Devoir 5 et 6: Analyse de texte et extension

Visseho Adjiwanou, PhD.

20 June 2023

NOTE
1. Ce devoir est à rendre au plus grand tard le **27 décembre**.
2. Vous le soumettez sur Moodle, une version RMarkDown et une version pdf.
3. Ce devoir est **individuel**. Cela ne vous empêche pas de poser des questions à vos ami-e-s. Mais...
4. Éviter le **plagiat**. Éviter de vous copier.

DEVOIR 5

Exercice 1: Le biais idéologique des journaux

Bien que certains défenseurs de la santé publique considèrent les cigarettes électroniques comme une aide efficace à l'arrêt du tabac, d'autres mettent en garde contre les risques potentiels, tels que les niveaux élevés de nicotine. Imaginez qu'un chercheur décide d'étudier l'opinion publique à l'égard des cigarettes électroniques en recueillant des messages Twitter sur les cigarettes électroniques et en effectuant une analyse des sentiments.

1. Quels sont les trois biais possibles qui vous préoccupent le plus dans cette étude?
2. Clark et al. (2016) mené une telle étude. D'abord, ils ont recueilli 850 000 tweets qui utilisaient des mots clés liés à l'e-cigarette de janvier 2012 à décembre 2014. Une inspection plus approfondie leur a permis de constater que bon nombre de ces tweets étaient automatisés (c.-à-d. non produits par des humains, c'est ce qu'on appelle des bots). Ils ont développé un algorithme de détection humaine pour séparer les tweets automatisés des tweets organiques (posté par des individus). En utilisant cet algorithme de détection humaine, ils ont trouvé que 80% des tweets étaient automatisés. Est-ce que cette conclusion change votre réponse à la partie (1)?
3. Quand ils ont comparé le sentiment dans les tweets organiques et automatisés, ils ont trouvé que les tweets automatisés étaient plus positifs que les tweets organiques (6,17 contre 5,84). Est-ce que cette conclusion change votre réponse à (2)?

Clark, Eric M., Chris A. Jones, Jake Ryland Williams, Allison N. Kurti, Mitchell Craig Norotsky, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "Vaporous Marketing: Uncovering Pervasive Electronic Cigarette Advertisements on Twitter." PLoS ONE 11 (7):e0157304. https://doi.org/10.1371/journal.pone.0157304.

Exercice 2

L'analyse de texte offre aux chercheurs un ensemble puissant d'outils pour extraire des informations générales à partir d'un grand nombre de documents.

Cet exercice est basé sur Gentzkow, M. and Shapiro, J. M. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

Nous analyserons les données des journaux à travers le pays pour voir quels sujets ils couvrent et comment ces sujets sont liés à leur parti pris idéologique. Les auteurs ont calculé une mesure de l'"inclinaison" d'un journal en comparant son langage aux discours prononcés par les démocrates et les républicains au Congrès américain.

Vous utiliserez trois sources de données pour cette analyse. Le premier, `dtm`, est une matrice documents-termes avec une ligne par journal, contenant les 1000 phrases - dérivées et traitées (stemmed and processed) - qui permettent le mieux d'identifier le locuteur comme républicain ou démocrate. Par exemple, "living in poverty" est une expression la plus fréquemment prononcée par les démocrates, tandis que "global war on terror" est une expression la plus fréquemment prononcée par les républicains; une expression comme "exchange rate" ne serait pas incluse dans cet ensemble de données, car elle est souvent utilisée par les membres des deux partis et est donc un mauvais indicateur d'idéologie.

Le deuxième objet, `papers`, contient des données sur les journaux sur lesquels `dtm` est basé. Les noms de lignes dans `dtm` correspondent à la variable `newsid` dans `papers`. Les variables sont

Name	Description
`newsid`	The newspaper ID
`paper`	The newspaper name
`city`	The city in which the newspaper is based
`state`	The state in which the newspaper is based
`district`	Congressional district where the newspaper is based (data for Texas only)
`nslant`	The "ideological slant" (lower numbers mean more Democratic)

Le troisième objet, `cong`, contient des données sur les membres du Congrès en fonction de leur discours politique, que nous comparerons à l'orientation idéologique des journaux des régions que ces législateurs représentent. Les variables sont :

Name	Description
`legname`	Legislator's name
`state`	Legislator's state
`district`	Legislator's Congressional district
`chamber`	Chamber in which legislator serves (House or Senate)
`party`	Legislator's party
`cslant`	Ideological slant based on legislator's speech (lower numbers mean more Democratic)

Question 0

Lisez l'article et présenter les limites de l'étude.

Question 1

Nous nous concentrerons d'abord sur l'inclinaison des journaux, que les auteurs définissent comme la tendance à utiliser un langage qui influencerait les lecteurs vers la gauche ou la droite politique. Chargez les données et présenter un graphique de distribution de `nslant` à partir des données `papers`, avec une ligne verticale à la médiane. Quel journal du pays a le plus grand penchant pour la gauche ? Qu'en est-il de la droite?

Question 2

Nous explorerons le contenu de ces journaux en utilisant le package `wordcloud`.

Chargez d'abord le package `wordcloud`. Créez un nuage de mots des premiers mots (au plus 20) à partir de la matrice documents-termes `dtm`. Quels ont été les principaux sujets d'actualité en 2005 lorsque ces données ont été collectées? Astuce: convertissez «dtm» en «matrice».

Maintenant, selectionner le dixième des journaux avec l'orientation politique la plus à gauche (la plus basse) et la plus à droite (la plus élevée). Créez deux nuages de mots indiquant les mots les plus couramment utilisés par chaque groupe de journaux (encore une fois, au plus 20 mots). En quoi leur langue diffère-t-elle ? Ont-ils quelque chose en commun ? Astuce: pour utiliser vos outils de selection/indexation habituels (dplyr), convertissez votre matrice dtm en une base de données à l'aide de la fonction `data.frame`.

Portez une attention particulière aux avertissements, car ils contiennent des informations importantes. Pour un bonus supplémentaire, voyez si vous pouvez les faire disparaître.

Question 3

Nous allons maintenant explorer la relation entre l'inclinaison politique des journaux et le langage utilisé par les membres du Congrès.

À l'aide de la base de données `cong`, calculez l'inclinaison moyenne par État séparément pour la Chambre et le Sénat. Utilisez maintenant `papiers` pour calculer l'inclinaison moyenne des journaux par état. Faites deux graphiques avec l'inclinaison Congessional sur l'axe des x et l'inclinaison du journal sur l'axe des y - un pour la Chambre, un pour le Sénat. Incluez la droite de régression dans chaque graphique: une rouge pour le Sénat et une verte pour la Chambre. Étiquetez vos axes, intitulez vos graphiques et assurez-vous que les axes sont les mêmes pour la comparabilité. Pouvez-vous conclure que les journaux sont influencés par le langage politique des élus ? Sinon, comment pouvez-vous interpréter les résultats?

Question 4

Nous allons maintenant examiner de plus près la relation entre l'inclinaison du Congrès et celle des médias au niveau du district, pour un État en particulier, le Texas. Pour ce faire, selectionner les deux bases de données de Texas uniquement, puis fusionnez-les par district et État, en ne conservant que les observations qui apparaissent dans les deux ensembles de données. Ensuite, produisez le même graphique qu'à la question 3 ci-dessus, mais au niveau du district (juste pour le Congrès - House). Que trouvez-vous? Selon vous, quels résultats sont les plus informatifs et pourquoi?

Question 5

Identifiez les termes les plus importants pour saisir les variations régionales dans ce qui est considéré comme digne d'intérêt - les termes qui apparaissent fréquemment dans certains documents, mais pas dans tous les documents. Pour ce faire, calculez le *terme fréquence-fréquence inverse du document (tf-idf)* pour chaque combinaison d'expression et de journal dans l'ensemble de données (pour cela, utilisez le package `tm` et l'objet `dtm` fourni à l'origine). On n'a pas vu le package `tm` en classe, essayer de répondre à la question avec `tidytext`.

Sélectionner à partir de la matrice transformée tf-idf que vous avez créée pour contenir le journal le plus proche de Princeton, le "Home News Tribune" d'East Brunswick, NJ. Imprimez les termes avec le plus grand tf-idf dans l'ordre décroissant. Quels sujets intéressent notre région, mais ne sont pas susceptibles de faire l'actualité nationale ?

Question 6

Regroupez tous les journaux de New Jersey sur leur mesure tf-idf. Appliquer l'algorithme k-means avec 3 clusters. Résumez les résultats en imprimant les dix termes les plus importants au centre de gravité de

chacun des groupes résultants, et montrez quels journaux appartiennent à chaque groupe. De quels sujets NJ se soucie-t-il ?

DEVOIR 6

Ecercice 3: Réseau de commerce international

La taille et la structure des flux commerciaux internationaux varient considérablement dans le temps. Cet exercice est basé en partie sur Luca De Benedictis and Lucia Tajoli. (2011). 'The World Trade Network.' *The World Economy*, 34:8, pp.1417-1454. Les données commerciales sont de Katherine Barbieri and Omar Keshk. (2012). *Correlates of War Project Trade Data Set*, Version 3.0. available at http://correlatesofwar.org.

Le volume de marchandises échangées entre les pays a augmenté rapidement au cours du siècle dernier, alors que les progrès technologiques ont réduit le coût du transport et que les pays ont adopté des politiques commerciales plus libérales. Parfois, cependant, les flux commerciaux ont diminué en raison d'événements perturbateurs tels que des guerres majeures et l'adoption de politiques commerciales protectionnistes. Dans cet exercice, nous explorerons certains de ces changements en examinant le réseau du commerce international sur plusieurs périodes. Le fichier de données `trade.csv` contient la valeur des exportations d'un pays vers un autre au cours d'une année donnée. Les noms et descriptions des variables dans cet ensemble de données sont:

Name	Description
`country1`	Country name of exporter
`country2`	Country name of importer
`year`	Year
`exports`	Total value of exports (in tens of millions of dollars)

Les données sont données pour les années 1900, 1920, 1940, 1955, 1980, 2000 et 2009.

Question 1

Nous commençons par analyser le commerce international comme un réseau dirigé non pondéré. Pour chaque année de la base de données, créez une matrice d'adjacence dont l'entrée (i, j) est égale à 1 si le pays i exporte vers le pays j. Si cette exportation est nulle, alors l'entrée est égale à 0. Nous supposons que les données manquantes, indiquées par `NA`, représentent un commerce nul. Présenter le graphique de la «densité du réseau», qui est définie au fil du temps comme suit,

$$\text{network density} = \frac{\text{number of edges}}{\text{number of potential edges}}$$

La fonction `graph.density` peut calculer cette mesure étant donné une matrice d'adjacence. Interpréter le résultat.

Question 2

Pour les années 1900, 1955 et 2009, calculez les mesures de centralité basées sur le degré (*degree*), l'intermédiarité (*betweenness*) et la proximité (*closeness*) (basées sur le degré total) pour chaque année. Pour chaque année, listez les cinq pays qui ont les valeurs les plus élevées de ces mesures de centralité. Comment évoluent les pays sur les listes au fil du temps ? Commentez brièvement les résultats.

Question 3

Nous analysons maintenant le réseau commercial international comme un réseau dirigé et pondéré dans lequel chaque lien a un poids non négatif proportionnel à son volume commercial correspondant. Créez une matrice de contengence pour ces données de réseau. Pour les années 1900, 1955 et 2009, calculez les mesures de centralité ci-dessus pour le réseau commercial pondéré. Au lieu du degré, cependant, calculez la *force du graphe (graph strength)*, qui dans ce cas est égale à la somme des importations et des exportations avec tous les nœuds adjacents. La fonction `graph.strength` peut être utilisée pour calculer cette version pondérée du degré. Pour l'intermédiarité et la proximité, nous utilisons la même fonction qu'auparavant, c'est-à-dire `closeness` et `betwenness`, qui peuvent gérer les graphiques pondérés de manière appropriée. Les résultats diffèrent-ils de ceux du réseau non pondéré ? Examinez les cinq premiers pays. Pouvez-vous penser à une autre façon de calculer la centralité dans ce réseau qui tient compte de la valeur des exportations de chaque pays ? Discutez brièvement.

Question 4

Appliquez l'algorithme PageRank au réseau commercial pondéré séparément pour chaque année. Pour chaque année, identifiez les 5 pays les plus influents selon cet algorithme. En outre, examinez comment le classement des valeurs de PageRank a changé au fil du temps pour chacun des cinq pays suivants: États-Unis, Royaume-Uni, Russie, Japon et Chine. Commentez brièvement les modèles que vous observez.

Ecxercice 4

https://cran.r-project.org/web/packages/wikipediatrend/wikipediatrend.pdf

Penney (2016) examiné si la publicité généralisée sur la surveillance NSA / PRISM (les révélations de Snowden) en juin 2013 était associée à une baisse soudaine et brusque du trafic vers les articles de Wikipédia sur des sujets qui soulèvent des problèmes de confidentialité. Si c'est le cas, ce changement de comportement serait compatible avec un effet paralysant résultant de la surveillance de masse. L'approche de Penney (2016) est parfois appelée une conception de série temporelle interrompue, et elle est liée aux approches décrites dans la section 2.4.3 de bitbybit de Salganik.

Pour choisir les mots-clés du sujet, Penney s'est référé à la liste utilisée par le département américain de la sécurité intérieure (DHS) pour le suivi et la surveillance des médias sociaux. La liste DHS catégorise certains termes de recherche dans un éventail de questions, à savoir `Préoccupation pour la santé` "Health Concern), `Sécurité des infrastructures` (Infrastructure Security) et `Terrorisme` (Terrorism). Pour le groupe d'étude, Penney a utilisé les 48 mots clés associés au `Terrorisme` (voir appendix table 8 de l'article). Il a ensuite agrégé le nombre de vues d'articles de Wikipedia sur une base mensuelle pour les 48 articles de Wikipédia correspondants sur une période de 32 mois, de début janvier à fin août 2014. Pour renforcer son argument, il a également créé plusieurs groupes de comparaison vues d'articles sur d'autres sujets.

Maintenant, vous allez reproduire et étendre Penney (2016). Toutes les données brutes dont vous aurez besoin pour cette activité sont disponibles sur Wikipedia. Ou vous pouvez l'obtenir à partir du paquet R wikipediatrend Meissner and R Core Team 2016. Lorsque vous écrivez vos réponses, veuillez noter quelle source de données vous avez utilisée. Cette activité vous donnera l'occasion de vous entraîner dans la recherche de données et de réfléchir aux expériences naturelles dans les sources de données volumineuses. Il vous permettra également de démarrer avec une source de données potentiellement intéressante pour les futurs projets.

1. Lisez Penney (2016) et reproduisez sa figure 2 qui montre les pages vues pour les pages `Terrorism` avant et après les révélations de Snowden. Interpréter les résultats.
2. Ensuite, reproduisez la figure 4A, qui compare le groupe d'étude (articles `Terrorism`) avec un groupe de comparaison en utilisant des mots clés classés sous `DHS & other agencies` dans la liste DHS (see appendix table 10 and footnote 139). Interpréter les résultats.

3. Dans la partie (2), vous avez comparé le groupe d'étude avec un groupe de comparaison. Penney a également comparé deux autres groupes de comparaison: les articles liés à la `Infrastructure Security` (tableau 3) et les pages populaires de Wikipédia (appendix table 12). Venez avec un groupe comparateur alternatif, et testez si les résultats de la partie (2) sont sensibles à votre choix de groupe de comparaison. Quel choix fait le plus logique? Pourquoi?

4. Penney a déclaré que les mots clés relatifs au `terrorism` ont été utilisés pour sélectionner les articles de Wikipédia parce que le gouvernement américain a cité le terrorisme comme une justification clé pour ses pratiques de surveillance en ligne. Pour vérifier ces 48 mots-clés liés au `terrorism`, Penney (2016) a également mené une enquête sur MTurk (vous, utiliser Facebook), demandant aux répondants de noter chacun des mots-clés en termes de problèmes gouvernementaux (Government Trouble), de respect de la vie privée (Privacy-Sensitive) et d'évitement (Avoidance) (appendix table 7 and 8). Répliquez l'enquête sur Facebook et comparez vos résultats.

5. D'après les résultats de la partie (4) et votre lecture de l'article, êtes-vous d'accord avec le choix des mots-clés de Penney dans le groupe d'étude? Pourquoi ou pourquoi pas? Sinon, que suggérez-vous plutôt?

Cette page démontre l'utilisation de wikipediatrend pour collecter les données de wikipedia: https://rpubs.com/aashishkpandey/Wikipedia-Trends

Penney, Jonathon. 2016. "Chilling Effects: Online Surveillance and Wikipedia Use." Berkeley Technology Law Journal 31 (1):117. https://doi.org/http://dx.doi.org/10.15779/Z38SS13.

Meissner, Peter, and R Core Team. 2016. "Wikipediatrend: Public Subject Attention via Wikipedia Page View Statistics." https://CRAN.R-project.org/package=wikipediatrend.

```
#install.packages("wikipediatrend")
library(wikipediatrend)
```

```
##
##   [wikipediatrend]
##
##   Note:
##
##     - Data before 2016-01-01
##       * is provided by petermeissner.de and
##       * was prepared in a project commissioned by the Hertie School of Governance (Prof. Dr. Simon Mu
##       * and supported by the Daimler and Benz Foundation.
##
##     - Data from 2016-01-01 onwards
##       * is provided by the Wikipedia Foundation
##       * via its pageviews package and API.
##
```

```
data(package="wikipediatrend")
```

```
## no data sets found
```

```
fl = wp_trend(                          # func wp_trend() builds a data object - mk
  "François_Legault",                               # search term is "Mary_Kom"
  from = "2011-01-01",
  to = Sys.Date())
```