



Travail de Bachelor 2010

**Filière Informatique de gestion**

**Entre Web 2.0 et 3.0 : opinion mining**

Etudiante : Carol Hermann

Professeur : Anne Le Calvé

## TABLE DES MATIERES

Table des illustrations.....	4
Resumé.....	7
1 Introduction.....	8
1.1 Objectifs généraux.....	8
1.2 Cahier des charges.....	9
1.2.1 Prototype.....	9
1.3 Gestion de projet.....	10
1.3.1 Planification et suivi de projet.....	13
1.3.2 Organisation du rapport.....	14
2 Etat de l'art.....	15
2.1 Concepts linguistique.....	15
2.1.1 Catégorisation de textes.....	15
2.2 Historique.....	18
2.3 Méthodes d'opinion mining.....	21
2.3.1 Méthodes Statistiques.....	21
2.3.2 Méthodes Symboliques.....	26
2.3.3 Méthodes Hybrides.....	27
2.4 Ressources.....	28
2.4.1 Librairies.....	28
2.4.2 Projets.....	33
2.4.3 Logiciels.....	35
2.4.4 Entreprises de services.....	38
2.4.5 Ressources Linguistiques.....	39
2.5 Les Challenges de l'opinion mining.....	40
2.6 Conclusion.....	41
2.6.1 Le futur.....	41
3 Evaluation.....	43
3.1 UIMA.....	43
3.1.1 SWOT.....	44
3.2 RapidMiner.....	45
3.2.1 SWOT.....	46
3.3 GATE.....	46
3.3.1 SWOT.....	48
Conclusion.....	48
4 GATE.....	49
4.1 Architecture.....	49
4.2 IDE.....	50
4.3 Plugins CREOLE.....	54
5 Prototype.....	59
5.1 Collecte d'informations.....	61
5.1.1 Structure des pages Internet.....	61
5.1.2 Plugin Comments Crawler.....	64
5.2 Traitement des données.....	67
5.2.1 Ressources standard.....	68

5.2.2 Ressources spécifiques .....	75
5.2.3 Lancement automatique .....	80
5.3 Evaluation et tests.....	84
5.3.1 Critères d'évaluations.....	84
5.3.2 Résultats .....	85
5.3.3 Améliorations .....	91
6 Conclusion.....	92
7 Déclaration sur l'honneur.....	92
8 Glossaire et liste des abréviations .....	93
9 Sources .....	95
9.1 Webographie .....	95
9.1.1 Etat de l'art.....	95
9.1.2 GATE .....	97
9.1.3 Prototype .....	97
Developpement.....	98
Prototype .....	98
10 Annexe .....	99
I Categories du corpus de Brown .....	99

## TABLE DES ILLUSTRATIONS

Figure 1 Planification Initiale.....	11
Figure 2 Estimation en pourcentage .....	12
Figure 3 Suivi du projet .....	13
Figure 4 Processus de traitement selon une méthode statistique .....	21
Figure 5 Schéma du classifieur selon SVM .....	23
Figure 6 Exemple de marge selon SVM.....	24
Figure 7 Traitement par SVM d'un cas non linéaire.....	24
Figure 8 Calculs des résultats d'une méthode statistique.....	25
Figure 9 Architecture du projet SEMPRES .....	34
Figure 10 Schéma d'un processus de traitement dans UIMA .....	43
Figure 11 Architecture de RapidMiner.....	45
Figure 12 Schéma d'un processus de traitement dans GATE.....	47
Figure 13 Architecture du Framework GATE.....	49
Figure 14 IDE GATE.....	50
Figure 15 Projet ANNIE.....	51
Figure 16 Processing Resources du projet ANNIE .....	52
Figure 17 Execution du projet ANNIE.....	52
Figure 18 Résultat du processus ANNIE .....	53
Figure 19 Console de gestion des plugins CREOLE .....	55
Figure 20 BootStrapWizard .....	56
Figure 21 Structure d'un plugin GATE .....	57
Figure 22 Classe du plugin après sa création par le Wizard .....	57
Figure 23 Fichier creole.xml après la création par le Wizard .....	58
Figure 24 Architecture des processus du prototype .....	60
Figure 25 : Page de commentaires d'hôtel.....	62

## Entre Web 2.0 et 3.0 : opinion mining

Figure 26 Bloc de commentaire complet .....	63
Figure 27 Bloc de navigation dans les pages de commentaires.....	64
Figure 28 Détail du fichier creole.xml pour le WebCrawler .....	65
Figure 29 Pipeline d'extraction .....	66
Figure 30 Liste des ressources utilisée par le pipeline de traitement linguistique .....	67
Figure 31 Résultat du Tokenizer.....	69
Figure 32 Résultat du Sentence Splitter.....	70
Figure 33 Résultat du POS Tagger de LingPipe.....	71
Figure 34 Paramétrage d'un gazetteer.....	72
Figure 35 Résultat du gazetteer standard d'ANNIE.....	72
Figure 36 Extrait du fichier SentiWordNet .....	73
Figure 37 Extrait du fichier SentiWordNet après transformation.....	73
Figure 38 Résultat de l'élément Gazetteer .....	74
Figure 39 Résultat des deux transducers .....	75
Figure 40 Exemple de mots possédant plusieurs annotations.....	76
Figure 41 Pattern des groupes de mots porteurs d'opinion .....	76
Figure 42 Résultat du POS extender .....	77
Figure 43 Résultat du Negation handler sur le verbe expect.....	78
Figure 44 Résultat du Opinion Calculator .....	79
Figure 45 Résultat du Mark Opinion Sentence .....	80
Figure 46 Code Java de chargement des modèles .....	81
Figure 47 Extrait du résultat xml d'un fichier annoté au format GATE .....	81
Figure 48 Extrait du fichier XML de résultat.....	83
Figure 49 Exemple de l'analyse d'un hôtel.....	86
Figure 50 Résultat de l'opinion .....	87
Figure 51 Evaluation des valeurs d'opinions.....	88
Figure 52 Valeurs des coefficients pour le calcul de la moyenne .....	89
Figure 53 Valeurs des coefficients pour le calcul de la moyenne par opinion.....	90

Entre Web 2.0 et 3.0 : opinion mining

## RESUME

L'opinion mining (ou fouille d'opinion en français) est un domaine en pleine expansion depuis l'émergence du web participatif (Web 2.0), toutes les sociétés étant demandeuses d'informations concernant les opinions des consommateurs, on peut se rendre compte du fort potentiel de ce domaine naissant en terme commercial.

Le premier objectif de ce travail est tout d'abord d'analyser l'état de l'art de l'opinion mining : les études passées et en cours, les méthodes de traitement possibles, ou encore les ressources disponibles.

Sur la base de cette analyse le deuxième objectif est de développer un prototype, basé sur un scénario de e-Tourism d'analyse de commentaires d'hôtels, qui implémente un processus d'opinion mining. Le logiciel GATE, véritable boîte à outil de traitement du langage naturel a été utilisé pour cette partie de développement.

Le prototype est constitué d'un premier processus d'extraction des données depuis des pages Internet et d'un second processus de traitement de ces données. Les résultats du traitement sont ensuite sauvés au format XML.

## 1 INTRODUCTION

L'opinion mining (ou fouille d'opinion en français) est un domaine qui a pris énormément d'importance depuis l'émergence du web participatif (Web 2.0). Tous les utilisateurs d'Internet ont dès lors la possibilité de donner leurs avis sur une infinité de sujets (produits, services, prestations d'entreprise) et d'en discuter sur des forums, blogs, ou des groupes de discussions. Toutes ces opinions sont cruciales pour les entreprises, tout d'abord afin de développer des tâches de veille (technologique, marketing, concurrentielle, sociétale) mais aussi pour détecter les consommateurs mécontents, déterminer la perception des gens sur de nouveaux produits ou se renseigner sur la réputation de leur société. Toutes ces données peuvent aussi servir aux acteurs du monde politiques afin de connaître les intentions des citoyens. Mais ces informations sont évidemment tout aussi utiles aux privés afin d'étudier le marché et ainsi pouvoir faire un choix parmi plusieurs produits similaire. Sans outils appropriés, il faudrait lire des centaines de rapports textuels, d'articles de journaux ou d'entrée dans des blogs pour pouvoir en extraire manuellement les éléments subjectifs. On se rend compte ainsi de tout l'intérêt de tels outils. Toutefois il est encore difficile pour les machines de pouvoir analyser ces données correctement. Toute la difficulté de tels traitements provient de la complexité du langage humain, ainsi que de la diversité des langues.

Récemment des recherches et des applications sur ce sujet ont commencés à être mises en place, aux vues du potentiel commercial énorme qu'ils peuvent engendrer. Et c'est donc sous ce terme d'opinion mining ou sentiment analysis que sont rassemblées toutes les méthodes qui vont permettre d'identifier les opinions dans un ensemble de textes.

### 1.1 OBJECTIFS GENERAUX

Les objectifs généraux de ce travail de Bachelor sont :

1. d'analyser l'état actuel des recherches, les solutions existantes sur l'opinion mining et de rédiger un état de l'art
2. sur la base de l'analyse, de développer un prototype qui intègre un processus de détection de l'opinion (positive, négative ou neutre) en s'appuyant sur un scénario, proposé par e-Tourism : l'analyse des évaluations d'hôtels provenant d'un site web.



## 1.2 CAHIER DES CHARGES

En début de projet, il a paru évident qu'un cahier des charges détaillé ainsi qu'une planification complète ne pouvait être définis avant d'avoir pu prendre en compte les résultats de la phase d'analyse. Le projet a donc été grossièrement divisé en trois parties (Analyse, Développement et Rédaction du rapport). Une fois terminée la phase d'analyse le cahier des charges concernant le prototype et la planification ont été repris et détaillés.

---

### 1.2.1 PROTOTYPE

Le prototype se présente sous la forme d'un projet Java intégrant le Framework du logiciel GATE (l'outil de traitement du langage naturel sélectionné durant la phase d'analyse).

Il permet, en lui donnant en paramètre l'url d'une page d'un hôtel du site [www.TripAdvisor.com](http://www.TripAdvisor.com), d'extraire tous les commentaires en langue anglaise. Chacun de ces commentaires est ensuite traité pour déterminer un score représentant la valeur d'opinion de ce commentaire. La moyenne de ces score va ensuite donner la valeur d'opinion de l'hôtel.

Les résultats de chaque hôtel sont regroupés ensuite dans un fichier XML pour en faciliter la lecture et l'utilisation.

A noter qu'un module java d'extraction de données d'une page web, résultat d'un précédent travail a été fourni en début de projet pour être intégré au prototype.

### 1.3 GESTION DE PROJET

Le projet à été découpé en trois phases distinctes : Analyse, Développement et Rédaction du rapport.

La phase d'analyse a été subdivisée en trois parties pour prendre en compte la ré-estimation de la phase de développement en cours de projet.

La partie Développement comprend une première étape d'extraction des données, le programme Java fournit pour le projet a permis de raccourcir le temps passé sur cette tâche mais pas de la supprimer complètement car des modifications ont du lui être apportées pour pouvoir s'adapter au site [www.tripadvisor.com](http://www.tripadvisor.com). Dans la phase de traitement pure des données textuelles, il a été décidé de n'effectuer que deux manipulations sur les commentaires, l'attribution d'une valeur au mot et le calcul de cette valeur au niveau du commentaire. Aux vues des premiers résultats, il a parut impossible d'arriver à des résultats satisfaisants sans prendre en compte la négation. Ce nouveau traitement a induit un dépassement du temps attribué à cette tâche durant le projet. Deux autres tâches font partie du développement : le traitement des données en sorties, soit la création d'un fichier XML récapitulatif de tous les commentaires pour un hôtel et la création d'un programme pour pouvoir se passer de l'interface graphique de GATE.

## Planification Initiale

	Heures
<b>TD Opinion Mining</b>	<b>343</b>
<b>Analyse</b>	<b>116</b>
Cahier des charges	5
Recherche	105
Modification du cahier des charges	6
<b>Developpement</b>	<b>182</b>
Extraction des données	30
Traitement des données	
Attribution d'une valeur d'opinion au mot	50
Calcul de la valeur d'opinion pour le texte	40
Stockage des résultats en xml	7
Lancement du processus automatique	30
<b>Analyse des résultats</b>	
Amélioration du processus	25
<b>Rapport</b>	<b>45</b>
Rédaction	42
Elements à rendre	3

Figure 1 Planification Initiale

Entre Web 2.0 et 3.0 : opinion mining

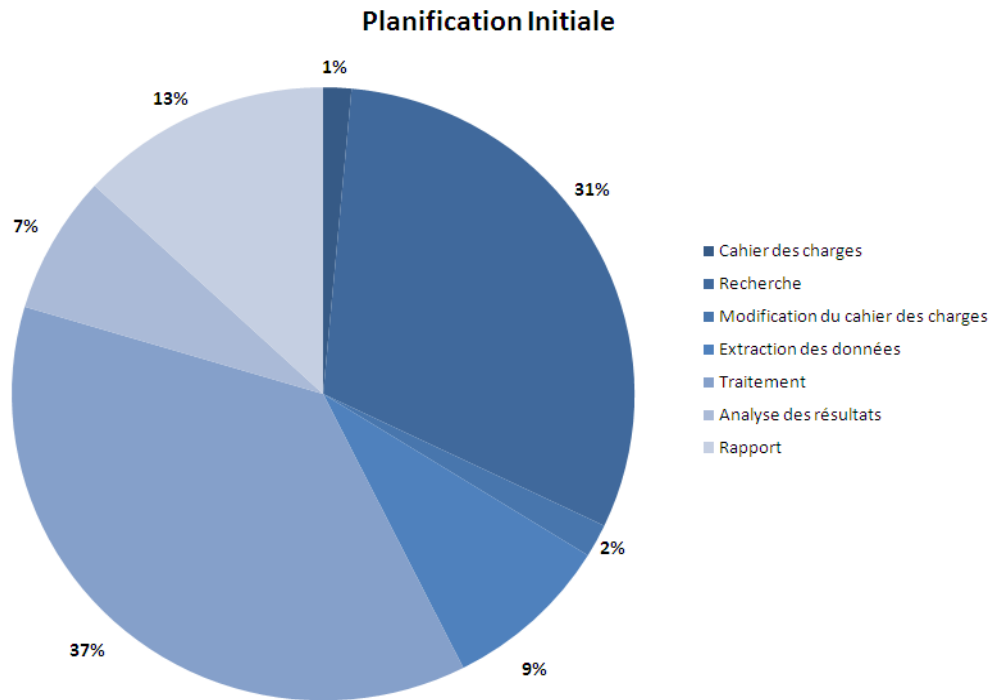


Figure 2 Estimation en pourcentage

Le schéma ci-dessus permet de se rendre compte de la part des différentes estimations sur le projet global.

1.3.1 PLANIFICATION ET SUIVI DE PROJET

Le projet s'est déroulé à temps partiel entre la semaine 53 de l'année 2009 et la semaine 28 de l'année 2010, et impliquait une charge de travail d'environ 360 heures.

Le tableau ci-dessous détaille le suivi des heures passées sur chaque étape durant ce laps de temps :

	Heures	Mois / Semaine																																	
		Decembre				Janvier				Février				Mars				Avril				Mai				Juin					Juillet				
		49	50	51	52	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
<b>TD Opinion Mining</b>	<b>343</b>																													<b>361</b>					
<b>Analyse</b>	<b>116</b>																													<b>117</b>					
Cahier des charges	5	5																																	5
Recherche	105	3	9		8	4	5		30	10	3	5	5	12	8																				106.5
Modification du cahier des charges	6				1																5														6.5
<b>Developpement</b>	<b>182</b>																													<b>201</b>					
Extraction des données	30	5																								3	24	5						37	
<b>Traitement des données</b>																																			
Attribution d'une valeur d'opinion au mot	50														11	12	10		20								10							63	
Calcul de la valeur d'opinion pour le texte	40																			20						8	8							36	
Stockage des résultats en xml	7																											5							5
Lancement du processus automatique	30																								14	2		2	10						28
<b>Analyse des résultats</b>																																			
Amélioration du processus	25																											20	10	2				32	
<b>Rapport</b>	<b>45</b>																													<b>43</b>					
Rédaction	42																											5	20	10	8			43	
Elements à rendre	3																																		0

Figure 3 Suivi du projet

Le rapport détaillé des heures hebdomadaire se trouvent en annexe sur le CD fournit avec ce rapport.

### 1.3.2 ORGANISATION DU RAPPORT

Le contenu du présent rapport se présente de la manière suivante. Dans le second chapitre nous expliquerons les notions générales liées à l'opinion mining, notamment l'aspect linguistique et les différentes méthodes utilisées pour la détection des sentiments. Il a pour but non seulement de résumer l'état actuel des recherches, de faire une synthèse des méthodes déjà proposées et d'envisager les implications de ces solutions, mais aussi de prendre position dans le choix des outils notamment pour la suite du rapport.

Dans ce second chapitre trois outils ont particulièrement été mis en lumière, le troisième chapitre va donc permettre de les étudier plus en détail afin de déterminer lequel est le plus adapté pour le développement du prototype.

Le quatrième chapitre explique de manière plus détaillée le fonctionnement du logiciel sélectionné. Pour se familiariser avec les aspects techniques, spécifiques au logiciel et qui seront utilisés dans le développement du prototype.

Ce développement ainsi que les résultats obtenus sont approfondis dans le chapitre cinq.

Le chapitre six conclut le rapport en faisant une synthèse des recherches et des résultats.

Une liste de termes techniques sont définis dans le glossaire, au chapitre 8 et finalement toutes les sources utilisées pour la rédaction des différents chapitres de ce rapport sont présentées dans le chapitre neuf.

## 2 ETAT DE L'ART

*“Sentiment analysis is a set of algorithms and tools for identifying and extracting a) features that express attitudes or opinions, b) attributes that indicate sentiment polarity, intensity, and other characteristics, and c) the topics those sentiments and attributes apply to.”* (Seth Grimes 2009, Stratégiste analytique chez AltaPlana Corporation et organisateur du Sentiment Analysis Symposium)

L'opinion mining est un domaine qui chevauche la linguistique, la statistique et l'informatique. Pour pouvoir saisir tous les atouts, une brève explication des termes linguistiques utilisés dans la détection de sentiments est présentée ci-dessous. Un historique des différentes recherches effectuées sur le sujet viendra poser les bases des trois types de méthodes existantes qui seront décrites par la suite. Les challenges qui jalonnent ces recherches seront mis en évidence dans la partie problématique. Cette analyse conclura finalement par l'actualité des recherches sur le sujet.

### 2.1 CONCEPTS LINGUISTIQUE

Afin de mieux comprendre la notion d'opinion il convient d'en donner tout d'abord une définition. L'opinion est une manière de penser sur un sujet ou un ensemble de sujets. C'est un jugement personnel que l'on porte sur une question, qui n'implique pas que ce jugement soit obligatoirement juste (Wikipedia).

De cette définition découle la première problématique de la détection d'opinion, c'est-à-dire de pouvoir déterminer la subjectivité du texte, sans prendre en compte sa partie objective ou factuelle, et là se situe toute l'ambiguïté de la fouille d'opinion. Selon le contexte, le mot « grand » peut tantôt être factuel (la taille d'un individu), tantôt exprimer une opinion (la qualité cinématographique d'un film). La prise en compte du contexte des mots a donc son importance pour pouvoir déterminer leur sens.

#### 2.1.1 CATEGORISATION DE TEXTES

La première approche de toutes les méthodes de fouilles d'opinion est de tenter de représenter une unité textuelle (un document, un paragraphe, une phrase) par un ensemble prédéfini de caractéristiques linguistiques - ou **traits** - puis d'utiliser la fréquence de ces traits pour décider de la catégorie d'un texte. Selon la problématique, les traits peuvent être variables et de l'attention portée au choix des meilleurs traits pour décrire l'unité textuelle va découler la précision des résultats. Les traits traditionnellement considérés appartiennent aux cinq niveaux d'abstraction linguistique suivants :

Entre Web 2.0 et 3.0 : opinion mining

**Le niveau Lexical**

Un texte peut être représenté par un vecteur de mots simples (ou plus précisément de lemmes), dont la fréquence des termes peut être considérée. Dans certains corpus, des recherches ont fait remarquer que la présence ou l'absence d'un mot semble efficace comme indicateur de subjectivité de même que la présence d'hapax ou de termes rares.

**Le niveau Morphosyntaxique**

Les informations sur la forme et la fonction d'un mot, notamment sa catégorie grammaticale (adjectif, verbe, nom,...) va permettre de façon brute de désambiguïser certains termes. En fouille d'opinion, les adjectifs sont particulièrement employés, mais les noms et les verbes sont également des indicateurs de subjectivité.

**Le niveau des Relation syntaxique**

Les différentes relations (coordination, subordination) entre les unités d'un texte sont considérées comme de bons traits, entre autre pour améliorer les résultats qui pouvaient être faussés par l'utilisation de tournures négatives ou diminutives.

**Le niveau Sémantique**

Des recherches sont en cours sur la manière de caractériser des expressions évaluatives complexes en considérant des classes sémantiques propres à l'évaluation : modalité évaluative (affect, appréciation, jugement), l'orientation axiologique (positif, négatif), intensité ou engagement du locuteur. Mais les outils informatiques prenant en compte ces recherches sont pour le moment assez rares.

**Le niveau Enonciatif**

La position d'un terme dans une unité textuelle (début, milieu ou fin d'un texte), peut potentiellement avoir un effet sur la façon dont ce terme affecte le sens général du texte.



### 2.1.1.1 LANGAGE EVALUATIF

L'évaluation peut se voir comme la « rupture de l'indifférence ». Le langage évaluatif exprime donc l'idée de préférence et d'opinion. Il incorpore différents éléments linguistiques qui peuvent permettre la détection et l'analyse de la subjectivité. Ces éléments découlent de la théorie de l'Appraisal qui suggère que les opinions et les sentiments en générale proviennent des évaluations cognitives personnelles que fait un individu face à un événement.

Le premier élément est le **vocabulaire de l'opinion** ou l'ensemble des mots qui expriment l'opinion (en général, les mots qui vont créer la rupture de l'indifférence).

Ce vocabulaire peut être subdivisé en catégories représentant la **modélisation**. La modélisation se définit comme l'inscription dans le discours de la prise d'attitude du locuteur à l'égard du contenu d'un énoncé. Les cinq catégories de modalité de l'évaluation sont :

- **l'opinion** : un fait présumé est évalué par le locuteur qui révèle du même coup son point de vue.
- le **jugement** (favorable ou défavorable) : le locuteur pose dans son énoncé, une action réalisée et juge si cet acte est bon ou mauvais.
- **l'accord** ou le **désaccord** : on présume qu'une demande a été adressée au locuteur qui va dire s'il adhère ou non à la vérité d'un propos tenu par un autre.
- **l'appréciation** (jugement intellectuel) : le locuteur évalue la valeur d'un fait en révélant ses propres sentiments selon son champ d'appréciation.
- **l'acceptation** ou le **refus** : on présume que qu'une demande d'accomplissement d'un acte a été adressée au locuteur qui peut répondre favorablement ou non à cette demande.

Dans chacune des ses modalités, la valeur peut être graduée selon sa force : soit du négatif au positif dans les cas de l'opinion, du jugement ou de l'appréciatif soit du plus ou moins pour l'accord/désaccord et acceptation/refus.

A ces modalités viennent se combiner trois types de vocabulaire de l'opinion :

- le vocabulaire **affectif** (l'ensemble des mots impliquant une réaction émotionnelle : effrayant, pleurer, séduire, étonnant)
- le vocabulaire **appréciatif** (impliquant un jugement de valeur positif ou négatif : intéressant, de qualité, trop cher, banalité)
- le vocabulaire **connoté** (mots possédant une signification affective en plus de son sens premier : force, liberté, original)

Entre Web 2.0 et 3.0 : opinion mining

Le deuxième élément du langage évaluatif est **le porteur de l'opinion**, c'est-à-dire la personne qui au travers le texte exprime son ressenti. Il peut s'agir de l'auteur du texte, ou bien d'une autre personne, lorsqu'un discours est rapporté. Le nombre de porteurs de l'opinion n'est d'ailleurs pas limité à un seul. Dans l'exemple de commentaires sur un site Internet, le porteur est souvent l'auteur lui-même.

Le troisième élément est le **thème de l'opinion** ou sur quel sujet, produit ou idée cette opinion est portée. Le thème peut être soit général, soit comporter des sous-thèmes. Dans le cas des critiques d'hôtel, celui-ci est bien le thème principal, puis chacune de ses caractéristiques (les chambres, l'accueil, etc..) sont autant de sous-thèmes.

Le dernier élément est l'**argumentation de l'opinion**. C'est la définition de l'organisation des différents arguments pour exprimer l'opinion. De manière générale une argumentation est mise en place pour justifier son point de vue ou réfuter un autre et ainsi essayer d'influencer autrui. Les arguments peuvent se présenter sous la forme de faits objectifs mais plus souvent aussi par des idées ou des croyances subjectives. Tous ses arguments peuvent être reliés par des connecteurs logiques (et, de plus, aussi, car, premièrement,..) qui sont aussi des indications de gradation de l'opinion. Ces connecteurs (qui relient non seulement les mots dans un phrase mais aussi les phrases dans un texte) peuvent être :

- de causalité (grâce à)
- de concession (malgré)
- d'opposition hypothétique (même si)
- de conséquence (provoque)
- de but (à seule fin)

## 2.2 HISTORIQUE

Même si les recherches en linguistiques existent depuis des années ce n'est véritablement qu'à partir des années 2000 (et l'essor d'Internet) que ces techniques ont été utilisées par l'informatique. Ci-dessous, par ordre chronologique, sont listés les chercheurs qui ont fait évoluer l'opinion mining, de telle manière que toutes les recherches actuelles se basent sur leurs résultats.

### **Banfield (1982)**

Dans son ouvrage « Unspeakable sentences : narration an representation in the language of fiction » Ann Banfield met en évidence les éléments du langage qui conduisent à la subjectivité d'un texte ou tout du moins d'une partie d'un texte.

### **Hatzivassiloglou et MckKeown (1997)**

Entre Web 2.0 et 3.0 : opinion mining

Ils sont les premiers à se pencher sur la classification des opinions. Dans leurs recherches ils mettent l'accent sur les adjectifs. Ils étudient plus précisément les phrases dans lesquels les adjectifs sont reliés entre eux par des conjonctions tels que « et » ou « mais » pour déterminer ainsi leur orientations l'un par rapport à l'autre.

**Wiebe (2000)**

Wiebe va affiner les remarques de Banfield en tentant de distinguer les phrases objectives des phrases subjectives. En annotant les adjectifs subjectifs d'un corpus et en leur assignant un score de un à trois, il va déterminer que pour chaque adjectif de score trois, vingt synonymes peuvent être découverts en utilisant des mesures de similarité ou le dictionnaire WordNet. Pour chaque adjectif l'orientation sémantique est ajoutée comme caractéristique. Ses résultats démontrent qu'une phrase est subjective à 55.8% si elle contient au moins un adjectif. Et si elle contient un adjectif noté 3 dans sa liste, le pourcentage monte à 71%.

**Pang (2002)**

Dans une étude qui permettait de classer les sentiments de commentaires de films, il est le premier à expérimenter l'apprentissage automatique. Cette méthode qui s'avérait bonne dans des cas de catégorisation de sujet, n'accomplissait pas d'aussi bons résultats pour la classification de sentiments. Il démontra aussi que la présence ou l'absence de mot pouvait être plus significatif que leur fréquence. Son travail a eu pour conséquence la création du corpus annotés de commentaire de films, qui est encore aujourd'hui souvent utilisé en tant que standard pour la recherche.

**Turney (2003)**

Pour ses recherches, Turney extrait d'un texte, des phrases de deux mots ou l'un est un adjectif et l'autre un adverbe. Il utilise ensuite le coefficient de corrélation et l'analyse sémantique latente pour mesurer les relations entre un mot et une liste de mots positifs ou négatifs. La somme des associations vers la liste des mots négatifs est soustraite à la somme des associations vers la liste positive. Enfin le commentaire est classé selon la moyenne de l'orientation des phrases qu'il contient.

**Hu & Liu (2004)**

Ils étudièrent les commentaires de produit sur des sites web afin de produire un résumé des déclarations positives et négatives des caractéristiques de ce produit. Pour obtenir ce résultat, ils vont identifier tout d'abord les caractéristiques discutées dans les commentaires en sélectionnant les mots apparaissant fréquemment, en supposant que les gens utilisent souvent les mêmes mots pour décrire la même caractéristique. Ils estimèrent aussi qu'une phrase porte une opinion seulement si elle contient à la fois une caractéristique et un adjectif (Pour cela ils utilisèrent une liste d'une trentaine d'adjectifs). Finalement chaque phrase est orientée selon la majorité des orientations des parties qu'elle contient.

**Kamp (2005)**

Entre Web 2.0 et 3.0 : opinion mining

Kamp lui aussi étudie les adjectifs dans la même logique que les recherches de Turney de 2003 à la différence qu'il utilise WordNet, le dictionnaire devenu un standard en tant que ressource linguistique, pour mesurer les « distances » sémantiques entre les adjectifs présents dans le texte. Il propose aussi trois nouvelles mesures : une mesure évaluative (good/bad), une mesure de puissance (strong/weak), et une mesure d'activité (active/passive)

**Esuli (2005)**

Esuli présente en 2005 une méthode qui améliore grandement les résultats obtenus jusqu'alors. Sa méthode se base sur la présomption que les termes avec des orientations similaires ont des gloses similaires. Il présente ainsi les termes d'un texte en tant que des vecteurs de gloses et leur ajoute un point calculé par tf-idf. Ses recherches ont eu pour résultat la création du dictionnaire SentiWordNet, une ressource lexical qui attribue à chaque sens des mots présents dans le dictionnaire WordNet, un score de positivité, de négativité et d'objectivité.

**Ku (2007)**

Ku a écrit un certain nombre d'ouvrages relatifs à l'opinion mining pour les langues chinoise et anglaise. Il illustre le fait que les opinions extraites des textes n'ont de l'importance que si elles concernent les sujets à évaluer. C'est pour cette raison qu'il met en place une détection de sujets avant tout traitement de texte, ensuite seulement il extrait les mots porteurs d'opinion. L'orientation des mots est ensuite calculée selon une formule qui prend en compte les caractéristiques de la langue.

**Ding (2007)**

Ses recherches améliorent les résultats de Hu et Liu de 2004 en attribuant un score d'orientation à chaque mot portant une opinion trouvé dans la phrase. Ce score prend aussi en compte l'orientation du mot le plus proche porteur lui aussi d'une opinion ainsi que la distance entre le mot et la caractéristique qu'il décrit. Dans ce cas un score moindre est donné aux mots les plus éloignés de la caractéristique.

## 2.3 METHODES D'OPINION MINING

Actuellement, deux types de méthodes sont utilisés dans l'opinion mining, les méthodes statistiques et les méthodes symboliques.

Une troisième manière de faire combine les deux précédente méthodes et est nommée hybride.

### 2.3.1 METHODES STATISTIQUES

Les méthodes statistiques appelées aussi classification supervisée ou encore basée sur corpus est le fait de regrouper les documents (ou les mots) dans deux classes (deux classes objectif/subjectif ou deux classes positif/négatif par exemple). Elles utilisent un corpus qui a été déjà été annoté manuellement au préalable, dans le but de le faire apprendre au système.

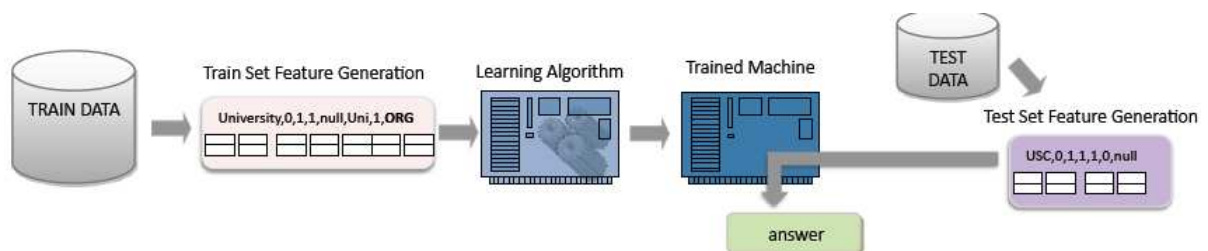


Figure 4 Processus de traitement selon une méthode statistique

La première étape dans une méthode statistique est la catégorisation d'un texte en le transformant en un "sac de mot" avec différentes caractéristiques (fréquence des mots, distances entre les sens, catégorie grammaticale des mots, taille des phrases, etc..). Une fois les caractéristiques définies et déterminées elles sont annotées manuellement (pour en valider la véracité) dans les documents du corpus d'entraînement.

Ce corpus est ensuite traité par un algorithme d'apprentissage automatique.

Après sa période d'apprentissage, un nouveau lot de données de test, catégorisées de la même manière que le lot d'entraînement, va être traité et, sur la base des résultats obtenus précédemment, ils vont être classés au bon endroit selon les éléments déjà rencontrés (ou qui s'en approche)

Les méthodes non supervisées donnent de bons résultats mais elles demandent un grand travail manuel d'annotation des documents au préalable et est très dépendante du domaine auquel appartiennent les

## Entre Web 2.0 et 3.0 : opinion mining

documents (un apprentissage sur des commentaires de films ne donnera pas de bon résultat pour un corpus de test contenant des commentaires sur des voitures par exemples). La principale raison étant qu'une même phrase peut avoir plusieurs sens dans différents domaines. Cependant pour les détections d'opinion dans un contexte multi-domaine, les recherches ont démontré que les traits qui sont de bons indicateurs de subjectivité dans au moins deux domaines sont considérés comme des traits indépendants au domaine.

Il existe de nombreux classifieur utilisant des algorithmes différents notamment :

- SVM (Support Vector Machine)
- K plus proches voisins
- Bayes naïf
- ...

Les résultats des études démontrent que la classification de texte avec SVM donnent les meilleurs résultats.

---

### 2.3.1.1 SVM

Le but de SVM est de trouver un classifieur qui va séparer les données en gardant la plus grandes distance entre cette séparation et les données. Ce classifieur est linéaire et est appelé hyperplan (H).

La figure 5 montre un schéma où chaque point représente un document et où l'hyperplan sépare les deux classes :

## Entre Web 2.0 et 3.0 : opinion mining

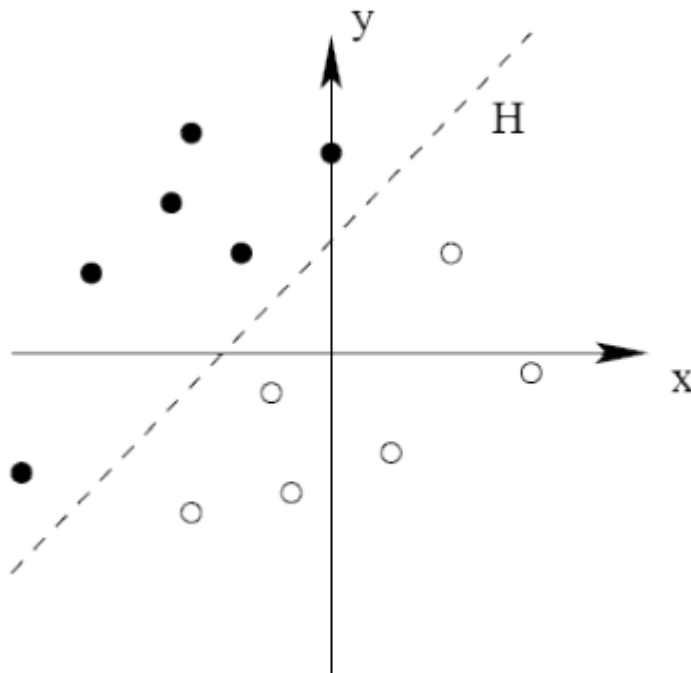


Figure 5 Schéma du classifieur selon SVM

Les points les plus proches de l'hyperplan sont nommés vecteurs de support. Comme il existe une multitude d'hyperplans valides, la propriété remarquable de SVM est de rechercher l'hyperplan le plus optimal en prenant en compte une marge. Cette marge, qui est la distance entre l'hyperplan et les données d'apprentissage les plus proches, doit être la plus grande possible.

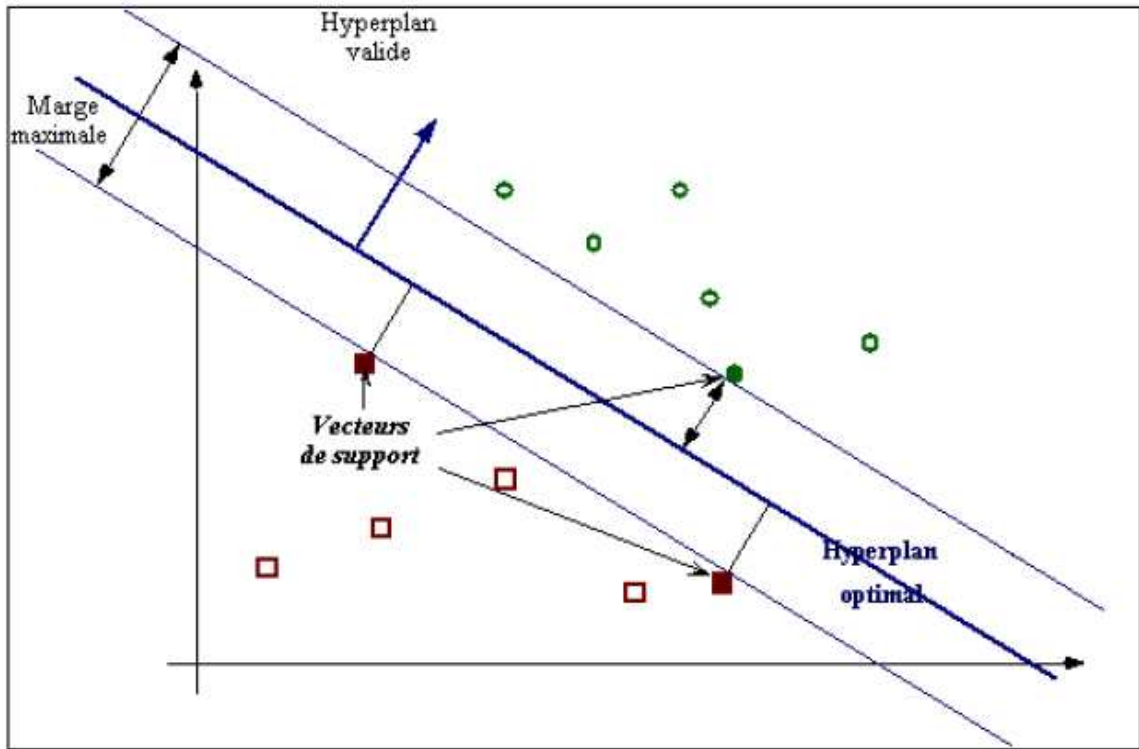


Figure 6 Exemple de marge selon SVM

Il se peut que l'hyperplan ne soit pas linéaire. Dans ce cas SVM va changer l'espace des données, en rajoutant une dimension, afin de permettre une séparation linéaire. Cette nouvelle dimension est appelé « espace de re-description ».

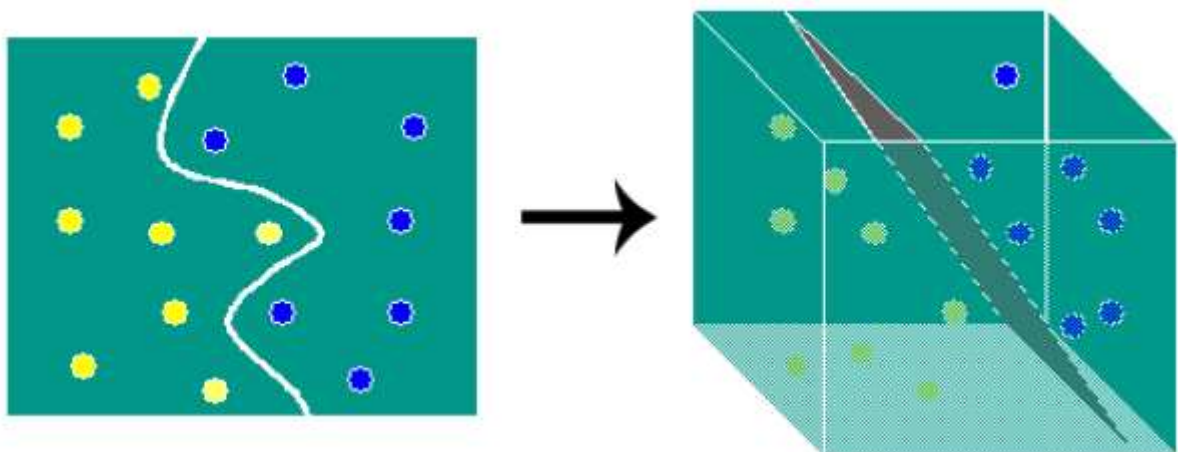


Figure 7 Traitement par SVM d'un cas non linéaire



## Entre Web 2.0 et 3.0 : opinion mining

Ainsi quelque soit les classes, SVM va toujours trouver un hyperplan séparateur entre les exemples. Et si deux dimensions ne suffisent pas, l'algorithme en rajoutera d'autres.

Le document sera ainsi classé d'après sa situation par rapport à l'hyperplan.

Il est possible aussi de traiter plus de deux classes : soit en les comparant deux à deux soit en recherchant l'hyperplan séparant chaque classe de toutes les autres.

### 2.3.1.2 MESURE DE PERFORMANCE

Dans les méthodes non supervisée, l'analyse des performances est toujours la même. Les résultats de la machine sont comparés aux résultats découverts par un être humain.

	Machine says yes	Machine says no
human says yes	tp	fn
human says no	fp	tn

**Table 1:** A confusion table

$$\text{Precision}(P) = \frac{tp}{tp+fp}; \quad \text{Recall}(R) = \frac{tp}{tp+fn}; \quad \text{Accuracy}(A) = \frac{tp+tn}{tp+tn+fp+fn}; \quad F_1 = \frac{2 \cdot P \cdot R}{P+R}$$

**Figure 8** Calculs des résultats d'une méthode statistique

La précision est calculée en en divisant le nombre de résultats corrects d'une des classes par la somme des résultats (correctes et incorrectes) de cette même classe. Une précision de 1 correspond à une réussite de 100% sur les résultats corrects, mais ne signifie pas que tous les résultats corrects aient été trouvés.

Le rappel est quand à lui le calcul du nombre de résultats corrects d'une classe divisé par la somme des résultats déterminés par l'être humain pour cette classe. Un rappel de 1 signifie que tous les éléments de la classes on été détectés mais pas que des éléments ne faisant pas partie de la classe y aient été intégrés.

L'exactitude des résultats et la somme des résultats corrects (toutes classes confondues) divisée par la somme de tous les résultats pour toutes les classes.

Les deux premiers calculs (la précision et le rappel) permette de mesurer le score F1 qui en statistique est le mesure de la précision d'un test.

### 2.3.2 METHODES SYMBOLIQUES

Les méthodes symboliques ou non supervisée ou encore basée sur dictionnaire ne font pas appels à l'apprentissage mais repèrent, évaluent et font une synthèse des expressions d'opinion dans un texte.

Lors de repérage, les termes d'opinion et les thèmes porteurs d'opinion sont extraits du texte. Ces termes peuvent être déterminés soit d'après des listes de vocabulaire ou des dictionnaires en ligne comme WordNet et son extension pour l'opinion : SentiWordNet.

La phase d'évaluation étiquète les mots extraits avec une certaine valeur (positive ou négative par exemple).

Enfin durant la synthèse, la somme des valeurs est calculée, pour permettre de trouver la valeur globale d'abord au niveau de la phrase puis au niveau du document.

Ces méthodes peuvent manquer d'informations contextuelles.

L'approche essentiellement symbolique semble, dans l'état actuel de son développement, mal adaptée à la tâche de caractérisation de textes dans leur intégralité et inter-domaine. Il apparaît qu'elle est plus adaptée à des analyses locales, à l'intérieur des textes, ce pour quoi elle était destinée initialement.

La mise en œuvre qui en découle constitue une chaîne d'analyses symboliques, s'appuyant sur des ressources lexico-sémantiques ainsi que sur les configurations textuelles caractéristiques du discours évaluatif.

#### 2.3.2.1 INDICES LINGUISTIQUES

Pour repérer dans le document les éléments à tenir compte, il existe plusieurs traitements linguistiques possible (ou TAL pour traitement automatique des Langues) :

- la **segmentation** du document en phrases, paragraphes, ou mots (communément appelé en anglais : Sentence splitter et tokenizer)
- la catégorisation des mots (en anglais **POS tagging** pour Part-Of-Speech tagging) détermine le mot comme étant un verbe, un adjectif, un nom propre, un adverbe, etc..
- la **lemmatisation** grammaticale ou autrement dit la récupération de la racine du mot dans le cas de verbe conjugué ou de mots au pluriel
- la détermination des **StopWords** , des mots qui n'ont pas d'intérêt au niveau grammatical ou dans le cas de l'opinion mining, non porteur d'opinion.
- l'attribution de pondération, en rajoutant une certaine valeur lors de la découverte de termes d'évaluation, de modificateurs d'intensité (véritablement, très, etc..), ou de négation.

---

### 2.3.3 METHODES HYBRIDES

Comme on peut le constater, les deux méthodes présentées ci-dessus ont l'une comme l'autre des limitations. Tous l'intérêt des méthodes dites hybrides est donc de combiner les points forts des méthodes supervisées et non supervisées.

Les méthodes hybrides, nommées aussi méthodes semi-supervisées prennent donc en compte tout le traitement linguistique des méthodes symboliques avant de lancer le processus d'apprentissage comme dans les méthodes statistiques.

Actuellement toutes les recherches font appel à cette manière de faire.

## 2.4 RESSOURCES

Internet propose un grand nombre d'outils liés à la recherche d'opinion de manière très générale, moteurs de recherche spécifique (gratuit ou non), logiciels de traitement de la langue, logiciel d'apprentissage machine, web services, projets de recherche universitaire, etc...

On retrouve dans ce chapitre, regroupées par catégories, les principales ressources liées à l'opinion mining.

---

### 2.4.1 LIBRAIRIES

---

#### 2.4.1.1 FREELING

<http://www.lsi.upc.edu/~nlp/freeling/>

FreeLing fait partie du projet OpenNLP. C'est surtout une librairie même si un simple programme est aussi fourni en tant qu'interface basique à cette librairie. Développé en C++, FreeLing est multiplateforme mais les utilisateurs déploient principalement le projet sous Linux, il existe une version pour Windows mais elle n'est pas supportée.

#### **Fonctionnalités :**

- Tokenisation de texte
- Splitter de phrase
- Analyse morphologique
- Traitement des suffixes
- Reconnaissance de mots
- Splitter de contractions
- Prédiction probabiliste de mots inconnus
- Détection d'entités
- Reconnaissance de dates, nombres, ratios ou monnaies
- POS tagging
- Classification des entités
- Annotations des sens et désambiguïsation basé sur WordNet

**Licence :** GNU General Public License

---

#### 2.4.1.2 LINGPIPE

<http://alias-i.com/lingpipe/>

Ling pipe est une suite de bibliothèques Java pour l'analyse du langage humain, qui est utilisée dans un certain nombre de projets commerciaux, académiques mais aussi gouvernementaux. LingPipe peut être utilisé directement mais des fonctionnalités supplémentaires peuvent être développées en Java. Un tutorial et de nombreux projets de démonstration existent à ce sujet.

Lingpipe peut aussi être intégré au logiciel GATE via un plugin.

##### **Fonctionnalités :**

- Détection des entités
- Liaison entre ces entités et des entrées dans une base de données
- Liaison entre les entités et des faits
- Classement des passages de texte par langage, encoding, genre, sujet ou sentiment
- Correction orthographique
- Regroupement des documents dans des topics implicites et recherche des axes significatifs
- POS Tagger

**Licence :** LingPipe propose différentes licences (Developer, Startup, etc..) toutes payantes exceptée la « Royalty Free » qui permet cependant une utilisation en production limitée.

---

#### 2.4.1.3 MALLET

<http://mallet.cs.umass.edu/>

Mallet est un package de classes Java pour le traitement du langage naturel et la classification de document, en partie développées par Andrew McCallum. Même s'il est mentionné dans de certaines études, et qu'il paraît être un outil satisfaisant Mallet ne propose pas beaucoup de documentations et n'est pas soutenu par une importante communauté.

##### **Fonctionnalités :**

- Importation de données
- Classification
- Annotation de séquence
- Modélisation de topic
- Optimisation de fonction
- Modèles graphiques

**Licence :** Common public Licence

#### 2.4.1.4 MINORTHIRD



**Methods for Identifying Names and Ontological Relationships in Text using Heuristics for Identifying Relationships in Data**

<http://sourceforge.net/apps/trac/minorthird/wiki>

Collection de classes java pour le stockage, l'annotation de texte, l'apprentissage, l'extraction d'entités et la catégorisation de texte, Minor Third est à l'origine développé par l'IPTO (the information processing Technology Office) de la DARPA (Defence Advanced Research Projects Agency), une agence du département de la Défense des Etats-Unis qui est chargée de la recherche et du développement de nouvelles technologies destinées à un usage militaire.

A la différence d'autres logiciels MinorThird est vraiment orienté apprentissage. Il permet d'analyser les performances des classifieurs. En tant qu'outil open source il est disponible à des fins de recherches mais aussi commerciales

**Fonctionnalités :**

- Annotation de textes
- Stockage et Manipulation de ces textes
- Tokenizer
- Méthodes d'apprentissage automatique
- Analyseur de résultats

**Licence :** BSD Licence

---

#### 2.4.1.5 NLTK

##### Natural Language ToolKit

<http://www.nltk.org/Home>

NLTK est une collection de modules Python, de données linguistique et de documentation pour la recherche et le développement de traitement du langage naturel et l'analyse de textes. Il fait partie du projet OpenNLP. C'est un outil incontournable pour les développeurs Python.

##### Modules inclus :

- Corpus reader : interfaces pour de nombreux corpora
- Tokenizers
- Lemmatisation
- POS Tagger
- Interprétation sémantique
- Intégration de Wordnet
- Plusieurs Classifieurs

**Licence:** Licence Apache

---

#### 2.4.1.6 RITA.WORDNET

<http://www.rednoise.org/rita/wordnet/documentation/index.htm>

RITA.Wordnet est une librairie Java pour manipuler les données du dictionnaire WordNet, développée par Daniel C. Howe. Elle fait partie d'un projet plus vaste qui propose des outils pour la "Generative Literature" (la production de texte par un ordinateur et non pas par un auteur). La philosophie de tous les modules du projet RITA est de proposer des APIs simples et intuitives.

##### Fonctionnalités :

- Permet un accès simplifié à l'ontologie WordNet
- Encapsule des fonctionnalités Jawbone/JWNL pour Java (deux autres librairies java pour WordNet)
- Propose des mesures de distances entre les termes de l'ontologie
- Supporte les différentes versions de WordNet

**Licence :** Creative Commons Licence (Attribution-NonCommercial-ShareAlike 3.0 United States)

---

#### 2.4.1.7 UIMA



<http://uima.apache.org/>

A l'origine créé par IBM Research, UIMA est publié en 2006 sur SourceForge et à présent disponible sur le site d'Apache Software Foundation. UIMA permet l'analyse de large volume d'informations non structurées afin d'en ressortir les données utiles pour l'utilisateur final.

Il propose une architecture logicielle qui spécifie des interfaces de composants, des structures de données, des patterns de conception et une démarche de développement pour décrire, créer et déployer des capacités d'analyse d'information. Son objectif est de permettre l'utilisation et la construction d'applications distribuées visant l'analyse de contenus multimédias non structurés.

UIMA supporte Java et C++

#### **Fonctionnalités :**

- Un Framework qui permet tous les traitements linguistiques basiques d'un texte
- Toute une gamme d'outils "externes" développés par la communauté
- Un environnement d'exécution des processus développé

**Licence :** Licence Apache



## 2.4.2 PROJETS

---

### 2.4.2.1 NACTEM



<http://www.nactem.ac.uk/>

The **National Centre for text Mining**

Ce centre est le premier centre de text mining public dans le monde. Il est géré par l'Université de Manchester en collaboration avec l'université de Tokyo. Il propose des services de text mining à la disposition de la communauté académique du Royaume-Unis. Notamment des logiciels, développés ou non par le centre (UIMA est présenté sur le site par exemple), des séminaires, des conférences, des tutoriaux ainsi que des publications sur le sujet.

---

### 2.4.2.2 OPENNLP

<http://opennlp.sourceforge.net/>

OpenNLP est un centre pour les projets open sources en lien avec le langage naturel.

Son premier but est d'encourager et de faciliter la collaboration entre les chercheurs et les développeurs sur leurs projets. Un de ses objectifs est aussi de donner plus de notoriété aux projets qu'il répertorie et d'augmenter leur interopérabilité.

Il propose aussi de nombreux outils NLP de traitement du langage naturel.

Voici une liste non exhaustive des projets qu'on retrouve sur le site du centre :

- OpenNLP tools : une collection d'outils de traitement de langue naturelle
- WordFreak : un outil java d'annotation linguistique
- ComLinToo : une suite d'outils Perl pour la linguistique computationnelle
- Ellogon : un composant de traitement de langue naturelle écrit en C, C++, Java, Tcl, Perl et Python
- Emdros : un engine de base de données pour le texte analysé ou annoté

## 2.4.2.3 SEMPRES PROJETS



[http://www.ofai.at/research/nlu/projects/nlproject\\_sempre.html](http://www.ofai.at/research/nlu/projects/nlproject_sempre.html)

L'OFAI (Austrian Research Institute for Artificial Intelligence) a plusieurs projets en cours dont notamment « Sempres » (Semantically Aware Profiling for recommenders) dont l'objectif est d'exploiter les informations factuelles et les opinions humaines disponibles sur Internet.

Le résultat de ce projet est l'implémentation d'une architecture constituée de pipelines de traitement et de construction d'ontologies. Le projet utilise GATE pour les deux pipelines (un pour les données factuelles et un autre pour les opinions) ainsi que plusieurs plug-ins. Il se base sur des commentaires du domaine de la musique et du cinéma. Les données sont majoritairement extraites de ressources telles que Wikipedia, IMDB et la polarité des opinions est obtenue grâce à des corpus comme le Multi-Domain Sentiment Data Set, le Polarity Dataset et SentiWordNet (des corpus souvent utilisés lors des recherches linguistiques)

Les différents modules supplémentaires inexistant dans GATE et des scripts sont développés en Python

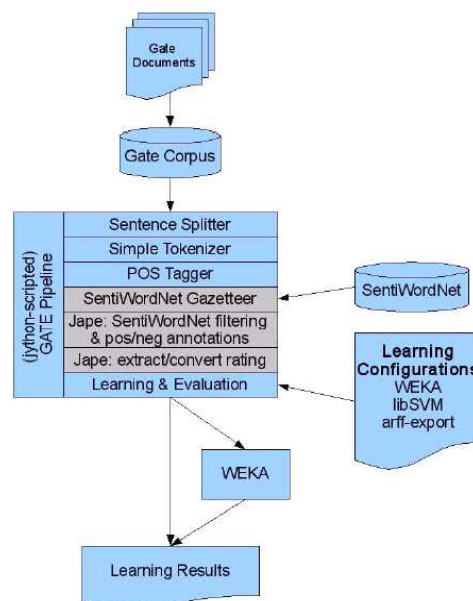
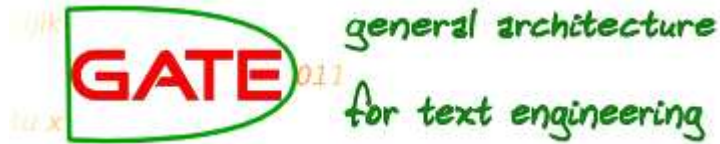


Figure 9 Architecture du projet SEMPRES

## 2.4.3 LOGICIELS

### 2.4.3.1 GATE



<http://GATE.ac.uk/>

GATE est un logiciel open source permettant le développement et le déploiement de composants qui vont traiter le langage humain :

- en proposant une architecture spécifique au traitement du langage
- en livrant, dans un Framework, les bibliothèques qui implémentent cette architecture et peuvent être utilisées pour encapsuler les traitements.
- en livrant un environnement de développement constitué d'outils graphiques.

Il est capable de résoudre presque tous les problèmes de traitement de texte. Il propose un IDE (GATE Developer), une application web (GATE Teamware), un Framework (GATE Embedded), et un processus standard (fourni à des fins d'apprentissage).

**Licence :** Creative Commons Licence (Attribution-NonCommercial-ShareAlike 3.0 Unported)

---

#### 2.4.3.2 RAPIDMINER



<http://rapid-i.com/>

Anciennement Yale, RapidMiner est le système open-source leader dans le data mining. Il est disponible seul, pour l'analyse de donnée mais aussi en tant qu'engine pour être intégrer dans d'autres applications. Son moteur d'apprentissage automatique se base sur Weka. Des centaines d'applications utilisant RapidMiner existent dans plus de 40 pays.

#### **Fonctionnalités :**

- Intégration de données
- Analyse ETL
- Analyse de données
- Reporting
- Une interface graphique pour la modélisation des processus, qui supporte la gestion d'erreur dès lors de la modélisation
- Transformation de meta data
- Un module de traitement de text

**Licence :** Rapid-I propose plusieurs types de licence mais la "community Edition" est gratuite

---

#### 2.4.3.3 WEKA



Waikato Environment for Knowledge Analysis

<http://www.cs.waikato.ac.nz/ml/weka/>

Weka est une collection de machines d'apprentissage pour des tâches de data mining développée par l'Université de Waikato en Nouvelle Zélande. Les algorithmes peuvent être soit appliqués sur un dataset, soit être appelé depuis du code java. Le logiciel contient des outils de visualisation, de pré-traitement, de classification, de clustering, et d'association selon des règles.

Ce logiciel est surtout utilisé pour l'apprentissage du data mining et pour la recherche. En terme de moteur d'apprentissage automatique, il est considéré comme un standard, d'ailleurs la plupart des logiciels de data mining open source l'utilise via des intégrations, comme c'est le cas pour GATE et RapidMiner par exemple.

Cependant Weka n'est pas un logiciel de traitement du langage naturel, par contre il peut être employé conjointement avec ceux-ci pour tester et visualiser les résultats.

**Licence :** GNU General Public License

#### 2.4.4 ENTREPRISES DE SERVICES

Toutes les entreprises listées ci-dessous proposent des services d'extraction et d'analyse automatique de l'opinion mais à l'exception du dernier, ces services sont tous payants.

Depuis l'essor des sites de réseaux sociaux, comme MSN, Facebook, Tweeter, pour n'en citer que quelques un, de nombreux nouveaux acteurs apparaissent sur le marché pour analyser toutes ces opinions avec plus ou moins de bon résultats. En effet, le langage n'étant pas une donnée constante, et encore moins dans le cadre de média sociaux, il semble assez illusoire de proposer des services "tout terrain" parcourant le web mondial (avec tout ce que cela implique de difficultés de langues et de cultures différentes). D'ailleurs les résultats sont pour le moment jugés assez insatisfaisants. Même si les scores obtenus dans des conditions bien précises peuvent atteindre 80% de réussite, l'assurance d'avoir une détection d'opinion sur des données correctes (sur les thèmes que l'on a véritablement recherchés) n'est en vérité pas très élevée et serait dans la pratique plus proche des 30%. De plus la majorité des sentiments ne pouvant pas être expressément identifiés sont classés dans les opinions neutres.

- **Connexor** (<http://www.connexor.eu/>)  
Société experte en linguistique moderne et en technologie logicielle, elle propose des services qui permettent d'extraire des informations d'une collection de documents, découvrir des mots-clés, reconnaître des noms et détecter des sentiments.
- **Interface SYBILLE (CELI-France)** (<http://www.celi-france.com/index.htm>)  
Société située à Grenoble qui propose de mesurer le degré de satisfaction des consommateurs sur Internet. L'interface SYBILLE a été développée pour le FODOP'08 (Atelier Fouille des données d'opinions de 2008), c'est un logiciel de détection de sentiment et de caractéristiques dans des forums qui implémente une méthode hybride et obtient des f-score entre 0.51 et 0.71
- **Sysomos** (<http://www.sysomos.com>), **ScoutLabs** (<http://www.scoutlabs.com>)  
Exemples type de ces nouvelles sociétés analysant les médias sociaux : Sysomos proposent de parcourir les milliards de conversations apparaissant sur les blogs, les forums, les sites sociaux et de pouvoir en extraire les influences clés ainsi que les opinions en temps réels, desquelles peuvent découler de nouvelles stratégies économiques.
- **SocialMention** (<http://socialmention.com/>)  
Moteur de recherche gratuit, SocialMention permet d'afficher les tendances pour un terme. A la manière de Google, il propose un service d'alerte par mail sur un sujet donné.  
Il fournit aussi des web services (permettant d'intégrer leur service dans d'autres applications) mais ils sont payants.

---

## 2.4.5 RESSOURCES LINGUISTIQUES

---

### 2.4.5.1 WORDNET

<http://wordnet.princeton.edu/>

WordNet est une base de données développée par des linguistes du laboratoire des sciences cognitives de l'Université de Princeton. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise.

Le système se présente sous la forme d'une base de données électronique qu'on peut télécharger (WordNet est distribué avec une licence spéciale très libérale, permettant de l'utiliser commercialement ou à des fins de recherche) sur un système local et y accéder à partir d'un programme à l'aide d'interfaces disponibles (Jawbone, JWNL ou encore RITA.Wordnet). GATE propose en standard un outil d'interface pour la version 1.6 de WordNet

---

### 2.4.5.2 SENTIWORDNET

<http://sentiwordnet.isti.cnr.it/>

Liée au projet WordNet, SentiWordNet est une ressource lexicale spécifique à l'opinion mining. Elle se présente sous la forme d'un fichier texte dans lequel sont ajoutés, pour chaque sens des mots présents dans WordNet, deux scores de sentiments entre 0 et 1: la positivité et la négativité (l'objectivité pouvant être calculée en soustrayant à un la somme des valeurs de positivité et de négativité). Un mot ayant plusieurs sens peut donc avoir plusieurs couples de valeurs. A noter aussi qu'aucun mot ne possède un score de 1 (que se soit positif ou négatif)

La version courante 3.0 est disponible en en faisant la demande auprès de ses développeurs.

## 2.5 LES CHALLENGES DE L'OPINION MINING

Comme on l'a vu tout au long de ce chapitre, l'opinion mining est certes un domaine encore jeune mais son exploitation est jalonnée de nombreux problèmes techniques qui malgré des améliorations depuis les débuts des recherches n'ont toujours pas été résolus de manière convaincante.

### **La détection des thèmes de l'opinion**

Sans connaître le thème sur lequel porte l'opinion, celle-ci n'a que peu d'importance. Et si l'opinion n'est pas attribuée au bon thème, les résultats en seront même complètement faussés.

### **L'extraction des caractéristiques (sous-thèmes de l'opinion) et le regroupement de synonyme**

La plupart des recherches actuelles se bornent à extraire les noms ou les couple noms/adjectifs. Cependant certains verbes peuvent être considérés comme des caractéristiques du thème de l'opinion mais ils sont difficiles à identifier. De plus, les gens ont tendances à utiliser des mots ou des phrases différentes pour décrire la même caractéristique. A cela vient s'ajouter le caractère implicite de certaines phrases, où des mots manquants ne permettent pas d'évaluer correctement ces caractéristiques.

### **La nature de langage**

Le registre de la langue tout d'abord rend plus difficile l'extraction des opinions sur des blogs ou des forums que dans un article de journal en ligne, en effet un registre plus familier, des phrases grammaticalement incorrectes ou des expressions locales empêchent l'analyse correcte de ces opinions.

Ensuite même dans un article ou un commentaire au registre de langage soutenu, une faute d'orthographe ne va pas permettre la reconnaissance du mot et peut engendrer des résultats erronés si la faute se situe sur un mot particulièrement porteur d'opinion.

Enfin le sarcasme, l'ironie, le cynisme ou le langage figuré comme la métaphore qui influencent le sens d'un texte sont pour le moment complètement ignorés, même si de toutes nouvelles recherches viennent de donner quelques résultats encourageants dans le domaine du sarcasme.

### **La classification de l'orientation de l'opinion**

Toute la problématique d'identifier des phrases ou des mots porteurs d'opinion découle du fait qu'il n'y a pas un nombre limités d'expressions utilisées pour l'exprimer. Même dans un unique domaine, un même mot peut être porteur ou non selon qu'il dépend de thèmes différents.

### **La force de l'opinion dans le temps**



Entre Web 2.0 et 3.0 : opinion mining

Si pour le moment les recherches tentent seulement repérer l'opinion de manière la plus correcte possible, aucunes pour l'instant se sont penchées sur le changement de la force de l'opinion (ou de l'opinion elle-même) dans le temps pour un même thème, notamment dans le fil de discussion dans un forum par exemple.

## 2.6 CONCLUSION

Le 13 avril 2010 s'est tenu à New York, le Sentiment Analysis Symposium. Organisé par Seth Grimes (Analyste chez AltaPlana Corporation), cet événement a rassemblé une centaine de personnes du monde de l'opinion mining. Quatre idées clés sont ressorties de cette journée. Idées qui peuvent aussi servir de résumé à l'état actuel de ce domaine en plein essor :

- Pour pouvoir analyser l'opinion, il faut un au préalable un véritable traitement linguistique. Sur la base de ce constat, se sont donc les méthodes hybrides qui vont rapidement devenir des standards, l'apprentissage automatiquement ne pouvant rien, en terme de résultat, si les données à apprendre n'ont pas été fortement traitées pour en déterminer le maximum d'informations.
- Les spécialistes du domaine n'envisagent pas une automatisation complète de l'opinion mining à moyen terme. L'image donnée durant le symposium est celle du détecteur de métaux. Il donne une information, mais il faut un humain pour creuser.
- Le taux de succès que l'on peut actuellement espérer d'une analyse de sentiment est de 80% et c'est aussi le taux moyen d'accord de deux juges humains, face à une expression écrite porteuse d'opinion. D'après les spécialistes, il est illusoire de vouloir aller plus loin.
- La demande des spécialistes mais aussi des industriels par rapport à ce domaine est la standardisation des mesures et des procédés. D'ailleurs une telle démarche s'amorce en ce qui concerne le vocabulaire

### 2.6.1 LE FUTUR

Jusqu'à présent les exemples classiques de détection des sentiments, ne prenaient en compte que les opinions centrées autour d'un produit ou d'un service. Un des futurs axes commerciaux du domaine va être de pouvoir aussi relier un thème générale a des sentiments et des opinions, car il peut arriver que la recherche d'opinion sur un produit en particulier ne génère pas un grand volume de résultats. Mais ce n'est pas parce qu'il n'y a pas de résultats sur un produit précis que les consommateurs n'ont pas d'opinions sur ce qu'ils attendent de ce produit vu d'une manière générale. Comme par exemple pour le secteur des assurances vie, où l'opinion générale au sujet de la sécurité des vols en avion pourrait être une information intéressante.

L'amélioration de la détection de la force d'une opinion est aussi un thème qui sera développé dans le futur, à la demande des managers qui ont besoin d'outils qui leur permettent de prendre une décision commerciale. Pour le moment l'opinion mining détermine seulement la positivité, la négativité ou la neutralité d'une opinion, mais il y a une grande différence entre "aimer" une marque et "adorer" une marque. Les sentiments ont des

Entre Web 2.0 et 3.0 : opinion mining

dimensions riches et nuancées qu'il faudra prendre en compte dans le futur pour des résultats vraiment exploitables.

### 3 EVALUATION

Dans le précédent chapitre nous avons découvert entre autres des outils permettant le développement d'un processus d'opinion mining. Trois d'entre eux nous semblaient particulièrement intéressants. Afin de déterminer lequel sélectionner pour la réalisation du prototype, une brève évaluation a été faite sur chacun d'entre eux pour mettre en évidence leurs points forts et leurs points faibles.

#### 3.1 UIMA

Le projet UIMA propose un Framework utilisable pour créer des pipelines (processus), dans lesquels seront traités des documents. A la fin du processus les résultats obtenus sont les documents et les annotations qui auront été ajoutés par les processus.

L'architecture d'un processus de traitement UIMA se présente de la manière suivante :

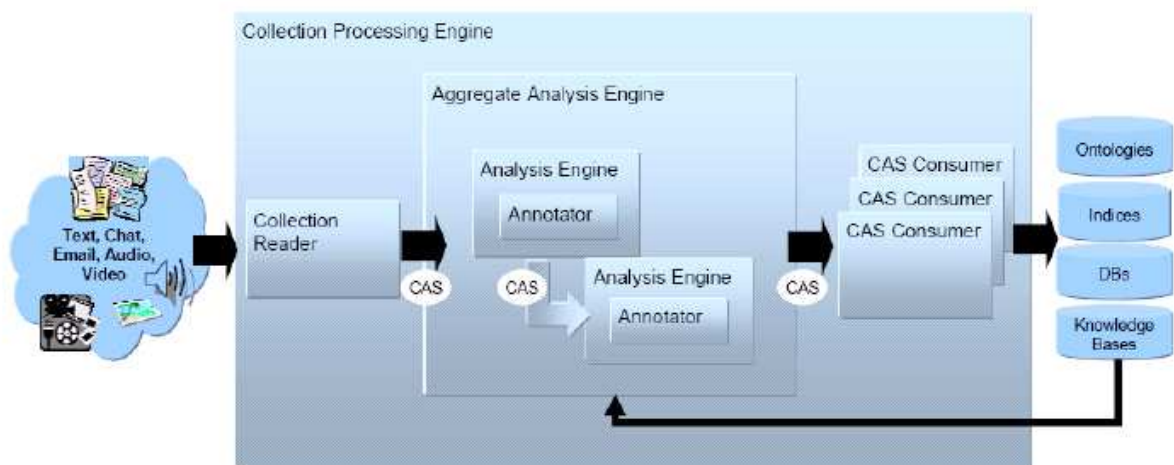


Figure 10 Schéma d'un processus de traitement dans UIMA

Des composants de bases sont définis pour supporter le cycle des développements des applications :

- Le collection Reader se connecte à une source d'informations et initialise le CAS (Common Analysis structure) avant l'analyse
- L'Aggregate Analyse Engine pouvant être constitué de plusieurs engines vont analyser les documents pour détecter certains attributs, et les Annotators qui contiennent les algorithmes d'analyses y sont encapsuler pour ajouter les interfaces nécessaire au déploiement dans le Framework UIMA

Entre Web 2.0 et 3.0 : opinion mining

- Les CAS Consumer interviennent à la fin du processus pour traiter les résultats d'analyse et les rendre exploitables par d'autres applications.

Tout comme GATE, des outils supplémentaires développés en Java peuvent être ajoutés. Mais à la différence de celui-ci, tout est développé en Java via des plugins spécifique d'Eclipse, aucune interface graphique ne permet au non développeur Java d'utiliser le logiciel.

### 3.1.1 SWOT

#### **Strengths (Force)**

Bonne documentation  
De nombreux exemples illustrés  
Une grande communauté (notamment française)  
Prend en charge non seulement des documents textuels mais aussi de l'audio et de la vidéo  
Des informations détaillées sur le temps de traitement de chaque composant

#### **Weaknesses (Faiblesse)**

De bonnes connaissances Java sont obligatoires  
Une liste d'outils standards limitée

#### **Opportunities (Opportunités)**

Aucune limitation

#### **Threats (Menaces)**

### 3.2 RAPIDMINER

RapidMiner est une suite de modules destinée principalement au domaine général du data mining. Un de ces modules RapidSentilyzer est cependant spécifique à la détection des sentiments. Mais ce module est payant. RapidMiner est un outil complètement orienté pour le développement d'applications commerciales. La société (Rapid-I) qui développe ce logiciel met d'ailleurs en place aussi de nombreux séminaire ou cours de formations.

Si on se limite à la version gratuite et communautaire de rapidMiner, la version 5 sortie en fin d'année 2009 est un outil complet pour le traitement, l'affichage et l'analyse de donnée. Avant la version 5 les outils spécifiques de text mining faisait partie des plugins, mais depuis la dernière version ils sont complètement intégrés au logiciel.



Figure 11 Architecture de RapidMiner

L'interface graphique est très intuitive et propose un outil de débogage en temps réel durant la phase de développement du processus.

RapidMiner utilise le moteur de Weka pour ces fonctionnalités d'apprentissage automatique.

### 3.2.1 SWOT

<p><b>Opportunities (Opportunités)</b></p> <p>Interfaces très intuitive Vues pour analyses très nombreuses</p>	<p><b>Weaknesses (Faiblesse)</b></p> <p>Outils de text mining limités Pas d'intégration d'outils personnels possibles Pas de tutorial propre au text mining</p>
<p><b>Strengths (Force)</b></p> <p>Environnement pour application commerciales Environnement de développement professionnel</p>	<p><b>Threats (Menaces)</b></p>

### 3.3 GATE

Le logiciel GATE, comme UIMA, propose un Framework, avec une architecture, des outils et des méthodes. Celui-ci est utilisé comme base pour le développement des composants. L'environnement de développement est une application autonome qui permet d'exécuter et de tester les composants et de modifier rapidement les paramètres de l'application.

GATE peut traiter toutes les sources de données statiques comme les lexiques ou les ontologies. Au niveau des documents, il peut prendre en charge plusieurs formats dont le XML, le RTF, le HTML et le SGML.

## Entre Web 2.0 et 3.0 : opinion mining

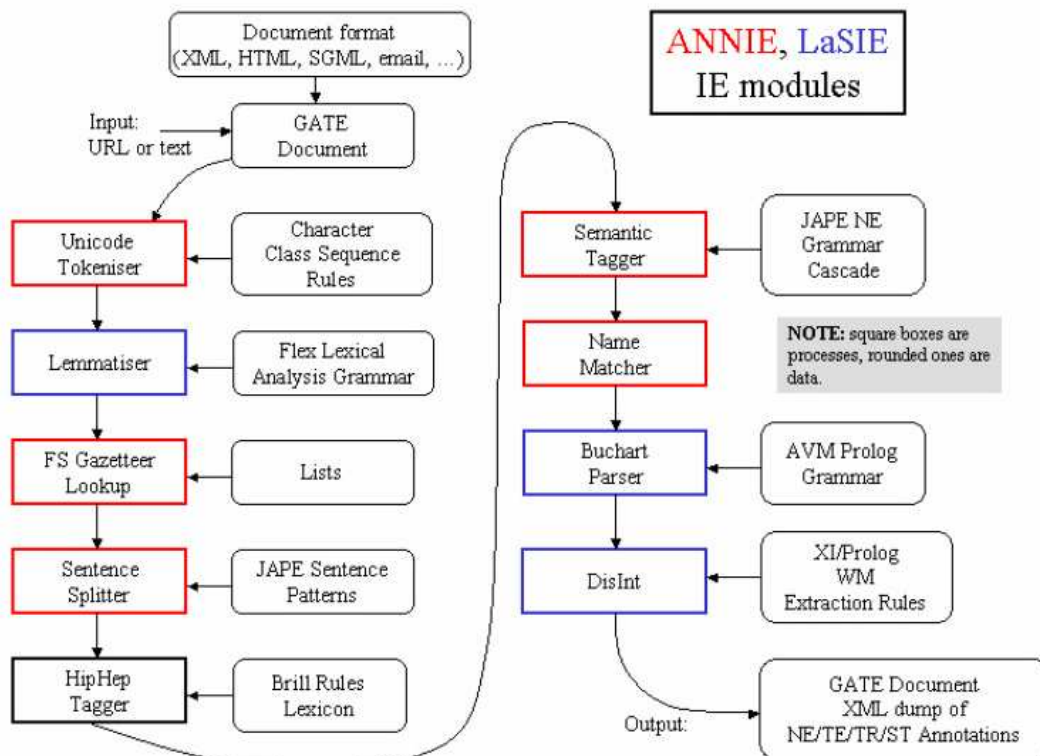


Figure 12 Schéma d'un processus de traitement dans GATE

Dès qu'un document est intégré dans le processus, il est séquentiellement traité par chaque élément de ce processus. Au final le résultat du donne un document XML qui contient le texte du document de départ plus les annotations.

GATE est soutenu par une importante communauté et implémente de nombreux modules développés par des groupes de recherche universitaires. Il propose des intégrations avec WEKA et Lingpipe par exemple.

Même si des projets de production, développé avec cet outils existent, il est avant tout utilisé pour la recherche et l'apprentissage. L'environnement graphique qu'il propose ainsi que son processus standard (ANNIE) permettent une prise en main rapide pour le néophyte en text mining.

Un point fort de GATE est son architecture extensible, même si en standard une liste conséquente d'outils existe déjà, on peut aisément étendre cette liste en créant ses propres outils en java. GATE propose même un assistant pour la création de nouveaux outils.

### 3.3.1 SWOT

#### **Strengths (Force)**

Bonne documentation  
De nombreux exemples illustrés  
Une grande communauté  
Un grand nombre d'outils  
Peut traiter un grand nombre de documents

#### **Weaknesses (Faiblesse)**

Même si non obligatoires des connaissances en Java sont un avantage  
Ne supporte que les formats de document textuels (HTML, RTF, E-mail, SGML, etc..) pas d'autres médias

#### **Opportunities (Opportunités)**

Ajout facilité de nouveaux outils  
Aucune limitation  
Accès simplifié pour les débutants

#### **Threats (Menaces)**

Aucunes données quand aux temps de traitement d'un corpus  
Demande beaucoup de mémoire

## CONCLUSION

Tous les outils décrits ci-dessus sont des logiciels robustes et qui donnent satisfaction à leurs utilisateurs. Cependant malgré sont interface intuitive RapidMiner a été mis de côté à cause de ses limitations en terme d'outils de traitement de la langue et l'impossibilité de pouvoir créer ses propres composants.

GATE et UIMA proposent des fonctionnalités très similaires. Malgré le fait que UIMA permet de traiter plus de média (l'audio et la vidéo), l'interface graphique de GATE a fait la différence dans le choix de ce logiciel. Sa plus grande rapidité de prise en main grâce à un processus standard d'apprentissage des concepts, nous a définitivement convaincu que GATE était l'outil le plus approprié au développement du prototype. Cependant ces deux logiciels ne sont pas exclusifs et il existe des modules qui permettent de les utiliser conjointement.



## 4 GATE

Dans ce chapitre nous verrons de manière plus détaillée les concepts techniques de GATE, qui vont être utilisés par la suite dans le développement du prototype.

### 4.1 ARCHITECTURE

Le Framework GATE constitué de classes Java se présente sous la forme de différentes couches ou ressources.

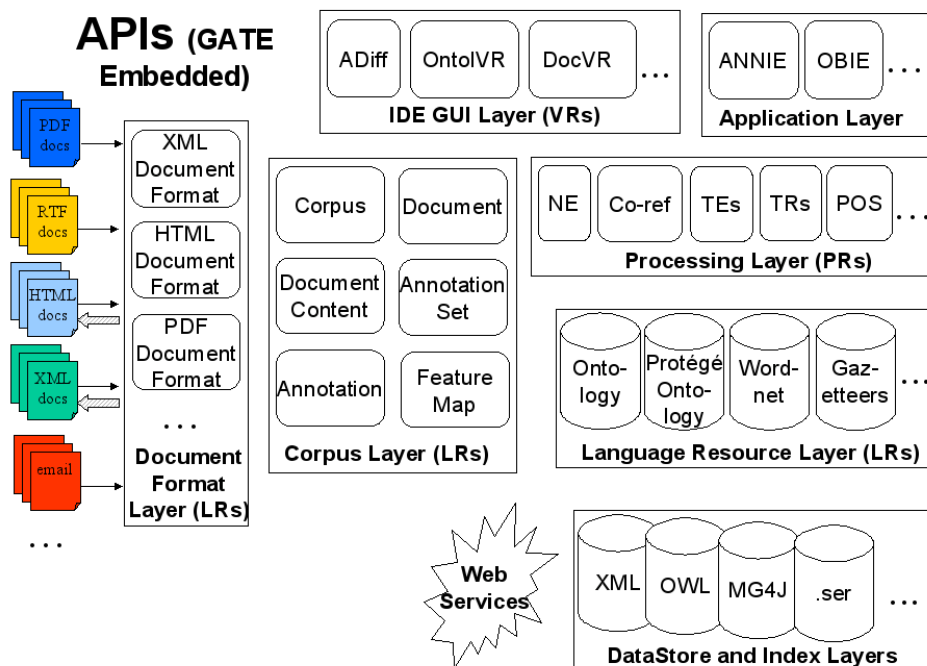


Figure 13 Architecture du Framework GATE

Deux types de classes sont importants pour les développeurs qui souhaitent étendre les fonctionnalités de GATE : les « Language Resource » et les « Processing Resource »

Les « **Language Resource** » (notés LR dans le schéma de la figure 13) sont tous les éléments qui vont permettre de stocker les informations textuelles à traiter (divers documents de différents formats, corpus, set d'annotations, annotations, liste pour les gazetteers, etc..)

## Entre Web 2.0 et 3.0 : opinion mining

Les « **Processing Resource** » (notés PRs) sont quant à eux les modules qui vont traiter les informations textuelles en utilisant les « Language Resource ».

Afin de pouvoir traiter un document selon ses besoins, un développeur java aura tout le loisir d'étendre la couche « Processing » en ajoutant de nouvelles « Processing Resource ».

## 4.2 IDE

Dans GATE le traitement d'un document consiste à rechercher un mot ou un groupe de mots et de l'annoter. Une annotation est définie par son type, son début, sa fin et possède un identifiant unique. A ces annotations peuvent être rajoutées des caractéristiques (features), les caractéristiques peuvent aussi être associées à un document ou à un corpus.

Au démarrage du logiciel, l'interface graphique se présente sous la forme d'un projet vide constituée de 4 types de ressource (Applications, Language Resources, Processing Resources et Datastores)

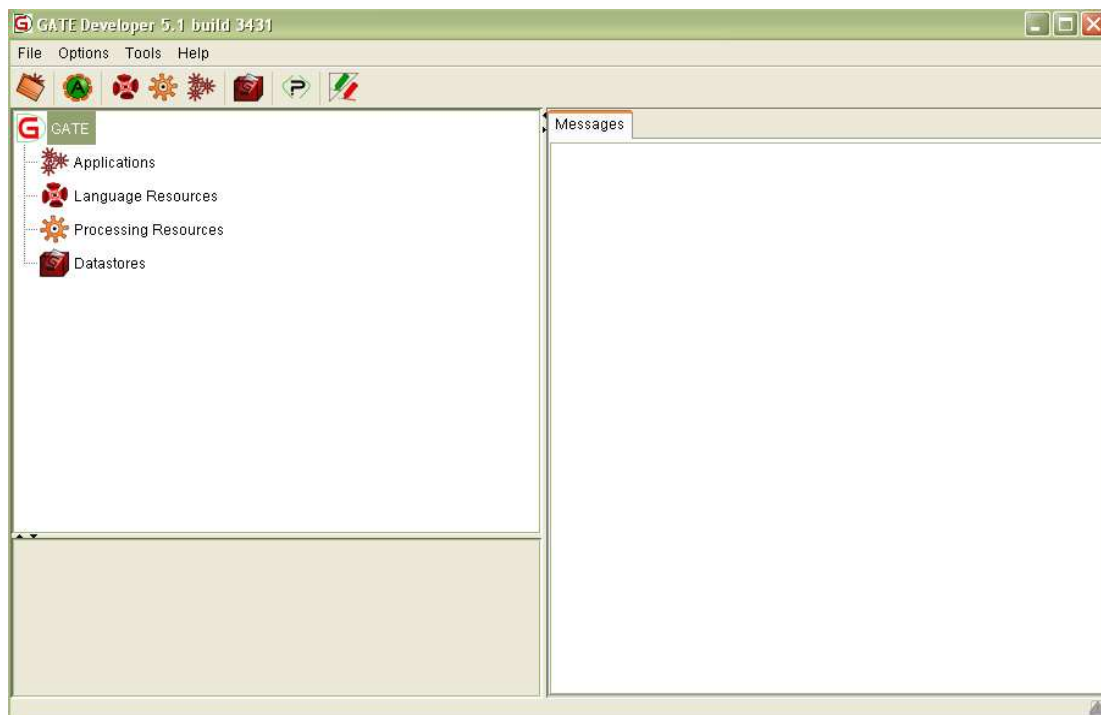


Figure 14 IDE GATE

En standard le projet ANNIE (qui fait partie de la couche Application dans l'architecture du Framework) est fourni pour débiter et servir de tutorial.

## Entre Web 2.0 et 3.0 : opinion mining

C'est un projet de type « Corpus » qui va donc effectuer un traitement - c'est-à-dire exécuter tous les éléments se trouvant dans la partie Processing Resources - sur chaque document présent dans le corpus.

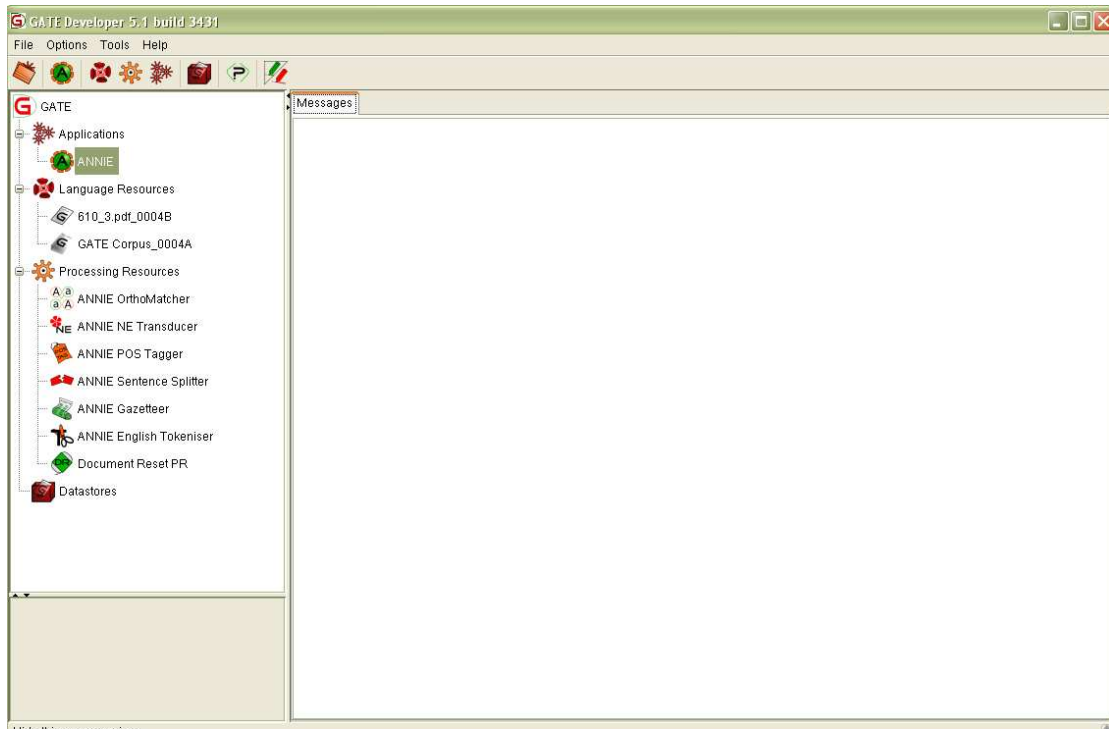


Figure 15 Projet ANNIE

Les processing resources du projet ANNIE sont les traitements basiques qui peuvent être effectués sur un document.

Entre Web 2.0 et 3.0 : opinion mining

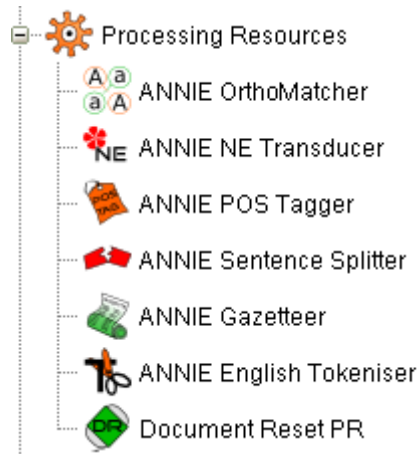


Figure 16 Processing Resources du projet ANNIE

Une fois les ressources sélectionnées dans le processus et le corpus qu'elles doivent traiter, choisis dans la liste, l'application peut être lancée.

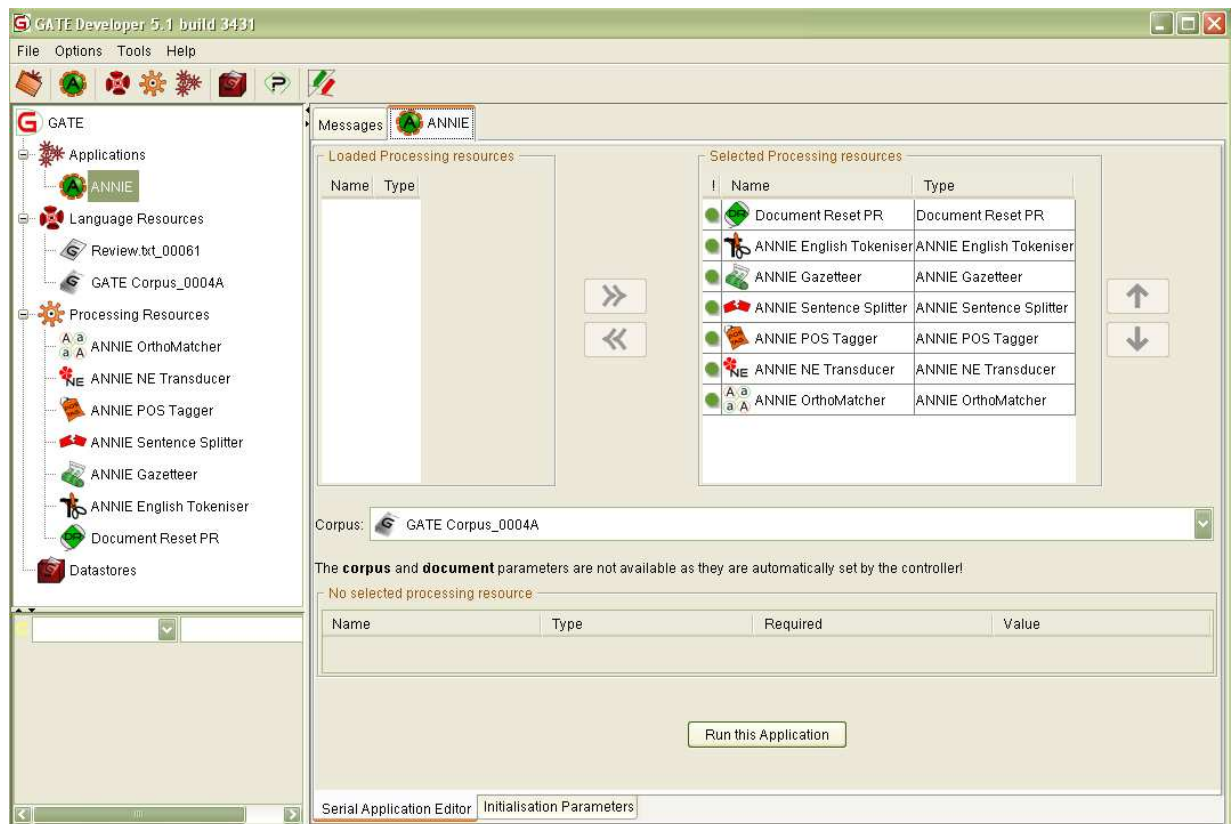


Figure 17 Execution du projet ANNIE

## Entre Web 2.0 et 3.0 : opinion mining

Lorsque GATE a terminé le traitement du processus, le résultat de celui-ci est visible en ouvrant le document.

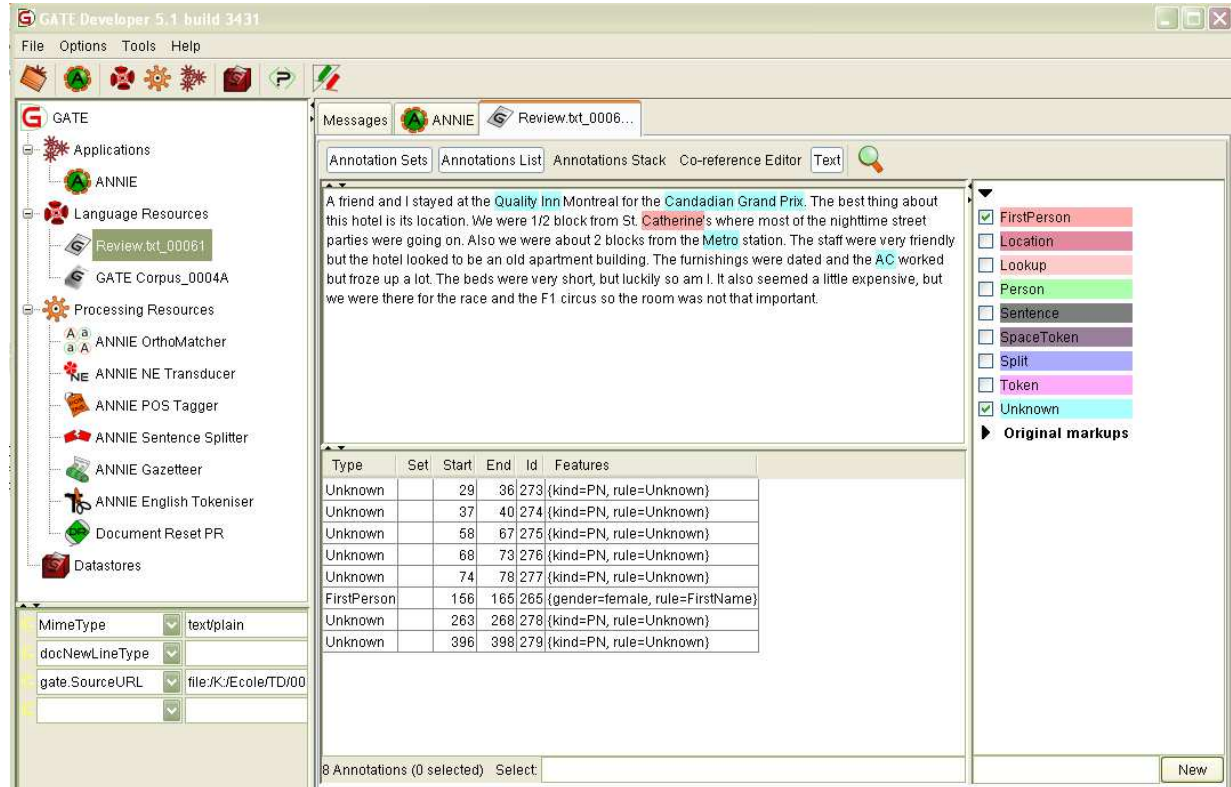


Figure 18 Résultat du processus ANNIE

Dans l'exemple de la figure 19, le processus à créé plusieurs sets d'annotations (FirstPerson, Location, Lookup,..) . En les sélectionnant on peut visualiser au moyen des couleurs quelles parties du texte ont été annotées. On peut aussi voire, dans le tableau inférieur, que les annotations possèdent des caractéristiques avec des valeurs (ici : kind, rule ou gender).

#### 4.2.1 GAZETTEER

Le gazetteer est un outil standard qui permet à GATE, en fournissant une ou plusieurs listes de mots, de retrouver ces mots en les annotant par la valeur "Lookup". On peut aussi rajouter une information à chaque liste de mot que le logiciel va appliquer à la caractéristique standard "MajorType" de l'annotation.

---

#### 4.2.2 JAPE

L'outil standard Transducer permet d'intégrer et d'exécuter un fichier JAPE dans le processus.

JAPE (pour Java Annotation Patterns Engine) est un langage spécifique qui permet entre autre de manipuler les annotations et les caractéristiques. La grammaire de ce langage consiste à appliquer une règle aux informations qui vérifient une condition. Un fichier JAPE peut effectuer des actions très complexes en intégrant du code Java.

---

### 4.3 PLUGINS CREOLE

---

#### 4.3.1 GESTION DES PLUGINS

Mise à part les « Processing Resources » standards du processus ANNIE, il en existe une multitude d'autres qui peuvent être intégrés au projet sous la forme de plugins. La gestion de ces plugins, étant une application en elle-même, se nomme CREOLE.

## Entre Web 2.0 et 3.0 : opinion mining

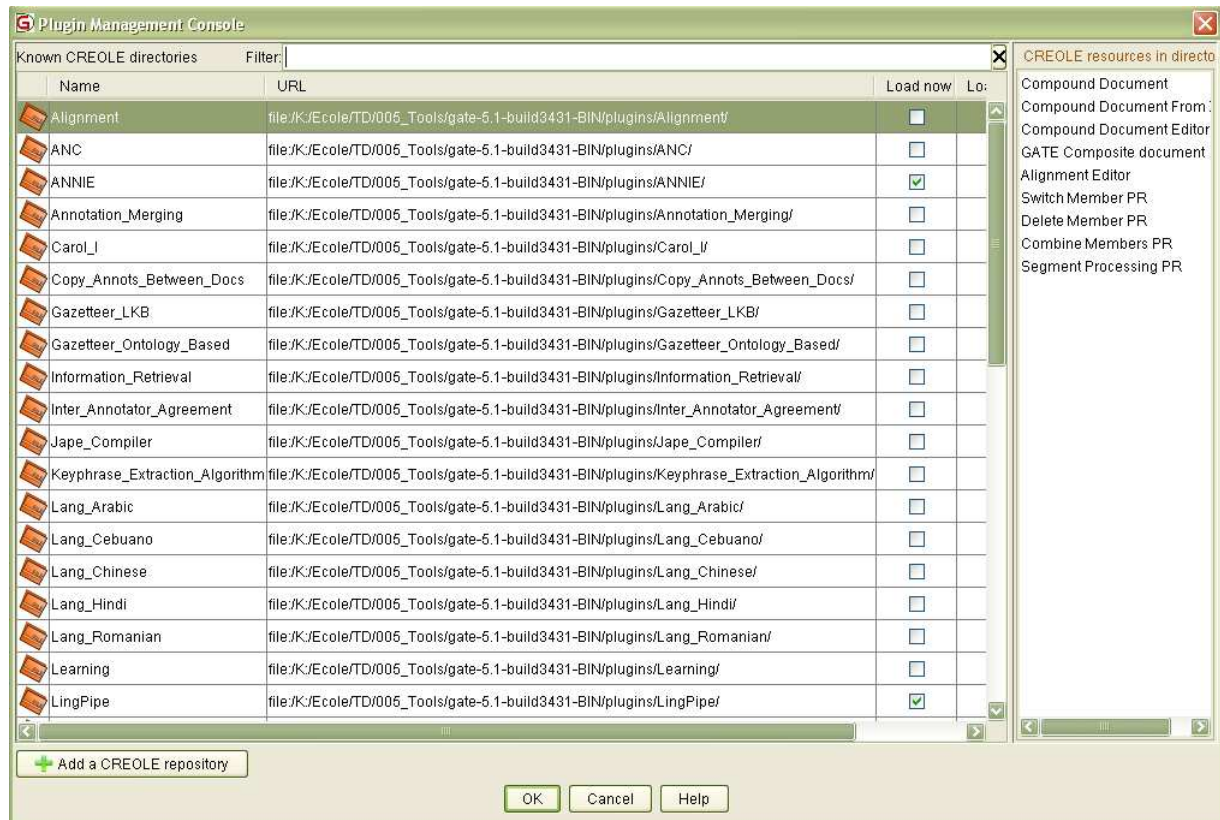
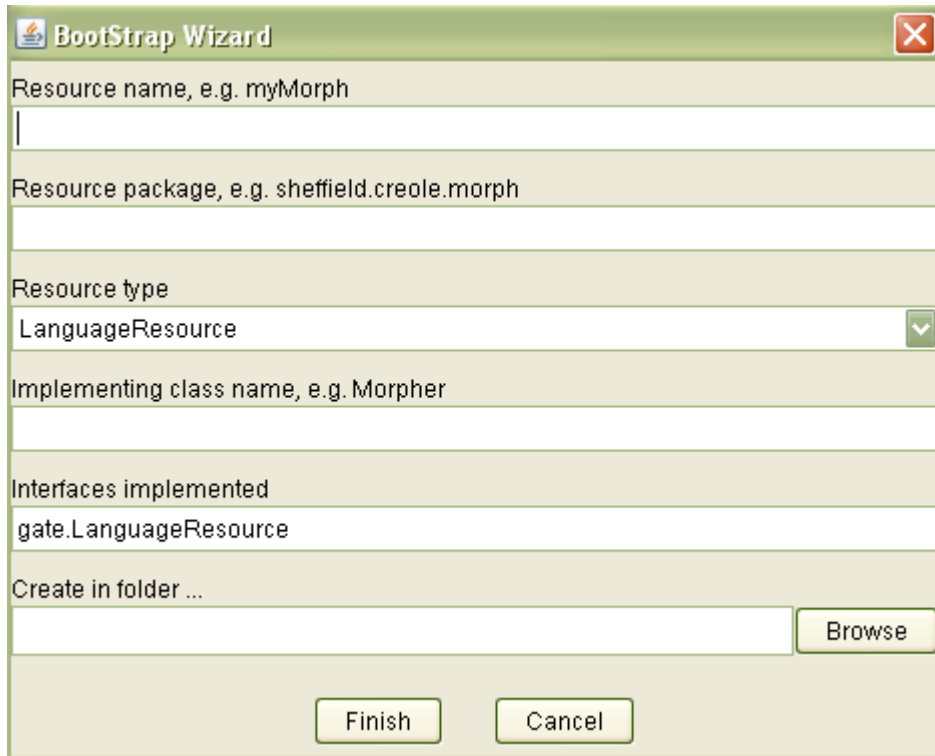


Figure 19 Console de gestion des plugins CREOLE

Ce gestionnaire permet aussi de rajouter de nouveaux plugins, développés pour des besoins spécifiques, pour autant que ceux-ci respectent la structure requise. Une autre application de GATE, nommée BootStrapWizard, permet de générer automatiquement la structure d'un plugin.

#### 4.3.2 CREATION D'UN PLUGIN



The image shows a 'BootStrap Wizard' dialog box with the following fields and controls:

- Resource name, e.g. myMorph: [Empty text field]
- Resource package, e.g. sheffield.creole.morph: [Empty text field]
- Resource type: [Dropdown menu showing 'LanguageResource']
- Implementing class name, e.g. Morpher: [Empty text field]
- Interfaces implemented: [Text field containing 'gate.LanguageResource']
- Create in folder ...: [Empty text field] with a 'Browse' button to its right.
- Buttons: 'Finish' and 'Cancel' at the bottom.

Figure 20 BootStrapWizard

Après avoir entré les informations du futur plugin (son nom, le nom du package, le type de ressources (Language ou Processing), etc..), un certains nombres de fichiers sont créées dans le répertoire sélectionné dans le Wizard.



## Entre Web 2.0 et 3.0 : opinion mining

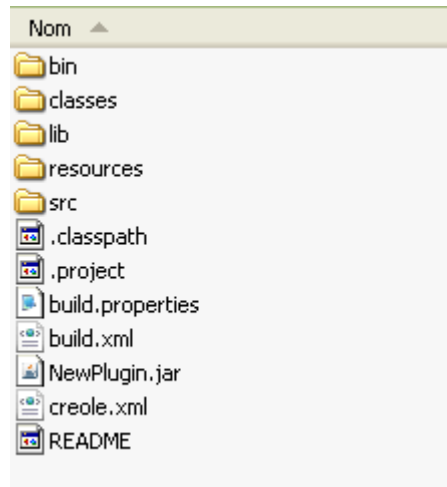


Figure 21 Structure d'un plugin GATE

Le projet n'a plus qu'à être importé dans Eclipse et configuré en intégrant la librairie GATE, pour pouvoir commencer le développement. La classe principale a déjà été créée, implémentant l'interface sélectionnée dans le Wizard (« LanguageResource » ou « ProcessingResource »).

```
package test;

import java.util.*;

/**
 * This class is the implementation of the resource NEWPLUGIN.
 */
@CreoleResource(name = "NewPlugin",
    comment = "Add a descriptive comment about this resource")
public class NewPlugin extends AbstractProcessingResource
    implements ProcessingResource {

} // class NewPlugin
```

Figure 22 Classe du plugin après sa création par le Wizard

Le fichier creole.xml va permettre au gestionnaire de reconnaître ce nouvel élément en tant que plugin. Directement après la création par le Wizard ce fichier est cependant réduit à sa plus simple utilisation.

## Entre Web 2.0 et 3.0 : opinion mining

```
<!-- creole.xml NewPlugin -->
<!-- $Id: creole.xml 9992 2008-10-31 16:53:29Z ian_roberts $ -->

<!--
This file just references the JAR file that contains the compiled resource.
Configuration is contained in the @CreoleResource annotation on
test.NewPlugin.
-->

<CREOLE-DIRECTORY>
  <JAR SCAN="true">NewPlugin.jar</JAR>
</CREOLE-DIRECTORY>
```

Figure 23 Fichier creole.xml après la création par le Wizard

Pour le prototype plusieurs "Processing Resource" spécifiques ont du être créés dans GATE. Afin de rassembler toutes les nouvelles ressources au même endroit, un unique projet de plugin a été créé qui va contenir toutes classes. Chacune des ses classes va correspondre à un "Processing Ressource".

## 5 PROTOTYPE

Dans ce chapitre nous verrons en détail l'architecture des processus de traitement dans GATE et les éléments qui le composent. Nous expliquerons aussi les méthodes externes au processus créent pour pouvoir exécuter ce processus en dehors de l'interface graphique du logiciel. Et pour finir, nous présenterons les résultats obtenus après des tests effectués en conditions réelles.

### Fonctionnement de l'application

Le prototype est constitué de deux étapes bien distinctes : l'extraction des données depuis le site TripAdvisor.com et le traitement de ces données.

- Le premier processus GATE (aussi appelé pipeline) constitué simplement des classes d'extraction va se charger d'extraire chacun des commentaires de l'hôtel dont l'url a été passé en paramètre, il va ensuite pour chaque commentaire créer un document texte et l'intégrer à un corpus.
- Le second pipeline va se baser sur le corpus créé par le premier processus pour traiter chaque document avec un certain nombre de plugins. Deux de ses plugins vont utiliser des données externes : la base de donnée de SentiWordNet ainsi qu'une liste de mot « StopWord ». Ces deux bases de données externes sont de simples fichiers texte.
- Une fois le traitement terminé toutes les annotations apportées aux documents GATE sont sauvegardées au format XML. Une classe Java (externe à GATE) va récupérer toutes les informations pertinentes de ces fichiers XML de résultat et créer un nouveau document unique lui aussi au format XML.

Entre Web 2.0 et 3.0 : opinion mining

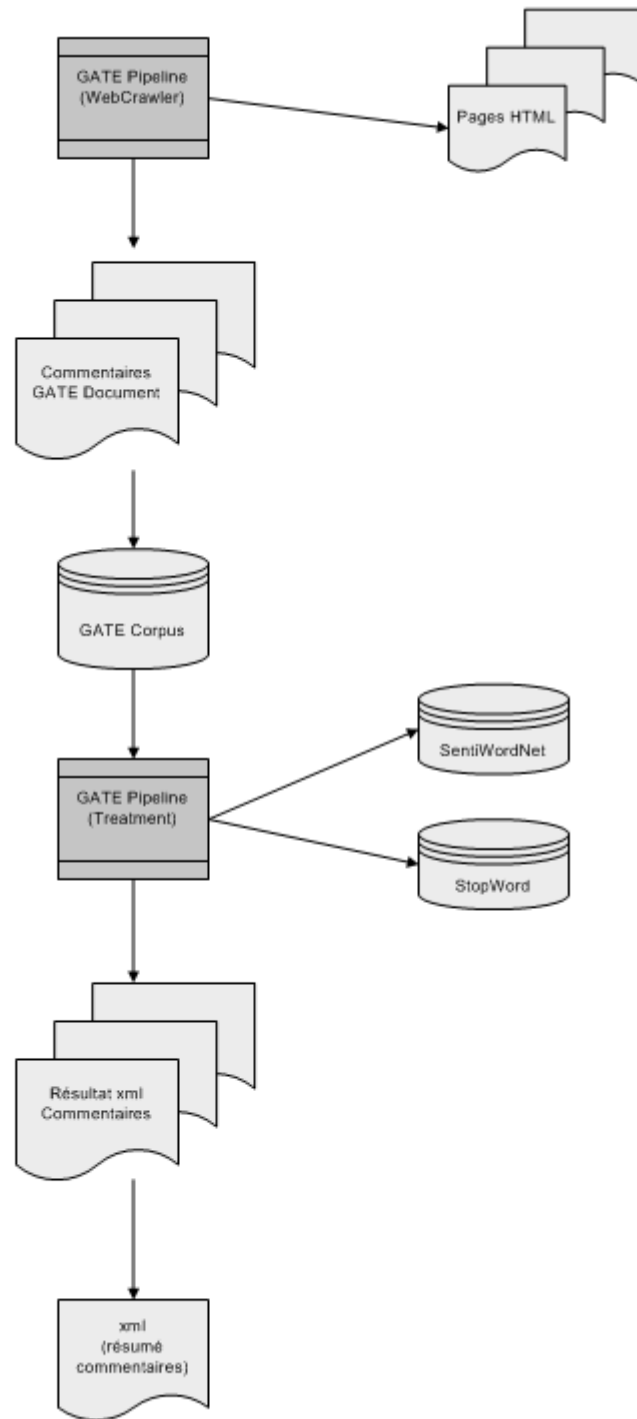


Figure 24 Architecture des processus du prototype

## 5.1 COLLECTE D'INFORMATIONS

Un précédent projet avait déjà eu pour but l'extraction de données depuis le site TripAdvisor. De ce projet a résulté un ensemble de classes java qui permettent :

- de parser la page en proposant des méthodes pour retrouver l'information. Ces méthodes reçoivent des noms de balises html ou des noms et des valeurs d'attributs et retournent soit le texte de la balise soit la valeur associée à l'attribut.
- d'extraire certaines informations des pages web au moyen des méthodes du point précédent.
- de créer des fichiers de type .owl contenant les données extraites.

Il est rapidement apparu que les informations déjà récupérées dans la classe d'extraction n'étaient pas suffisantes (le commentaire n'était pas présent) et que la création d'un fichier .owl n'avait pas d'intérêt pour le prototype. Il a donc été décidé de garder seulement le parser et de modifier la classe d'extraction pour la faire correspondre exactement au besoin du présent projet.

Comme le souligne l'auteur des classes Java dans un des commentaires, l'extraction des données est tributaire de la structure des pages html traitées. En effet les méthodes du parser récupèrent les informations sur la base des noms des balises ou des attributs de ces balises. Or cette structure a depuis été modifiée sur le site de TripAdvisor. La majorité de ces modifications concernent le nom ou la valeur des attributs. Cependant la récupération des pages suivantes (tous les commentaires ne se trouvent pas sur une seule page), a impliqué une modification du parser lui-même car la structure HTML de cette partie de la page avait changé.

---

### 5.1.1 STRUCTURE DES PAGES INTERNET

TripAdvisor est un site de recommandation pour tout ce qui concerne les voyages. Il propose des recherches par ville, région, etc. afin de trouver des hôtels, des vols, des restaurants ou des activités.

En ce qui concerne la catégorie des hôtels (celle qui nous intéresse ici), les utilisateurs peuvent y laisser des commentaires. Du point de vue HTML, la sélection d'un hôtel nous amène sur une première page (ex : [http://www.tripadvisor.fr/Hotel\\_Review-g155032-d631714-Reviews-Le\\_Relais\\_Lyonnais-Montreal\\_Quebec.html](http://www.tripadvisor.fr/Hotel_Review-g155032-d631714-Reviews-Le_Relais_Lyonnais-Montreal_Quebec.html))

## Entre Web 2.0 et 3.0 : opinion mining

Page d'accueil → Canada → Québec → Montréal → Hôtel Montréal

## Le Relais Lyonnais



[Photos de l'hôtel](#)  
[Services et équipements](#)



1595 rue St Denis, Montréal, Québec H2X 2K3 , Canada  
 +1 (514) 448-2999 | Site internet de l'hôtel | Contactez l'hôtel

### Vérifier les prix et la disponibilité

Arrivée : 18/7/2010      Départ : 25/7/2010      Adultes : 2

**VOIR LES PRIX**

Booking.com     ebookers.ch  
 Expedia.de

Ouvre une fenêtre pour chaque offre. Veuillez désactiver les programmes de blocage des fenêtres pop-up.

### Avis d'authentiques voyageurs

#### Classer les avis par type de voyage et par notes

##### ▸ Tout (115)

Professionnel (12)  
 En amoureux (72)  
 En famille (12)  
 Escapades entre amis (4)  
 Voyage solo (10)

**96% des voyageurs conseillent**

115 avis

Excellent		101
Très bon		9
Moyen		4
Mauvais		0
Épouvantable		1

1-10 sur 115

« 1 2 ... 12 »

Trier par [ Date ▼ ] [ Note ]

Français en 1er

#### “magique”



prubou  1 contribution  
 Saguenay

13 juin 2010 | Type de voyage : En couple

Cette hotel ns a charmé par son allure très contemporain , plancher de bois chocolat, lit très douillet , une salle de bain hors pairs munie d'une grande douche en verre et d'une propreté très rare. Chocolats fins sur la table de chevet ainsi qu'une cafetière espresso pour votre réveil. Bay Window ouvrant et petit balcon pour dégusté un bon... suite

Figure 25 : Page de commentaires d'hôtel

Sur cette première page on retrouve le nom et les informations de l'hôtel, un bloc d'informations permettant de connaître l'avis général de tous les commentaires et ensuite les dix premiers commentaires.

## Entre Web 2.0 et 3.0 : opinion mining

Mais les commentaires visibles sur cette première page ne sont pas complets, ils représentent seulement les premières lignes. Un lien (sur le mot « suite » ou sur le titre du commentaire) permet d'accéder à une nouvelle page (ex : [http://www.tripadvisor.fr/ShowUserReviews-g155032-d631714-r67380839-Le\\_Relais\\_Lyonnais-Montreal\\_Quebec.html#CHECK\\_RATES\\_CONT](http://www.tripadvisor.fr/ShowUserReviews-g155032-d631714-r67380839-Le_Relais_Lyonnais-Montreal_Quebec.html#CHECK_RATES_CONT)).

Cette nouvelle page affiche sensiblement les mêmes informations cependant on y retrouve le commentaire complet.

“ magique ”

**Le Relais Lyonnais**



prubou  1 contribution  
Saguenay

13 juin 2010 | Type de voyage : En couple

Cette hotel ns a charmé par son allure très contemporain , plancher de bois chocolat, lit très douillet , une salle de bain hors pairs munie d'une grande douche en verre et d'une propreté très rare. Chocolats fins sur la table de chevet ainsi qu'une cafetière espresso pour votre réveil. Bay Window ouvrant et petit balcon pour dégusté un bon digestif et comme si cela n'était pas suffisant les propriétaires vs accueillent pour le petit déjeuner avec un sourire et une chaleur que vous n'êtes pas prêts d'oublier. Mais encore..... que dire du petit déjeuner !!! une cuisine au goût de provence tellement raffinée qu'elle est digne d'un 5 étoiles mais encore mieux tout cela accompagné d'un personnel avenant, souriant et de qualifié. Alors je vous recommande cet hôtel et qu'importe ;faite un exception pour vivre des moments parfaits sous le magnétisme de ses propriétaires et la magie en plein cœur de Montréal. Merci à cet grande hôtel de m'avoir reçu comme ci j'étais la grande star de l'heure.

**Mes notes pour cet hôtel**

 Rapport qualité / prix	 Service
 Chambres	 Literie
 Emplacement	
 Propreté	

**Date du séjour** juin 2010

**But du voyage** Vacances

**J'ai voyagé avec** Avec son époux(se) / partenaire

**Membre depuis** 13 juin 2010

**Recommanderiez-vous cet hôtel à un ami ?** Oui

Cet avis est l'opinion subjective d'un membre de TripAdvisor et non de TripAdvisor LLC.

Cet avis vous a-t-il été utile ?

PAS UTILE  UTILE

[Voir le profil](#) | [Envoyer un message](#)

[Signaler un avis](#)

**VOIR LES PRIX**

Prix moyen\* :

**96 € - 173 €**

(toute l'année)

- ▶ Services et équipements
- ▶ Photos de l'hôtel

Figure 26 Bloc de commentaire complet

Chaque bloc de commentaire est composé d'une balise HTML <div> dans laquelle on retrouve à chaque fois les mêmes informations, pour autant qu'elles aient été saisies par le commentateur.

## Entre Web 2.0 et 3.0 : opinion mining

Pour chacun des ses blocs de commentaires la classe d'extraction a été modifiée pour pouvoir récupérer : la note (de 1 à 5 basée sur une image représentant une suite de ronds remplis ou vides), le pseudonyme du commentateur, la date du commentaire et le commentaire lui-même.

Le site TripAdvisor est un site multilingue qui possède plusieurs domaines ([www.tripadvisor.fr](http://www.tripadvisor.fr), [www.tripadvisor.com](http://www.tripadvisor.com), etc..). Les commentaires peuvent être saisis dans plusieurs langues et sont affichés ainsi à l'écran mais pourtant dans le source de la page HTML, les commentaires qui ne sont pas dans la langue du domaine ne font pas parti de la page html (ils sont en fait appelés au moyen de fonctions javascript). Donc depuis le domaine [www.tripadvisor.fr](http://www.tripadvisor.fr) il n'est pas possible de pouvoir extraire les commentaires écrit en anglais.

Comme il a été précisé dans le cahier des charges que seuls les commentaires en langues anglaise sont traités ensuite par le processus GATE. Les URLs du site TripAdvisor passées au prototype doivent donc être celles du domaine en langue anglaise ([www.tripadvisor.com](http://www.tripadvisor.com)).

La deuxième problématique rencontrée est la modification de la structure html pour la navigation entre les pages Internet. Sur le site, les commentaires sont affichés au nombre de cinq par pages. Une barre de navigation est présente avant le premier commentaire et après le dernier. Selon le nombre de pages de commentaire, on voit s'afficher les liens vers les pages précédentes ou suivantes.

Un lien vers la page suivante est toujours présent sur toutes les pages (visuellement c'est l'icône ») mais aucune méthode existante du parser ne permettait à l'origine de récupérer cette information.

111-115 sur 115



Figure 27 Bloc de navigation dans les pages de commentaires

La méthode existante de récupération de tous les liens de la page a donc été modifiée afin de pouvoir passer en paramètres la valeur de l'attribut « class » du lien, qui va servir de critère de recherche, en effet ce lien « Page suivante » possède une valeur unique pour l'attribut « class ».

### 5.1.2 PLUGIN COMMENTS CRAWLER

Un plugin CREOLE de type « Processing Resource » a été créé pour encapsuler les classes d'extraction et de passage modifiées.

Pour pouvoir intégrer le résultat de cette extraction au processus GATE, de fonctionnalités supplémentaires ont été intégrées à ce plugin en plus de l'appel aux classes. Une fois l'extraction effectuée le plugin va, pour chacun des blocs de commentaire anglais trouvés, créer un document GATE contenant le texte du commentaire et auquel il va rajouter, en tant que caractéristiques les autres informations récupérées (le nom de l'utilisateur, la date, la note sur 5 ainsi qu'une note en pourcent calculée depuis la note sur 5). Pour finir il intègre chaque document dans un corpus GATE.



## Entre Web 2.0 et 3.0 : opinion mining

Cette nouvelle ressource doit donc être définie comme telle dans le fichier creole.xml

```
<RESOURCE>
  <NAME>TD - TripAdvisor Comment Crawler</NAME>
  <JAR>OpinionMining.jar</JAR>
  <CLASS>tb_resource.opinionmining.TACrawler</CLASS>
  <COMMENT>Extract the comments of the web page</COMMENT>
  <PARAMETER NAME="root" RUNTIME="true" COMMENT="The Starting root for the crawl" OPTIONAL="false">java.lang.String</PARAMETER>
  <PARAMETER NAME="docPath" RUNTIME="true" COMMENT="The Path to store the documents" OPTIONAL="false">java.lang.String</PARAMETER>
  <PARAMETER NAME="outputCorpus" COMMENT="The corpus to insert the document" RUNTIME="true" OPTIONAL="false">gate.Corpus</PARAMETER>
  <ICON>binoculars.gif</ICON>
</RESOURCE>
```

Figure 28 Détail du fichier creole.xml pour le WebCrawler

Les balises Name, Jar, Class et Comment sont utiles à l'application CREOLE pour correctement intégrer le nouveau plugin.

Les trois paramètres décrits dans le fichier XML sont :

- **root** : l'URL de l'hôtel sur [www.tripadvisor.com](http://www.tripadvisor.com)
- **docPath** : le chemin du répertoire dans lequel les fichiers texte correspondant aux commentaires vont être créés.
- **outputCorpus** : le nom du corpus dans lequel doivent être intégrés les fichiers textes.

Ces trois paramètres sont modifiables au runtime comme l'indique l'attribut Runtime=true et sont obligatoires via l'attribut Optional = false.

Après avoir intégré le nouveau plugin au projet grâce au manager CREOLE et sélectionné la ressource dans le pipeline (processus) d'extraction, les trois paramètres peuvent être configurés et le plugin est prêt à être testé dans l'interface graphique de GATE.

## Entre Web 2.0 et 3.0 : opinion mining

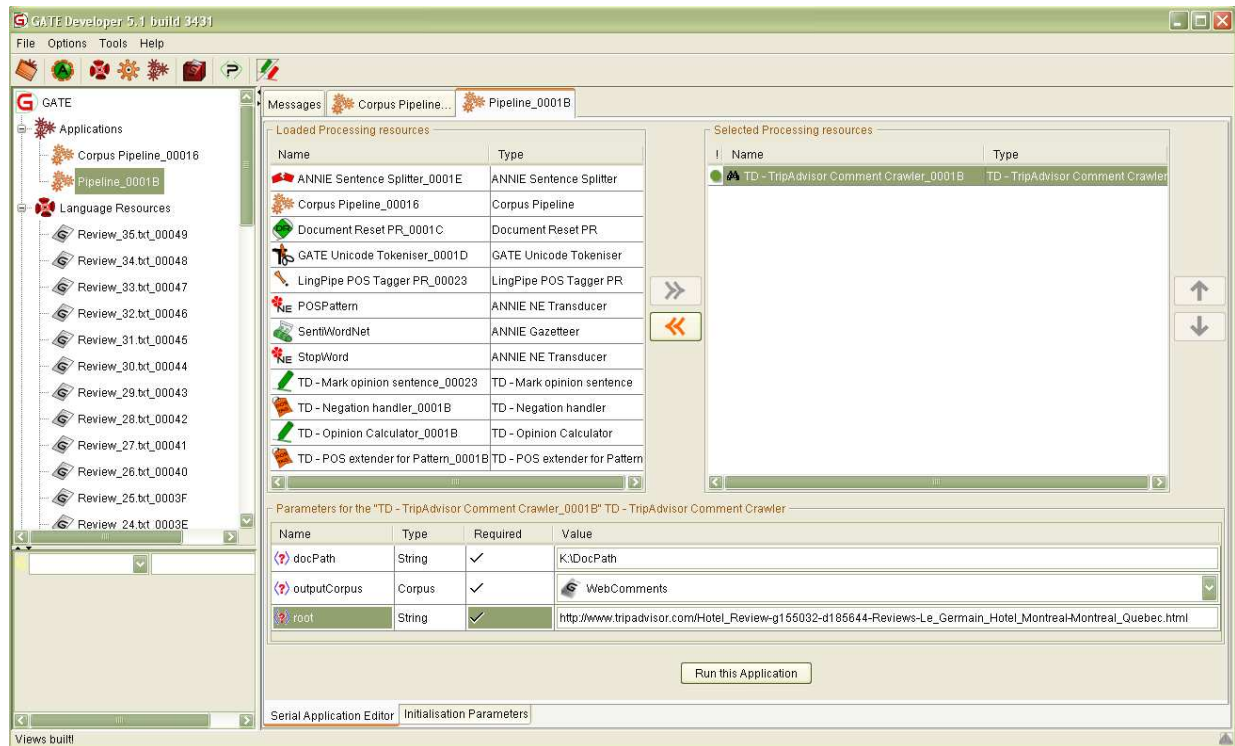


Figure 29 Pipeline d'extraction

Pour pouvoir être réutilisé par la suite le modèle du processus est sauvegardé au format .gapp (format propre à GATE).

## 5.2 TRAITEMENT DES DONNEES

Après le processus d'extraction, tous les documents se retrouvent dans le corpus GATE et vont pouvoir être traités par le second pipeline.

A la différence du pipeline d'extraction précédent ce deuxième processus est de type « Corpus pipeline », c'est-à-dire qu'il va effectuer le traitement pour chaque ressource présente dans le corpus qui lui est donné.

Le traitement effectué sur un document implique un ensemble de ressources, certaines sont présentes en standard dans GATE d'autres ont dû être créées spécifiquement pour le projet.

L'ordre d'exécution de ces différentes ressources est important car les éléments peuvent utiliser en partie les résultats des éléments précédents pour pouvoir exécuter leur traitement.

Enfin de la même manière que le pipeline d'extraction des données, le modèle à la fin est sauvegardé manuellement au format gapp.











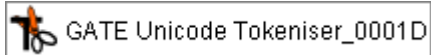
	GATE Unicode Tokeniser_0001D
	ANNIE Sentence Splitter_0001E
	LingPipe POS Tagger PR_00023
	SentiWordNet
	NE POSPattern
	NE StopWord
	TD - POS extender for Pattern_0001B
	TD - Negation handler_0001B
	TD - Opinion Calculator_0001B
	TD - Mark opinion sentence_00023

Figure 30 Liste des ressources utilisées par le pipeline de traitement linguistique

## 5.2.1 RESSOURCES STANDARD

### 5.2.1.1 TOKENIZER



L'élément GATE Unicode Tokeniser va s'occuper de séparer le texte en mots (Token) et espaces (SpaceToken) et de les annoter en conséquence. Il rajoute aussi à ces deux annotations quelques caractéristiques.

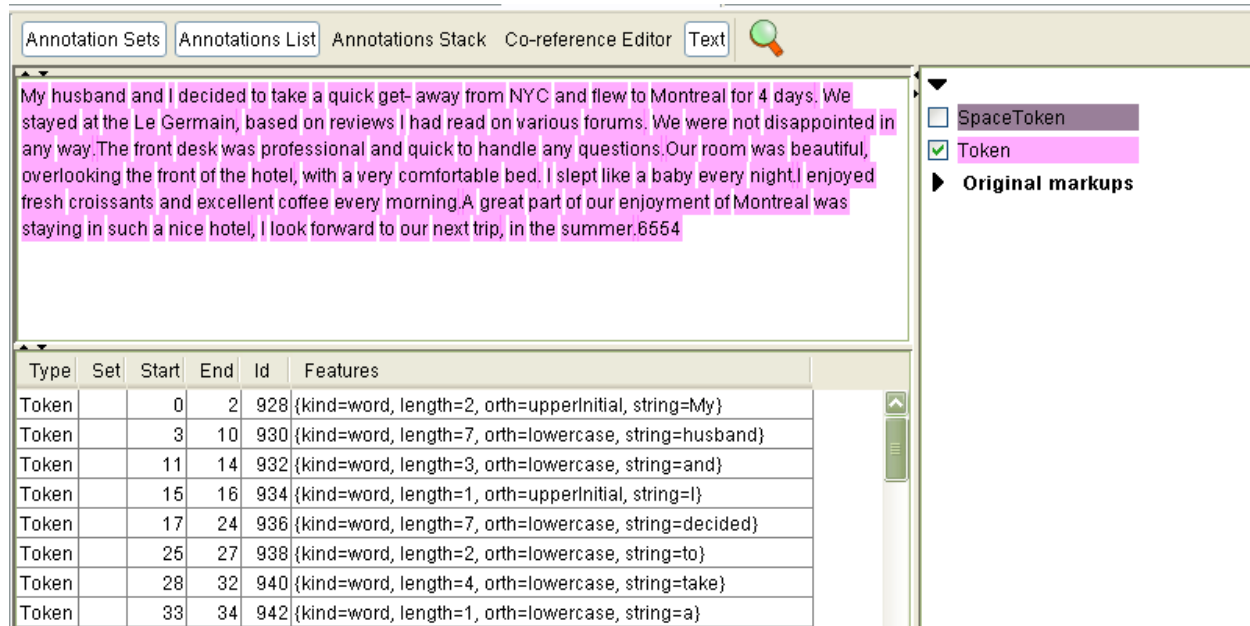
Pour les Token :

- kind = peut-être soit un mot (word), un nombre (number) ou un signe de ponctuation (punctuation)
- length = le nombre de lettre du token
- orth = détermine si le mot est en minuscule (lowercase), commence par une majuscule (upperInitial) ou et en majuscule (allCaps) ou un mélange (mixedCaps)
- string = reprend la valeur du token (le mot annoté)

Pour les SpaceToken :

- kind = majoritaire de type (space)
- length = la longueur de l'espace

## Entre Web 2.0 et 3.0 : opinion mining



Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

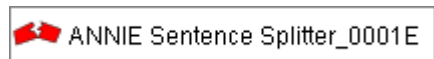
My husband and I decided to take a quick get-away from NYC and flew to Montreal for 4 days. We stayed at the Le Germain, based on reviews I had read on various forums. We were not disappointed in any way. The front desk was professional and quick to handle any questions. Our room was beautiful, overlooking the front of the hotel, with a very comfortable bed. I slept like a baby every night. I enjoyed fresh croissants and excellent coffee every morning. A great part of our enjoyment of Montreal was staying in such a nice hotel, I look forward to our next trip, in the summer.6554

Type	Set	Start	End	Id	Features
Token		0	2	928	{kind=word, length=2, orth=upperInitial, string=My}
Token		3	10	930	{kind=word, length=7, orth=lowercase, string=husband}
Token		11	14	932	{kind=word, length=3, orth=lowercase, string=and}
Token		15	16	934	{kind=word, length=1, orth=upperInitial, string=I}
Token		17	24	936	{kind=word, length=7, orth=lowercase, string=decided}
Token		25	27	938	{kind=word, length=2, orth=lowercase, string=to}
Token		28	32	940	{kind=word, length=4, orth=lowercase, string=take}
Token		33	34	942	{kind=word, length=1, orth=lowercase, string=a}

SpaceToken  
 Token  
 Original markups

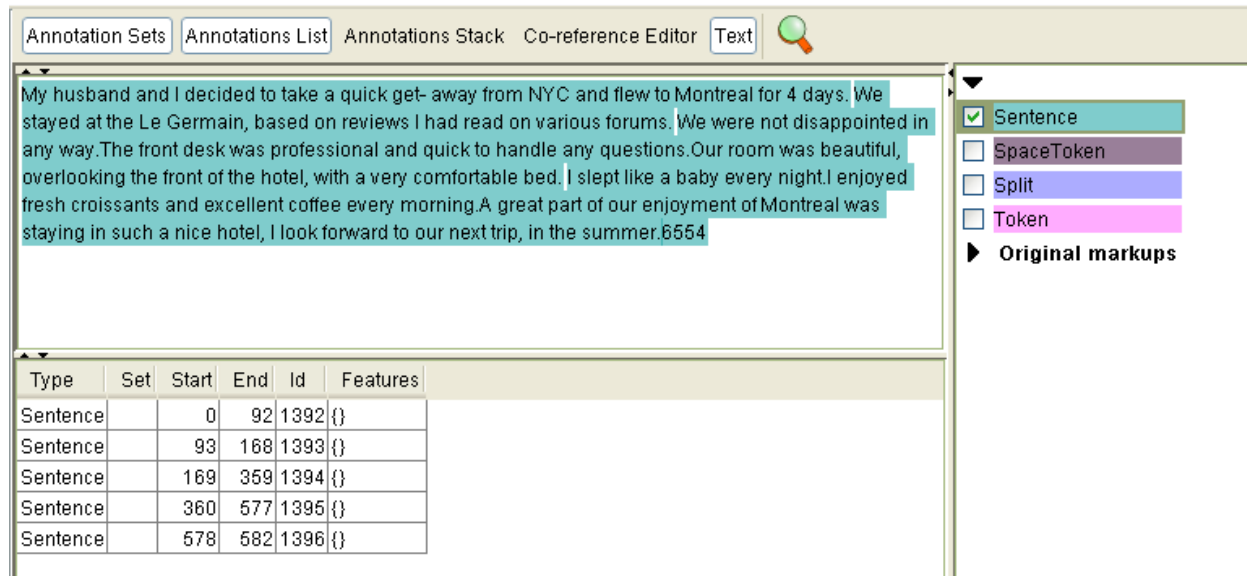
Figure 31 Résultat du Tokenizer

## 5.2.1.2 SENTENCE SPLITTER



Cet élément segmente le texte en phrase et les annote avec la valeur Sentence sans caractéristique. Les séparations entre les phrases sont annotées avec la valeur Split.

## Entre Web 2.0 et 3.0 : opinion mining




My husband and I decided to take a quick get-away from NYC and flew to Montreal for 4 days. We stayed at the Le Germain, based on reviews I had read on various forums. We were not disappointed in any way. The front desk was professional and quick to handle any questions. Our room was beautiful, overlooking the front of the hotel, with a very comfortable bed. I slept like a baby every night. I enjoyed fresh croissants and excellent coffee every morning. A great part of our enjoyment of Montreal was staying in such a nice hotel, I look forward to our next trip, in the summer.6554

Type	Set	Start	End	Id	Features
Sentence		0	92	1392	{}
Sentence		93	168	1393	{}
Sentence		169	359	1394	{}
Sentence		360	577	1395	{}
Sentence		578	582	1396	{}

Figure 32 Résultat du Sentence Splitter

## 5.2.1.3 POS TAGGER

 LingPipe POS Tagger PR\_00023

Il existe dans GATE plusieurs POS Taggers dont l'utilité est de rajouter la caractéristique Category à l'annotation Token. (Il faut donc que l'élément tokenizer aie été exécuté.)

La liste des catégories d'un mot (adjectif, prénom, verbe, etc..) utilisée par cet élément se base sur les POS tags déterminé par le corpus de Brown (pour un extrait des catégories de ce corpus voire l'annexe I).

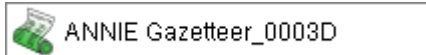
Le LingPipe POS Tagger gère correctement les premiers mots d'une phrase en majuscule à la différence du POS Tagger standard du projet ANNIE qui les considère comme des noms propres.

## Entre Web 2.0 et 3.0 : opinion mining



Figure 33 Résultat du POS Tagger de LingPipe

## 5.2.1.4 GAZETTEER



Le gazetteer standard d'ANNIE permet d'annoter avec la valeur Lookup, des mots présents dans le texte que le processus retrouve dans une ou plusieurs listes fournies. Le composant est paramétré en standard avec un fichier contenant une soixantaine de listes à vérifier (se sont de simples fichiers texte), chacune des listes peut-être paramétrée avec un terme qui sera ajouté à l'annotation en tant que caractéristiques majorType.

## Entre Web 2.0 et 3.0 : opinion mining

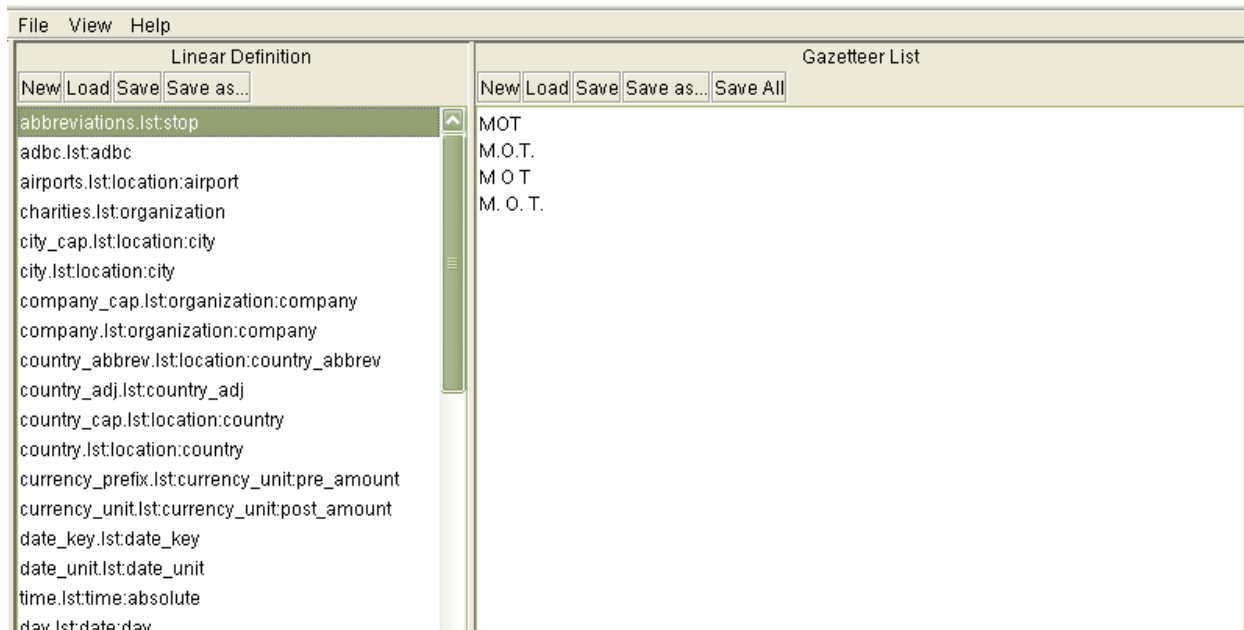


Figure 34 Paramétrage d'un gazetteer

Dans l'exemple de la figure 35, Montreal a été trouvé dans la liste city qui comme indiqué ci-dessus rajoute les valeurs : location et city à la caractéristique majorType

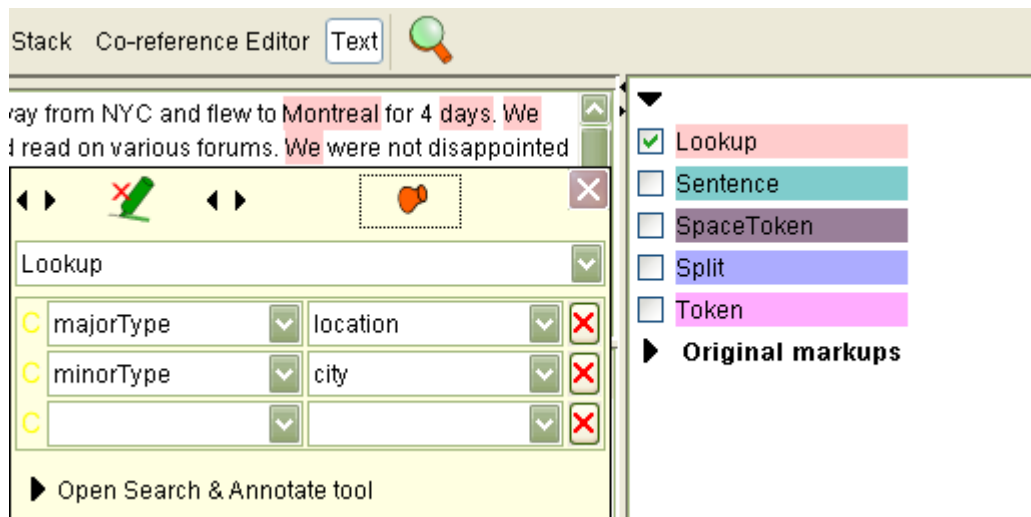


Figure 35 Résultat du gazetteer standard d'ANNIE

Cet élément peut aussi être paramétré pour rechercher d'autres valeurs à intégrer aux caractéristiques en plus du majorType, pour autant que la liste respecte un certain format.

Dans notre projet le gazetteer ne contiendra que deux listes : SentiWordNet.lst et StopWord.lst

#### 5.2.1.4.1 SENTIWORDNET



## Entre Web 2.0 et 3.0 : opinion mining

La phase d'analyse a permis de découvrir cette intéressante ressource linguistique qui détermine l'opinion d'un mot présent dans le dictionnaire WordNet. Ce fichier est le point central du prototype car tous les calculs qui suivront se baseront sur les valeurs récupérées de ce fichier.

SentiWordNet se présente sous la forme suivante :

```
a 00002843 0 0 basiscopic#1 facing or on the side toward the base
a 00002956 0 0 abducting#1 abductor#1 especially of muscles; drawing away from the midline of the body or from an adjacent part
a 00003131 0 0 adductive#1 adducting#1 adductor#1 especially of muscles; bringing together or drawing toward the midline of the body or toward an adjacent part
a 00003256 0 0 nascent#1 being born or beginning; "the nascent chicks"; "a nascent insurgency"
a 00003553 0 0 emerging#2 emergent#2 coming into existence; "an emergent republic"
a 00003700 0.25 0 dissilient#1 bursting open with force, as do some ripe seed vessels
a 00003829 0.25 0 parturient#2 giving birth; "a parturient heifer"
a 00003939 0 0 dying#1 in or associated with the process of passing from life or ceasing to be; "a dying man"; "his dying wish"; "a dying fire"; "a dying civilization"
a 00004171 0 0 moribund#2 being on the point of death; breathing your last; "a moribund patient"
```

Figure 36 Extrait du fichier SentiWordNet

Chaque ligne du document correspond à une entrée dans le dictionnaire WordNet, les deux premières colonnes correspondent à l'identifiant du mot dans le dictionnaire, la troisième colonne est la valeur positive du mot, la quatrième colonne sa valeur négative. Dans la cinquième colonne on retrouve le mot ainsi que des synonymes et sa définition. Comme un même mot peut avoir plusieurs sens (synset), le numéro de ce sens est accolé au mot avec le caractère #.

Pour pouvoir utiliser cette ressource dans le processus de traitement GATE, il a fallu tout d'abord transformer ce fichier texte. Pour ce faire l'outil de transformation de fichier texte : awk, a été utilisé.

Après transformation le fichier SentiWordNet ne contient plus que 4 colonnes séparées par le caractère |. Le mot se trouve dans la première colonne les trois autres colonnes contiennent les valeurs de positivité, de négativité et le numéro du sens.

```
dissilient|p=0.25|n=0|positionSynset=0
parturient|p=0.25|n=0|positionSynset=1
dying|p=0|n=0|positionSynset=0
moribund|p=0|n=0|positionSynset=1
last|p=0|n=0|positionSynset=4
abridged|p=0|n=0|positionSynset=0
shortened|p=0|n=0|positionSynset=3
cut|p=0|n=0|positionSynset=2
half-length|p=0|n=0|positionSynset=1
potted|p=0|n=0|positionSynset=2
```

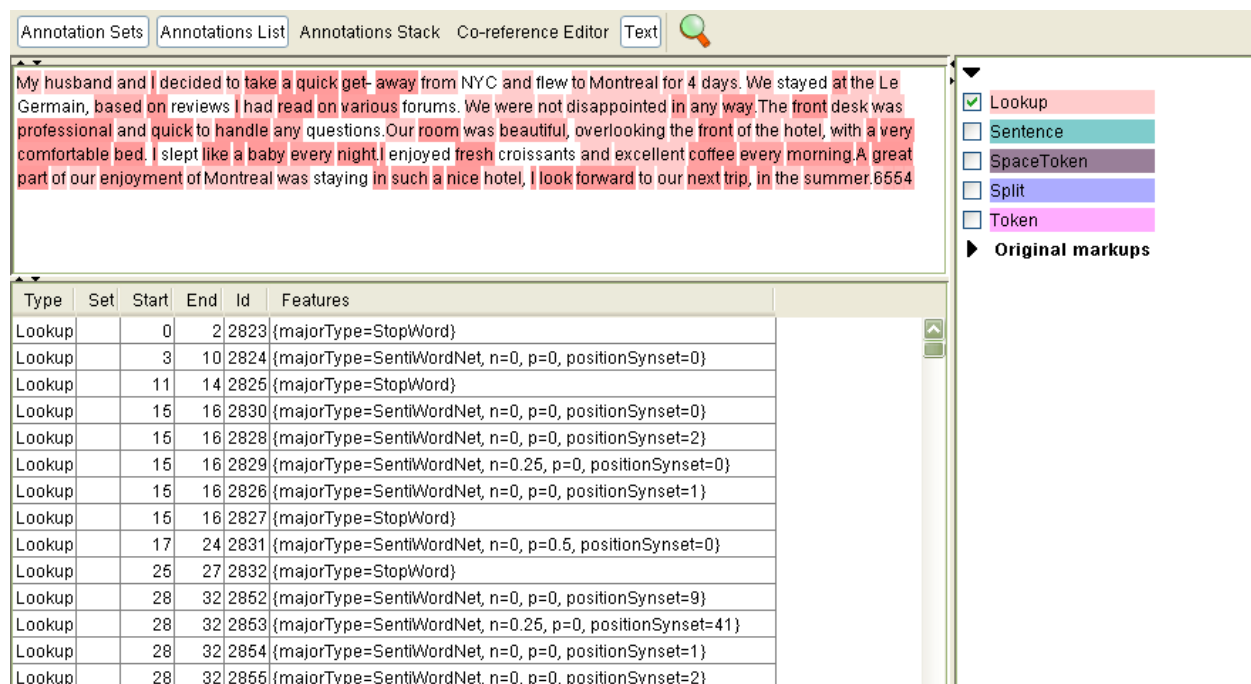
Figure 37 Extrait du fichier SentiWordNet après transformation

De cette manière, paramétré correctement le gazetter va rajouter en plus de la caractéristique majorType, les caractéristiques et leur valeur p, n et positionSynset à l'annotation Lookup correspondant au mot.

## Entre Web 2.0 et 3.0 : opinion mining

## 5.2.1.4.2 STOPWORD

En plus de la liste SentiWord, le gazetteer du processus va aussi contenir une liste de mot qui, n'ayant aucune importance en terme grammaticale mais surtout en terme d'opinion, ne devront jamais être traités durant le processus. Pour cela ils auront la caractéristiques majorType définie avec la valeur StopWord

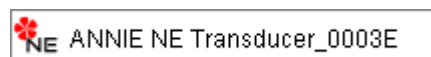


The screenshot shows the Gazetteer tool interface. At the top, there are tabs for 'Annotation Sets', 'Annotations List', 'Annotations Stack', 'Co-reference Editor', and 'Text'. The main text area contains a paragraph about a trip to Montreal, with several words highlighted in red. To the right, there is a legend with checkboxes for 'Lookup' (checked), 'Sentence', 'SpaceToken', 'Split', and 'Token'. Below the text area is a table with the following columns: Type, Set, Start, End, Id, and Features.

Type	Set	Start	End	Id	Features
Lookup		0	2	2823	{majorType=StopWord}
Lookup		3	10	2824	{majorType=SentiWordNet, n=0, p=0, positionSynset=0}
Lookup		11	14	2825	{majorType=StopWord}
Lookup		15	16	2830	{majorType=SentiWordNet, n=0, p=0, positionSynset=0}
Lookup		15	16	2828	{majorType=SentiWordNet, n=0, p=0, positionSynset=2}
Lookup		15	16	2829	{majorType=SentiWordNet, n=0.25, p=0, positionSynset=0}
Lookup		15	16	2826	{majorType=SentiWordNet, n=0, p=0, positionSynset=1}
Lookup		15	16	2827	{majorType=StopWord}
Lookup		17	24	2831	{majorType=SentiWordNet, n=0, p=0.5, positionSynset=0}
Lookup		25	27	2832	{majorType=StopWord}
Lookup		28	32	2852	{majorType=SentiWordNet, n=0, p=0, positionSynset=9}
Lookup		28	32	2853	{majorType=SentiWordNet, n=0.25, p=0, positionSynset=41}
Lookup		28	32	2854	{majorType=SentiWordNet, n=0, p=0, positionSynset=1}
Lookup		28	32	2855	{majorType=SentiWordNet, n=0, p=0, positionSynset=2}

Figure 38 Résultat de l'élément Gazetteer

## 5.2.1.5 TRANSDUCER



Les transducer utilisent des fichiers JAPE pour exécuter leur traitement, ces fichiers peuvent être très complexes et même intégrer du Java. Les deux transducers utilisés dans le pipeline restent cependant très simples. Les deux vont rajouter un nouveau set d'annotation au texte.

Le POSPattern va se baser sur l'annotation Token pour faire ressortir ceux dont la catégorie grammaticale est potentiellement porteur d'opinion (les noms les adjectifs, les verbes), en laissant de côté les conjonctions de

## Entre Web 2.0 et 3.0 : opinion mining

coordination, les chiffres, les articles, etc.. Il va aussi reprendre la caractéristique category et la copier dans la nouvelle annotation "TokenPattern".

Le StopWord va quand à lui se baser sur les Lookup et annoter les mots repérés comme faisant partie de la liste StopWord du gazetteer pour créer l'annotation "StopWord".

Ces deux nouvelles annotations vont faciliter les traitements dans la suite du processus, car il est plus aisé de retrouver les mots sur la base d'annotations plutôt que sur la base des caractéristiques.

Type	Set	Start	End	Id	Features
StopWord		0	2	4733	{rule=Stop}
TokenPattern		3	10	4683	{category=nn}
StopWord		11	14	4734	{rule=Stop}
StopWord		15	16	4735	{rule=Stop}
TokenPattern		17	24	4684	{category=vbd}
StopWord		25	27	4736	{rule=Stop}
StopWord		28	32	4737	{rule=Stop}
TokenPattern		28	32	4685	{category=vb}

Figure 39 Résultat des deux transducteurs

## 5.2.2 RESSOURCES SPECIFIQUES

Comme les traitements qui vont suivre ne pouvaient être exécutés par des éléments standards de GATE, une nouvelle ressource de type « Process Resource » a à chaque fois été ajoutée au plugin « Opinion Mining » précédemment créée.

### 5.2.2.1 POS EXTENDER FOR PATTERN

A ce stade du processus tous nos mots porteurs d'opinions ont été décelés par le gazetteer, néanmoins certains de ces mots, bien que possédant une orientation positive ou négative, ne peuvent pas être pris en

Entre Web 2.0 et 3.0 : opinion mining

considération. Soit parce que le mot est aussi annoter avec la valeur StopWord, soit parce que le mot ne fait pas partie d'un groupe de mot potentiellement lié à l'opinion. Par exemple dans le début de phrase « As one who travels half the year,... », le mot one a énormément de sens dans SentiWordNet est à donc était annoté avec la valeur lookup à plusieurs reprise, toutefois ce début de phrase ne peut pas être évalué comme subjectif et la valeur d'opinion du mot one ne devrait pas être prise en compte.

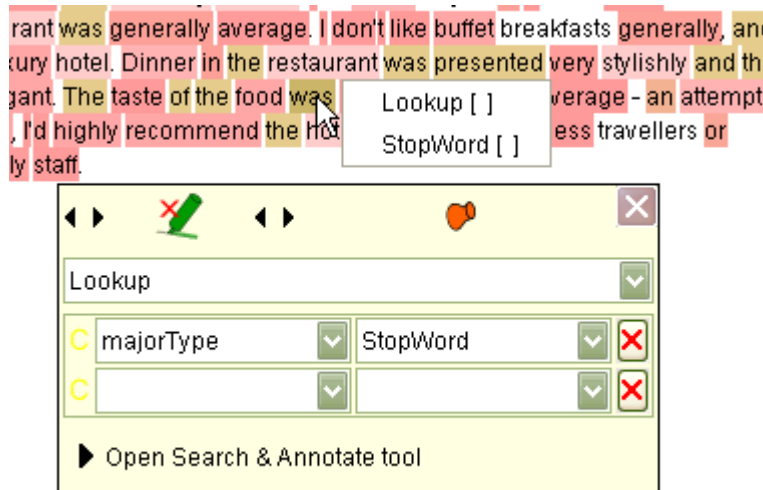


Figure 40 Exemple de mots possédant plusieurs annotations

Pour spécifier avec plus de détail quels mots, ayant une opinion, doivent être intégrés dans le calcul de l'opinion du texte, une nouvelle annotation va être ajoutée sur la base des modèles déterminés par Bin Liu (2010). Le but en est d'extraire du texte les groupes de mots qui sont plus sujet à être porteur d'opinion.

	<b>First word</b>	<b>Second word</b>	<b>Third word (Not Extracted)</b>
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Figure 41 Pattern des groupes de mots porteurs d'opinion

Les catégories affichées dans la figure 41 correspondent aux tags décrits dans le corpus de Brown (voire annexe I), JJ correspondant globalement aux adjectifs, NN aux noms, RB aux adverbes et VB aux verbes.

Chaque élément annoté précédemment par la valeur TokenPattern est contrôlé pour déterminer s'il peut appartenir à un des patterns de mots-ci-dessus. Les StopWord sont ainsi volontairement évité.

## Entre Web 2.0 et 3.0 : opinion mining

Ce sont seulement ses mots-ci, annotés "POSPattern" qui seront ensuite pris en compte dans le calcul de la valeur d'opinion du texte.

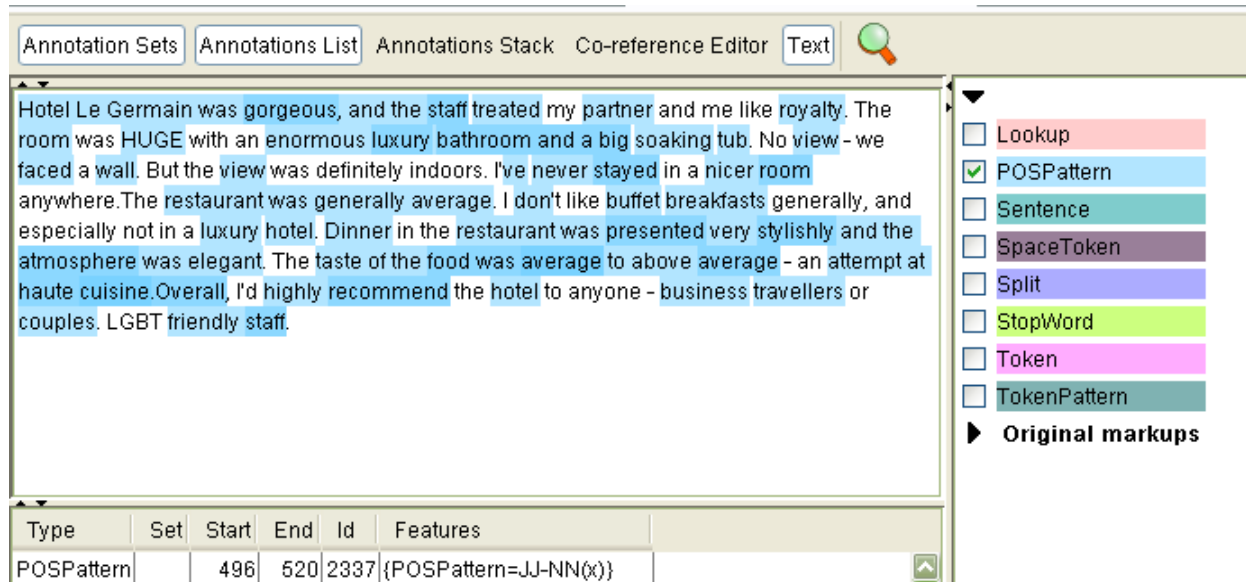


Figure 42 Résultat du POS extender

### 5.2.2.2 NEGATION HANDLER

Ce nouvel élément va parcourir les Token à la recherche des mots : no, not, never et la lettre t (pour la négation des verbe être et avoir en anglais) et inverser les valeurs d'opinion. La valeur positive est remplacée par la valeur négative et inversement. Dans le même temps toutes les annotations Lookup de ces 4 mots négatifs sont supprimés afin de ne plus les prendre en compte dans les futurs calculs d'opinion.

Une caractéristique "modified\_Negation" à la valeur true est rajouté aux mots ainsi traités.

## Entre Web 2.0 et 3.0 : opinion mining

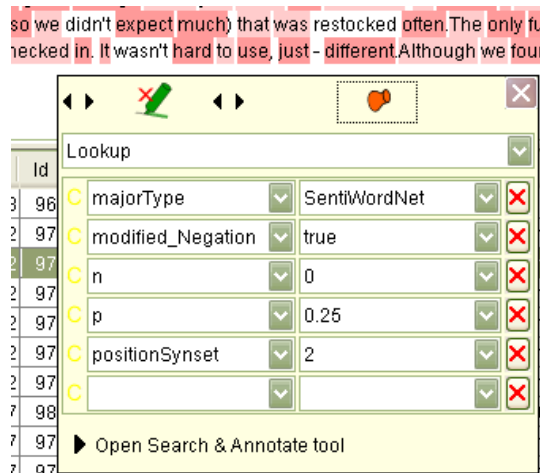


Figure 43 Résultat du Negation handler sur le verbe expect

## 5.2.2.3 OPINION CALCULATOR

Comme on a déjà pu le voir dans les exemples précédents, un mot peut avoir plusieurs sens et chacun de ces sens possède des valeurs d'opinion différentes. L'objectif de cette nouvelle ressource est de calculer la valeur d'opinion unique de chaque mot. VERMA S., BHATTACHARYYA P. (2008), dans leur rapport, propose trois manières de calculer cette valeur unique.

En posant :

$Pos(W_k)$  = Positive score given by SentiWordNet to the  $k^{th}$  Sense of  $W$

$Neg(W_k)$  = Corresponding negative score

$K$  = Number of senses of word

- Le premier calcul est la moyenne des valeurs positives et négatives :

$$Score(W) = [\sum_K Max(Pos(W_K), Neg(W_K))] / K$$

- Le second calcul est le maximum de tous les sens (il se base sur la présomption que les gens expriment leurs sentiments avec force et qu'ils emploient des mots avec une haute valeur de polarité) :

$$Score(W) = Max_K [Max(Pos(W_K), Neg(W_K))]$$

## Entre Web 2.0 et 3.0 : opinion mining

- Le troisième calcul est le poids moyen de tous les sens (dans SentiWordNet les sens sont donné selon leur fréquence, il est donc raisonnable de penser que les mots les plus fréquents peuvent avoir une pondération plus élevée par rapport aux moins fréquents) :

$$Score(W) = \frac{\sum_k weight_k + \max(pos(w_k), neg(w_k))}{\sum_k weight_k}$$

Avec :

$$weight_k = 1 - \frac{\text{position of } W \text{ in } k^{\text{th}} \text{ synset}}{\text{number of words in } k^{\text{th}} \text{ synset}}$$

Pour faciliter les références à ces calculs tout au long du présent rapport et des résultats, ces trois calculs ont été appelé "average" (pour la moyenne), "max" (pour le maximum des sens) et "weight" (pour le poids moyen des sens)

Chaque mot annoté par la valeur POSPattern (précédemment effectuée par la ressource POS Pattern extender) est prise en compte dans les calculs. Si le mot possède des valeurs d'opinion les trois calculs sont effectués et une nouvelle annotation « SentimentScore » est créée avec en caractéristique les trois valeurs trouvées.

A nouveau les stopWord ne sont pas pris en compte.

Type	Set	Start	End	Id	Features
SentimentScore		17	24	850	{averageMa
SentimentScore		35	40	851	{averageMa

Figure 44 Résultat du Opinion Calculator

#### 5.2.2.4 MARK OPINION SENTENCE

## Entre Web 2.0 et 3.0 : opinion mining

Après avoir annoté les mots, la dernière ressource va quand à elle annoter les phrases et insérer des caractéristiques au document. A titre d'indication purement visuelle, elle va ajouter trois nouvelles annotations selon l'orientation des phrases (positive, négative ou neutre) sur la base du calcul max présent dans l'annotation SentimentScore (Si ce chiffre est supérieur à .0, la phrase est positive, s'il est égal à 0, neutre et inférieur à 0, négative)

La moyenne de chacun des trois calculs : max, average et weight est calculée pour le document. Comme cette valeur peut être comprise entre -1 et 1. Elle est transformée pour ne plus avoir que des valeurs possibles en 0 à 1 (et ainsi pouvoir être comparé à la note en pourcent donné sur le site TripAdvisor)

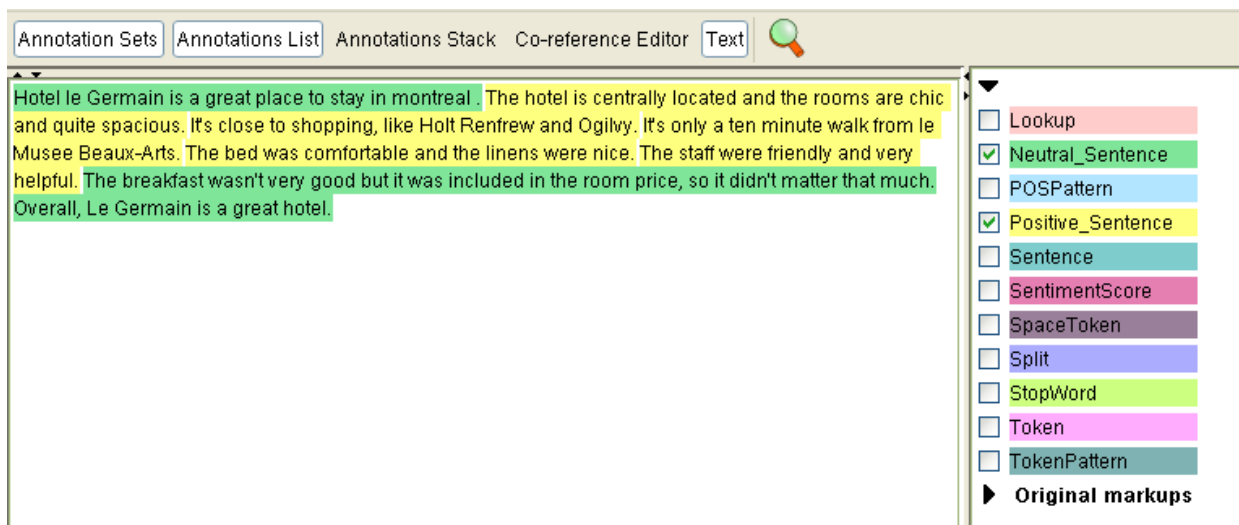


Figure 45 Résultat du Mark Opinion Sentence

### 5.2.3 LANCEMENT AUTOMATIQUE

A ce point du développement du prototype le processus de traitement est terminé, néanmoins l'exécution n'est possible que dans le logiciel GATE et les résultats sont seulement visible document par document.

Afin de ne pas avoir à ouvrir l'interface de GATE pour lancer le traitement. Le Framework du logiciel prévoit aussi des méthodes qui permettent la création et l'exécution des processus. Une de ces méthodes : « loadObjectFromFile » permet de charger un processus sur la base de son fichier modèle au format gapp.

Comme chacun de nos pipelines a été sauvegardés manuellement dans ce format, la classe Java "BatchPipelineExec" créer spécifiquement pour les deux processus va donc recevoir en paramètre les deux fichiers gapp préalablement sauvegardés pour les charger et les exécuter l'une après l'autre.



## Entre Web 2.0 et 3.0 : opinion mining

```

// load the second saved application
CorpusController application2 =
    (CorpusController)PersistenceManager.loadObjectFromFile(gappFile2);

//set the previously created Corpus
System.out.println("Execute the language treatment application ...") ;
application2.setCorpus(outputCorpus) ;

application2.execute();

System.out.println("Language Treatment finished ...") ;

```

Figure 46 Code Java de chargement des modèles

Dans la figure 46 qui est un extrait de la classe Java, on peut voir la création de l'application grâce au fichier .gapp grâce à la méthode loadObjectFromFile, une fois le modèle charger, il suffit ensuite de lancer la méthode execute() va démarrer le traitement selon les ressources définies dans le modèle.

Les résultats des deux pipelines, c'est-à-dire les documents-commentaires annotés, sont sauvegardés de manière standard GATE au format XML physiquement dans un répertoire.

La structure de ce fichier XML est la suivante :

```

<?xml version='1.0' encoding='windows-1252'?>
<GateDocument>
<!-- The document's features-->

<GateDocumentFeatures>
<Feature>
  <Name className="java.lang.String">weightDoc</Name>
  <Value className="java.lang.String">0.7100</Value>
</Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
<Feature>
</GateDocumentFeatures>
<!-- The document content area with serialized nodes -->

<TextWithNodes><Node id="0" />We<Node id="2" /> <Node id="3" />have<Node id="7" /> <Node id="8" />stayed<Node id="
<!-- The default annotation set -->

<AnnotationSet>
<Annotation Id="1" Type="Token" StartNode="0" EndNode="2">
<Feature>
  <Name className="java.lang.String">length</Name>
  <Value className="java.lang.String">2</Value>
</Feature>
<Feature>

```

Figure 47 Extrait du résultat xml d'un fichier annoté au format GATE

## Entre Web 2.0 et 3.0 : opinion mining

Dans la balise racine <GATEDocument>, on trouve trois balises enfants :

- <GATEDocumentFeatures> : qui va lister chacune des caractéristiques du document. Les caractéristiques sont définies par leur nom et leur valeur.
- <TextWithNodes> : le texte du commentaire dans lequel sont intercalées des balises <Node> l'attribut id sert de référence pour le positionnement des annotations.
- <AnnotationSet> : qui contient toutes les annotations. On y retrouve les spécifications d'une annotation, son id, son nom, son nœud de début et son nœud de fin (qui font références à la balise <Node> du point ci-dessus). Les caractéristiques des annotations sont affichées dans les balises <Feature> définie comme au niveau du document par un nom et une valeur.

---

### 5.2.3.1 PUBLICATION DES RESULTATS

La lecture de ces résultats au format GATE n'étant pas des plus aisée et comme les informations des annotations ne sont pas toutes d'un grand intérêt, une dernière classes Java, complètement externe à GATE cette fois-ci, va parcourir le répertoire contenant tous ces fichiers XML et récupérer les caractéristiques des documents pour en faire un fichier XML final structuré de la manière suivante :

## Entre Web 2.0 et 3.0 : opinion mining

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<GateDocuments>
  <Document>
    <weightDoc>0.6360</weightDoc>
    <averageDoc>0.6425</averageDoc>
    <date>31.05.2010</date>
    <ratingPercent>0.800</ratingPercent>
    <userName>columbiakat</userName>
    <maxDoc>0.6095</maxDoc>
  </Document>
  <Document>
    <weightDoc>0.7100</weightDoc>
    <averageDoc>0.7390</averageDoc>
    <date>13.05.2010</date>
    <ratingPercent>1.000</ratingPercent>
    <userName>abbeystein</userName>
    <maxDoc>0.6695</maxDoc>
  </Document>
  <Document>
    <weightDoc>0.6905</weightDoc>
    <averageDoc>0.7210</averageDoc>
    <date>07.05.2010</date>
    <ratingPercent>1.000</ratingPercent>
    <userName>Jennifer1258</userName>
    <maxDoc>0.6770</maxDoc>
  </Document>
</GateDocuments>
```

Figure 48 Extrait du fichier XML de résultat

Une balise document représente chaque commentaire, elle contient les cinq caractéristiques vraiment intéressantes : le calcul du poids (weightDoc), le calcul de la moyenne (averageDoc), la date du commentaire, la note donnée par l'utilisateur sur le site mais en pourcent, le nom de l'utilisateur et le calcul du maximum (maxDoc).

## 5.3 EVALUATION ET TESTS

Les données de tests concernent les commentaires de 31 hôtels (2354 commentaires). Comme on se le rappelle la note du commentaire (de 1 à 5) laissée sur le site TripAdvisor avait été transformée en pourcentage dès son extraction, et les calculs du max, de l'average et du weight, pouvant aller de -1 à 1 ont été transformés au moment de leur insertion des les caractéristiques du document (cf. 5.2.2.4). Nous nous retrouvons donc avec 4 valeurs, ayant la même échelle, pour chaque commentaire. Pour l'analyse des résultats, on a présumé que la note donnée par l'utilisateur à l'hôtel, reflétait correctement ce qu'il exprimait dans son commentaire. Et pour cette raison elle est utilisée comme valeur de contrôle.

Les marges qui déterminent la positivité ou la négativité d'un commentaire sont les suivantes :

- 0.3 et en dessous le commentaire est négatif
- 0.3 à 0.7 le commentaire est neutre
- 0.7 et au-dessus le commentaire est positif

S'il avait été logique de penser que les hôtels mal notés auraient eu plus de commentaires, ce n'est pas le cas en vérité. La batterie de tests contient donc majoritairement des hôtels notés positivement ou neutres.

Sur les 31 hôtels, 17 sont notés positivement, 12 sont neutres, et 2 sont notés négativement.

### 5.3.1 CRITERES D'EVALUATIONS

Pour analyser les résultats, le fichier final du processus, c'est-à-dire le résumé de tous les commentaires est traité au moyen d'Excel et les différents calculs sont comparés à la note donnée par l'utilisateur.

Pour chacune des trois valeurs d'opinion découvertes durant le processus de traitement (le max, l'average et le weight), le coefficient de corrélation et le coefficient de coordination ont été calculés.

#### 5.3.1.1 COEFFICIENT DE CORRELATION

Le coefficient de corrélation, une mesure symétrique qui varie entre -1 (relation négative parfaite) et +1 (relation positive parfaite). Il prend la valeur de 0 s'il n'y a pas de relation linéaire entre les deux variables.

Le signe de la relation nous indique le sens de la relation.

La valeur absolue du coefficient de corrélation indique l'intensité de la relation linéaire entre les variables, les valeurs absolues les plus grandes indiquant des relations plus fortes.

Entre Web 2.0 et 3.0 : opinion mining

**Avantage**

Le coefficient de corrélation donne une bonne estimation de l'ampleur de la relation linéaire entre deux variables quantitatives.

---

**5.3.1.2 COEFFICIENT DE DETERMINATION**

Mathématiquement parlant, le coefficient de détermination vaut le carré du coefficient de corrélation. Sa valeur indique le pourcentage de variation expliquée par la relation entre les deux variables. Comme sa valeur est le carré d'un nombre variant de -1 à 1, le coefficient varie de 0 et 1.

**Avantage**

Le coefficient de détermination à l'avantage de nous donner le pourcentage de la relation entre les deux variables

**Inconvénient**

Cependant la mesure est toujours positive, ce qui nous fait perdre le sens de la relation.

---

**5.3.2 RESULTATS**

Chaque hôtel a été analysé et ses résultats ont été reportés dans un fichier de résultat global.

## Mount Royal Hotel & Hostel

[http://www.tripadvisor.com/Hotel\\_Review-g60763-d557499-Reviews-Mount\\_Royal\\_Hotel\\_Hostel-New\\_York\\_City\\_New\\_York.html](http://www.tripadvisor.com/Hotel_Review-g60763-d557499-Reviews-Mount_Royal_Hotel_Hostel-New_York_City_New_York.html)

	TA Ratings	Weight	Max	Average
Moyenne	0.365	0.529	0.522	0.523
Coefficient de corrélation	1.000	0.546	0.562	0.528
Coefficient de détermination	100%	30%	32%	28%
Positivité	neutre	neutre	neutre	neutre

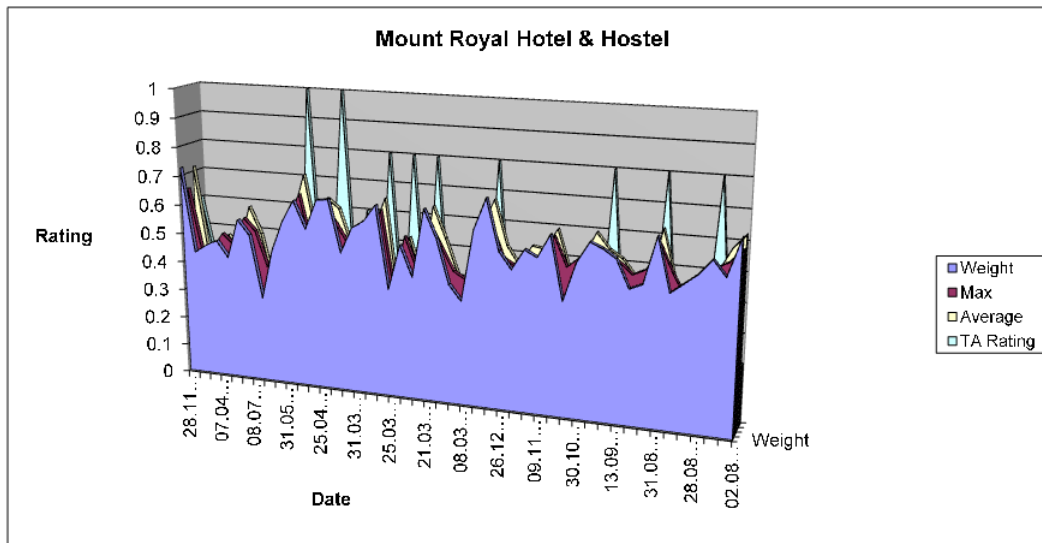


Figure 49 Exemple de l'analyse d'un hôtel

Tous les résultats par hôtel se trouvent sur le CD fournit en annexe à ce rapport.

Le premier résultat de la figure 50, qui met en relation les opinions détectées entre les trois calculs et la note attribuée sur le site semble démontrer que le prototype donne de bons résultats. Sur les 31 hôtels, la méthode de calcul la moins bonne (le max) donne tout de même un résultat qui s'approche à 61% de l'opinion fixée par le commentateur. Le weight a 71% de réussite et le meilleur calcul (l'average) atteint même 81% de résultats corrects.

Entre Web 2.0 et 3.0 : opinion mining

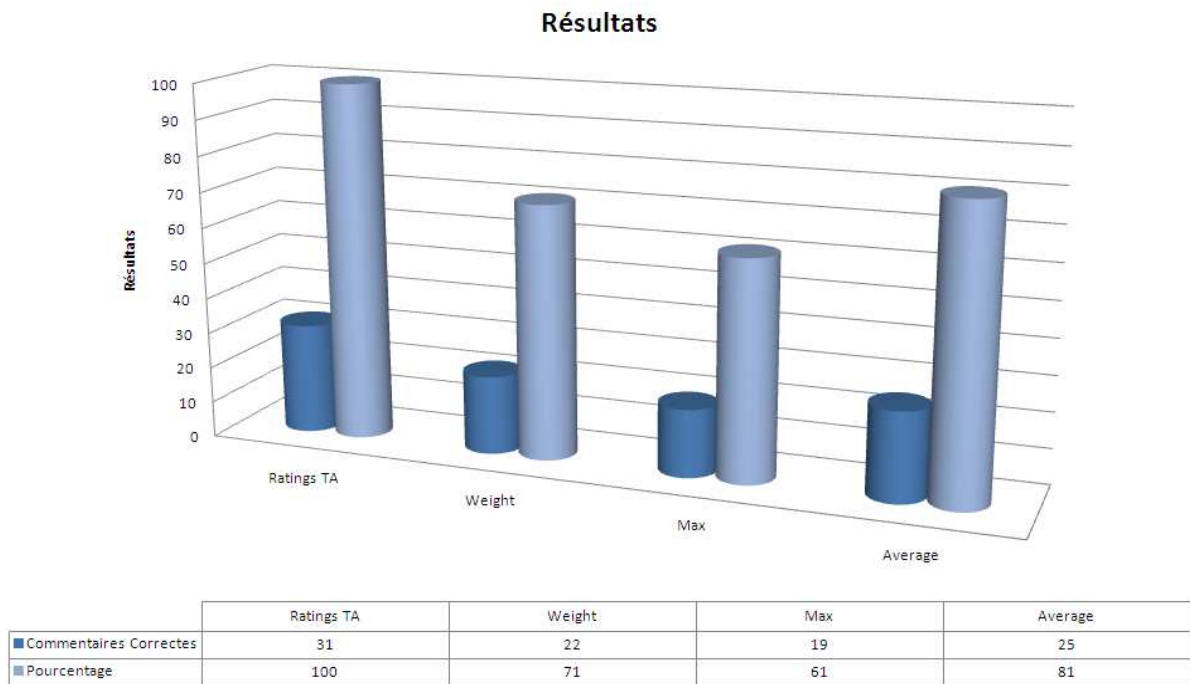


Figure 50 Résultat de l'opinion

Cependant la figure 51, qui montre les valeurs d'opinion (les réelles valeurs entre 0 et 1 calculées) pour chaque hôtel démontre toute la problématique de la validité de résultat.

Sur ce schéma on peut observer que les valeurs d'opinion entre les trois calculs sont très proches même si le calcul de l'average donne effectivement les meilleurs résultats. Cette valeur est tout de même très éloignée de la valeur d'opinion générée sur la note du commentaire.

## Entre Web 2.0 et 3.0 : opinion mining

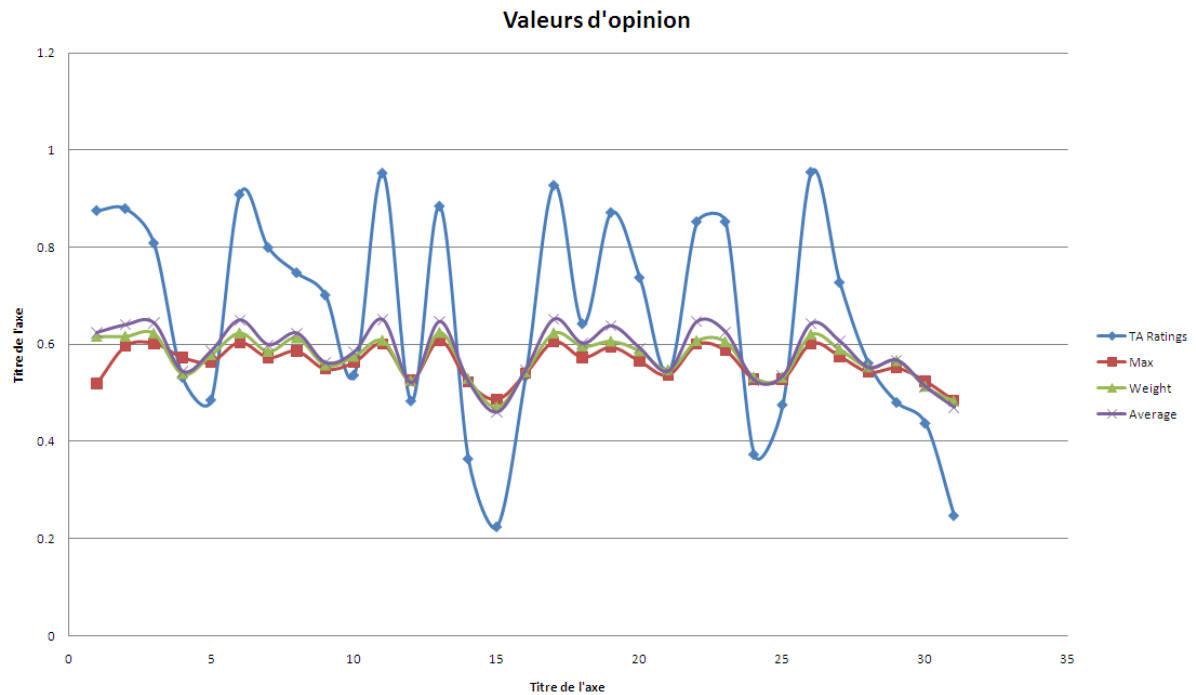


Figure 51 Evaluation des valeurs d'opinions

Cette différence peut cependant s'expliquer par le fait que dans SentiWordNet aucun des mots n'a une valeur strictement négative ou positive de un. Donc aucun des commentaires ne peut avoir une valeur maximale ou minimale et la valeur d'opinion des calculs sera forcément moins forte que celle de la note (si la note du commentateur était de 1 ou de 5).

Pour pouvoir, malgré cette différence de force, quand même valider les résultats, on va prendre en compte les deux coefficients : de corrélation et de détermination et ainsi pouvoir juger de relation entre les valeurs d'opinion.

Comme le calcul de l'average donne les meilleurs résultats c'est sur ce calcul que s'est portée l'analyse.



## Coefficients du calculs "Average"

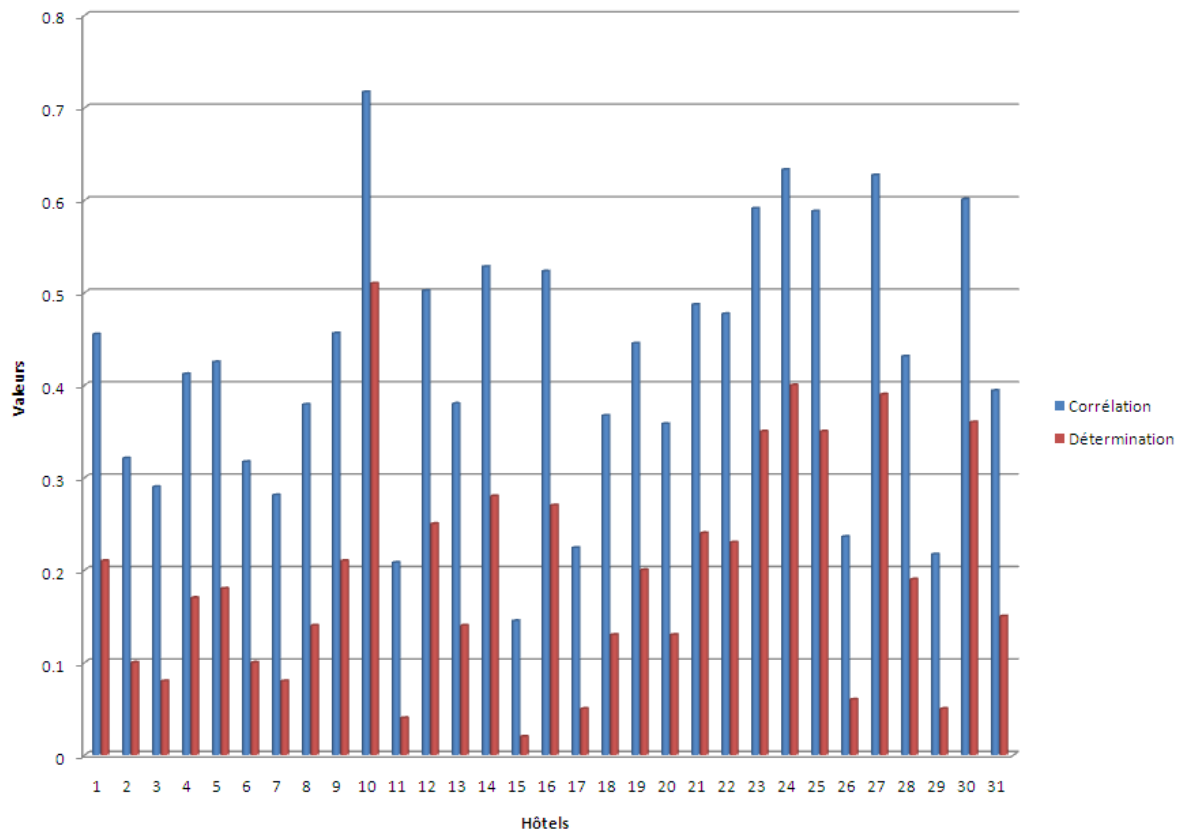


Figure 52 Valeurs des coefficients pour le calcul de la moyenne

Dans la figure 52, on peut observer avec le coefficient de corrélation que tous les chiffres sont positifs cela permet déjà d'affirmer que toutes les valeurs ont une relation qui va dans le même sens. Néanmoins l'intensité de cette relation n'est ni très élevée (la valeur maximum de 1 représentant une relation parfaite), ni très constante. La moyenne de ces coefficients de coordination pour l'ensemble des hôtels est de 0.42.

Le coefficient de détermination qui est plus facile à appréhender car il donne une valeur en % est lui aussi très bas avec une moyenne pour tous les hôtels testés de 20%.

Le dernier schéma de la figure 53 confirme que ce pourcentage est cependant légèrement supérieur dans les commentaires neutres mais catastrophique quand il s'agit de juger les commentaires négatifs.

## Entre Web 2.0 et 3.0 : opinion mining

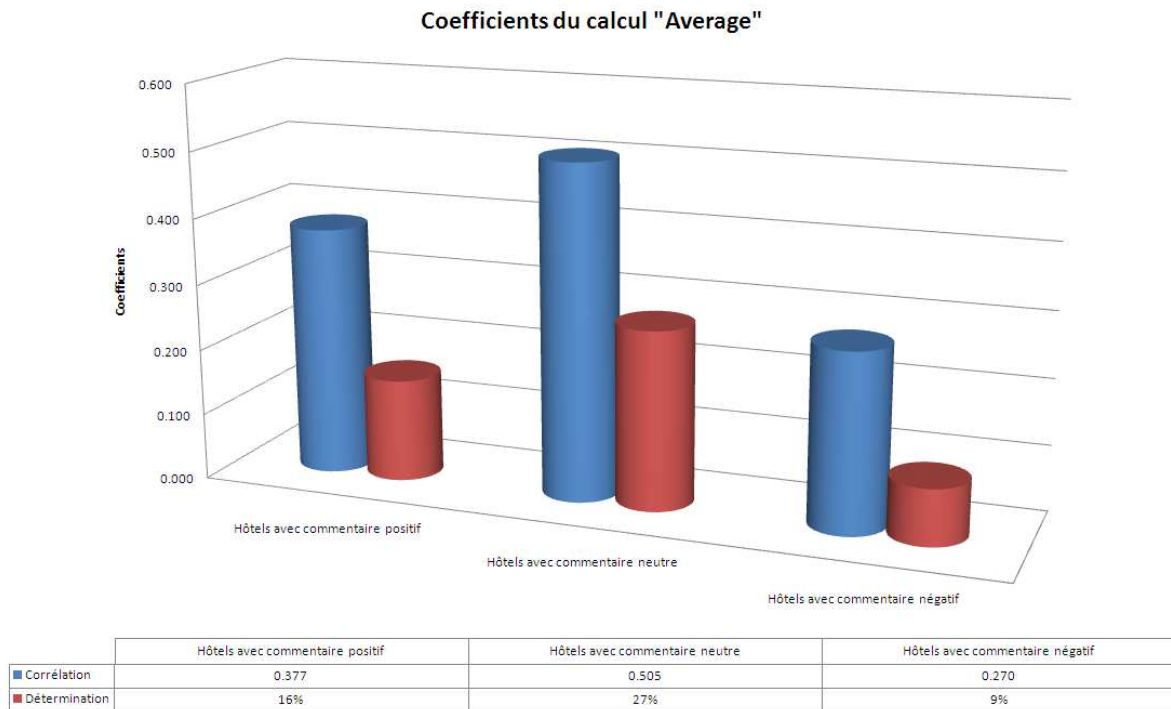


Figure 53 Valeurs des coefficients pour le calcul de la moyenne par opinion

## ANALYSE DETAILLÉE

Pour expliquer les mauvais résultats obtenus dans le calcul de la valeur d'opinion, quelques commentaires ayant une très forte différence entre la valeur calculée et celle notée, ont été analysés plus finement au moyen de l'interface graphique de GATE et les points ci-dessous sont rapidement apparus :

- Des mots ont des sens étonnants dans SentiWordNet. Par exemple le terme "impeccable" n'a que deux sens dans SentiWordNet et ils sont les deux négatifs à plus de 60%, une telle polarité dans une phrase qui contiendrait peu de mots, va changer perceptiblement la valeur d'opinion du commentaire. SentiWordNet étant un fichier généré par des machines, il n'est évidemment pas exempt d'imperfections.
- Durant le processus, aucune lemmatisation n'a été effectuée, ce sont les mots présents tels quels dans le document qui sont recherchés dans SentiWordNet. Les termes qui contiennent donc des extensions par rapport à un lemme ne seront pas trouvés dans le dictionnaire. L'exemple du mot "nicest" est flagrant, de même que tous les adjectifs superlatifs, il n'a pas d'instance dans le dictionnaire SentiWordNet, et donc le processus ne tient pas en compte un mot qui serait pourtant un lourd porteur d'opinion.
- Les mots mal orthographiés empêchent de la même manière que le point ci-dessus de retrouver les mots dans le dictionnaire SentiWordNet.

Entre Web 2.0 et 3.0 : opinion mining

- Le processus ne prend pas non plus en compte les relations entre les patterns d'opinion qui ont été définis. La gestion de la négation s'effectue par exemple seulement sur le mot qui suit le terme négatif, si la négation porte sur deux adjectifs reliés par une conjonction de coordination, le deuxième adjectif n'est pas traité correctement.

---

### 5.3.3 AMÉLIORATIONS

Le prototype intégrant les traitements qu'il était possible de développer durant le temps attribué au travail de Bachelor, il n'est évidemment pas optimal. La première des améliorations possibles serait d'intégrer l'apprentissage machine dans le traitement. Le résultat de l'état de l'art mettait en évidence le fait que se sont les méthodes semi-supervisées qui retournent les meilleurs résultats. La mise en place de cet apprentissage est toutefois longue car il faut créer un corpus d'entraînement manuellement et celui-ci doit être suffisamment conséquent pour que les données soient réellement valides.

Cependant la partie supervisée n'est pas suffisante, comme exprimé durant le Sentiment Analysis Symposium, la partie de traitement linguistique est très importante et un travail le plus soigné possible doit être effectué sur les textes pour obtenir une désambiguïsation des termes la plus complète possible. Au niveau du processus, cela se traduirait par des modules de gestion des relations grammaticales, et de la prise en compte des lemmes plutôt que des termes lors de la recherche dans SentiWordNet.

## 6 CONCLUSION

Nous avons pu voir tout au long de ce rapport que l'opinion mining était encore un domaine jeune, en cours d'étude. Grâce au web 2.0 et son aspect participatif, ce domaine est malgré tout promis à un avenir florissant. Certaines méthodes sont d'ors et déjà mise en place par des sociétés pour surfer sur la vagues des réseaux sociaux et de tous leurs échanges d'opinions, mais leurs résultats sont soumis à des critiques, portant surtout sur la validité des résultats. Des résultats qui comme ceux qui découlent du prototype peuvent sembler correctes sans analyses plus approfondies.

Si pour le moment aucun outil ne permet une détection complète de l'opinion sans développement, les logiciels existants, se basant sur une déjà grande expérience dans le domaine du text mining, sont de bonne facture.

La suite logique dans l'évolution de ce domaine va être la standardisation : du vocabulaire des règles, et des critères d'évaluation. Dès règles grammaticales adaptées à un traitement informatique, et ce dans aussi dans d'autres langues que l'anglais, vont être mises en place. Déjà des ressources similaires à WordNet apparaissent pour le français (Wolf).

Enfin si on imagine facilement toutes les limites liées à ce domaine découlant de la grande complexité du langage, certaines études prouvent cependant que des résultats probants sont possibles même sur une thématique aussi difficile à détecter que le sarcasme par exemple.

## 7 DECLARATION SUR L'HONNEUR

Je déclare, par ce document, que j'ai effectué le travail de Bachelor ci-annexé seul, sans autre aide que celles dûment signalées dans les références, et que je n'ai utilisé que les sources expressément mentionnées. Je ne donnerai aucune copie de ce rapport à un tiers sans l'autorisation conjointe du RF et du professeur chargé du suivi du travail de Bachelor, y compris au partenaire de recherche appliquée avec lequel j'ai collaboré, à l'exception des personnes qui m'ont fourni les principales informations nécessaires à la rédaction de ce travail.

## 8 GLOSSAIRE ET LISTE DES ABREVIATIONS

**Awk** : (pour Alfred Aho, Peter Weinberger et Brian Kernighan, ses créateurs) est un langage de traitement de lignes, disponible sur la plupart des systèmes Unix et sous Windows avec Gawk. Il est principalement utilisé pour la manipulation de fichiers textuels, pour des opérations de recherches, de remplacement et de transformations complexes. (Wikipedia)

**Classifier** : Représenter une unité textuelle par un ensemble prédéfini de caractéristiques linguistiques (traits) et d'utiliser la fréquence de ces traits pour décider de la catégorie d'un texte (VERNIER M., MONCEAUX L., DAILLE B., DUBREIL)

**Corpus** : Ensemble de documents regroupés dans une optique précise. En linguistique, le terme est lié au développement des systèmes informatiques, en particulier à la constitution de bases de données textuelles. On parle de corpus pour désigner l'aspect normatif de la langue, sa structure et son code en particulier (Wikipedia)

**GATE**: General Architecture for Text Engineering

**Glose** : une glose est un commentaire linguistique ajouté au texte expliquant un mot étranger ou dialectal, un terme rare. Actuellement glose renvoie à l'explication de ce mot. (Wikipedia)

**Hapax** : un mot qui n'a qu'une seule occurrence dans la littérature. Il peut être un mot rare mais aussi une erreur (Wikipedia).

**JAPE** : Java Annotation Patterns Engine

**Lemme**: (ou lexie, item lexical) est l'unité autonome constituante du lexique d'une langue. C'est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire et dite de forme canonique (ex : pour les mots, le singulier, pour les verbes, l'infinitif. (Wikipedia)

**Lexique** : En linguistique, le lexique d'une langue constitue l'ensemble de ses lemmes ou, d'une manière plus courante mais moins précise, l'ensemble de ses mots. Dans l'usage courant, on utilise plus facilement le terme vocabulaire. (Wikipedia)

**LSA** : (pour Latent Semantic Analysis, Analyse sémantique latente en français) est un procédé de traitement des langues naturelles, dans le cadre de la sémantique vectorielle, qui permet d'établir des relations entre un ensemble de documents et les termes qu'ils contiennent. (Wikipedia)

**Machine learning** : (en français **Apprentissage automatique** ou **Apprentissage artificiel**) consiste dans le développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. L'apprentissage automatique est un champ d'étude de l'intelligence artificielle. (Wikipedia)

**Morphosyntaxe** : Etude de la forme des mots et de leur syntaxe (dictionnaire.reverso)

**NER** : Name Entity Recognition

Entre Web 2.0 et 3.0 : opinion mining

**NLP (TALN en français)** pour **Natural Language Processing (Traitement automatique du langage naturel en français)** : discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. (Ex : la traduction automatique, la correction orthographique, la reconnaissance vocal, etc...) (Wikipedia)

**Objectif** : ce qui existe en soi, indépendamment du sujet pensant (atlf)

**Ontologie** : ensemble structuré des termes et concepts représentant le sens d'un champ d'informations (métadonnées, espace de noms, éléments d'un domaine de connaissances) et qui décrit les concepts et les relations liés à un domaine de connaissance particulier, tout en spécifiant le vocabulaire de ce domaine et sa sémantique. (Wikipedia)

**Opinion** : Manière de penser sur un sujet ou un ensemble de sujets, jugement personnel que l'on porte sur une question, qui n'implique pas que ce jugement soit obligatoirement juste. (atlf)

**Opinion Mining (ou Sentiment Analysis)** : (**fouille d'opinion** ou **détection des sentiments** en français) large domaine du NLP, qui vise à déterminer l'attitude d'un orateur ou d'un auteur au sujet d'un certain thème. (Wikipedia)

**Part-of-speech Tagging (POS Tagging)** : (en français **étiquetage des rôles grammaticaux** ou **étiquetage morphosyntaxique**) est le processus qui permet d'identifier pour chaque mot sa classe morphosyntaxique à partir de son contexte de connaissance lexicales (Ex : préposition, nom commun masculin pluriel, ponctuation, verbe participe passé singulier, etc..) (technolanguage)

**PMI** : (pour **Pointwise Mutual Information** en français **coefficient de corrélation**) est une mesure d'association utilisée dans les théories de l'information et les statistiques. Si applique ce calcul à deux variable, le résultat 0 exprime le fait que leux variables sont indépendantes, un résultat positif, que les deux variables sont positivement corrélées alors qu'un résultat négatif, que les deux variables sont négativement corrélées. (Wikipedia)

**Sémantique** : branche de la linguistique qui étudie les signifiés (le concept). L'analyse sémantique se différencie de l'analyse lexicale, en s'intéressant au mot pour le mot. Alors que l'analyse lexicale va s'intéresser à celui-ci dans la phrase entière, en relation avec d'autres mots compléments. (Wikipedia)

**Stemming (Lemmatisation en français)** : Processus morphologique permettant de regrouper les variantes d'un mot (Wikipedia)

**Subjectif** : ce qui est propre à un sujet déterminé, qui ne vaut que pou lui seul (atlf)

**TAL** : Traitement automatique des Langues

**Text mining** : Ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.(Wikipedia)

**TF-IDF** : (de l'anglais term frequency-inverse document frequency) est une méthode de pondération. Cette mesure statistique permet d'évaluer l'importance d'un mot par rapport à un document extrait d'un corpus. Le poids augmente proportionnellement en fonction du nombre d'occurrences du mot dans le document et varie également en fonction de la fréquence du mot dans le corpus. (academic)

**Théorie de l'Appraisal** (ou **Théorie de l'évaluation cognitive**) : Théorie selon laquelle l'émotion est le fruit des évaluations cognitives que l'individu fait au sujet d'un événement (qu'il soit interne ou externe) ou d'une situation. Cette approche suppose que, pour comprendre les émotions, il est tout d'abord nécessaire de comprendre les évaluations que l'individu fait au sujet des événements de son environnement. (Wikipedia)

## 9 SOURCES

### 9.1 WEBOGRAPHIE

#### 9.1.1 ETAT DE L'ART

##### 9.1.1.1 CONCEPTS

BING LIU. (2009) "Opinion Mining and Search" <<http://www.cs.uic.edu/~liub/opinion-mining-and-search.pdf>>

FERRARI S., MATHET Y., CHARNOIS T., LEGALLOIS D. (2008) "Retour d'expérience sur deux approches" <[http://dominique-legallois.6mablog.com/public/Analyse\\_d\\_opinion\\_-\\_discours\\_valuatif\\_et\\_classification\\_de\\_documents.pdf](http://dominique-legallois.6mablog.com/public/Analyse_d_opinion_-_discours_valuatif_et_classification_de_documents.pdf)>

STAVRIANOU A., CHAUCHAT J-H. (2008) "Opinion Mining Issues and Agreement Identification in Forum Texts" <<http://www.lirmm.fr/~mroche/FODOP08/ArticlesFODOP08/Article5.pdf>>

Deft'09 (2009) Actes de l'atelier de clôture de la cinquième édition du Défi Fouille de Textes <<http://deft09.limsi.fr/index.php?id=1&lang=fr>>

ZIGHED D.A., VENTURINI G. (sous la direction) (2009). *Fouille de données d'opinions*. Revue des Nouvelles Technologies de l'Information, Cépaduès-Edition. <<http://cepadues.com/docs/910/RNTI-E-17.pdf>>

BO PANG, LEE L. (2008) "Opinion Mining and Sentiment Analysis" <<http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf>>

GILLOT S. (2010) "Fouille d'opinions" <[ftp://ftp.irisa.fr/local/caps/DEPOTS/BIBLIO2010/Gillot\\_Sebastien.pdf](ftp://ftp.irisa.fr/local/caps/DEPOTS/BIBLIO2010/Gillot_Sebastien.pdf)>

GARDIN P. (2009) "Application de la théorie de l'Appraisal à l'analyse d'opinion" <[http://majecstic2009.univ-avignon.fr/Actes\\_MajecSTIC\\_RJCP/MajecSTIC/articles/1272.pdf](http://majecstic2009.univ-avignon.fr/Actes_MajecSTIC_RJCP/MajecSTIC/articles/1272.pdf)>

##### 9.1.1.2 HISTORIQUE

PRABOWO R., THELWALL M. (2009) "Sentiment Analysis: A combined Approach"

<<http://www.cyberemotions.eu/rudy-sentiment-preprint.pdf>>

---

#### 9.1.1.3 METHODES

MOHAMADALLY H., FOMANI B. (2006) "SVM : Machines à Vecteurs de Support ou Séparateurs à vastes Marges" <[http://georges.gardarin.free.fr/Surveys\\_DM/Survey\\_SVM.pdf](http://georges.gardarin.free.fr/Surveys_DM/Survey_SVM.pdf)>

VERNIER M., MONCEAUX L., DAILLE B., DUBREIL E (2009) "Catégorisation des évaluations dans un corpus de blogs multi-domaine" <[http://hal.archives-ouvertes.fr/docs/00/40/54/07/PDF/RNTI\\_VernierMonceauxDailleDubreil.pdf](http://hal.archives-ouvertes.fr/docs/00/40/54/07/PDF/RNTI_VernierMonceauxDailleDubreil.pdf)>

---

#### 9.1.1.4 RESSOURCES

ESULI A., SEBASTIANI F. (2006) "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining" <<http://nmis.isti.cnr.it/sebastiani/Publications/LREC06.pdf>>

RAPID-I *RapidMiner 4.4 User Guide, Operator Reference, Developer Tutorial*

<<http://switch.dl.sourceforge.net/project/yale/1.%20RapidMiner/4.4/rapidminer-4.4-tutorial.pdf>>

JUNG B., PIRKER H. (2009) "SEMPRE Sentiment Classification with GATE"

<<http://www.ofai.at/~bernhard.jung/projects/semprer/tr-sentiment-classification.pdf>>

KOZAREVA Z. (2010) "CS544: NER with Weka" <[http://www.isi.edu/natural-language/teaching/cs544/kozareva\\_ner\\_weka.pdf](http://www.isi.edu/natural-language/teaching/cs544/kozareva_ner_weka.pdf)>

ES-SALIHE M., BOND S. (2006) "Étude des frameworks UIMA, GATE et OpenNLP" <[http://www.crim.ca/fr/r-d/technologie\\_internet/documents/Etude-UIMA-GATE-OpenNLP.pdf](http://www.crim.ca/fr/r-d/technologie_internet/documents/Etude-UIMA-GATE-OpenNLP.pdf)>

---

#### 9.1.1.5 CHALLENGE

BING LIU (2010) "Sentiment Analysis : A Multi-Faceted Problem" <<http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf>>

---

#### 9.1.1.6 CONCLUSION

Le blogger LINGWAY (21.04.2010) <<http://lebloglingway.blogspot.com/2010/04/sentiment-analysis-symposium-lingway-y.html>>

JENNIFER ZAINO (2010) "The Future of Sentiment Analytics" pour *Semantic Web*

<[http://www.semanticweb.com/news/the\\_future\\_of\\_sentiment\\_analytics\\_157586.asp](http://www.semanticweb.com/news/the_future_of_sentiment_analytics_157586.asp)>



---

### 9.1.2 GATE

DAVIS B. (2010) *GATE 4.0 Tutorial* <<http://GATE.ac.uk/wiki/quick-start/GATE-Tutorial-2010-GATE4.0.pdf>>

MAYNARD D. (2004) "Introduction to ANNIE" <<http://GATE.ac.uk/sale/talks/annie-tutorial.ppt>>

---

### 9.1.3 PROTOTYPE

SAGGION H., FUNK A. (2009) "Extracting Opinions and Facts for Business Intelligence" <<http://GATE.ac.uk/sale/rnti-09/final-version/Saggion-Funks-OM-09.pdf>>

CUNNINGHAM et al. (2010) "Developing Language Processing Components with GATE Version 5 (a User Guide)" <<http://GATE.ac.uk/sale/tao/tao.pdf>>

THAKKER D., OSMAN T., LAKIN P. (2009) "GATE JAPE Grammar Tutorial" <<http://GATE.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf>>

VERMA S., BHATTACHARYYA P. (2008) "Incorporating Semantic Knowledge for Sentiment Analysis" <<http://www.cse.iitb.ac.in/~pb/papers/icon09-sa.pdf>>

BING LIU (2010) "Sentiment Analysis and Subjectivity" <<http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf>>

STEFANO BACCIANELLA, ANDREA ESULI, FABRIZIO SEBASTIANI. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204. <<http://sentiwordnet.isti.cnr.it/>>

DEVELOPPEMENT

---

PROTOTYPE

CRETTON F Projet Java "WebCrawler"

## 10 ANNEXE

### I CATEGORIES DU CORPUS DE BROWN

Le corpus de Brown répertorie 226 catégories, seules les plus pertinentes sont listées dans ce tableau.

## Entre Web 2.0 et 3.0 : opinion mining

Tag	Description
JJ	adjective
JJ\$	adjective, genitive
JJ+JJ	adjective, hyphenated pair
JJR	adjective, comparative
JJR+CS	adjective + conjunction, coordinating
JJS	adjective, semantically superlative
JJT	adjective, superlative
NN	noun, singular, common
NN\$	noun, singular, common, genitive
NN+BEZ	noun, singular, common + verb "to be", present tense, 3rd person singular
NN+HVD	noun, singular, common + verb "to have", past tense
NN+HVZ	noun, singular, common + verb "to have", present tense, 3rd person singular
NN+IN	noun, singular, common + preposition
NN+MD	noun, singular, common + modal auxiliary
NN+NN	noun, singular, common, hyphenated pair
NNS	noun, plural, common
NNS\$	noun, plural, common, genitive
NNS+MD	noun, plural, common + modal auxiliary
NP	noun, singular, proper
NP\$	noun, singular, proper, genitive

## Entre Web 2.0 et 3.0 : opinion mining

Tag	Description
JJ	adjective
JJ\$	adjective, genitive
JJ+JJ	adjective, hyphenated pair
JJR	adjective, comparative
JJR+CS	adjective + conjunction, coordinating
JJS	adjective, semantically superlative
JJT	adjective, superlative
NN	noun, singular, common
NN\$	noun, singular, common, genitive
NN+BEZ	noun, singular, common + verb "to be", present tense, 3rd person singular
NN+HVD	noun, singular, common + verb "to have", past tense
NN+HVZ	noun, singular, common + verb "to have", present tense, 3rd person singular
NN+IN	noun, singular, common + preposition
NN+MD	noun, singular, common + modal auxiliary
NN+NN	noun, singular, common, hyphenated pair
NNS	noun, plural, common
NNS\$	noun, plural, common, genitive
NNS+MD	noun, plural, common + modal auxiliary
NP	noun, singular, proper
NP\$	noun, singular, proper, genitive

## Entre Web 2.0 et 3.0 : opinion mining

<b>RB</b>	adverb
<b>RB\$</b>	adverb, genitive
<b>RB+BEZ</b>	adverb + verb "to be", present tense, 3rd person singular
<b>RB+CS</b>	adverb + conjunction, coordinating
<b>RBR</b>	adverb, comparative
<b>RBR+CS</b>	adverb, comparative + conjunction, coordinating
<b>RBT</b>	adverb, superlative
<b>VB</b>	verb, base: uninflected present, imperative or infinitive
<b>VB+AT</b>	verb, base: uninflected present or infinitive + article
<b>VB+IN</b>	verb, base: uninflected present, imperative or infinitive + preposition
<b>VB+JJ</b>	verb, base: uninflected present, imperative or infinitive + adjective
<b>VB+PPO</b>	verb, uninflected present tense + pronoun, personal, accusative
<b>VB+RP</b>	verb, imperative + adverbial particle
<b>VB+TO</b>	verb, base: uninflected present, imperative or infinitive + infinitival to
<b>VB+VB</b>	verb, base: uninflected present, imperative or infinitive; hyphenated pair
<b>VBD</b>	verb, past tense
<b>VBG</b>	verb, present participle or gerund
<b>VBG+TO</b>	verb, present participle + infinitival to
<b>VBN</b>	verb, past participle
<b>VBN+TO</b>	verb, past participle + infinitival to
<b>VBZ</b>	verb, present tense, 3rd person singular