangiles (Matthew, Mark, Luke and John) sont identifiés avec une marge de 1-3 versets



- C1 : Jesus, discipl, Peter, John, Simon. Les personnages principaux des évangiles
- C2 : ask, temple, sit, pharise, whosoever, ship, heal. Les principaux actes et rencontres de Jésus.

La Bible : Les évangiles (2)



- Mise en évidence de la différence entre les évangiles synoptiques (Matthew, Mark and Luke) celui de John : CMI cumulé pour C2 pour un évangile synoptique ($> 25 \times 10^{-4}$) comparé à celui de John ($< 5 \times 10^{-4}$)
- O La généalogie de Jésus présente uniquement dans les évangiles de Matthew et Luke
- O La passion et résurrection du Christ commune aux quatre évangiles

La Bible : les 66 livres

Hiérarchie agglomérative vers 66 segments/intervalles

- 13 fins de livres identifiées parfaitement
- avec une tolérance ± 5 versets : +8 fins de livres identifiées

Sur-segmentation de la Genèse (et de l'Exode) : Différentes histoires de différents personnages de l'origine du monde jusqu'à Abraham, Jacob et Joseph avec un vocabulaire spécifique à ces deux livres (peu présent dans la suite de la Bible)

La Bible . . . dézoomée

Vers les grandes parties de la Bible et le modèle nul

- Segmentation parfaite Ancien/Nouveau Testament
- Segmentation parfaite dans l'ancien testament :
  Pentateuque, Livres historiques – Livres Poétiques et Prophètes

Plan

1 Modèles en grilles : principes

2 Expérimentations / Applications

3 Conclusion & Perspectives

Conclusion & perspectives

Bilan préliminaire

- Vers la segmentation exploratoire de texte
- Modèles en grilles : Méthode sans paramètre utilisateur
- Outils d'exploration / exploitation du résumé
- Critères de visualisation des caractéristiques du résumé

Bientôt au prochain épisode

- Validation expérimentale sur des données artificielles
- Expérimentations sur d'autres textes/livres
- Expérimentations sur d'autres séquences de mots (tweets, résumés d'articles d'une conférence, . . .)
- Comparaisons avec des méthodes concurrentes

Merci . . .

. . . pour votre attention