Opinion mining et Sentiment analysis

Fouille d'opinions et analyse des sentiments

Ricco Rakotomalala

Savoir ce que pensent les « gens » (électeurs, clients, concurrents, etc.) est fondamental pour la prise de décision.

Les enquêtes d'opinions constituent une manière de les récolter.

Avec le Web 2.0, ces informations sont disponibles à profusion, sous forme d'avis sur les sites de vente en ligne (accompagnée d'une note d'ailleurs), de discussions dans les médias et réseaux sociaux (blogs, wikis, twitter, facebook, etc.).

Ces écrits sont de nature différente de ceux des professionnels (pour lesquels la trame et les critères sont explicites – ex. Essai Z3 2.8), ils intègrent une dimension émotionnelle et sont peu codifiés (non cadrés par un questionnaire)

Ex. Avis sur AMAZON concernant l'Intégrale de Spirou N°9

La fouille d'opinions, en particulier à partir de données des réseaux sociaux, est un excellent substitut nettement moins coûteux des enquêtes d'opinions.

- Evaluation des produits, d'une politique, d'une personnalité
- En appui des systèmes de recommandations (ex. ne pas proposer des produits qui ont des mauvaises notes)
- Analyse de la popularité, des tendances
- Positionnements par rapport à un sujet délicat (ex. mariage pour tous)

Mais les réseaux sociaux permettent d'aller plus loin….

- Susciter des réactions (ex. loi travail)
- Identifier des leaders d'opinions et/ou des spammeurs d'opinions
- Détecter des communautés (ex. Affaire Brown à Ferguson)

1. Fouille d'opinions – Cadre général
2. Analyse des sentiments
3. Analyse des sentiments sur Twitter
4. Plus loin avec la fouille d'opinions
5. Bibliographique

Cadre général de la fouille d'opinions

FOUILLE D'OPINIONS

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop . . .

(1) relate un fait objectif. (2), (3) et (4) expriment une opinion subjective plutôt positive ; (5) et (6) une opinion négative.

L'entité iPhone en général est le sujet de (2) ; (3), (4) et (6) sont relatifs respectivement aux aspects « touch screen », « voice quality » et « price » de l'iPhone ; « me » est le sujet de (5).

« I » (je) est le titulaire des opinions (2), (3) et (4) ; pour (5) et (6), il s'agit de « mother ».

Entité est la cible de l'opinion (objet, produit, évènement, personne…)

Aspect de l'entité ciblée par l'opinion. Parfois, l'opinion peut porter sur l'entité en général (ex. *nice phone*).

Holder (titulaire) de l'opinion c.-à-d. celui qui l'exprime. Important à distinguer lorsqu'ils sont plusieurs et que des phénomènes de communautés peuvent apparaître (coopérant ou s'opposant)

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the (price of the) phone was too expensive, and wanted me to return it to the shop . . .

Orientation de l'opinion, qui peut être de polarité positive ou négative. Une opinion peut être *régulière* (une appréciation) ou *comparative* (en comparaison avec une autre entité). Seul le premier cas nous intéresse ici.

Revoir avis sur AMAZON concernant l'Intégrale de Spirou N°9

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the (price of the) phone was too expensive, and wanted me to return it to the shop . . .

Une opinion est un quintuplet $(e_i,a_{ij},o_{ijkl},h_k,t_l)$, où e_i est l'entité, a_{ij} est un des aspects de e_i, h_k est le titulaire, o_{ijkl} est son orientation (polarité), et t_l est la date (time) où elle a été exprimée. Parce que bien sûr, une opinion peut être fluctuante dans le temps. !

Tous les éléments ne sont pas strictement nécessaire. Par ex., on peut omettre le titulaire si l'on souhaite faire des statistiques sur un grand nombre de documents ; ignorer la date si on travaille à fenêtre temporelle fixée ; etc.

Dans certains cas, il peut être intéressant d'ajouter une granularité supplémentaire pour approfondir les analyses (ex. sexe, âge ou CSP des titulaires, etc.)

1. Extraction des entités et regroupement. Identification des entités, regroupement des éventuels synonymes.
2. Extraction des aspects. Association avec les entités, regroupement des éventuels synonymes.
3. Identification du titulaire, datation. Un titulaire exprime une opinion à une date donnée, qui peut être déterminante dans l'analyse.
4. Détermination de l'orientation de l'opinion. Elle peut être positive, neutre, ou négative. Parfois, il est nécessaire de faire la distinction au préalable entre une expression objective (fait) et subjective (opinion).
5. Enumération de l'ensemble des tuples (e_i,a_{ij},o_{ijkl},h_k,t_l) dans le corpus suite aux étapes ci-dessus.

L'analyse peut être simplifiée souvent. Ex. tweets consécutifs à la publication du projet de la loi travail. Titulaires = Institution vs. Autres ; Entité = loi travail ; Aspect = aspect général ; Date = Suite à la publication du projet de loi jusqu'à son vote.

sans compter les modèles pré-entraînés sur des grands corpus qui font référence aujourd'hui (ex. sur Hugging Face)

Focus sur l'orientation des opinions

ANALYSE DES SENTIMENTS

L'analyse des sentiments s'intéresse à l'orientation d'une opinion par rapport à une entité ou à un aspect d'une entité. On parle de *polarité*, elle peut être positive, neutre, ou négative.

Nous positionnons l'analyse au niveau du document (*document level sentiment*). L'individu statistique est le document. On aurait pu descendre d'un cran et décomposer au niveau des phrases (*sentence level sentiment*), un document pouvant être constitué de plusieurs phrases (*ex.* J'aime les histoires. Mais l'emballage n'est pas terrible.)

dans ce cas, on peut aussi considérer que phrase = document, question de découpage… toujours la question de "l'individu statistique"

Nous considérons que le titulaire de chaque document est unique. Mais il peut y avoir des titulaires différents d'un document à l'autre, plusieurs documents peuvent avoir le même titulaire.

Le niveau de sentiment « neutre » peut être une position intermédiaire entre « positif » et « négatif ». Il peut être également significatif d'un énoncé objectif (fait), auquel cas il convient de discerner ce qui relève de l'opinion (subjective) ou non dans un premier temps (voir plus loin…)

énoncé objectif vs. opinion subjective

Apprentissage statistique. Les documents sont étiquetés manuellement par un expert (ex. -1, 0 , +1). Nous rentrons dans le cadre bien connu de la catégorisation de textes, nous utilisons les différentes techniques de machine learning. La démarche est rigoureuse, mais l'étiquetage expert est un véritable goulot d'étranglement, il peut être aussi bruité (l'expert n'est pas infaillible…).

Approche la plus simple, la plus populaire….

Utilisation d'un thésaurus de sentiments. Des polarités sont associés à des termes ou à des phrases complètes. La polarité d'un document peut être alors calculée à partir de la somme des polarités des termes (ou des phrases) qui la compose. En pratique, ce n'est pas toujours facile. Il y a la synonymie à gérer, un même terme peut avoir des polarités différentes selon les domaines (ex. « épais » dans la mode et dans le catch n'a pas vraiment la même tonalité…), il faut traiter convenablement la négation (ex. il n'est pas rapide), l'ironie, les sarcasmes, etc.

"remonte dans ton arbre".... pris individuellement, les termes ne sont en rien négatifs, le tout…

Nous disposons d'un corpus étiqueté.

Sentiment	Text
neg	is so sad for my APL friend.....
neg	I missed the New Moon trailer...
pos	omg its already 7:30 :O
neg	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11..
neg	i think mi bf is cheating on me!!! T_T
neg	or i just worry too much?
pos	Juuuuuuuuuuuuuuuussssst Chillin!!
neg	Sunny Again Work Tomorrow :-| TV Tonight
pos	handed in my uniform today . i miss you already
pos	hmmmm.... i wonder how she my number @-)
neg	I must think about positive..

Avec la représentation « bag-of-words » (et outils associés : stemming, etc.), nous pouvons appliquer la démarche de la catégorisation de textes.

La méthode Naive Bayes (modèle d'indépendance conditionnelle) est souvent mise en avant dans ce contexte (ex. Coelho & Richert, chapitre 5), mais d'autres méthodes supervisées peuvent être utilisées.

Liu (page 416) rapport une technique très simple due à Dave et al. (2003) utilisée dans le contexte de l'analyse des critiques de produits.

Pour une cible binaire (1/0), le score d'un terme t_j est égal à

$$score(t_j) = \frac{p(t_j/1) - p(t_j/0)}{p(t_j/1) + p(t_j/0)}$$

score ici = sorte de sur-représentation du terme dans la classe 1

Nombre de fois où le terme t_j apparaît dans les documents de la classe 1, divisé par le nombre total de termes apparaissant dans les documents de la classe 1.

Pour évaluer la « positivité » d'un document d, nous calculons :

$$eval(d) = \sum_j score(t_j)$$

Sur l'ensemble des termes qui compose le document d

Pour classer un document (affectation à la classe 1 ou 0) :

$$Si\ \text{eval(d)} > 0\ \text{Alors class(d)} = 1\ \text{sinon}\ 0$$

A chaque terme peut être associé un degré de positivité (ou négativité). SentiWordNet propose un thésaurus décrivant la polarité d'une liste prédéfinie de termes.

Polarité

Exemples

# POS	ID	PosScore	NegScore	SynsetTerms	Glossary
a	3700	0.25	0	dissilient#1	bursting open with force, as do some ripe seed vessels
a	3829	0.25	0	parturient#2	giving birth; "a parturient heifer"
a	5599	0.5	0.5	unquestioning#2 implicit#2	being without doubt or reserve; "implicit trust"
a	5718	0.125	0	infinite#4	total and all-embracing; "God's infinite wisdom"

Permet d'identifier le synset sur WordNet (base de données lexicale). « a » veut dire « adjectif »; ID est son numéro dans la base.

Termes du synset, nombre de sens associé aux termes, appartenant à l'ensemble de synonymes (**synset** – groupes de mots interchangeables pour un sens particulier)

Cette étape nécessite de connaître la catégorie du terme ('a' : adjective, 'n' : noun, 'v' : verb, 'r' : adverbe). Le **part of speech tagging** (analyse morphosyntaxique) permet d'organiser le termes d'une phrase en une structure arborescente (voir Cours introductif) qui permet d'identifier leur rôle.

Exemples

Je **livre** du vin.
C'est un bon **livre**.
Voici une **livre** de pain.

A l'instar de SentiWordNet, nous pouvons construire un lexique des sentiments avec des polarités spécifiques à un domaine (Baccianella et al., 2010).

1. Nous choisissons un noyau de termes d'opinions (*opinion words*) reconnus comme positifs et négatifs dans le domaine (SentiWordNet utilise les synsets).

2. Ces noyaux sont étendus en analysant leurs synonymes et antonymes (que WordNet peut fournir), en en veillant à ne pas dépasser un certain **rayon** qui est un paramètre clé du dispositif.

3. A partir de ces ensembles de termes (qui sont étiquetés donc), nous construisons un classifieur où les variables prédictives sont constituées par les glossaires (avec une représentation bag of words par exemple).

4. Le classifieur ainsi élaboré permet d'attribuer un degré de positivité / négativité aux autres termes.

Twitter, un contexte privilégié pour l'analyse des sentiments

ANALYSE DES SENTIMENTS SUR TWITTER

Twitter est un outil de microblogage qui permet à tout un chacun de communiquer via un message limité à 140 caractères. Les utilisateurs (@) peuvent interagir entre eux, il est possible de définir des sujets avec (#). Il est incontournable aujourd'hui dans la stratégie de communication des décideurs (ex. Le Monde – Stratégie Trump)

Pourquoi des analyses sur des Tweets ?

- Les documents sont brefs et de longueur équivalentes. Un document est souvent focalisé sur un aspect d'une entité, et est associé une et une seule orientation.
- On y exprime souvent des avis. Nous avons bien des opinions subjectives souvent.
- Les auteurs (titulaires) sont clairement identifiables avec @. Les éventuels interlocuteurs aussi. On peut voir les communautés émerger (ex. Paris plage)
- Les sujets (entités) sont clairement identifiables avec #.
- Mises à jour fréquentes, dynamisme et réactivité des acteurs. On peut détecter des tendances par rapport aux auteurs / communautés, par rapport aux sujets.
- Les messages se prêtent à une multitude d'analyses (ex. Tweet Sentiment Visualization)… mais, attention, ils ne permettent pas de tout faire (ex. présidentielles)

Le nettoyage du texte joue un rôle très primordial. Le format induit notamment l'utilisation de raccourcis et de smiley qui deviennent partie intégrante de la langue et qu'il faut savoir gérer, la normalisation est importante, il faut aussi tenir compte des liens (http).

Exemple : 10 tweets avec le sujet #macron extraits le 17/01/2017 (96 jours avant le 1er tour des présidentielles 2017).

```
library(twitteR)

#compte pour l'extraction from dev.twitter.com
api_key <- ""
api_secret <- ""
token <- ""
token_secret <- ""

#créer une connexion avec  Twitter
setup_twitter_oauth(api_key, api_secret, token,
token_secret)

#récupération des tweets
tweets <- searchTwitter("#macron",n=10,lang="fr")
print(tweets)
```

```
 [1] "RT @PorcherThomas: #Macron se soucie des agriculteurs
alors qu'il était ds la commission #attali qui a donné les
pleins pouvoirs à la grand…"
 [2] "RT @SansPartiFixe: \xed\xed Les Fillonistes sont
aveuglés par leur pseudo \"Victoire\" à la primaire et ne
voient pas arriver le duel #Macron vs #ML…"
 [3] "RT @CyrDici: @L_oranaise_ \nET C EST #macron LA CRÉATURE
DES BANQUIERS QUE LES JOURNALISTES VEULENT NOUS IMPOSER
?\nhttps://t.co/flSDVWTDBY …"
 [4] "RT @CliveTwo: \"Je te nommerai première
ministre...\"\n\n#macron #royal https://t.co/3Cb2H6schm"
 [5] "RT @AjgmHenry: Qui peut encore dire que le projet de
#Macron est vide ou qu'il est une bulle médiatique ? De nvlles
propositions chaque jou…"
 [6] "RT @geigerie: #macron ne veut pas répondre sur ses
soutiens financiers. #QuiFinanceMacron ?
https://t.co/AVoAze8WD1"
 [7] "RT @CliveTwo: \"Je te nommerai première
ministre...\"\n\n#macron #royal https://t.co/3Cb2H6schm"
 [8] "RT @AntoineBisco: Ne jamais oublier que #Macron à été
choisi par #Attali . https://t.co/dnigGfEeIK"
 [9] "RT @Daarjeeling: @rvan92colombes la position de #Macron
= casse-tête. Il a été ministre de Hollande, dont il défend le
bilan : autant à gau…"
[10] "RT @CerdagneFrance: D'après @BFMTV , #macron aurait,
hier, multiplié les pains. Ministre, il avait déjà multiplié
les bus... et les faillit…"
```

cf. le livre PDF en ligne… même si c'est obsolète, les idées restent d'actualité

Dans son livre « Mining the Social Web », Russell (2013) décrit un ensemble de *recettes* en Python pour accéder à différents types d'analyses.

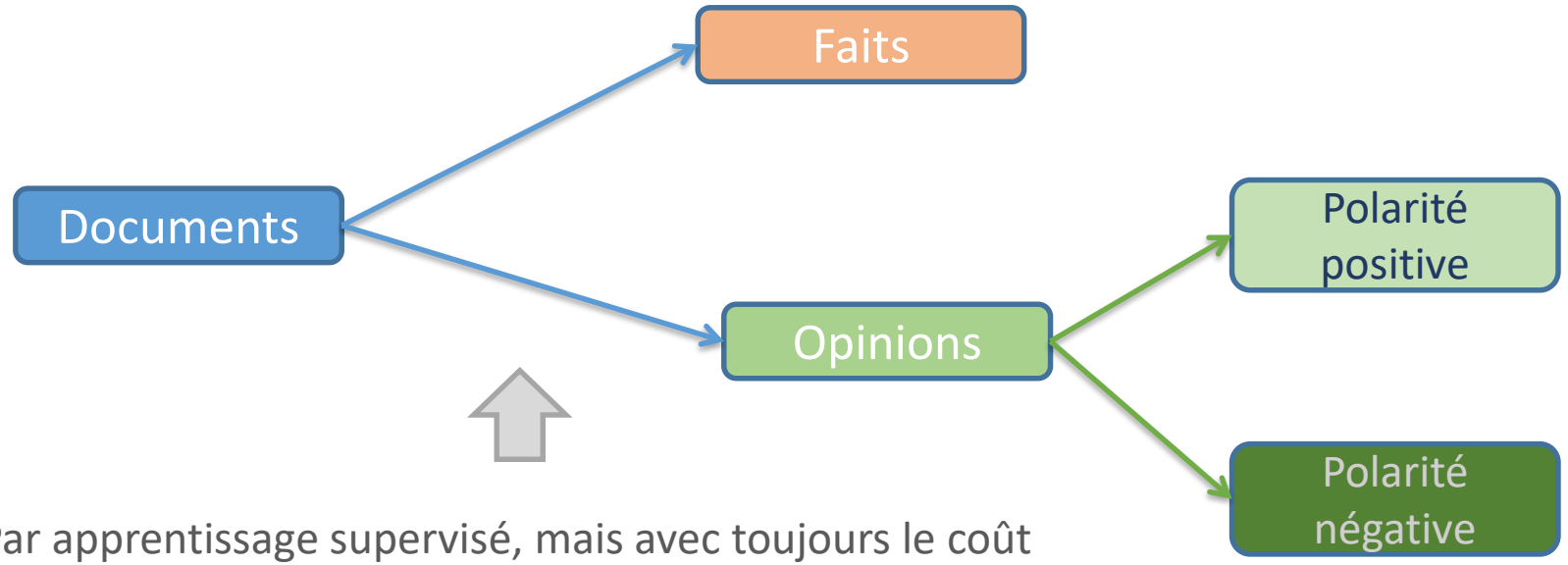
Des API existent également pour d'autres types de réseaux sociaux : Facebook, LinkedIn, Google+. Voir le même ouvrage.

cf. tuto sur LinkedIn par ex. http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#3075570128898707548

Part II. Twitter Cookbook

PLUS LOIN AVEC LA FOUILLE D'OPINIONS

Avant de s'intéresser à la polarité, il peut être intéressant déjà d'identifier si le document correspond à une opinion subjective ou à un fait objectif. On aurait une analyse en deux temps.



1. Par apprentissage supervisé, mais avec toujours le coût lié à l'étiquetage manuel
2. En utilisant un lexique. La catégorie lexicale joue un rôle important : les adjectifs et/ou les adverbes sont plus fréquents dans les documents exprimant une opinion.
3. Des formules peuvent annoncer une opinion (Opinion and Statements).

Plutôt qu'une polarité (+) ou (-), une note (*rating*) est associée au document

Ex. Blake & Mortimer – Tome 24 – Sur AMAZON

★☆☆☆☆ **Une histoire à oublier ...**

Par [redacted] MEMBRE DU CLUB DES TESTEURS le 9 janvier 2017

Format: Relié

Il n'y a rien dans cet opus qui puisse rappeler la grande époque de Blake et Mortier ... scénario plat et ennuyeux ... même Olrik a préféré rester planqué dans sa cellule plutôt que de tenter de gesticuler dans une intrigue pour le moins étriquée ... il faut vous reprendre ... où vous allez définitivement nous ennuyer !

★★★★★ **BLAKE et MORTIMER Le testament de William S**

Par Client d'Amazon le 8 janvier 2017

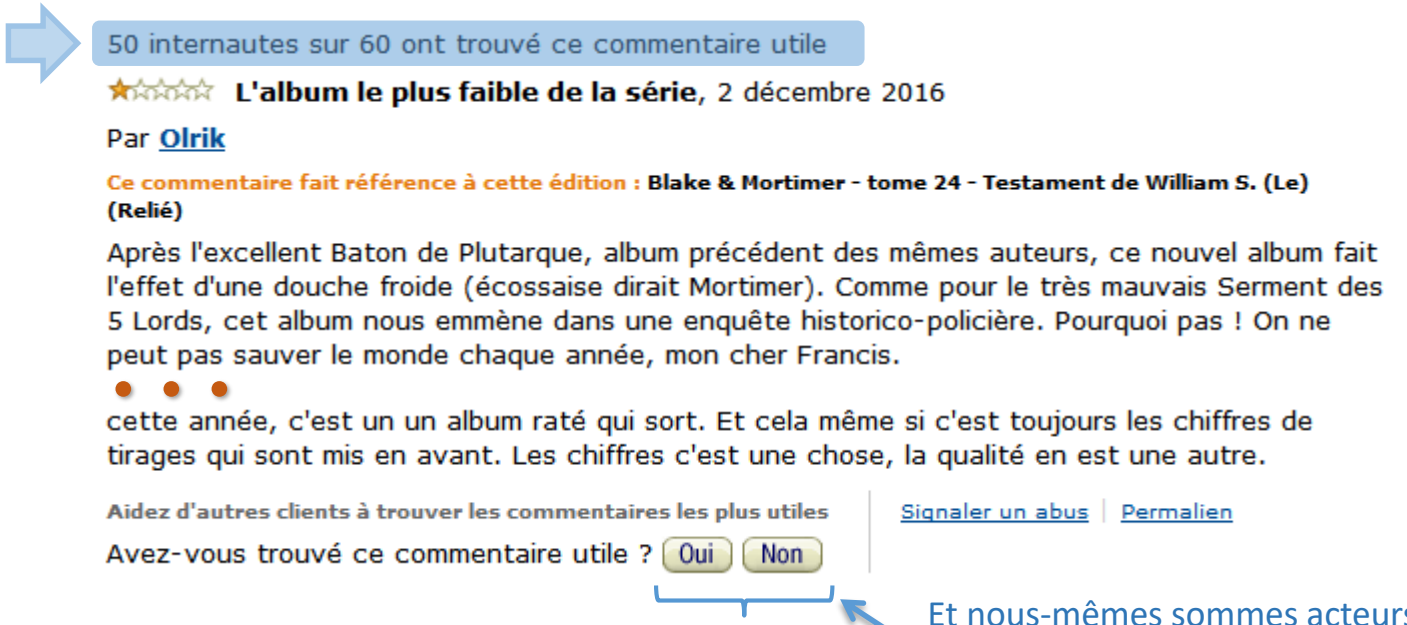
Format: Relié | Achat vérifié

j'ai choisi ce livre pour compléter ma collection .Bien écrit , coloré et bien dessiné .Je recommande sa lecture. Très bonne série .

• • •

La cible peut être considérée comme quantitative ou, plus subtilement, qualitative ordinale. Nous sommes dans le cadre de la régression.

L'opinion s'exprime sous la forme d'un intérêt que l'on peut porter à un document. Pour un document, nous disposons de plusieurs évaluations (plusieurs observations). La variable cible s'exprime sous la forme d'une proportion.



Et nous-mêmes sommes acteurs de l'enrichissement de la base.

cf. vidéo TD7, feu "Rég. logistique"

Des techniques (ex. régression logistique) peuvent prendre en compte ce genre de situation où pour une expression de la description (*covariate pattern*), nous disposons de plusieurs observations de la cible (ex. Rég. Logistique, page 65).

Les enjeux sont fort, des agitateurs peuvent venir perturber sciemment la perception d'un produit, d'une personnalité… en l'encensant ou en la dénigrant (fake opinions, fake reviews,…). On pourrait évoquer les trolls également.

cf. Google : "rue89 troll captcha"

On distingue généralement deux type de stratégies de spamming.

1. Spammeur individuel. Le malandrin agit seul, avec une ou plusieurs identités.

2. Groupes de spammeurs coordonnés. Les brigands agissent en groupe, de manière plus ou moins coordonnée, pour contrôler et orienter la perception d'un sujet.

On peut toujours utiliser les techniques supervisées : mais l'étiquetage manuel des documents reste un goulot d'étranglement ; et l'affaire devient plus compliquée lorsqu'on a plusieurs acteurs (les données viennent en grappes).

forme de "clustering"

Recherche des régularités : lorsque l'acteur est seul ou lorsque l'action est coordonnée, des éléments de langages récurrents apparaissent (termes d'opinions, …). En recherchant les similitudes entre les documents, nous pouvons mettre en évidence des groupes (ou la répétition d'un acteur). Une très forte homogénéité doit attirer notre attention.

cf. tutoriels - Détection des outliers sous Python

Recherche d'anormalité (outliers) : lorsque des documents ou des notes s'écartent significativement des autres, on peut s'en inquiéter. Il faut bien sûr que ces observations correspondent à une minorité.

individu statistique est modifié + suivi dans le temps

Analyse des comportements des titulaires : Au-delà des documents, le comportement des acteurs est une source d'indentification des spammeurs (ex. un acteur se concentre sur un seul type de produit ; un acteur donne systématiquement des notes très basses ou très hautes ; un produit concentre un très grand nombre d'évaluations ; …).

Ouvrages et articles

- Coelho L.P., Richer W., « Building Machine Learning Systems with Python », 2nd Edition, Packt Publishing, 2015.
- Liu B., Lei Z., « A survey of Opinion Mining and Sentiment Analysis », in Aggarwal C., Zhai C., « Mining Text Data », chapter 13, Springer, 2012.
- Liu B., « Web Data Mining - Exploring Hyperlinks, Contents and Usage Data », Springer, 2008.
- Pang B., Lee L., « Opinion mining and Sentiment Analysis », in Foundations and Trends in Information Retrieval, 2(1-2), p. 1-135, 2008.
- Russell M.A., « Mining the Social Web – Data Mining Facebook, Twitter, Linkedin, Google+, Github, and more », O'Reilly, 2013.