Projet « Fouille de Textes » – Classification d'e-mails

Masters ECD - SISE

Nous recevons des e-mails tous les jours. Après les avoir consultés, nous essayons de les classer dans des dossiers (« folder » en anglais) selon le thème que l'on pourrait leur associer (ex. travail, amis, etc.). Pour certains messages, nous n'avons pas de catégorie précise. Nous les mettons alors dans un dossier « fourre-tout » qui, au fil des années, enfle de manière démesurée. De plus, les catégories elles-mêmes ne sont pas figées et peuvent évoluer au cours du temps (ex. le dossier « amis » peut être subdivisé en « amis encore » et « amis plus vraiment », etc.). A force de vouloir bien faire, on se retrouve souvent avec toute une série de dossiers dont on n'identifie même plus la teneur.

Dans une entreprise, la classification des messages est un enjeu crucial. Les collaborateurs reçoivent un grand nombre d'e-mails professionnels. Les classer de manière harmonisée est primordial. Les individus doivent disposer d'une perception cohérente de la nature des informations transmises au sein de la société.

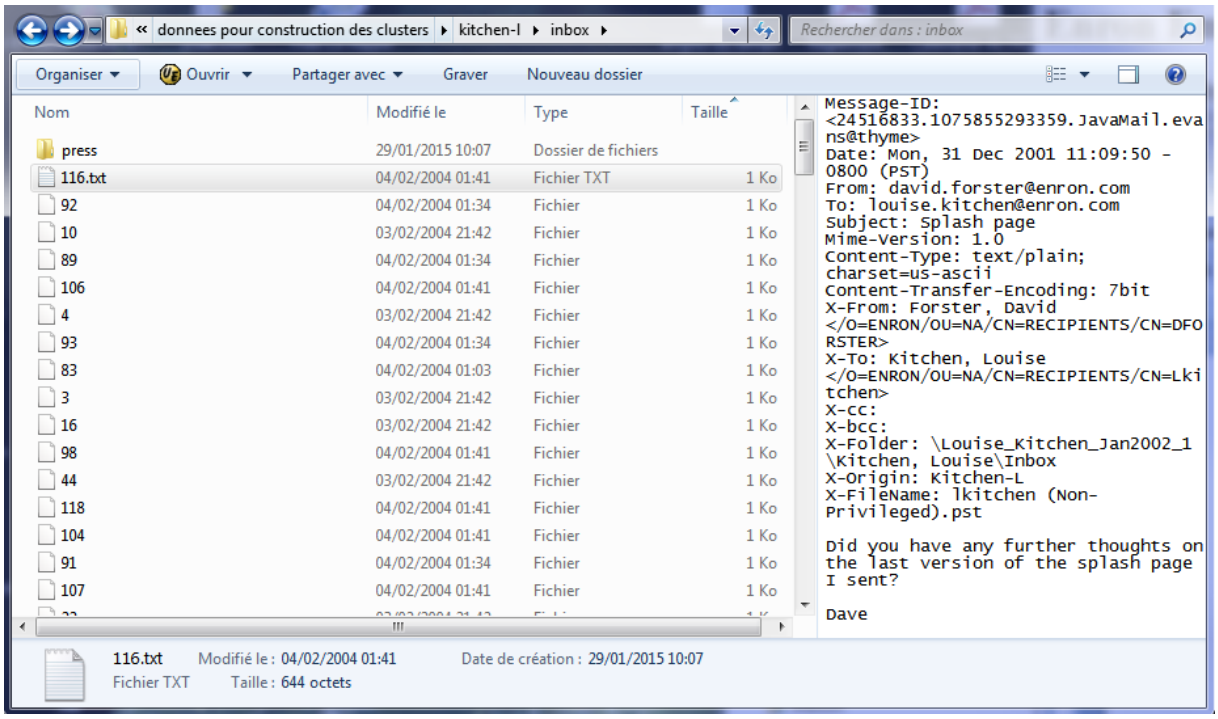
La base de données « Enron Email Dataset »[1] recense les messages de l'encadrement senior de la société Enron[2]. Nous étudions les e-mails de 7 collaborateurs dans notre projet (beck-s, farmer-d, kaminski-v, kitchen-l, lokay-m, sanders-r, williams-w3). Les messages sont de natures diverses. Les utilisateurs les ont classés eux-mêmes dans des dossiers selon la perception qu'ils en ont. Voici par exemple les dossiers de « kitchen-l » :

Comme on peut le voir, « kitchen-l » ne s'est pas vraiment foulé. D'autres en revanche sont particulièrement méticuleux : « beck-s » a ainsi réparti ses messages sur 114 dossiers.

Dans chaque dossier, un message correspond à un fichier au format texte. Voici le contenu du message n°116 dans le dossier « inbox » de « kitchen-l ».

[1] https://www.cs.cmu.edu/~./enron/

[2] http://fr.wikipedia.org/wiki/Enron



Objectif du projet

L'objectif du projet est de construire une typologie des e-mails reçus en s'appuyant sur les messages de ces collaborateurs d'Enron. Toute votre étude doit pouvoir être retracée dans un programme R que vous remettez avec votre rapport au format PDF.

Remarque : Les données « Enron Email Dataset » sont publiques. D'autres chercheurs ont travaillé dessus. Vous pouvez vous appuyer sur leurs publications pour faire avancer votre travail. Il faut simplement les citer clairement et sans ambiguïté dans votre rapport.

a. Construction de la typologie

A partir des données disponibles, construire une typologie des e-mails qui vous paraît la plus adaptée. Le nombre de groupes (clusters) est laissé à votre appréciation. Vous devez justifier votre choix dans votre rapport.

Les groupes doivent être interprétés c.-à-d. nous devons comprendre clairement le pourquoi de chaque groupe (ex. les termes – au sens text mining - associés, etc.) et ce qui les différencie.

A priori (c'est à vous de voir), il n'est pas utile de tenir compte de la pré-classification en dossiers effectuée par les utilisateurs. Il n'est peut-être pas nécessaire non plus (à vous de voir également) de tenir compte de tous les messages dans la construction de la typologie (ex. est-ce que les messages envoyés sont pertinents pour réaliser une classification des messages reçus ? peut-être qu'une stratégie d'échantillonnage est judicieuse pour dépasser les problèmes de volumétrie ?).

b. Déploiement sur de nouveaux messages

Votre programme doit comporter une partie « déploiement » c.-à-d. il doit être possible de charger un message contenu dans un fichier texte et de l'associer à un des clusters existants. Le détail de la

règle d'affectation doit être affiché dans la sortie écran. Cela peut être le degré d'appartenance à chaque cluster, cela peut être également les termes clés qui ont permis le rattachement du message au cluster. L'utilisateur doit en tout cas disposer de toutes les informations utiles lui permettant de comprendre la raison de cette affectation.

La méthode d'affectation aux clusters que vous avez adoptée doit être détaillée dans le rapport.

Votre programme R doit être suffisamment souple pour que le correcteur puisse le lancer sur un fichier-message quelconque. Documentez votre code à cet effet.

A rendre

Vous devez rendre dans un fichier archive à votre NOM et PRENOM (ex. benzema_karim.zip). Il comporte au moins :

Un document PDF retraçant les différentes étapes de votre travail.

Autant que possible, argumentez votre démarche. Si vous avez tenté plusieurs approches (ou plusieurs types de descripteurs et de pondération) avant de fixer votre choix définitif, retracez - même de manière succincte - les différents essais et ce qui a motivé leur abandon.

On s'attend à trouver dans votre rapport des descriptions détaillées et argumentées concernant : (1) la préparation des fichiers et le processus de nettoyage des données que vous avez effectués (ex. les messages qui n'ont pas été pris en compte ; les parties des messages qui ne sont pas pris en compte, etc.) ; (2) le système de pondération adopté et éventuellement la sélection des termes réalisée ; (3) la méthode de classification automatique (clustering, topic model…) mise en œuvre pour élaborer les groupes ; (4) la description des groupes ; (5) le mode d'affectation aux groupes des nouveaux messages.

Le programme R utilisé pour la réalisation du projet, incluant la partie déploiement.

Il doit être suffisamment clair, bien architecturé, et abondamment commenté. Le correcteur doit pouvoir retracer les résultats annoncés dans le rapport. La partie déploiement doit être opérationnelle.