Clustering de textes et graphe de députés

Achille

Juin - Juillet 2020

1 Introduction

1.1 Quoi

Appliquer des méthodes de clustering , de représentation de textes et de comparaisons entre textes courts qui sont ceux de l'assemblée nationale, (en déduire des liens entre députés).

1.2 Pourquoi

Explorer et répertorier des méthodes utilisables sur ce type de données (savoir de quelle manière les députés+8+++++++++5 partagent des arguments etc.)

1.3 Contexte

1.4 Données

2 Clustering de textes

L'objectif de cette section est d'analyser un ensemble de méthodes pour établir des clusters sur les textes. Nous avons choisi des méthodes qui s'appuient sur le champs lexical d'une lettre, révélateur du sujet abordé. À ce jour, toutes nos méthodes utilisent 2 étapes cruciales : mettre le document sous forme vectorielle, de préférence de dimension faible 3, puis appliquer des techniques classiques de clustering sur ces représentations vectorielles 3.5.

3 Représentation vectorielle des documents et réduction de dimension

3.1 Premières représentations

Il est d'abbord nécessaire de partir d'un matériau de base, une première représentation vectorielle de haute dimension, avant de réduire les dimensions dans la section 3.4. Ces premières matrices sont sous la forme suivante : chaque ligne représente un document et chaque colonne représente un mot.

3.1.1 TF

La représentation la plus intuitive pour représenter le champs lexical d'un texte est la matrice Term-Frequency, constituée d'entiers indiquant combien de fois un certain mot apparaît dans un certain document.

3.1.2 TF-IDF

Le Term Frequency - Inverse Document Frequency ...

3.1.3 SIF

Le Smooth inverse Frequency ...

3.2 Méthodes de réduction de dimension

3.2.1 Analyse en composante principale

L'analyse en composante principale

3.2.2 Analyse Sémantique Latente

L'analyse Sémantique Latente ... effectue une SVD , Division en Valeurs Singulières, sur la matrice des TF-IDF.

3.2.3 Allocation de Dirichlet Latente

L'Allocation de Dirichlet Latente (LDA) s'appuie sur la matrice des Term-Frequency. On commence par choisir la dimension k qui veut représenter concrètement le nombre de thèmes pouvant être abordés. La forme vectorielle de dimension k finalement obtenue pour chaque document représentera à quel point le document correspond à chaque thème généré. Les documents sont représentés par des distributions de probabilités (**RANDOM MIXTURE**) sur les thèmes. Les thèmes sont représentés par des distributions de probabilité pour chaque mot. Le LDA suppose que les documents ont tous été générés de la façon suivante :

Choisir le nombre de mots $N \sim$ Poissson(ξ)
Choisir les probabilités de chaque thème $\theta \sim$ Dir(α)
for chaque mot w_n, n allant de 1 à N **do**
  Choisir un thème $Z_n \sim$ Multionomial($\alpha, 1$)
  Choisir un mot w_n avec une probabilité $P(w_n|z_n, \beta)$, une probabilité multinomiale dépendante du thème Z_n.
end for

k est la dimension de la distribution de Dirichlet utilisée, donc du paramètre α et donc du nombre de thèmes possibles. β est une matrice $V \times k$ représentant pour chaque mot la probabilité d'apparaître dans chaque thème (avec V le nombre de mots). α sont les paramètres, changeant pour chaque document, de laloi de Dirichlet générant θ, le vecteur associant à chaque thème sa probabilité.

On peut donc en déduire à présent une manière d'inférer ces paramètres, chose plus compliquée que nous verrons peut-être. Cela se déroule de la manière suivante : Au bout d'un moment ça converge.

Algorithm 1 Inférence des paramètres

Initialiser les valeurs de θ
for chaque document d **do**
  for chaque mot w du document **do**
    Calculer la probabilité $P(t|d)$ que le document d soit assigné au thème t
    Calculer $P(w|t)$ la probabilité que le thème t soit assigné au mot w
    On attribue alors le thème t au produit de ces deux probabilités.
  end for
end for

3.2.4 Factorisation en Matrices non Négatives

Que cela signifie

3.2.5 Uniform Manifold Approximation for Projection and Representation

3.3 Représentation visuelle

3.4 Distances

3.5 Clustering avec les représentations obtenues

4 Annexes calculatoires

4.1 Méthodes

4.1.1 K-means demo

Si on est dans un minimum local pour la fonction de la somme des distances intra-classe :

$$\sum_{i \in I} \sum_{r \in R} \sum_{j \in D} \mu_{r,i}(x_{i,j} - c_{r,j})^2$$

avec :

- I : Ensemble des points
- R : Ensemble des clusters
- D : Dimensions
- $\mu_{r,i}$: 1 ou 0 selon que le point i appartient ou non au cluster r
- $x_{i,j}$: Valeur dans la dimension j du point i
- $c_{r,j}$: Valeur dans la dimension j du centre du cluster r

alors on a sa dérivée qui est nulle. On calcule donc sa dérivée par rapport à une dimension j' du centre d'un cluster r', donc par rapport à $c_{r',j'}$:

$$\frac{d}{dc_{r',j'}} \sum_{i \in I} \sum_{r \in R} \sum_{j \in D} \mu_{r,i}(x_{i,j} - c_{r,j})^2 = 0$$

Cette dérivée est nulle pour tous les termes $c_{r,j}$ avec $r \neq r'$ et $j \neq j'$. On obtient donc la simplification suivante :

$$\frac{d}{dc_{r',j'}} \sum_{i \in I} \mu_{r,i}(x_{i,j'} - c_{r',j'})^2 = 0$$

$$\frac{d}{dc_{r,j}} \sum_{i \in I} \mu_{r,i}(x_{i,j} - c_{r,j})^2 = 0 \text{ en remplacant } c_{r',j'} \text{ par } c_{r,j} \text{ pour simplifier les notation}$$

$$\sum_{i \in I} -2\mu_{r,i}(x_{i,j} - c_{r,j}) = 0$$

$$\sum_{i \in I} \mu_{r,i}(x_{i,j} - c_{r,j}) = 0$$

$$\sum_{i \in I} \mu_{r,i} x_{i,j} = \sum_{i \in I} \mu_{r,i} c_{r,j}$$

$$\sum_{i \in I} \mu_{r,i} x_{i,j} = (\sum_{i \in I} \mu_{r,i}) c_{r,j}$$

$$\sum_{i \in I} \mu_{r,i} x_{i,j} = |r| c_{r,j} \text{ avec } |r| \text{ nombre de points de r}$$

$$c_{r,j} = \frac{\sum_{i \in I} \mu_{r,i} x_{i,j}}{|r|}$$

$$= \frac{\sum_{i \in I_r} x_{i,j}}{|r|}$$

4.2 Trouver le bon nombre de clusters

4.2.1 Somme des variances intra-groupes

$$W = \sum_{r \in R} \sum_{i \in I_r} \mu_{r,i} ||x_i - c_r||^2$$

Ou, en version blabla :

$$W = \sum_{r \in R} \sum_{i \in I_r} \sum_{j \in D} \mu_{r,i} ||x_{i,j} - c_{r,j}||^2$$

4.2.2 Calinski-Harabasz

`https://scikit-learn.org/stable/modules/clustering.html`

Rapport entre la variance inter-groupes et des variance intra-groupes.

La variance inter-groupes est la somme des distances entre le centres des clusters et le centre global, pondéré par le nombre de points de chaque cluster. Elle mesure la dispersion des clusters.

$$V = \sum_{r \in R} |r| \|c_r - c\|$$

avec c le centre global.

La variance intra-groupes est la somme des distances des points d'un cluster avec le centre de ce cluster, divisé par le nombre de points de ce cluster. Elle mesure la dispersion d'un cluster.

$$W_r = \sum_{r \in R} \frac{1}{\|I_r\|} \sum_{i \in I_r} \|x_i - c_r\|$$

Pour obtenir l'indice de Calinski-Harabasz, on divise la variance inter-groupes par la somme des variances intra-groupes. Le tout est ocefficienté par le rapport entre la différence entre le nombre de points et le nombre de clusters désirés et le nombre de clusters moins 1.

$$S_{CH} = \frac{(n - |R|)V}{(|R| - 1) \sum_{r \in R} W_r}$$

Plus l'indice est grand, meilleure est la classification, car cela signifie que les clusters sont plus dispersés enre eux et que les points d'un clusters sont moins dispersés.

4.2.3 Davies-bouldin

L'indice calcule la somme sur chaque cluster de : le maximum pour tous les autres clusters de la somme de la distance moyenne avec leur centre des deux clusters et la distance entre les centres de deux clusters.

$$S_{DB} = \frac{1}{|R|} \sum_{r \in R} \max_{r' \neq r} \left(\frac{\bar{\delta_r} + \bar{\delta_{r'}}}{d(c_r, c_{r'})} \right)$$

avec

$$\bar{\delta_r} = \frac{1}{|r|} \sum_{i \in I_r} d(x_i, c_r)$$

la distance moyenne des points d'un cluster à son centre.

Plus l'indice est petit, plus la classification est bonne. Cela signifie en effet que la distance des points avec leur centre réduit alors que celle entre les clusters augmente.

4.3 Comparer et combiner les clusters

Nous sommes confrontés à un grand nombre de méthodes de clustering, et il s'agit à présent de savoir comment les combiner et les comparer entre eux.

4.3.1 Adjusted Rand Index

Adjusted rand score avec sklearn

4.4 Autres techniques de clustering

Les k-means, qui s'appuyent sur la distance euclidienne (voir 4.1.1), ne permetent pas d'appréshender des formes non convexes. Comme il n'y a aucune raisnon que les documents traités soient sous forme convexe, il est donc nécessaire de faire appel à d'autres techniques de clustering.

4.4.1 Clustering Hiérarchique

4.4.2 Spectral Clustering

Cette méthode, résumée et comparée aux kernel k-means dans [1], s'appuie sur la matrice des affinités, qui peut donc être calculée avec n'importe quelle mesure de similarité.

[2] [**?**]

References

[1] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 551–556, New York, NY, USA, 2004. Association for Computing Machinery.

[2] Manning. Sentence classification in nlp. `https://freecontent.manning.com/sentence-classification-in-nlp/`, 2019.