

ANALYSE ^{DES} SENTIMENTS

avec WordStat



» Qu'est-ce que l'analyse automatique de sentiments?

Avec l'avènement des médias sociaux et l'abondance des informations textuelles circulant sur le web, de plus en plus de spécialistes s'intéressent aux opinions énoncées par les internautes. Que ce soit dans un contexte sociologique, de marketing digital, de service clients ou de communication, l'analyse automatique des sentiments consiste à tenter de détecter à partir des commentaires recueillis la tonalité exprimée par l'émetteur en se basant sur le repérage de mots positifs, négatifs et parfois neutres.

Bien que très en demande, cette approche souffre de limite et à l'heure actuelle il n'existe aucun outil d'analyse de sentiments « prêt à l'emploi » (out-of-the-box) en mesure de faire une analyse fiable de verbatims clients de toutes provenances.

En effet, outre la complexité du langage naturel et en particulier celui en usage dans les médias sociaux, l'absence de prise en compte du contexte général ne permet pas d'interpréter facilement le sentiment véhiculé par le message.

Ainsi, le fait que certains termes changent de connotation selon le contexte particulier dans lequel ils sont utilisés illustre bien cette dernière difficulté (par ex. : le mot « frais » pour qualifier la qualité d'un aliment ou désigner des coûts bancaires).

Bref, la tâche est ardue, et pour pallier à cette difficulté et effectuer une analyse de sentiments adéquate, vous aurez besoin de développer votre propre dictionnaire de sentiments contenant un vocabulaire approprié associé à l'expression des sentiments positifs et négatifs spécifique à votre domaine d'application.

Dictionnaire WordStat pour pour les sentiments

Nous avons développé le « WordStat Sentiment Dictionary » pour l'analyse des sentiments des textes en langue anglaise. Bien que nous ne disposions pas présentement d'équivalent pour la langue française, les mêmes principes de développement et de validation devraient être appliqués pour assurer la fiabilité des résultats. Ce dictionnaire a été conçu en combinant des mots négatifs et positifs des trois dictionnaires suivant : « Harvard IV », « Regressive Imagery » (Martindale, 2003) et le « LIWC » (Pennebaker, 2007). L'utilitaire WordStat pour construire des dictionnaires a été utilisé par la suite pour accroître la liste de mots, en identifiant automatiquement les synonymes potentiels, les mots reliés ainsi que toutes les formes fléchies. Le dictionnaire comporte plus de 9533 mots négatifs et 5537 mots positifs. Le sentiment n'est pas mesuré avec ces deux listes, mais plutôt au moyen de deux ensembles de règles permettant la prise en compte des négations qui les précèdent. Par exemple, le sentiment négatif est mesuré en utilisant les deux règles suivantes :

- Mots négatifs non précédés d'une négation (non, pas, jamais) à moins de quatre mots dans la même phrase.
- Mots positifs précédés d'une négation à moins de quatre mots dans la même phrase.

Le sentiment positif est mesuré de la même façon en recherchant des mots positifs non précédés d'une négation ainsi que des termes négatifs après une négation. Cependant, nos propres expériences suggèrent que cette dernière règle a moins de valeur prédictive et pourrait même légèrement détériorer la mesure des sentiments. Il peut toutefois exister des situations où une telle règle pourrait aider à prédire les sentiments positifs. Nous avons donc décidé de conserver cette dernière règle et de laisser à l'utilisateur le soin de décider de sa pertinence.

Utilisation recommandée

NOUS RECOMMANDONS DE NE PAS D'UTILISER CE DICTIONNAIRE TEL QUEL. Pour obtenir une mesure la plus précise possible, nous vous recommandons vivement de personnaliser ce dictionnaire en appliquant la procédure suivante :

1. Retirer les mots spécifiques à un domaine - Identifier et supprimer les mots fréquents qui peuvent être spécifiques à votre domaine d'intérêt et qui n'ont pas de connotations positives ou négatives. La révision de tous ces mots peut prendre beaucoup de temps. Le moyen le plus efficace serait d'appliquer ce dictionnaire à un grand ensemble de documents de votre domaine et d'identifier les mots qui apparaissent fréquemment. Vous devez ensuite utiliser la fonction « mot-clé en contexte » de WordStat pour évaluer comment ces mots sont utilisés.

2. Identifier les erreurs de prédictions - Si vous avez un ensemble de documents qui ont déjà été classés comme positifs ou négatifs, ou qui contiennent des scores de satisfaction ou tout autre indicateur de sentiment, nous suggérons d'utiliser la fonctionnalité de tableau croisé de WordStat pour évaluer la corrélation entre mots positifs et négatifs fréquents et ces indicateurs. À partir d'une telle liste, prêtez une attention particulière à tout mot qui semble inversement lié à la prédiction attendue. À l'aide de la fonction « mot-clé en contexte », examinez comment ces mots sont utilisés. S'ils sont généralement précédés d'une négation (à quatre mots ou moins), vous pouvez conserver ces mots dans le dictionnaire, car WordStat contient des règles pour les prendre en compte.

3. Ajouter les mots et les phrases des sentiments spécifiques au domaine - Souvent, il y a des mots spécifiques à votre domaine qui sont utilisés pour désigner des aspects ou des caractéristiques positifs ou négatifs. Par exemple, si vous produisez des téléphones intelligents, des éléments comme « empreinte », « friture » ou « qualité sonore » peuvent être fortement associés à des commentaires positifs ou négatifs. Pour les constructeurs automobiles, les expressions « angle mort », « plastique dur » ou toute mention de « bruit », de « vent » ou de « jambes » peut également être liée à des opinions négatives sur un modèle spécifique de voiture. Si vous avez accès à une collection d'évaluations positives et négatives, un moyen simple d'identifier ces mots spécifiques au domaine serait de corréler les mots et groupes de mots les plus fréquents avec les scores de satisfaction et d'identifier ceux qui prédisent fortement les scores négatifs et positifs. Il existe cependant un piège à éviter lorsqu'on choisit ces prédicteurs en fonction de leur corrélation élevée avec les scores de satisfaction : la mesure de sentiment obtenue peut devenir insensible aux changements. Par exemple, si beaucoup de gens se plaignent de la mauvaise qualité sonore d'un téléphone cellulaire, alors l'expression « qualité sonore » sera probablement prédictive de commentaires négatifs. Si, en réaction à ces évaluations, le fabricant publie une nouvelle version avec une qualité sonore améliorée, tous les nouveaux commentaires positifs concernant cette qualité sonore améliorée peuvent être mal classés comme négatifs. Ce manque de sensibilité aux changements est un défaut potentiel de l'analyse automatique de sentiments par apprentissage machine qu'il est possible d'éviter en combinant l'apprentissage machine au vocabulaire contrôlé du dictionnaire de sentiments.

Pour un web-démo sur vos données
ou pour plus d'informations :

(+1) 514 899-1672 ou

sales@provalisresearch.com

