ARNAUD SOULIER

MATHIEU REMY

POUR LE 08-09/03/2018

PROJET TALN : COMPTE-RENDU INTERMEDIAIRE

Analyse de sentiment et génération de tweets

PROBLEMATIQUES

L'analyse de sentiment est un champ du traitement du langage naturel, qui consiste à déterminer la tonalité émotionnelle d'un mot, une phrase ou un texte. Cette tonalité peut être plus ou moins négative ou positive, mais également plus ou moins objective ou subjective.

D'autre part, l'enjeu de la génération de tweets est de créer des phrases courtes bien formées et porteuses de sens, dans le but d'être par la suite publiées sur Twitter.

APPROCHE CHOISIE

La première partie de notre projet consistera à créer une application qui, à partir d'un sujet (mot-clé), ira récupérer des tweets, pour réaliser une analyse d'opinions : c'est-à-dire déterminer si un sujet est perçu positivement ou négativement en ce moment par les internautes, par exemple en lui assignant une note.

Dans la seconde partie, nous créerons un module qui ira chercher la tendance (hashtag) en top-tweet la plus positive du moment, pour générer un tweet positif à son sujet et le mettre en ligne.

L'objectif final est donc de concevoir un agent intelligent qui cherchera à suivre les tendances et à se conformer relativement aux autres sur Twitter.

TECHNIQUES EXISTANTES

Cette bibliographie rassemble diverses informations, modules et outils qui pourront s'avérer utiles dans la conception de notre projet. Nous avons principalement accès notre recherche sur le traitement de l'anglais (mais pas que) et le langage Python, qui dispose d'un large panel de modules pour les domaines qui nous intéressent.

RECUPERATION ET PUBLICATION DE TWEETS

- **Twitter developer**: https://developer.twitter.com/en.html
  Site web officiel où il faudra de créer un compte développeur, qui conditionne l'accès aux API (car nécessité de s'authentifier sur le réseau pour accéder aux données). Nous utiliserons le service gratuit, qui fournit un accès limité aux données, mais qui s'avère suffisant ici.
- **Tweepy**: http://www.tweepy.org/
  C'est une API python de Twitter populaire et user-friendly. Elle sera utile pour interagir avec Twitter : récupérer des stream de tweets, la trend list du moment, mettre en ligne un tweet, etc.
- **Twitter-python**: https://github.com/bear/python-twitter
  Une alternative à l'API précédente.

ANALYSE DE SENTIMENT

- **NLTK**: https://www.nltk.org/
  Une boîte à outils complète pour Python (voir **TextBlob**).
- **TextBlob**: http://textblob.readthedocs.io/en/dev/index.html
  Une autre boîte à outils, basée sur NLTK, davantage user-friendly. Elle contient comme fonctionnalités : **des dictionnaires de toute sorte (dont WordNet), le parsing, les n-grammes, l'extraction de groupes nominaux, les principaux classifieurs (naïve bayésienne, arbre de décision, etc.), la traduction, la tokenisation**, et bien d'autres choses.
- **Vader**: https://github.com/cjhutto/vaderSentiment
  Un module et lexique adapté à l'analyse de sentiment sur les réseaux sociaux.
- **LVF**: http://rali.iro.umontreal.ca/rali/?q=fr/lvf
  Un dictionnaire des verbes français.
- **FondamenTAL** : http://talep.lif.univ-mrs.fr/FondamenTAL/
  Une page listant de nombreuses ressources en français.
- **Dicovalence**: https://www.ortolang.fr/market/lexicons/dicovalence
  Un autre dictionnaire des verbes français.
- **FEEL**: http://www.lirmm.fr/~abdaoui/FEEL.html
  Un dictionnaire recensant les polarités de 14000 mots distincts en français.

- **WordNet**: http://wordnetweb.princeton.edu/perl/webwn
  Une base de données lexicale où les mots sont triés en « synsets ».
- **SentiWordNet**: http://sentiwordnet.isti.cnr.it
  Extension de WordNet ajoutant leur polarité aux synsets.
- **JeuxDeMots** : http://www.jeuxdemots.org/jdm-accueil.php
  Un dictionnaire d'association lexicale en français.

GENERATION DE TWEETS

- **NLTK** : voir plus haut.
- **Chaînes de Markov** : https://fr.wikipedia.org/wiki/Cha%C3%AEne_de_Markov
  Un article Wikipédia sur cet outil mathématique qui pourrait être utile.
- **NumPy**: http://www.numpy.org/
  Un module efficace pour la gestion des matrices en Python.
- **TFLearn**: http://tflearn.org/
  Un module de Machine Learning basé sur **TensorFlow** et plus user-friendly que ce dernier. L'ayant déjà utilisé par le passé, nous l'envisageons comme une piste intéressante pour la génération de textes.

COMPARAISON DES OUTILS EXISTANTS

Cette section rend à la fois compte des articles que nous avons consultés lors de la recherche d'informations, ainsi que des projets similaires existants dans le domaine. Certains pourraient sembler redondants à première vue, mais nous n'avons listés que ceux qui nous ont paru significatifs (présence d'une ou plusieurs informations inspirantes ou utiles).

ANALYSE DE SENTIMENT

- **Analyse de sentiment Twitter** : http://www.agroparistech.fr/ufr-info/membres/cornuejols/Teaching/Master-AIC/PROJETS-M2-AIC/PROJETS-2016-2017/analyse-de-sentiments(lambert-bellard-lorre-kouki).pdf
  Un article résumant assez bien l'état de la recherche.
- **Sentiment140** : http://www.sentiment140.com/
  Un logiciel d'analyse de sentiment sur Twitter relatif à des sujets, des produits ou des marques. Il utilise des algorithmes de Machine Learning (plus précisément : la Distant Supervision).

- **ResTS**: http://www.aclweb.org/anthology/F12-3018
  Un article à propos du logiciel ResTS, qui réalise des résumés d'opinions sur des produits commerciaux à partir de Twitter et **SentiWordnet**. On y trouve un descriptif détaillé des techniques employées (équations mathématiques et pseudo-codes).
- **Best free tools** : https://www.softwareadvice.com/resources/free-twitter-sentiment-analysis-tools/
  Un article présentant une pléiade de logiciels d'analyse de sentiment sur Twitter.
- **Sentiment analysis in Python** : http://blog.aylien.com/build-a-sentiment-analysis-tool-for-twitter-with-this-simple-python-script/
  Un article décrivant une implémentation en Python.
- **Step-by-Step Twitter Sentiment Analysis**: http://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis/
  Un autre article d'analyse de sentiment sur Twitter, présentant une implémentation utilisant la classification naïve bayésienne (du module NLTK).
- **Twitter Sentiment Analysis using Python**: https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
  Une autre implémentation, utilisant **TextBlob**.
- **Azure Stream Analysis**: https://docs.microsoft.com/fr-fr/azure/stream-analytics/stream-analytics-twitter-sentiment-analysis-trends
  Un article sur l'analyse de sentiment avec Microsoft Azure, la plateforme de cloud-computing au succès grandissant (dispensable mais intéressant).
- **Sentiment Viz** : https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
  Un autre exemple logiciel (avec une interface).
- **Another Twitter sentiment analysis with Python** :
  https://towardsdatascience.com/another-twitter-sentiment-analysis-bb5b01ebad90
  Un "article-fleuve" en 11 parties présentant de très nombreuses techniques.
- **How to Build the Trump Twitter Sentiment Analysis Dashboard**:
  https://hackernoon.com/visualizing-trump-7e1cb5e55a78
  Un analyseur de sentiment utilisant **PubNub** et **IBM Watson**.
- **Sentiment Analysis of Today's Tweets**:
  https://medium.com/@olafusimichael/sentiment-analysis-of-todays-tweets-about-president-buhari-using-python-s-vader-and-sentiwordnet-c4a42e8c748f
  Utilisation intéressante de **Vader** et **SentiWordNet**.
- **ADVANSE** : http://talnarchives.atala.org/ateliers/2015/DEFT/deft-2015-long-009.pdf
  Un article complet sur un analyseur de tweets français.

GENERATION DE TWEETS

- **Génération automatique de textes** : https://interstices.info/jcms/int_63524/generation-automatique-de-textes
  Un article général sur le sujet.
- **Conversational UI Principles** : https://medium.com/swlh/conversational-ui-principles-complete-process-of-designing-a-website-chatbot-d0c2a5fee376
  Un article sur le fonctionnement des chatbots.

- **Génération de textes**: https://en.wikipedia.org/wiki/Natural_language_generation et https://fr.wikipedia.org/wiki/G%C3%A9n%C3%A9ration_automatique_de_textes
  Les articles Wikipédia sur le sujet.
- **SimpleNLG** : https://github.com/simplenlg/simplenlg
  Un générateur de textes réalisé en Java, intéressant à décortiquer.
- **SimpleNLG-EnFr** : http://www-etud.iro.umontreal.ca/~vaudrypl/snlgbil/snlgEnFr_english.html
  Extension de SimpleNLG pour le français.
- **poetry-generator**: https://github.com/schollz/poetry-generator
  Un générateur de poèmes.
- **Générateur automatique de textes** : http://www.samszo.univ-paris8.fr/Generateur-automatique-de-textes
  Un projet universitaire achevé.
- **Natural Language Generation :** http://www.inf.ed.ac.uk/teaching/courses/nlg/
  Un cours universitaire.

CAHIER DES CHARGES DU PROTOTYPE

A priori, nous allons réaliser notre prototype en Python, étant donné la variété de modules disponibles dans ce langage. Le projet étant ambitieux et divisé en deux parties, Python nous permettra de passer plus de temps sur la partie algorithmique, plutôt que de s'enliser dans des problématiques de récupération ou de structuration de données. Par ailleurs, nous ne nous intéresserons qu'à la partie anglophone de Twitter.

L'utilisation du programme se fera directement dans la console, sauf si nous avons suffisamment de temps pour réaliser une petite interface graphique très minimaliste et modeste. Nous avons conscience de l'étendu du travail qu'il faudra réaliser, mais nous sommes déterminés à le mener à son terme.

DONNEES

- **Twitter** : La première tâche à réaliser sera d'implémenter la récupération de n tweets à partir d'un mot-clé au moyen d'une API. Ensuite nous allons devoir filtrer (**stop-words**) et structurer ces données (**TextBlob**) pour l'usage qui va en être fait. Il faudra également que notre programme soit capable de récupérer la trend-list du moment, et aussi de publier un simple tweet.

- **Les dictionnaires** : Le choix des dictionnaires va s'avérer significatif. Nous allons devoir prendre en compte l'aspect grammatical (**WordNet** et **NLTK**), tenir compte de la spécificité du langage employé sur Twitter (Prise en compte des émoticônes, de

l'argot, des mots grossiers, des fautes d'orthographe, etc.), et choisir un dictionnaire pour la polarité des mots (**SentiWordNet, Vador**).

ALGORITHMES

Pour cette partie du projet, nous devrons implémenter quatre modules :

- **L'analyse de sentiment appliqué à un tweet anglais** : Il va falloir réaliser un analyseur pour les phrases qui, à partir de la polarité des mots et de leur pondération, nous fournisse un score émotionnel (% ou indice), qui prennent en compte la spécificité de ce qu'est un tweet (court, lié à un ou plusieurs hashtags, un lieu géographique, dans un langage particulier, utilisant l'ironie ou non, etc.).

- **L'analyse d'une tendance liée à un mot-clé** : Notre programme doit pouvoir, à partir d'un mot-clé donné par l'utilisateur (terme ou hashtag en anglais), récupérer une liste de tweets du moment sur le sujet et appliquer l'analyseur de sentiment sur cette liste. Le but étant de calculer un score émotionnel pour ce mot-clé.

- **L'analyse de la Trend List** : Après avoir récupérer le top-tweet, nous aurons besoin d'un module qui réalisera l'analyse de sentiment de ces différents hashtags, dans le but de les trier du plus positif au plus négatif. Nous garderons en mémoire les tweets du hashtag le plus positif, pour les utiliser dans le module suivant.

- **Le générateur de tweet** : Après avoir déterminé la tendance du moment la plus positif, le générateur devra créer un tweet positif sur le sujet, puis le mettre en ligne. Nous ne sommes pas réellement satisfaits des informations que nous avons pu trouver sur ce sujet, nous comptons donc approfondir nos recherches davantage. A priori, nous utiliserons les chaînes de Markov ou bien des techniques de Machine Learning.

FONCTIONNALITES ET INTERFACE UTILISATEUR

Le programme comportera deux parties :

- **L'analyseur de tendance** : A partir d'un mot-clé en anglais donné par l'utilisateur, il renverra un score émotionnel (% ou indice) lié à l'analyse des tweets du moment. Si vous avons le temps, nous réaliserons l'affichage d'un nuage de points sur un graphique pour représenter visuellement le résultat.

- **Le bot « suiveur et optimiste »** : Cette partie de notre programme exploitera l'intégralité des fonctions implémentées. L'application devra réaliser, à la demande, l'analyse de la Trend List Twitter du moment pour déterminer le sujet populaire le plus positif, et ensuite générer et publier un tweet positif sur ce sujet. Le déroulement des différentes étapes sera affiché au fur et à mesure sur la sortie.