REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

RAPPORT DU PROJET DE FIN D'ÉTUDES

Pour obtenir le diplôme de

Post-Graduation Spécialisé en Big Data et Calcul Intensif

Préparé par :

Abdelkrim HAMANI, Hamza KERMICHE

Sous la direction de :

Dr. Ahcen BENDJOUDI (Encadreur)
Dr. Abdelbaset KABOU (Co-encadreur)

Analyse des Données des Médias Sociaux avec les Technologies Big Data

Soutenue le: 25/07/2019

Devant le jury composé de :

Dr. **Omar NOUALI**	CERIST	Président
Mr. **Nadir BOUCHAMA**	CERIST	Examinateur
Mlle. **Sera BOUHENNI**	ESI	Examinateur
Dr. **Abdelbaset KABOU**	CERIST	Encadrant

Année Universitaire : **2017 - 2018**

Analyse des données des médias sociaux avec les technologies Big Data

Abdelkrim HAMANI, Hamza KERMICHE

Mémoire soumis en vue de l'obtention du diplôme de
Post-Graduation Spécialisé

Alger

Juin 2019

ملخص

أصبح تحليل بيانات الوسائط الاجتماعية، وخاصة تويتر، مجالا للدراسة والنشاط التجاري ذو الأهمية المتزايدة وذلك في الكثير من المجالات: السياسية، الاجتماعية، الامنية، التجارية، الخ. في هذا السياق، يمثل تحليل المشاعر مجالا بحثيا واعداً للمؤسسات المهتمة بآراء الأشخاص المعبر عنها في تويتر. موضوع هذا المشروع يندرج ضمن هذا السياق. الهدف يتمثل في توفير إطار عام يتيح تحديد الرأي العام للمستخدمين بوضوح فيما يتعلق بموضوع معين، وذلك باستخدام تقنيات البيانات الكبيرة (Big Data)، مطبقة على مجالات مختلفة كالأمن، السياحة، الصحة العامة، اوالتجارة، الخ. تم ادراج حالة استخدام بشكل مفصل اين تم استخدام ميزات مكتبية (StandfordNLP) المصممة لمعالجة اللغة الطبيعية (NLP) بعد جمع البيانات من خلال واجهة تويتر لبرمجة التطبيقات (TwitterAPI). أظهرت النتائج أن وسائل التواصل الاجتماعي بشكل عام ، وتويتر على وجه الخصوص ، يمكن أن تمثل مصدر لبيانات جِد مفيدة تساعد في الحصول على مؤشرات توجيهية عامة، بالغة الاهمية لاتخاذ قرارات استراتيجية أفضل.

الكلمات المفتاحية: البيانات الكبيرة، الشبكات الاجتماعية، تويتر، تحليل المشاعر

Abstract

The analysis of social media data, especially Twitter, has become a field of study and a business activity of growing importance. It can derive value related to the different domains, political, social, security, commercial, etc. In this context, sentiment analysis is still a promising research discipline for organizations interested in the opinions of people expressed on Twitter. The subject of this thesis is in line with the previous topic. It consists in providing a general framework making it possible to clearly identify the general opinion of the users with regard to a given subject, using the Big Data techniques, applied to the different domains eg, Security, Tourism, Public Health, Trade, etc. A use case is well detailed where after collecting the data through the Twitter API, the analysis is done using the features of the StanfordCoreNLP library, designed for natural language processing (NLP). The results show that social media in general, and Twitter in particular, can effectively serve as a source of extremly usefull data to obtain key indicators for better decision-making or strategy guidance.

Keywords : Big data, Social networks, Twitter, Sentiment Analysis

Résumé

L'analyse des données des médias sociaux, en particulier les données de Twitter, est devenue un domaine d'étude et une activité commerciale d'une importance grandissante. Elle permet de tirer de la valeur liée au différents domaines, politique, social, sécuritaire, commercial, etc. Dans ce contexte, l'analyse des sentiments est un domaine qui montre un intérêt grandissant dans nos sociétés car plusieurs organismes dans différents domaines s'interessent de plus en plus aux opinions des gens exprimés sur Twitter. Le sujet de ce mémoire s'inscrit dans cet thématique. Il consiste à proposer un framework générale permetant d'identifier clairement l'opinion générale des utilisateurs par rapport à un sujet donné et ce en utilisant les techniques du Big Data, appliquées aux différents domaines e.g, Sécurité, Tourisme, Santé Publique, Commerce, etc. Un cas d'utilisation a été bien détaillé où après avoir collecté les données grâce à l'API Twitter, l'analyse a été faite en utilisant les fonctionnalités de la bibliothèque StanfordCoreNLP, conçue pour traitement du langage naturelle. Les résultats obtenus démontrent que les médias sociaux en général et Twitter en particulier, peuvent effectivement servir de source de données permettant d'obtenir des indicateurs clés pour une meilleure prise de décision ou orientation de stratégies.

Mot Cléfs : Big data, Réseaux sociaux, Twitter, Analyse de Sentiment

Table des matières

Table des matières vi

Table des figures ix

Liste des tableaux x

Introduction 11

I Background 13

1 Big Data 14

1.1 Notions de base et caractéristiques des Big Data 14

1.1.1 Introduction 14

1.1.2 Disciplines participant au Big Data 16

1.1.3 Métiers du Big Data 17

1.1.4 Sources des Big Data 18

1.1.5 Big Data analytics : concept et outils 18

1.1.5.1 Apache Hadoop 19

1.1.5.2 Apache Spark 20

1.1.5.3 Apache Storm 22

1.1.5.4 Apache Flink 22

1.1.5.5 Apache Samza 23

2 Réseaux Sociaux 25

2.1 Définition et Notions de base 25

2.2 Twitter 26

2.2.1 Contenu des tweets 26

2.2.2 Metadonnées des Tweets et des utilisateurs 27

2.2.2.1 Metadonnées des Tweet . . . 27
2.2.2.2 Metadonnées de l'utilisateur . . . 28
2.3 L'API de Twitter . . . 29
2.3.1 Les classes d'API Twitter . . . 30

II Etat de L'art 32

3 Twitter Big Data analytics 33
3.1 Introduction . . . 33
3.2 Méthodes et types d'analyse des tweets . . . 33
3.2.1 Analyse Thématique . . . 34
3.2.2 L'analyse des sentiments . . . 34
3.3 Approches d'analyse de sentiments . . . 35
3.3.1 Approches basées sur le lexique . . . 36
3.3.2 Approches basées sur le Machine Learning . . . 37
3.4 Applications . . . 40
3.4.1 Domaine de la santé . . . 40
3.4.2 Questions politiques . . . 41
3.4.3 Phénoménes sociaux . . . 42
3.4.3.1 le harcèlement . . . 42
3.4.3.2 Les communautés de gangs et leurs membres . . . 42
3.4.4 Trafic routier . . . 42
3.4.5 Prévision d'anomalie et de contenu malveillants . . . 43
3.4.6 Prévision de la popularité . . . 44

III Proposition 46

4 Cas d'utilisation : Analyse de sentiment sur Twitter 47
4.1 Introduction . . . 47
4.2 Choix des outils . . . 48
4.2.1 La collecte des tweets . . . 49
4.2.2 L'analyse des sentiments . . . 49
4.2.3 La visualisation . . . 50
4.3 Architecture . . . 50
4.3.1 Spark Context . . . 51
4.4 Résultats et discussion . . . 55

Conclusion 56

References 58

Annexe 62