Intégration de données connectées

PHIK

Deleglise Hugo
Des Courières Kevin
Jendari Ikram
Tith Ponnaka

Sommaire

Introduction

Lors de notre entrée dans la filière MIASHS, nous avons pris conscience d'un fait. Cette même chose qui, depuis plusieurs années, provoque l'ouverture de filières MIASHS partout en France et dans le monde : Nous entrons dans l'ère du numérique.

Dans tous les secteurs, les données – autrefois physiques – sont transformées en données numériques et des flots de données toujours plus importants demandent à être traités chaque jour. Pour répondre à ce besoin dont l'amplitude augmente de manière exponentielle, des individus doivent être formés pour savoir traiter ces données de façon rapide, précise et efficace.

Ces données peuvent être de tout type : photos, vidéos, coordonnées GPS, métadonnées, ou encore du texte.

Ce sont les données textuelles qui nous intéressent ici. Comme pour les autres formats de données, le besoin en termes d'analyses textuelles se fait de plus en plus pressant. Que ce soit les publicitaires, les entreprises voulant évaluer la réception d'un produit ou encore au niveau de la recherche en psychologie/sociologie, il est important pour eux de pouvoir rapidement discerner un état d'esprit, l'/les émotion(s) dominante(s)dans des textes ou dans des messages. L'utilisation de logiciels pour traiter automatiquement ces grandes quantités de textes et avoir une idée des tendances, avis, retours d'un sujet semble alors nécessaire.

Hugo DELEGLISE et Ponnaka TITH avaient étudié cette question durant leur 3ème année de licence et avaient pu développer une version bêta d'un logiciel d'analyse d'émotions de textes.

Nous nous proposons cette année d'améliorer 3 aspects de ce logiciel :

- Appliquer ce que nous avons vu en cours d'intégration de données connectées, à savoir la récupération « propre » de données web sans stockage sur la machine.

- Développer un code entièrement écrit en Javascript.

- Parfaire les caractéristiques et fonctionnalités de notre logiciel en faisant un travail que nous n'avions pas pensé à faire l'année passée : se concerter avec les individus cibles de notre logiciel sur leurs besoins et leurs attentes.

Voici les parties du LEAN CANVAS qui devaient être remplies :

Segment de clientèle:

Pour des raisons personnelles (idéologiques) et pratiques (nous pouvons facilement entrer en contact avec des professeurs de Psychologie/sociologie) nous décidons de nous concentrer sur les besoins des:

- chercheurs en psychologie, sociologie
- associations

Problèmes :

- Difficultés à évaluer rapidement un ressenti moyen accompagnant un ensemble de messages
- Difficultés à évaluer un retour sur investissement (pour une association)

Solutions à la disposition des utilisateurs:

- Des logiciels assez superficiels n'évaluent que la polarité positive/négative d'un message.
- Des logiciels très complets destinés à des experts en info/stat (comme le ''SentiCompass'' de Wang & Al)

I. ITW's

Nous décidons d'interviewer principalement des chercheurs en Psychologie/Sociologie qui sont les cibles de notre programme.

Les questions seront en général les suivantes:

« *La capacité à évaluer rapidement une réponse émotionnelle relative à de grandes quantités de textes pose-t-elle problème ?*»
« *Un programme de traitement automatique de ces données serait-il utile ?*»
« *Quelles sources de données sont le plus souvent utilisées ?»*
« *Quelles émotions serait-il important de quantifier ?*»
« *Quelles fonctionnalités pourraient être judicieuses ?*»

ITW 1: *Pascale MOLINER, professeur en Psychologie sociale à l'université Montpellier 3.*

L'interview se fait par téléphone.

Il ressort de cette interview que ce type de programme lui serait utile, aussi bien pour traiter automatiquement de longues séries d'entretiens que des données de forums. Voici les principales idées qui ressortent de cet entretien :

- Les modèles émotionnels utilisables sont souvent constitués de plus d'émotions négatives que d'émotions positives, il est donc important d'avoir une information qui ne porte que sur l'aspect positif/négatif d'un mot.

- Il serait également judicieux de rendre dynamique le logiciel émotionnel, de sorte qu'un chercheur puisse y rajouter des mots en fonction de ses recherches.

- Créer une fonctionnalité permettant l'analyse textuelle phrase par phrase afin de voir l'évolution de longs messages.

- Relier notre programme au logiciel R pour qu'il soit possible d'effectuer des analyses plus poussées.

ITW 2:

Madame BLANC, professeure en Psychologie du développement à l'université Montpellier 3.

L'interview se fait dans son bureau à l'UM3.

Nos concepts et idées lui ont plu, en particulier le fait de pouvoir voir l'évolution d'une émotion dans le temps. Elle nous a présenté le logiciel d'analyse d'émotions qu'elle utilise pour ses recherches qui se nomme « emotex», mais celui-ci ne traite que l'aspect positif/négatif d'un mot et non les émotions à proprement parler comme notre logiciel. Elle nous a ensuite fait savoir que pour ses recherches, les sources à traiter sont en général des textes stockés en brut sur son ordinateur comme des extraits de livres ou bien des articles de presse.

L'utilisation de l'outil serait encore plus intéressante si les utilisateurs pouvaient utiliser une ontologie autre que celle fournie, la leur par exemple.

Elle nous a enfin dit qu'elle nous mettrait en relation avec une étudiante en thèse qui pourrait avoir besoin de cet outil.

ITW 3:

Monsieur ESTELLON, professeur de Psychanalyse à l'Université Montpellier 3.

Il a été croisé dans la rue sur la place Albert 1er, l'interview se fait sur place.

Cette discussion ne nous indiquera pas de nouvelles fonctionnalités, mais nous donnera surtout des idées sur les limites de notre programme. Il nous fait savoir que ce type d'outils ne lui serait pas utile pour soigner ses patients. En effet, les statistiques donnent une idée de tendances globales, mais chaque individu est unique, possède son propre dictionnaire mental et doit donc être écouté indépendamment des autres. C'est d'ailleurs ce sujet qui oppose depuis toujours les psychanalystes et les cliniciens quantitativistes. D'après lui, chaque discours peut s'inscrire dans une complexité que ce type de logiciel ne peut élucider : Nous pouvons pleurer de joie, rire de

nervosité et il arrive qu'une détresse extrême se cache dans un discours joyeux aux premiers abords … Nous devons donc avoir conscience que ce programme doit être utilisé avec précaution dans certains cas.

ITW 4 :

Madame Syssau-Vaccarella, professeure de Psychologie cognitive à l'Université Montpellier 3, spécialiste des émotions.

L'interview se fait dans son bureau à l'UM3.

Lors de cet entretien – très fructueux–, c'est surtout la partie dictionnaire qui sera abordée. En effet, madame Syssau-Vaccarella a elle-même participé à la création de dictionnaires émotionnels, elle est donc au courant de ce qui se fait en la matière. Après nous avoir écoutés sur le fonctionnement et les composants de notre logiciel, elle nous a donné plusieurs conseils d'optimisation et d'adéquation aux besoins des chercheurs:

- Notre dictionnaire caractérise un mot par 6 émotions en plus de sa valence positive ou négative. Cela est un plus car ce qui se fait dans le domaine se limite généralement à la valence. Cependant, il serait encore plus novateur et utile à leur profession d'étendre le dictionnaire à une palette plus large d'émotions pour gagner en précision dans l'analyse automatique de textes. Par exemple, notre modèle ne contient que « joie » comme émotion positive, mais cette joie est-elle source d'extase, d'admiration, de confiance, d'amour ou encore de sérénité ?

- Le fait que l'émotion à laquelle renvoie un mot soit codée en 0/1 (présence/absence) est un peu simpliste, il serait bienvenu de trouver un moyen pour nuancer cela.

- Il faudrait étendre notre dictionnaire à d'avantage de mots et savoir traiter les adverbes en particulier. Les adverbes quantifient la puissance d'un verbe ou d'un mot situé plus loin dans la phrase. Exemple : « Je suis très déçu ». Ce traitement des adverbes n'a jamais été fait et serait un vrai plus d'après elle.

- Le traitement des émoticônes est important, en particulier si les textes proviennent de réseaux sociaux. Elle nous a conseillé de nous adresser à Rachel Panckhust, une chercheuse en linguistique qui a travaillé sur la signification des émoticônes

- Pouvoir utiliser leur propre ontologie émotionnelle (comme l'a fait remarque Madame Blanc)

Pour améliorer les points précédents, elle nous donne quelques pistes :

- Nous devrions nous intéresser à d'autres modèles émotionnels, en particulier celui de Russell.

- Pour connaître la représentation mentale moyenne d'un mot dans une population, une technique consiste à questionner directement la population. Mais cela doit être fait selon des critères précis pour être valable. Elle nous en donnera les techniques plus tard si besoin.

- Il serait intéressant de nous intéresser à l'analyse sémantique latente qui établit des relations entre un ensemble de textes et les termes qu'ils contiennent, dans le but d'étudier la co-ocurrence entre les mots.

- Elle nous a donné un dictionnaire émotionnel dont elle est l'auteure, entièrement crée par des analyses de populations. Il peut être intéressant de comparer nos deux dictionnaires car les méthodes de créations utilisées sont complémentaires.

II. Logiciel

(a) Technologies

- Code en JavaScript
- Correction orthographique avec le dictionnaire « dictionnaire »
- Lemmatisation programmée par nos soins
- Vectorisation des mots lemmatisés avec le dictionnaire émotionnel « feel »
- Interface en HTML et CSS

(b) Fonctionnalités

Dans un premier temps, nous indiquons au logiciel le pseudo dont nous souhaitons traiter les tweets. A noter que le logiciel fonctionne seulement sur les tweets en français.

Les tweets récupérés seront caractérisés selon 6 émotions qui sont la joie, la colère, la tristesse, le dégoût, la surprise et la peur. Nous aurons également une valence moyenne (positif/négatif) pour chacun d'eux.

Pour chacune des émotions une moyenne globale des tweets sera calculée et une liste déroulante nous permettra de voir tous les tweets en détail ainsi que le pseudo de l'auteur et des 6 émotions.

Conclusion

Malgré le temps que ce sujet demandait et le nombre de difficultés qui se sont posées, notre logiciel donne des résultats exploitables, ce qui est satisfaisant en soi. En effet, avec plus ou moins d'efficacité, toutes nos boites sont fonctionnelles, ce qui nous permet d'effectuer l'action prévue : quantifier les émotions de tweets.

En revanche, si nous avions eu plus de temps, nous aurions pu intégrer :

- Un graphique pour avoir quelque chose de visuel et plus rapide à analyser
- Des statistiques plus complexes
- Traiter les hashtags
- Optimiser le traitement du langage
- Rendre le dictionnaire dynamique