Master Informatique
Parcours Données et Connaissances
Projet Inter-Promo 2019 : Le Campus du Futur

Contents

Tâche 3 : Analyse de satisfaction des utilisateurs sur les technologies IoT

1 Missions de la Tâche 3

Cette tâche a pour objectif de mesurer le degré de satisfaction des utilisateurs sur les technologies IoT en générale (Internet des objets, villes intelligentes, bâtiments intelligents, économie d'énergie, etc.). Nous nous focalisons sur des avis écrits provenant de trois sources différentes :

- des tweets d'utilisateurs, extraits de Twitter,
- des articles de presses (LeMonde, NewYork Times, etc.).
- des utilisateurs (personnels et étudiants) de l'université Paul Sabatier qui ont exprimé leurs avis via un sondage en ligne.

A partir de ces sources, nous nous intéressons aux questions suivantes :

- Les utilisateurs sont-ils pour, contre, ou indifférents à l'usage de ces technologies dans la vie courante.
- Quelles sont les émotions exprimées (peur, optimisme, surprise, joie, etc.) ?
- Quelles sont les thèmes les plus fréquemment discutés autour des IoT.
- Ces avis/émotions/thèmes sont-ils dépendants du genre du corpus (tweets, presse, sondage), de la localisation géographie (France, USA, etc.), du genre de l'utilisateur, de la langue du corpus (français, anglais), ou encore de phénomènes/événements externes (comme la météo du jour ou la survenu de catastrophes naturelles).

1.1 Objectifs

Pour répondre aux questions précédentes, quatres tâches sont proposées :

- Tâche 3.1 ([$M1\_4$]) : Analyse exploratoire des conversations sur Twitter sur l'usage des technologies IoT
- Tâche 3.2 ([$M1\_5$]) : Génération de résumé rétrospectif des commentaires des utilisateurs sur Twitter sur l'usage des IoT.
- Tâche 3.3 ([$M2\_3$]) : Moteur d'accès personnalisé aux tweets sur les IoT.
- Tâche 3.4 ([$M2\_4$]) : Analyse multi-genres et multi-lingues du degré de satisfaction des utilisateurs sur les technologies IoT.

1.2 Données à utiliser

Les groupes auront à disposition les corpus suivants, disponibles en téléchargement FTP depuis des liens fournis sur la page web du projet :

- Un corpus de tweets en langue anglaise relatifs au #Iot, #InternetOfThing (fichier IoT-Tweet.csv). Pour chaque tweet, nous avons : son id, le sentiment/opinion qu'il véhicule (positif/négatif/neutre), le pays d'origine du tweet et le genre de l'utilisateur qui a posté le tweet (femme/homme).

- Un corpus de tweets relatifs aux économies d'énergie et aux changements climatiques annotés en émotions (peur, joie, ennui, colère, etc.) (fichier EmotionDeft2015.zip). Cette archive contient des fichiers au format arff[1].
- Un corpus de tweets en langue anglaise annoté en émotions (fichier Emotion-Tweets.zip)
- Un corpus de titres d'articles de presse en anglais annoté en émotions (fichier Emotion-News.zip)
- Des fichiers csv issus d'un sondage LimeSurvey où les personnels et étudiants de l'UPS ont donné leurs avis sur les IoT (ces fichiers seront à disposition des étudiants fin janvier, une fois le sondage clôturé).

1.3 Dépendances entre les groupes

- Mutualisations groupe (M1 Tâche 3.1) - groupe (M1-Tâche 3.2) : comme indiqué dans la partie descriptive des tâches, chacune des étapes de collecte du corpus (Etape 1) et d'indexation (Etape 2) doit être réalisée de façon mutualisée selon une organisation à définir entre les groupes. En ce sens, les réalisations de chaque groupe doivent être complémentaires en vue d'atteindre l'objectif commun de chaque étape.
- Interactions groupes (M1 Tâche 3.1, Tâche 3.2)- groupe (M2, Tâche 3.3 et Tâche 3.4). Les deux groupes de M1 fournissent au groupe M2 : a) le corpus de tweets complété; a) l'index structuré et non structuré de la collection. Les groupes M1-M2 s'entendent sur les formats d'échange. Les autres interactions M1-M2 sont indiquées dans le descriptif des tâches : a) thèmes LDA en interactions avec le groupe 3.1; b) les tweets pertinents à une requête en interaction avec le groupe 3.2.
- Dans la tâche 2 : Réflexion à mener avec les groupes de la tâche 2 pour prendre en compte le feedback (degré de satisfaction) des utilisateurs dans le processus d'optimisation du confort.

2 Tâche 3.1 : Analyse exploratoire des conversations sur Twitter sur l'usage des technologies IoT

Groupe de travail concerné : [$M1\_4$].

Enseignant référent pour cette tâche : Lynda Tamine-Lechani

Mots-clés : Indexation de textes, analyse de graphes, analyse thématique LDA, statistique descriptive.

Sujet adossé au TIR (S8)

L'objectif de cette tâche est de produire un tableau de bord qui donne un aperçu des tendances émergeant des commentaires des utilisateurs sur les technologies IoT, exprimés sur Twitter. Ce tableau de bord permet de répondre principalement aux questions suivantes :

- Quelles sont précisément les différents sujets de conversation en lien avec les IoT ?

[1] https://waikato.github.io/weka-wiki/arff_stable/

- Existe-t-il des groupes d'utilisateurs particulièrement actifs sur ces sujets ?
- Quelle est la répartition géographique des utilisateurs échangeant sur des sujets IoT ?
- Comment évolue ces sujets au cours du temps en termes d'utilisateurs actifs, d'opinions ?
- Existe-il des relations entre l'intensité/avis exprimés lors de ces échanges et d'autres événements (eg., Rassemblement Cop21, périodes d'élections etc.)

La liste des questions est donnée à titre indicatif et il est même souhaitable que le groupe soit force de propositions. Voici les principales étapes de réalisation :

- Etape 1 : ***Constitution du corpus de Tweets***. **Etape mutualisée avec la Tâche 3.2**. Cette étape consiste à alimenter le corpus fourni avec les textes des tweets en utilisant le hastag # IoT avec l'API Twitter. Vous devriez collecter/filtrer les tweets, retweet, rely, etc. durant la période de référence.
- Etape 2 : ***Indexation du corpus de Tweets***. **Etape mutualisée avec la Tâche 3.2**. Cette étape consiste à générer à partir des fichiers json les index sur : a) le contenu structuré : hashtag, indication (tweet, retweet, reply), date, localisation, utilisateur, avis, etc. ; b) contenu non structuré : texte du tweet (mots).
- Etape 3 : ***Analyse thématique des sujets de conversation***. Cette étape consiste à : a) construire le graphe social qui traduit les fils de discussion (via les tweets, retweets, reply) selon un algorithme adapté (**articles fournis en lien avec le projet TIR**); b) construire des composantes connexes du graphe; c) pour chaque composante, identifier les sujets en utilisant une analyse thématique basée sur le LDA; d) donner une étiquette sur chaque sujet généré par le LDA en faisant une annotation manuelle validée par des accords inter-annotateurs.
- Etape 4 : ***Analyse statistique descriptive des conversations***. Cette étape consiste à générer des tableaux et graphiques qui donnent un aperçu visuel des caractéristiques du corpus de tweets. Des exemples d'analyses : répartition des sujets selon le nombre d'utilisateurs, la répartition géographique, polarité des avis (positifs, négatifs), hashtags; évolution temporelle des échanges (nb users, nb reply, etc.); corrélation des avis avec des événements extérieurs, caractérisation des groupes d'utilisateurs/sujets, etc. (à enrichir).

3 Tâche 3.2 : Génération de résumé rétrospectif des commentaires des utilisateurs sur Twitter sur l'usage des IoT

Groupe de travail concerné : [$M1\_5$].

Enseignant référent pour cette tâche : Lynda Tamine-Lechani

Mots-clés : Indexation de textes, algorithmes de recherche basés graphes (eg., PageRank), algorithmes de diversification thématique et temporelle (eg., MMR).

Sujet adossé au TIR (S8)

L'objectif de cette tâche est de produire à partir d'une liste de mots clés qui décrit un sujet d'intérêt donné, un résumé rétrospectif thématique et temporel des tweets des utilisateurs sur les technologies IoT, exprimés dans le corpus. Deux méthodes de construction de résumés seront implémentées et évaluées. Ce résumé est extractif (comme vu en cours) est constitué d'une liste de tweets issus du corpus. Le résumé est sensé :

- Présenter les tweets les plus pertinents au sujet d'intérêt
- Être caractérisé par une diversité thématique (sujets ou sous-sujets différents, **interaction possible avec la Tâche 3.1**) en priorité, diversité des utilisateurs (homme/femme), diversité d'opinions (positif/négatif) en second lieu
- Représenter une frise chronologique, en ce sens que l'ordre temporel est à préserver d'un tweet au tweet suivant dans la liste

Voici les principales étapes de réalisation :

- Etape 1 : ***Constitution du corpus de Tweets***. **Etape mutualisée avec Tâche 3.1**. Cette étape consiste à alimenter le corpus fourni avec les textes des tweets en utilisant le hastag # IoT avec l'API Twitter. Vous devriez collecter/filtrer les tweets, retweet, reply, etc. durant la période de référence du corpus et en considérant les utilisateurs de référence connus dans le corpus.
- Etape 2 : ***Indexation du corpus de Tweets***. **Etape mutualisée avec Tâche 3.1**. Cette étape consiste à générer à partir des fichiers json les index sur : a) contenu structuré : hashtag, indication (tweet, retweet, reply), date, localisation, utilisateur, avis, etc. ; b) contenu non structuré : texte du tweet (mots).
- Etape 3 : ***Génération du résumé***. Cette étape consiste à appliquer deux types de méthodes de résumés et adaptées aux tweets(**articles fournis en lien avec le projet TIR**): a) une méthode vectorielle basée sur la caractéristiques de contenu (**en interaction avec le groupe M2**) et de temps et b) une méthode basée graphes qui est une variante du pageRank combinant le contenu et temps et c) des adaptations à ces algorithmes de votre part pour traiter la diversité des utilisateurs, des avis etc.
- Etape 4 : ***Evaluation de la qualité des résumés***. Cette étape consiste à évaluer expérimentalement les résumés produits par chacune des méthodes implémentées (**articles fournis en lien avec le projet TIR**). Pour cela, vous serez amenés à : a) constituer un petit corpus de résumés annoté; b) évaluer puis comparer les performances de chaque méthode de génération de résumés.

4 Tâche 3.3 : Moteur d'accès personnalisé aux tweets sur les IoT

Groupe de travail concerné : [$M2$_3]

Enseignant référent pour cette tâche : Lynda Tamine-Lechani

Mots-clés : Modèle utilisateur, modèles d'accès personnalisé, apprentissage automatique (classification supervisée, non supervisée, réseaux de neurones)

L'objectif de cette tâche est de réaliser un système d'accès aux tweets qui fonctionne aussi bien en mode pull qu'en mode push. Le mode pull consiste à produire à partir d'une liste de mots clés qui décrit une requête donnée : a) les tweets pertinents associés à cette requête (**en interaction avec le groupe M1 Tâche 3.2**) et adaptés au profil de l'utilisateur; b) les utilisateurs actifs/influenceurs sur ce sujet. Le mode push consiste à recommander à l'utilisateur (en lien avec son modèle/profil) : a) des tweets pertinents et b) des utilisateurs à suivre (to fellow). Le corpus

qui vous est fourni est annoté des informations liées au genre, localisation et avis général. Il est à rappeler que dans un contexte réel, ces données sont généralement absentes et donc à prédire.

Voici les principales étapes de réalisation :

- Etape 1 : ***Construction et validation des profils utilisateurs***. Cette étape consiste à modéliser, implémenter le profil multidimensionnel de chaque utilisateur puis évaluer sa qualité. Les dimensions du profils sont : a) la dimension thématique à construire partir de ses tweets. Plusieurs possibilités : utiliser des W2vec, Doc2vec, LDA (**interaction possible avec le groupe M1 Tâche 3.1 qui génère les sujets des conversations/tweets**); b) prédire le genre des utilisateurs, le sentiment/avis et la localisation. Pour cela utiliser au mieux : i) des classifieurs SVM eg., pour la détection du genre; des réseaux de neurones profonds pour la prédiction du sentiment/avis utilisant un lexique; eg., un classifieur bayésien ou modèle de langue pour la prédiction de la localisation. Toute autre solution, de préférence basée sur une méthode d'apprentissage automatique, est la bienvenue.

- Etape 2 : ***Réalisation du mode d'accès Pull***. Cette étape consiste à sélectionner : a) les tweets pertinents en lien avec le sujet et le profil utilisateurs. Privilégier l'utilisation de modèles d'accès basés sur l'apprentissage (learning to rank) basés sur SVM, rankboost par exemple en le combinant avec un appariement/filtrage avec le profil construit à l'étape 1; b) les influenceurs sur le sujets associés à la requête en utilisant une méthode basée graphes. Plusieurs possibilités pour la représentation des sujets de la requête : classification k-means de word embeddings des tweets résultats, sujets LDA fournis (**interactions avec le groupe M1 tâche 3.1**), projection sur une ontologie (eg., wikipédia, wordnet), à vos propositions.

- Etape 3 : ***Réalisation du mode d'accès Push***. Cette étape consiste à recommander a) des tweets pertinents pour l'utilisateur en accord avec son profil. Pour cela indiquer le principe de simulation des tweets pertinents (1/0) puis utiliser de préférence une méthode de recommandation de filtrage collaboratif basé sur un perceptron multi-couches; b) des utilisateurs à suivre en utilisant des informations issues du graphe social et des profils.

- Etape 4 : ***Evaluation des performances de modes d'accès pull et push***. Cette étape consiste à évaluer expérimentalement les performances des sorties de chaque mode de système au mieux, du mode d'accès push au minima. Pour cela, vous serez amenés à : a) construire une vérité terrain en indiquant les principes de simulation retenus; b) décrire et implémenter une méthodologie d'évaluation (apprentissage-test, cross-validation) en indiquant la méthode évaluée, les paramètres qui impacteraient les résultats ; c) évaluer les performances de chaque méthode par variation des paramètres retenus en b).

5 Tâche 3.4 : Analyse de satisfaction multi-genres et multi-lingues

Groupe de travail concerné : [$M2\_4$]

Enseignant référent pour cette tâche : Farah Benamara

Mots-clés : Analyse d'émotions, traduction automatique, reconnaissance d'entités nommées, web sémantique, apprentissage automatique

Cette tâche a pour objectif d'analyser les émotions exprimées par des utilisateurs sur les technologies IoT en général selon deux perspectives : (a) analyse multi-lingues sur des corpus en

anglais et en français, (b) analyse multi-genres sur des corpus de tweets, de news et de commentaires en ligne. Ces analyses permettront de capter les divers types d'émotions (comme la peur, l'étonnement, le pessimisme, l'optimisme, etc.) ressentis par les internautes sur l'usage des IoT dans la vie quotidienne.

Voici les principales étapes à réaliser :

- Etape 1 : ***Collecte et stockage du corpus***. **Etape mutualisée avec la Tâche 3.1 et 3.2**.
  - Corpus de tweets : Etendre le corpus en langue anglaise fournis en collectant des tweets relatifs à d'autres mots clés (ou hashtags) comme *smart buildings, smart cities, climate change, energy saving, global warming*, etc. **(interactions avec le groupe M1 tâche 3.1)**. Faire de même en collectant des tweets en languge française[2].
  - Corpus de news: A partir d'API dédiées, collecte de titres de presse relatifs aux IoT en anglais et en français[3].
  - Stockage des corpus collectés dans un format à définir (csv, xml, base de données noSQL type documents, etc.)
- Etape 2 : ***Analyse exploratoire du corpus***. **Etape mutualisée avec la Tâche 3.1 et 3.2**. Pour chaque genre de corpus et pour chaque langue, extraire les termes les plus fréquents (unigrams/bigrams/trigrams) et construire le nuage de mots correspondants. Selon le type de pré-traitement effectué (suppression mots vides, lemmatisation, etc.), observez et analyser les résultats selon le genre du corpus et la langue **(interactions avec le groupe M1 tâche 3.1)**
- Etape 3 : ***Détection des émotions***. Pour chacun des corpus collecté, concevoir un système d'apprentissage automatique qui attribue à chaque message du corpus une catégorie d'émotions parmi un ensemble de catégories prédéfinies. Plusieurs modèles doivent être proposés, comme par exemple :
  - Modèles d'apprentissage supervisé (type SVM, Random Forest, etc.) à base de sac de mots (réutilisation des termes extraits à l'étape 2), de LDA (interaction avec le groupe M1 Tâche 3.1), de lexiques d'émotions, ou alors d'autres traits pertinents.
  - Modèle à base de réseaux de neurones avec représentation vectorielle par plongement de mots.

  Les corpus IoT que vous avez collectés ne sont pas annotés en émotions. Pour réaliser le système de détection, vous devez alors entraîner vos modèles sur des corpus annotés, puis les tester sur le corpus IoT. Vous devez donc prévoir d'annoter manuellement un corpus de test (environ 500 instances du corpus IoT), afin dévaluer les performances de vos modèles[4].

  Voici les principales étapes à effectuer :

  - Concevoir un modèle sur les corpus en anglais en utilisant des corpus d'apprentissage annotés manuellement en émotions (cf. plus bas). Tester ensuite vos modèles sur le corpus IoT.

[2]Vous pouvez pour cela utiliser l'API twitter. Cependant, cette dernière offre des possibilités de collecte réduite (nombre de tweets retourné par requête faible, lenteur, etc.). Plusieurs *tweet scraper* existent, à vous de trouver celui qui convient.

[3]Voici quelques API disponibles : https://en.wikipedia.org/wiki/Lit_of_news_media_APIs https://newsapi.org/ https://webhose.io/, https://rapidapi.com/contextualwebsearch/api

[4]Pour l'annotation, vous pouvez utiliser des outils en ligne tels que Brat ou WebAnno. Vous devez également prévoir de calculer les accords inter-annotateurs (mesure Kappa)

    - Pour les tweets : utiliser les corpus du fichier EmotionTweet.zip
    - Pour les news headlines : utiliser les corpus du fichier AffectiveText.zip
  - Concevoir le même modèle pour le français.
    - Pour les tweets : Entraîner vos modèles sur le corpus EmotionDeft2015, puis testez le sur le corpus français IoT.
    - Pour les news : Utiliser des API de traduction automatique (type Google translate ou Bing) pour créer des corpus en français manuellement annotés en émotion. Le corpus EmotionNews sera traduits de l'anglais vers le français. Construire un modèle de détection des émotions en utilisant les corpus traduits comme données d'entraînement, puis appliquer le modèle sur le corpus de news IoT en français.

  Evaluer les performances de vos différents modèles d'apprentissage et effectuer une analyse qualitative des résultats obtenus, tout en dégageant les différences en terme d'émotions ressentis par langue et par genre de corpus.

- ***Etape 4 : Extraction des entités nommées et entity linking***. L'objectif est d'extraire la listes des principaux acteurs privés ou publics qui ont un lien avec les technologies IoT. Il faudra concevoir un système qui : (a) Détecte si un mot ou un groupe de mot est une entité nommée ou non, (b) puis identifie le type de l'entité (entreprises, organisation, personnalité politique, etc.). Ce système devra fonctionner pour chaque langue et chaque type de corpus. Une fois les entités nommées extraites et désambiguisées, identifier les émotions exprimées sur chacune de ces entités[5].

- ***Etape 5 : Détection des émotions sur le corpus de sondage***. Appliquer les modèles appris précédemment pour détecter les émotions exprimées par le personnels/étudiants de l'UPS via le sondage en ligne : construction du nuage de mots clés, les avis exprimés sont-ils positifs/négatifs **(mutualisation avec la tâche 3.1 et 3.3)**, quels est le type des émotions exprimées, etc.

[5]Plusieurs outils existent pour cette tâche, il conviendra de les tester et de retenir celui qui retourne les meilleurs résultats. Voici quelques outils à tester : GATE NLP Framework: ANNIE et TwitIE pipeline, Stanford NER, TextRazor. A vous d'en trouver d'autres