INSTITUT FRANCOPHONIE INTERNATIONAL (IFI)

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

TRAVAIL PERSONNEL ENCADRE (TPE)

Resume automatique de texte

RAPPORT DE SOLUTION PROPOSES

Rédigé par :

Hugues Kanda

Encadrant

Ho Tuong Vinh

PROMOTION 22

Année Académique 2018– 2019

I. INTRODUCTION

Mon travail est basé sur le problème pratique de résumer automatique de texte oriente vers la prévention, l'éradication et détection des maladies ravageuses déclenchées sur les réseaux sociaux. De jour en jour, la masse d'informations textuelles sous forme électronique ne cesse d'augmenter, que ce soit sous forme de documents accessibles sur internet, dans les bases de données des entreprises et des gouvernements. Il devient de plus en pins difficile d'accéder aux informations intéressantes sans l'aide d'outils spécifiques.

Dans ce contexte, il est nécessaire de pouvoir accéder au contenu des textes par des moyens rapides et efficaces. C'est la fonction d'un résumé qui constitue un moyen efficace et éprouvé pour représenter le contenu des textes, et ainsi permettre un accès rapide à leur contenu.

La technique du résumé automatique qui permet de choisir un ou plusieurs documents utilement est une réponse à ce besoin.

II. SOLUTION PROPOSEE

Le résumé automatique de textes est une discipline du traitement automatique de langues (TAL) qui a pour objectif de compresser les documents textuels.

Depuis lors, de nombreux travaux ont été publiés pour traiter le problème de la résume automatique de texte.

En général, il existe deux méthodes différentes pour le résumé automatique: **l'extraction et l'abstraction**.

La méthode de résumé de texte extractif fonctionne en sélectionnant un sous-ensemble de mots, phrases ou phrases existantes du texte original pour former un résumée, elle utilise une approche statistique pour sélectionner des phrases ou des mots-clés importants à partir du document. Les phrases extraites tendent à être plus longues que la moyenne. Les informations conflictuelles peuvent ne pas être présentées avec précision.

La méthode de résumé de texte abstrait génère une phrase à partir d'une représentation sémantique, puis utilise des techniques de génération de langages naturels pour créer un résumé plus proche de ce qu'un humain pourrait générer. Un tel résumé pourrait contenir des mots qui ne sont pas explicitement présents dans l'original. Il consiste à comprendre le texte original et à le répéter en moins de mots. Il utilise une approche linguistique pour comprendre le texte original, puis génère un résumé.

Dans le cas de notre travail nous allons nous concentrai en utiliser la méthode de résumé de texte extractif pour les problèmes suivants : actuellement les réseaux sociaux (Facebook, Twitter, etc.) sont devenus les sources génératrices des données et examinera principalement les approches concernant l'extraction de phrases, la synthèse de domaines spécifiques et les méthodes de synthèse multi-documents.

Dans notre travail, nous comptons exploiter ces sources afin de détecter et résumer de manière automatique tous les cas liés aux maladies qui font rage en Afrique à savoir : Ebola, méningite, malaria, et d'autres pestes et fléaux.
Notre solution consiste à implémenter la méthode extractif fonctionne en sélectionnant un sous-ensemble de mots, phrases ou phrases existantes du texte original pour former un résumée et construire un prototype pour le résumé de texte.

Par ce fait, nous aimerions performer l'analyse textuelle basée sur l'usage d'apprentissage automatique afin d'alerter les gouvernements, institutions ou organisations qui s'en occupent dans le souci de prendre des promptes décisions qui vont impacter positivement les mesures préventives, mobilisation de personnel, de corps médical, de ressource humains, et le déploiement le personnel dans le pays affecte.

III. **OUTILS A UTILISER**

Pour s'initier au résume automatique de texte, il existe des outils et aides pratiques en ligne. Mais quel est l'outil le plus adapté pour mon travail? Cela va dépendre de la langue et de la méthode TALN que nous souhaitons utiliser. Parmi les outils open-source les plus connus.

Concernant la plateforme technique pour la mise en œuvre des fonctionnalités, nous avons utilisé comme :

- Logiciels : RStudio

  RStudio est un environnement de développement intégré (IDE) pour R. Il comprend une console, un éditeur de mise en évidence syntaxique qui prend en charge l'exécution directe de code, ainsi que des outils de traçage, d'historique, de débogage et de gestion d'espace de travail.

  RStudio est disponible en versions open source et commerciales et fonctionne sur le bureau (Windows, Mac et Linux) ou dans un navigateur connecté à RStudio Server ou RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS et SUSE Linux).

  Algorithm : l'allocation de Dirichlet latente (LDA), Traitement de language naturel (OpenNLP), Tm, Monkylearn, Google Machine Learning API, Tidytextmining.

  L'allocation de Dirichlet latente (LDA) est un modèle statistique génératif qui permet à des ensembles d'observations d'être expliqués par des groupes non observés qui expliquent pourquoi certaines parties des données sont similaires.

openTraitement de language naturel est une collection d'outils de traitement du langage naturel comprenant un détecteur de phrases, un tokenizer, un pos-tagger, un analyseur syntaxique superficiel et complet et un détecteur d'entité nommée, utilisant le package Java Maxent pour former et utiliser des modèles d'entropie maximum.

Tm fournit un cadre d'exploration de texte complet pour R, donne un aperçu détaillé et présente des techniques pour les méthodes d'analyse à base de comptage, le regroupement de texte, la classification texte et les noyaux de chaîne.

Monkylearn fournit une interface wrapper aux services d'apprentissage automatique pour l'analyse de texte, c'est-à-dire, la classification et l'extraction.
Google Machine Learning API
Tidytextmining fournit des moyens d'extraction de texte pour le traitement de texte et l'analyse des sentiments à l'aide de dplyr, ggplot2 et d'autres outils propres.

IV. Présentation du chronogramme

Le travail réalisé s'est déroulé en plusieurs étapes, en fonction des dates, nous les avons évalués en nombre de jours. A ces étapes correspondent des tâches, portant chacune un numéro de précision. Le tableau ci-dessous présente un état caractéristique du déroulement de notre travail.

N^0	Nom de la tâche	Début	Fin
1	Remise et recueil d'information Relatifs au thème	2/juin/2018	9/juin/2018
2	Installation de logiciel RStudio sous Windows et Linux	10/juin/2018	17/juin/2018
3	Apprentissage de manipulation de logiciel	18/juin/2018	25/juin/2018
4	Installation de package et Library	26/juin/2018	2/juillet/2018
5	Créations des AP key Facebook et Twitter	3/juillet/2018	9/juillet/2018
6	Connection de l'ID et key secret de Facebook, Twitter	10/juillet/2018	17/juillet/2018

	dans le logiciel RStudio		
7	Collection de donne Facebook et Twitter	18/juillet/2018	24/juillet/2018
8	Exploitation des données	25/juillet/2018	31/juillet/2018
9	Rassemblement des outils et début réalisation de prototype, Début de Rapport de rédaction	1/aout/2018	7/aout/2018
10	Test des algorithmes et correction des erreurs	8/aout/2018	22/aout/2018
11	Fin de rédaction de rapport et validation du prototype	23/aout/2018	18/septembre/2018

V. Conclusion

Nous avons présenté le problème actuel en science, celui de se tenir à jour, malgré une quantité grandissante de publications. Il existe plusieurs techniques de résumé automatique de texte, tel que présenté dans ce travail. Nous avons énuméré des techniques pour identifier le type de discours des phrases de la prévention, l'éradication, détection et comment construire des résumés extractifs. Aussi, nous avons vus des techniques de résumé multiples et d'extraction de méta-information. Certains travaux de résume automatique demandent un traitement spécial : pouvoir mesurer leurs impacts face au résume automatique de texte.

Notre participation à ces travaux nous a permis de développer des techniques intéressantes pour extraire de l'information à extraire dans les réseaux sociaux et prédire la prévention, l'éradication et détection, afin d'alerter les gouvernements, institutions ou organisations qui s'en occupent dans le souci de prendre des promptes décisions qui vont impacter positivement les mesures préventives, mobilisation de personnel, de corps médical, de ressource humains, et le déploiement le personnel dans le pays affecte. Nous croyons pourvoir ajouter aux techniques existantes lors de notre maitrise.

Référence :

https://www.rstudio.com/products/rpackages/
https://cran.r-project.org/web/packages/available_packages_by_name.html
https://crunch.kmi.open.ac.uk/w/index.php/Tutorials
http://jamaity.org/publication/modele-de-chronogramme/
https://www.slideshare.net/jamaity_tn/grdr-outil-planification-exemple-de-chronogramme
http://www.institut-numerique.org/9-chronogramme-du-travail-de-these-523964956abb5