



Opinion Mining Master SETI-AIC

Chloé Clavel, Telecom-ParisTech

<https://clavel.wp.imt.fr/>





Introduction



Introduction

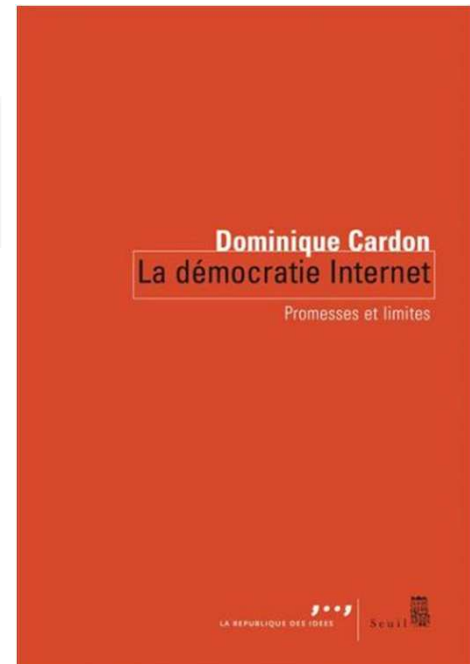
- **Différentes dénominations**
 - *Opinion extraction, opinion mining, sentiment analysis, subjectivity analysis, affect sensing, emotion detection*
- **Les applications**
 - L'analyse des réseaux sociaux
 - L'interaction humain-agent : ex: chatbot

Données Sociales et analyse d'opinions

- **Les données sociales:**
 - Expressions des citoyens et des médias sur le web
- **Contexte :**
 - Renouvellement des possibilités de critique et d'action via Internet



Lecture : « La démocratie
Internet »
Dominique Cardon





Données Sociales et analyse d'opinions

■ Enjeux :

- Analyse des tendances sociétales
- Analyse des opinions des citoyens sur les candidats lors des élections
- Analyse des critiques de films (movie reviews)
- Analyse des opinions des internautes sur un produit/Analyse de l'e-reputation d'une marque, d'un produit
- Identifier les clients cibles/systèmes de recommandation
- Évaluer le succès de campagne de communication



Données Sociales et analyse d'opinions

■ Disciplines impliquées :

- La sociologie :
 - analyse qualitative/manuelle/sociologique sur des corpus de taille réduite sélectionnés pour former un panel d'études
- L'informatique :
 - développement de méthodes d'analyse automatique de gros corpus

L'interaction humain-agent/robot

■ Agents artificiels & Robotique

- Analyser et reproduire les comportements humains pour interagir socialement avec l'homme.
- Agents conversationnels animés,
- Robots & « affective avatar »



[Pelachaud, 2005]



Robots AIBO & KISMET [Breazeal et Aryananda, 2002]



Nao
(Aldebaran Robotics)

https://www.youtube.com/watch?v=Ea_ytY0UDs0 Luc Steels - BREAKING THE WALL TO LIVING ROBOTS. How Artificial Intelligence Research Tries to Build Intelligent Autonomous Systems

Interaction humain-agent: LiveChat et relation client



L'interaction humain-robot

- **Le robot Berenson au quai Branly**
 - « *Les visiteurs ont été invités à observer le comportement de Berenson et à interagir avec lui, contribuant ainsi à définir les critères d'appréciations esthétiques de ce robot amateur d'art. »*





Terminologie et modèles théoriques



Détection d'opinions : enjeux et difficultés

- **Aller au-delà d'une distinction positif/négatif**
 - les opinions sont des phénomènes subjectifs dont l'analyse dépend :
 - De la situation dans laquelle s'exprime l'opinion
 - De la personne qui exprime l'opinion (ex: les tweets)
 - Phénomènes liés au sentiment/opinion
 - *Émotion, opinion, sentiment, humeur, attitude, positionnement interpersonnel, traits de personnalité, affect, jugement, appréciation*



Détection d'opinions : enjeux et difficultés

- **A choisir en fonction de l'application:**
 - Bien définir ce que l'on cherche à détecter !
 - Ex: les concepts de satisfaction/mécontentement/attentes des enquêtes de satisfaction ne sont pas pertinents pour l'analyse des corpus web ou des centres d'appels
 - S'appuyer sur des modèles théoriques issus de la psychologie ou de la sociologie



Terminologie

■ La typologie de Scherer:

- **Emotion**: Phénomène bref, réaction physiologique, évaluation d'un événement majeur (stimulus)
- **Humeur (Mood)**: diffus, sans cause, faible intensité, longue durée
- **Positionnement interpersonnel (interpersonal stance)**: positionnement affectif vis-à-vis d'une autre personne dans une interaction
- **Attitudes**: durable, croyances colorées affectivement, disposition envers des objets et des personnes
- **Traits de personnalité** : dispositions stables liées à la personnalité, tendances comportementales typiques

■ Exercice : attribuer les exemples ci-dessous aux classes ci-dessus :

- Sympathique
- Maussade
- Méprisant
- Jaloux
- Triste

Terminologie et applications

- **Exemples d'application et terminologie associée selon la typologie de Scherer :**

- Détecter lorsque l'utilisateur est énervé dans un système de dialogue humain-machine
 - -> émotion



- Détecter lorsque l'étudiant est désorienté, ennuyé ou frustré dans des systèmes de e-learning
 - -> émotion



Terminologie et applications

- Détecter des personnes déprimées pour des robots dans le cadre de l'assistance aux personnes âgées
 - -> humeur



- Détecter des comportements amicaux ou hostiles dans des conversations
 - -> positionnement interpersonnel
- Détecter des personnalités plutôt extraverties ou introverties pour des *Serious games* d'entraînement aux entretiens d'embauche
 - -> traits de personnalité



Modèles théoriques utilisés en sentiment analysis

- **Théorie de l'évaluation adaptée pour le TAL [Martin and White, 2005]**
 - Une expression évaluative (porteuse de subjectivité) est définie par :
 - une *source* qui exprime ...
 - ex. le locuteur
 - ... une *évaluation* sur ...
 - Type d'évaluation : affect, jugement ou appréciation
 - Affect : réaction personnelle, référence à un état émotionnel (bonheur, etc.) (e.g., 'I am very angry with you')
 - Jugement : attributions de qualités (capacité, ténacité) à des personnes en fonction de principes normatifs (e.g., 'your cruelty is well-known')
 - Appréciation : évaluation de choses (produit, processus) (e.g., 'I find that this T-shirt is ugly')
 - Polarité : positif/négatif
 - Intensité
 - ... une *cible*
 - Situation, produit ou personne



Modèles théoriques utilisés en sentiment analysis

- **Théorie de l'évaluation adaptée pour le TAL [Martin and White, 2005]**
 - Ex. « la facture est trop chère »
 - Client qui exprime une appréciation sur un produit
 - Avantage :
 - permet de distinguer des expressions d'opinions de différentes personnes sur différentes cibles
 - permet de distinguer les expressions d'affect et de jugement



Normes et W3C

- **Encore peu de choses sur les opinions et les sentiments**
 - <http://www.w3.org/community/sentiment/> :Linked Data Models for Emotion and Sentiment Analysis Community Group
- **Bien définies pour les émotions :**
 - Emotion Markup Language
<http://www.w3.org/TR/2014/REC-emotionml-20140522/>



Méthodes d'analyse d'opinions



Enjeux et difficultés :

EXO : La critique est elle positive ou négative?

souligner les expressions correspondant à l'expression d'une opinion. Paraissent-elles plutôt positives ou négatives de manière générale?

- **“This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.”**
- **« Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised. »**



Enjeux et difficultés

- “This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”
- Well as usual Keanu Reeves is nothing special, but surprisingly, the **very talented** Laurence Fishbourne is **not so good** either, I was surprised.



Détection d'opinions : enjeux et difficultés

- **Traitement de la négation (“Ce film n’est pas bien”) et des intensifieurs (“Ce film est très bien”)**
- **Identification de la cible de l’opinion**
 - « Je suis satisfait des contacts que j’ai eus avec le service client mais pas des tarifs pratiqués »
 - Concepts détectés
 - Opinion : satisfaction
 - Thématiques: contact et prix
 - Enjeu : pouvoir détecter automatiquement ce sur quoi porte l’opinion
 - Résolution d’anaphore : “il les adore”



Détection d'opinions : enjeux et difficultés

- **Utilisation de la métaphore**
 - 'réchauffement climatique' et 'changement climatique'
[Ahmad et *al.* 2011]
- **Utilisation du contexte :**
 - phrases précédentes, personnalité du locuteur, contexte d'interaction



1^e type de méthode : Détection de mots clés

- **Keyword spotting** : l'approche la plus naïve mais aussi la plus accessible et économe
- **Principe** :
 - Le texte est classé dans la catégorie d'opinions correspondant à la présence de mots clairement associés à une opinion ou une émotion
 - « je suis **content** » => positif
- **Limites** :
 - Ne traite pas la négation
 - « je ne suis pas **content** » => positif
 - Ignore les mots qui sont implicitement positifs ou négatifs
 - « le réchauffement climatique »



2^e type de méthode : Affinité lexicale

■ Principe :

- Assigner aux différents mots une probabilité d'appartenance à une catégorie d'opinion ou d'émotion
 - Ex : « réchauffement » est assigné à la classe négative avec une probabilité de 75%
- Ces probabilités sont apprises sur des corpus annotés

■ Limites :

- Opère au niveau du mot et non au niveau de la phrase (ne traite pas la négation, ni le contexte sémantique)
 - Ex tiré de [Moilanen 2007] « The senators supporting(+) the leader(+) failed(-) to praise(+) his hopeless(-) HIV(-) prevention program.”
- Les probabilités apprises dépendent fortement du corpus d'apprentissage et donc du domaine du corpus



Lexiques d'opinions en anglais

- SentiWordNet <http://sentiwordnet.isti.cnr.it/>
 - Repose sur Wordnet : base de données lexicales
 - Principe : ensemble de synonymes les *synsets*
 - Version anglaise : <http://wordnetweb.princeton.edu/perl/webwn>
 - Version française : Wordnet Libre du Français (WOLF) : <http://alpage.inria.fr/~sagot/wolf.html>

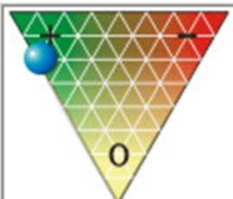
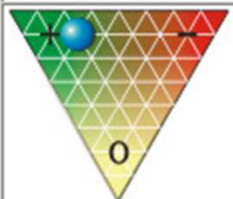
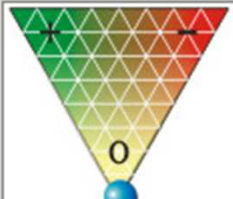


Lexiques d'opinions en anglais

- SentiWordNet <http://sentiwordnet.isti.cnr.it/>
 - Principe : ajouter à chaque synset un score positif, un score négatif ET un score d'objectivité compris entre 0 et 1
 - [estimable(J,3)] “may be computed or estimated”
Pos 0 Neg 0 Obj 1
 - [estimable(J,1)] “deserving of respect or high regard”
Pos .75 Neg 0 Obj .25

Lexiques d'opinions en anglais

- SentiWordNet

estimable	Search word	<input checked="" type="radio"/> show position
Adjective 3 senses found.		
 <p>P = 0.75, N = 0, O = 0.25</p>	<u>estimable(1)</u> <i>deserving of respect or high regard</i>	
 <p>P = 0.625, N = 0.25, O = 0.125</p>	<u>honorable(5)</u> <u>good(4)</u> <u>respectable(2)</u> <u>estimable(2)</u> <i>deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</i>	
 <p>P = 0, N = 0, O = 1</p>	<u>computable(1)</u> <u>estimable(3)</u> <i>may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</i>	

[main page](#)

(c) Andrea Esuli 2005 - andrea.esuli@isti.cnr.it

Lexiques d'opinions en anglais

■ Wordnet affect

- Sélection d'un sous-ensemble de wordnet
- Étiquette affective + valence

Etiquette affective	Exemples de synsets associés
<i>Emotion</i>	nom ANGER#1, verbe FEAR#1
<i>Mood</i>	nom ANIMOSITY#1, adjectif AMIABLE#1
<i>Trait</i>	nom AGGRESSIVENESS#1, adjectif COMPETITIVE#1
<i>Cognitive State</i>	nom CONFUSION#2, adjectif DAZED#2
<i>Physical State</i>	nom ILLNESS#1, adjectif ALL IN#1
<i>Edonic Signal</i>	nom HURT#3, nom SUFFERING#4
<i>Emotion-Eliciting Situation</i>	nom AWKWARDNESS#3, adjectif OUT OF DANGER#1
<i>Emotional Response</i>	nom COLD SWEAT#1, verbe TREMBLE#2
<i>Behaviour</i>	nom OFFENSE#1, adjectif INHIBITED#1
<i>Attitude</i>	nom INTOLERANCE#1, nom DEFENSIVE#1
<i>Sensation</i>	nom COLDNESS#1, verbe FEEL#3

Tiré de [https://www.proxem.com/Download/Research/BDL-CA07-WordNet et son ecosysteme-Francois Chaumartin.pdf](https://www.proxem.com/Download/Research/BDL-CA07-WordNet%20et%20son%20ecosysteme-Francois%20Chaumartin.pdf)



Lexiques d'opinions en anglais

- LIWC (Linguistic Inquiry and Word Count) Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX
- **Home page:** <http://www.liwc.net/>
- **2300 mots, >70 classes**
- **Version française :** http://sites.univ-provence.fr/wpsycle/outils_recherche/liwc/FrenchLIWCDictionary_V1_1.dic



Lexique d'opinions en français

- **Emotaix en français**
 - http://sites.univ-provence.fr/~wpsycle/outils_recherche/outils_recherche.html#emotaix

LIWC français

Tableau 1

Les 80 descripteurs analysés par le LIWC2007 version anglaise (extrait de Pennebaker et al., 2007 ; NB: entre parenthèses l'effectif de radicaux présents dans le dictionnaire anglais).

Processus linguistiques	Processus psychologiques	Préoccupations personnelles	Dimensions du langage oral	Ponctuation
Total de mots	Processus sociaux (465)	Travail (327)	Consentement (30)	Total
Mots par phrase	Famille (64)	Accomplissement (186)	Phatiques (8)	Points
Mots du dictionnaire	Amis (37)	Loisirs (229)	Remplisseurs (9)	Virgules
Mots de plus de 6 lettres	Humains (61)	Maison (93)		Doubles points
Total de mots fonctionnels (464)	Processus affectifs (915)	Argent (173)		Points virgules
Total des pronoms (116)	Émotions positives (406)	Religion (159)		Points d'interrogation
Pronoms personnels (70)	Émotions négatives (499)	Mort (62)		Points d'exclamation
1 ^{er} personne du singulier (12)	Anxiété (91)			Tirets
1 ^{er} personne du pluriel (12)	Colère (184)			Guillemets
2 ^e personne (20)	Tristesse (101)			Apostrophes
3 ^e personne du singulier (17)	Processus cognitifs (730)			Parenthèses
3 ^e personne du pluriel (10)	Perspicacité (195)			Autres ponctuations
Pronoms impersonnels (46)	Causation (108)			
Articles (3)	Divergence (76)			
Verbes (383)	Tentative (155)			
Verbes auxiliaires (144)	Certitude (83)			
Verbes au passé (145)	Inhibition (111)			
Verbes au présent (169)	Inclusion (18)			
Verbes au futur (48)	Exclusion (17)			
Adverbes (69)	Processus perceptifs (273)			
Prépositions (60)	Vue (72)			
Conjonctions (28)	Audition (51)			
Négations (57)	Toucher (75)			
Quantificateurs (89)	Processus biologiques (567)			
Nombres (34)	Corps (180)			
Jurons (53)	Santé (236)			
	Sexualité (96)			
	Alimentation (111)			
	Relativité (638)			
	Mouvement (168)			
	Espace (220)			
	Temps (239)			

A. Piat et al. / Psychologie française 56 (2011) 145–159

3^e type de méthodes : règles sémantiques

(manque|~negation-patt|(il/#NEG/y/avoir/~negation-patt))/(#PREP_DE)?/ (conseil|contact|~services-lex)

« manque de qualité de service »



« il n'y a vraiment pas eu de contact », ...

Concept
INSATISFACTION

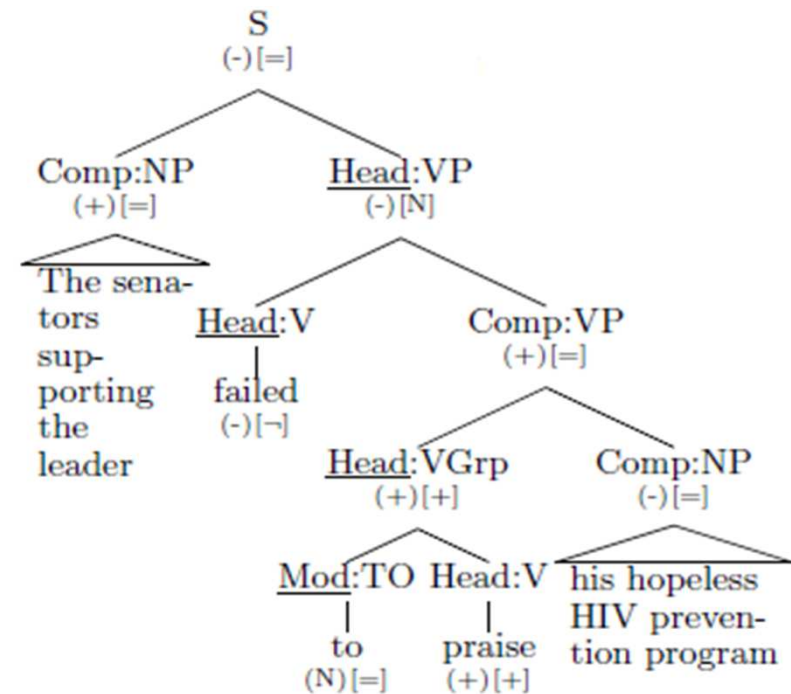
✦ Principe :

- Lexique de sentiment (ex : SentiWordNet)
- Règles d'extraction [Moilanen 2007] [Taboaba et al.][SenticPatterns]

3^e type de méthodes : règles sémantiques

- **Approche compositionnelle [Moilanen 2007] :**
 - Représentation de la phrase sous forme de constituants

« The senators supporting the leader
failed to praise his hopeless HIV
prevention program »

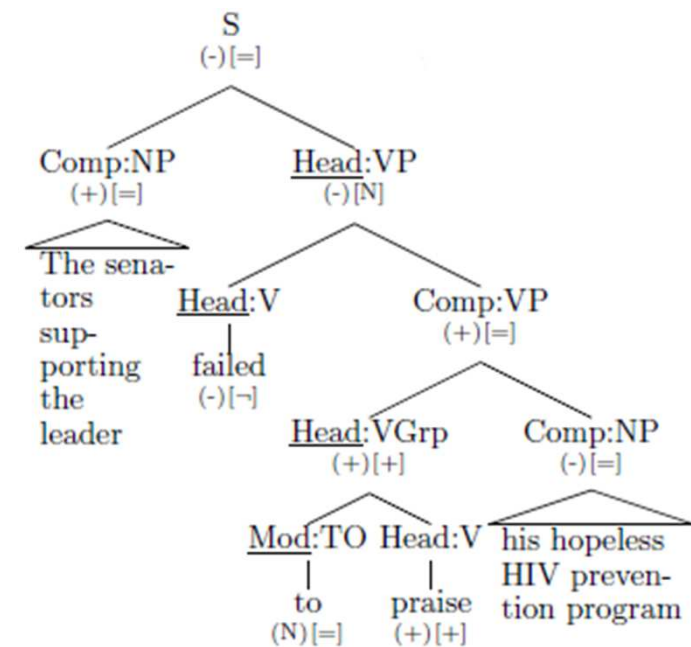


- Calcule la polarité globale d'un constituant de sortie à partir des constituants d'entrée

3^e type de méthodes : règles sémantiques

■ Approche compositionnelle [Moilanen 2007] :

- Règles de propagation : la polarité d'un constituant neutre est “effacée” par celle d'un constituant non neutre
 - $\{(+)(N)\} \rightarrow (+)$
 - $\{(-)(N)\} \rightarrow (-)$
- Règles d'inversion : $(+) \rightarrow (-)$; $(-) \rightarrow (+)$ pour gérer par exemple la négation
- Règles de résolution de conflits de polarité : lorsque les deux polarités sont conflictuelles à différents niveaux de la structure syntaxique





3^e type de méthodes : règles sémantiques

■ Approche compositionnelle

- Utilisée pour distinguer affect/judgment/appreciation
 - Recognition of affect, judgment and appreciation in Text – Neviarouskaya et al., COLING 2010
 - ‘I feel highly unfriendly attitude towards me’ -> Affect
 - ‘The shop assistant’s behavior was really unfriendly’ -> Judgment
 - ‘Plastic bags are environment unfriendly’ -> Appreciation

Affect : réaction personnelle, référence à un état émotionnel (bonheur, etc)

Jugement : attributions de qualités (capacité, ténacité) à des personnes en fonction de principes normatifs

Appréciation : évaluation de choses (produit, processus)



3^e type de méthodes : règles sémantiques

- **Taboaba et al. : Lexicon-Based methods for sentiment analysis**
- **Principe:**
 - Attribue une SO (Semantic Orientation) entre -5 et 5 aux adjectifs, noms, verbes et adverbes



EXO : SO value entre -5 et 5

Monstruosity

Masterpiece

Hate

Disgust

Relish

Endear

Fabricate

Delay

Inspiration

Inspire

Determination

Sham




Table 1

Examples of words in the noun and verb dictionaries.

Word	SO Value
monstrosity	−5
hate (noun and verb)	−4
disgust	−3
sham	−3
fabricate	−2
delay (noun and verb)	−1
determination	1
inspire	2
inspiration	2
endear	3
relish (verb)	4
masterpiece	5

- **Taboaba et al. : Lexicon-Based methods for sentiment analysis**
- **Principe:**
 - Gestion des intensifieurs : modification de la SO


Table 3
Percentages for some intensifiers.

Intensifier	Modifier (%)
slightly	−50
somewhat	−30
pretty	−10
really	+15
very	+25
extraordinarily	+50
(the) most	+100

EXO :

Si *sleazy* a une SO de 3, quelle est la SO de *somewhat sleazy* ?

Si *excellent* a une SO de 5, quelle est la SO de *most excellent* ?

- 
- **Taboaba et al. : Lexicon-Based methods for sentiment analysis**
 - **Principe:**
 - Gestion de la négation :
 - switch negation pour les cas simples (good(+3), not good(-3))
 - Recherche de la négation dans les cas plus compliqués
 - Ex: « Nobody gives a good performance in this movie »
 - Gestion des « Irrealis blocking »: ex: « would »
 - « This should have been a great movie » (SO = 3 -> SO = 0)



3^e type de méthodes : règles sémantiques

- **Avantage:**

- modèles plus fins intégrant les propriétés intrinsèques des expressions de sentiment et d'opinion
- Rendent possible l'identification de la cible et de la source de l'opinion et l'implémentation des modèles théoriques (ex: modèle de Martin and White)

- **Inconvénient:**

- Modèles peu génériques – faible interopérabilité



4^e type de méthodes :
machine learning



2 Types de tâches:

- **Classer, catégoriser les documents en thèmes, en opinions, etc.**
 - La catégorisation ou classification supervisée
 - Ex : SVM (support vector machines), classifieur bayésien naïf



2 Types de tâches:

- Repérer des expressions
 - Ex: détection d'entités nommées

[Localité d'Ukraine] menace les livraisons de gaz à l' UE
. affaire Madoff contient encore de nombreuses zones d
de l' UE sous l'il de **Paris** [Communes de France] . La
tionnisme de **Nicolas Sarkozy** [Chef d'État] . Avec l'
ment culturel . La **Russie** [Pays] a cessé de fournir
ent] n' a pas à craindre pour ses approvisionnements .
le de l' occupation américaine en **Irak** [Pays] . Le
ourées entre jeunes et policiers . Des engins incendiaires

Tirée de <http://www.tal.univ-paris3.fr/plurital/travaux-2009-2010/bao-2009-2010/MarjorieSeizou-AxelCourt/webservices.html>



Catégorisation – les deux phases

■ Phase 1 – l'apprentissage

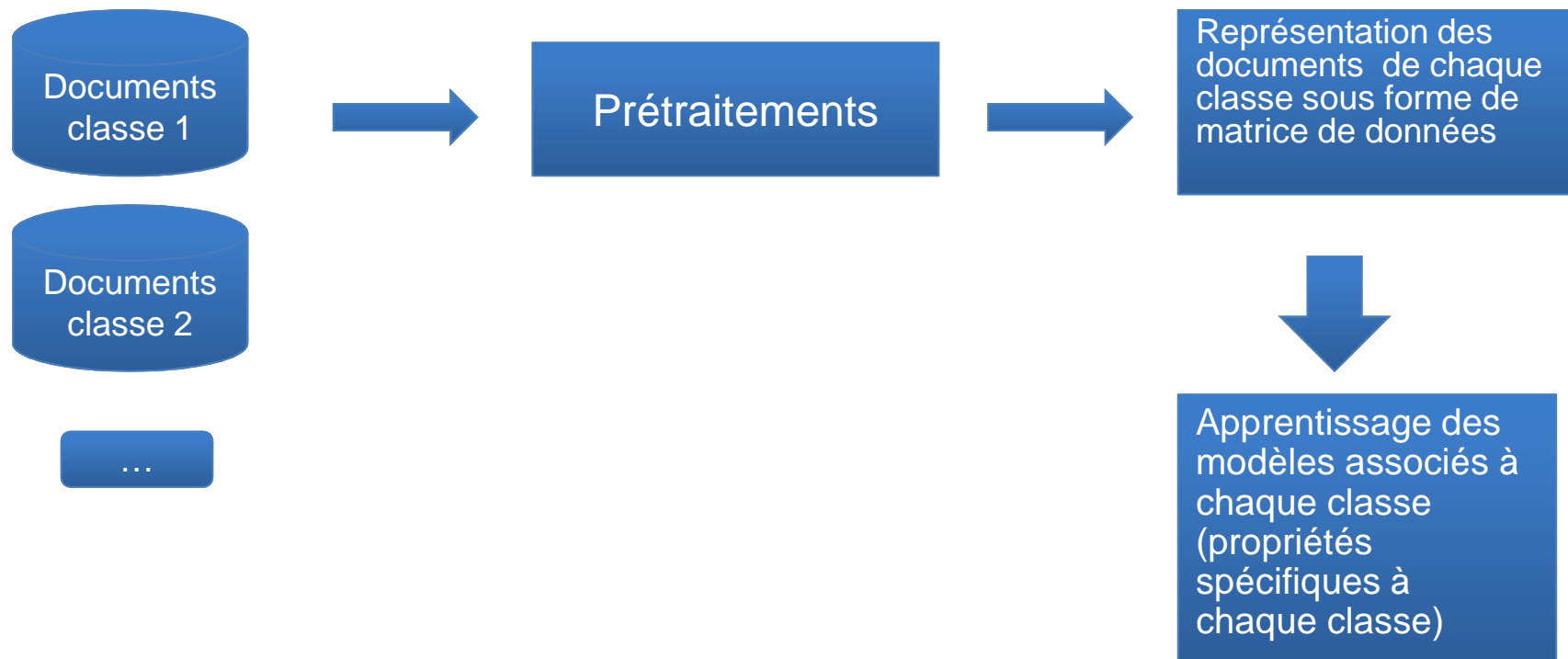
- Corpus d'apprentissage = ensemble de documents textuels annotés en opinions
 - Annotation : chaque document est associé à une classe :
 - Ex1. Corpus d'articles de journaux : le thème de l'article (international, politique, sciences, sports, etc).
 - Ex2. Corpus de critiques de films : la note donnée par l'internaute (1 à 5)
- Objectif : Apprendre à partir des données du corpus les caractéristiques communes à chaque classe

■ Phase 2 – le test/la classification/la décision

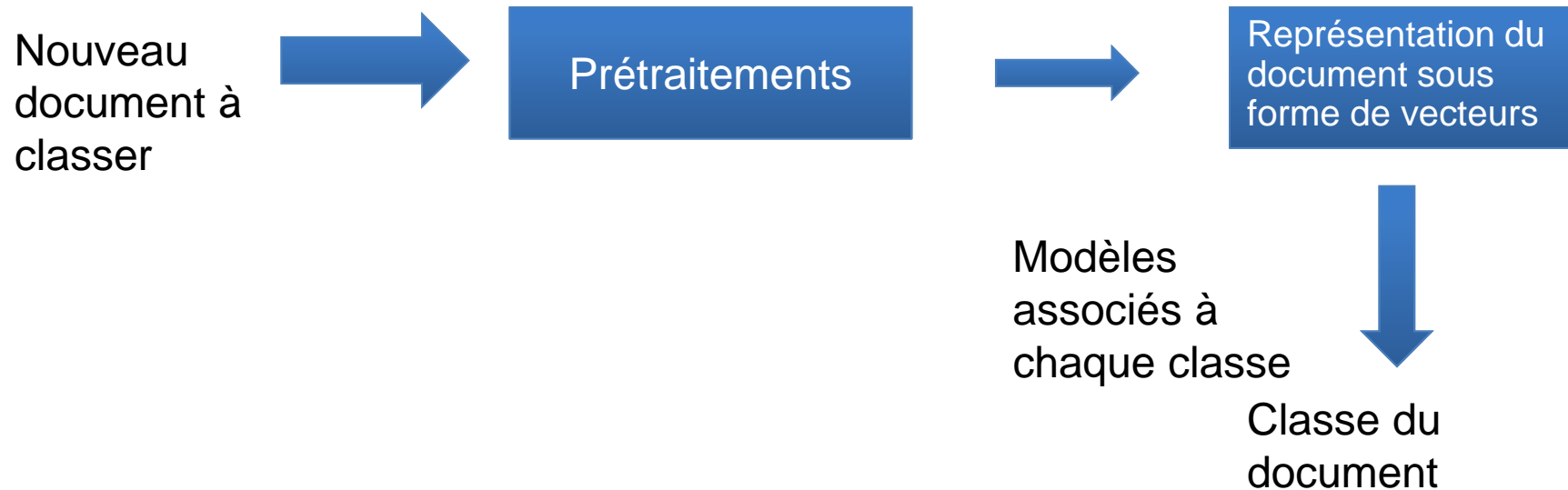
- À chaque nouveau document en entrée du système est attribuée automatiquement une classe

Catégorisation – phase 1 : l'apprentissage

■ Apprentissage des classes



Catégorisation - phase 2 : la décision





Les étapes préalables à l'analyse de données textuelles

1. Segmentation du texte en unités lexicales :

- mots et phrases

2. Le traitement lexical :

- déterminer les informations lexicales associées à chaque mot isolément (règles morphologiques et dictionnaire)

3. Le traitement syntaxique :

- Désambigüiser en fonction du contexte, extraire les relations grammaticales que les mots et les groupes de mots entretiennent entre eux
 - Analyse morpho-syntaxique
 - Chunking

Ex: « Le compteur intelligent Linky sera déployé à Paris en 2013. »

Le/Compteur/Intelligent/Etc.

1.

Le : déterminant masculin singulier ou pronom personnel masculin singulier

2.

3.

Le : déterminant masculin singulier



Prétraitements

- **Segmentation en mots / tokenization : choix des mots à considérer**
 - Filtrage des signes (ponctuation, dates)
 - Filtrage des anti-mots (stop words) à partir d'une liste de mots
 - Mots de liaisons et d'articulation du texte car peu de pouvoir discriminant
 - Filtrage des hapax
 - termes qui sont très peu fréquents dans le corpus
 - Peuvent correspondre à des mots mal orthographiés

Prétraitements

- **Segmentation en mots / tokenization : choix des mots à considérer**
 - Regrouper des termes autour de leur racine ou de leur lemme
 - Racinisation (stemming) : tronquer certains suffixes
 - Lemmatisation (après une analyse morphosyntaxique)

Xerox

[10/1996, 10/1997]

La	le
petite	petit
ferme	ferme
du	de=le
père	père
Fouchard	Fouchard
se	se
trouvait	trouver
au sortir du	au sortir de=le
défilé	défilé
.	.

+DET_SG
+ADJ2_SG
+NOUN_SG
+PREP_DE
+NOUN_SG
+NOUN_INV
+PC
+VERB_P3SG
+PREP
+NOUN_SG
+SENT

Lemme

POS : Part Of Speech

Prétraitements

■ choix des mots à considérer

- Grouper les mots en n-grammes
 - Ex: considérer tous les couples de mots (bigrammes, trigrammes)

$P(\text{Le président François Holland a présenté ses vœux}) = ??$

2-grammes		3-grammes	
$P(\text{le} < s >)$	1.3941	$P(\text{le} < s >)$	1.3009
$P(\text{président} \text{le})$	1.7206	$P(\text{président} < s >, \text{le})$	1.3844
$P(\text{François} \text{président})$	2.4011	$P(\text{François} \text{le}, \text{président})$	2.2343
$P(\text{Holland} \text{François})$	0.3444	$P(\text{Holland} \text{président}, \text{François})$	0.1158
$P(\text{a} \text{Holland})$	1.0458	$P(\text{a} \text{François}, \text{Holland})$	0.9839
$P(\text{présenté} \text{a})$	2.7520	$P(\text{présenté} \text{Holland}, \text{a})$	2.5205
$P(\text{ses} \text{présenté})$	2.0150	$P(\text{ses} \text{a}, \text{présenté})$	1.5563
$P(\text{vœux} \text{ses})$	2.5941	$P(\text{vœux} \text{présenté}, \text{ses})$	1.7149
$P(< /s > \text{vœux})$	1.4140	$P(< /s > \text{ses}, \text{vœux})$	1.2823
=	15.6819	=	13.0930
$\Rightarrow PP =$	55.2625	$\Rightarrow PP =$	28.4956

© Cours Modèle
de langage
Alexandre
Allauzen

- Ex: regrouper les termes appartenant au même syntagme



Représentation du document sous forme de matrices de données

- **1 doc = 1 vecteur (a_1, \dots, a_N) de longueur N (le nombre de mots dans l'ensemble des textes)**
 - où a_i = nombre d'occurrences du mot i dans le texte
 - où a_i = TFIDF du mot i dans le texte
 - TFIDF (Term Frequency Inverse Document Frequency) = mesure statistique utilisée pour évaluer la représentativité d'un terme/mot par rapport à un document dans une collection de textes
 - La représentativité du terme augmente proportionnellement au nombre de fois où le terme apparaît dans un document (TF), mais il est pondéré par sa fréquence dans l'ensemble du corpus (IDF)
- **Base de documents = matrices terme/document**



Calcul de TF-IDF

■ Formule TF-IDF du mot w dans le document d

$$\begin{aligned} TFIDF(w, d) &= TF_{w, d} \cdot IDF_{w, d} \\ &= TF_{w, d} \cdot \left(\log_2 \frac{N}{DF_w} \right) \end{aligned}$$

- N : le nombre total de documents dans le corpus
- TF : Term Frequency
 - nombre d'occurrences de w dans le document considéré (on parle de « fréquence » par abus de langage).
 - Variantes :
 - fréquences booléennes: $tf(w, d) = 1$ si w dans d , 0 sinon
 - logarithmically scaled frequency: $tf(w, d) = 1 + \log f(w, d)$, ou 0 si $f(w, d)$ est 0;
- DF : Document Frequency
 - nombre de documents contenant le mot w



Calcul TF-IDF

■ Exercice 1

- Ex 1 : La base contient 1000 documents, calculer la TF-IDF du mot « compteur » dans le document d, sachant que le document d contient 3 fois le mot compteur et que 70 textes contiennent également le mot « compteur »
- $TF\text{-}IDF(\text{« compteur »}, d) = ?$



Calcul TF-IDF

■ Exercice 1

- Ex 1 : La base contient 1000 documents, calculer la TF-IDF du mot « compteur » dans le document d, sachant que le document d contient 3 fois le mot compteur et que 70 textes contiennent également le mot « compteur »

- $$\text{TF-IDF}(\text{« compteur »}, d) = 3 \cdot \left(\log_2 \frac{1000}{70} \right) = 11,5$$



Calcul TF-IDF

■ Exercice 2

- Le mot « compteur » apparaît toujours 3 fois dans le document d mais apparaît cette fois dans 900 documents
- $\text{TF-IDF}(\text{« compteur »}, d) = ?$



Calcul TF-IDF

■ Exercice 2

- Le mot « compteur » apparaît toujours 3 fois dans le document d mais apparaît cette fois dans 900 documents

- $$\text{TF-IDF}(\text{« compteur »}, d) = 3 \cdot \left(\log_2 \frac{1000}{900} \right) = 0.45$$

=> Le poids du mot compteur dans le document est moins important

Représentation des mots sous forme de vecteurs sémantiques

- **Objectif** : fournir une représentation des mots sous forme de vecteurs qui capturent les relations sémantiques entre les mots
- **Exemple d'outil** : word2vec de Google <https://code.google.com/p/word2vec/>
 - Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.
 - Technique inspirée du deep-learning qui permet d'améliorer les performances des méthodes de classification de documents (incluant la classif d'opinions) ou d'extraction d'information.
- **Voir aussi le papier de Stanford**:
 - Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.
 - Technique inspirée de la LDA : Latent Dirichlet Allocation (modèle de topic probabilistique)

	romance	romance	romance
	love	charming	screwball
romantic	sweet	delightful	grant
	beautiful	sweet	comedies
	relationship	chemistry	comedy



Exemple de méthodes de classification supervisée

■ Le classifieur Bayésien naïf (Naive Bayes Classifier)

- principe général du classifieur Bayésien= à classer, choisir la classe c qui maximise $P(c | o)$
 - étant donné une observation o
 - par exemple ici o = le document

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c | o)$$

- Loi de Bayes, et le fait que $P(o)$ est constant pour toute classe, on obtient :

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c | o) = \underset{c}{\operatorname{argmax}} \frac{P(o | c)P(c)}{P(o)} = \underset{c}{\operatorname{argmax}} P(o | c)P(c)$$



Exemple de méthodes de classification supervisée : Le classifieur Bayésien naïf

$$\hat{c} = \arg \max_c P(c | o) = \arg \max_c \frac{P(o | c)P(c)}{P(o)} = \arg \max_c P(o | c)P(c)$$

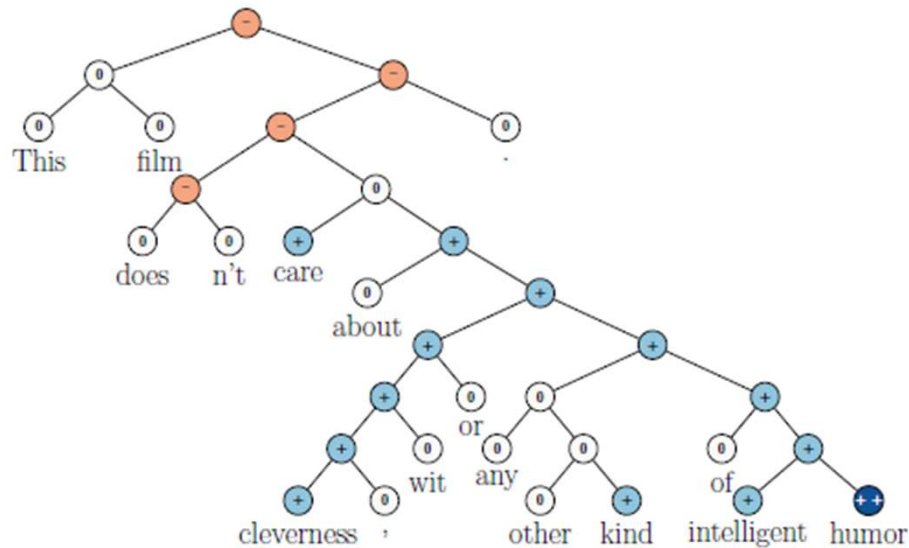
- Naïf : hypothèse d'indépendance forte entre les caractéristiques de l'observation
 - $o = \text{doc}$ et (m_1, \dots, m_N) les mots du document o
 - $P(o/C) = P(m_1, \dots, m_N/C) = \prod_{i=1}^N P(m_i/C) \rightarrow$ passer en log

$$\hat{c} = \arg \max_{c \in \mathbb{R}} [\log(P(c)) + \sum_{i=1}^N \log(P(m_i/c))]$$

- Apprentissage sur un ensemble de documents
 - Estimation de $p(c)$ et de $p(m_i/c)$
 - $P(c)$ = nombre de docs dans la classe C /nombre total de docs
 - $P(m_i/c)$ = fréquence du mot i dans la classe C

Réseaux de neurones et deep learning

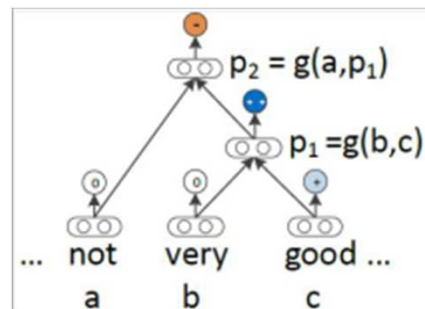
- Remise au goût du jour des réseaux de neurones avec l'émergence du deep learning
 - Utilisation des réseaux récurrents tensoriels
 - permettent de prendre en compte la structure d'une phrase.



- ✧ exemple d'utilisation des réseaux récurrents
 - REF : R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA : Association for Computational Linguistics, October 2013, pp. 1631? 1642.

Réseaux de neurones et deep learning

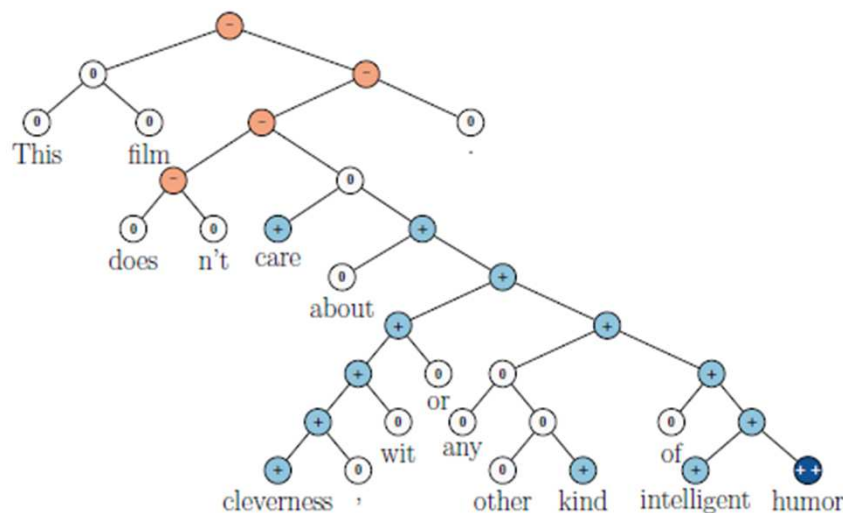
- Utilisation des réseaux récurrents tensoriels
 - Représentation de la phrase par un arbre (utilisation du parseur de Stanford)
 - On applique les récursivement les fonctions d'activation:



- Apprentissage : apprentissage de la fonction g du passage au parent dans l'arbre binaire de représentation la phrase

Méthodes statistiques et machine learning

- Méthodes récentes
 - Réseaux de neurones récurrents et deep learning
 - Ex: Socher, R., Perelygin, A., & Wu, J. (2013). Recursive deep models for semantic compositionality over a sentiment treebank.



- Base de données Sentiment treebank : annotation pour fournir la structure nécessaire à l'application d'un modèle récurrent
- phrases de critiques de films
parsées avec le parseur de Stanford -> arbre qui représente la phrase.
- Annotation des nœuds de l'arbre en (-, +, 0)



Evaluer les performances

- **Dans la tâche de classification d'un document en une classe c**
 - Précision : (le nombre de fois où le système a attribué correctement la classe c) / (le nombre de fois où il a attribué la classe c)
 - Rappel : (le nombre de fois où le système a attribué correctement la classe c) / (le nombre de fois où il aurait dû l'attribuer)
 - F-mesure moyenne harmonique pondérée de la précision et du rappel = $2 \times (P \times R) / (P + R)$

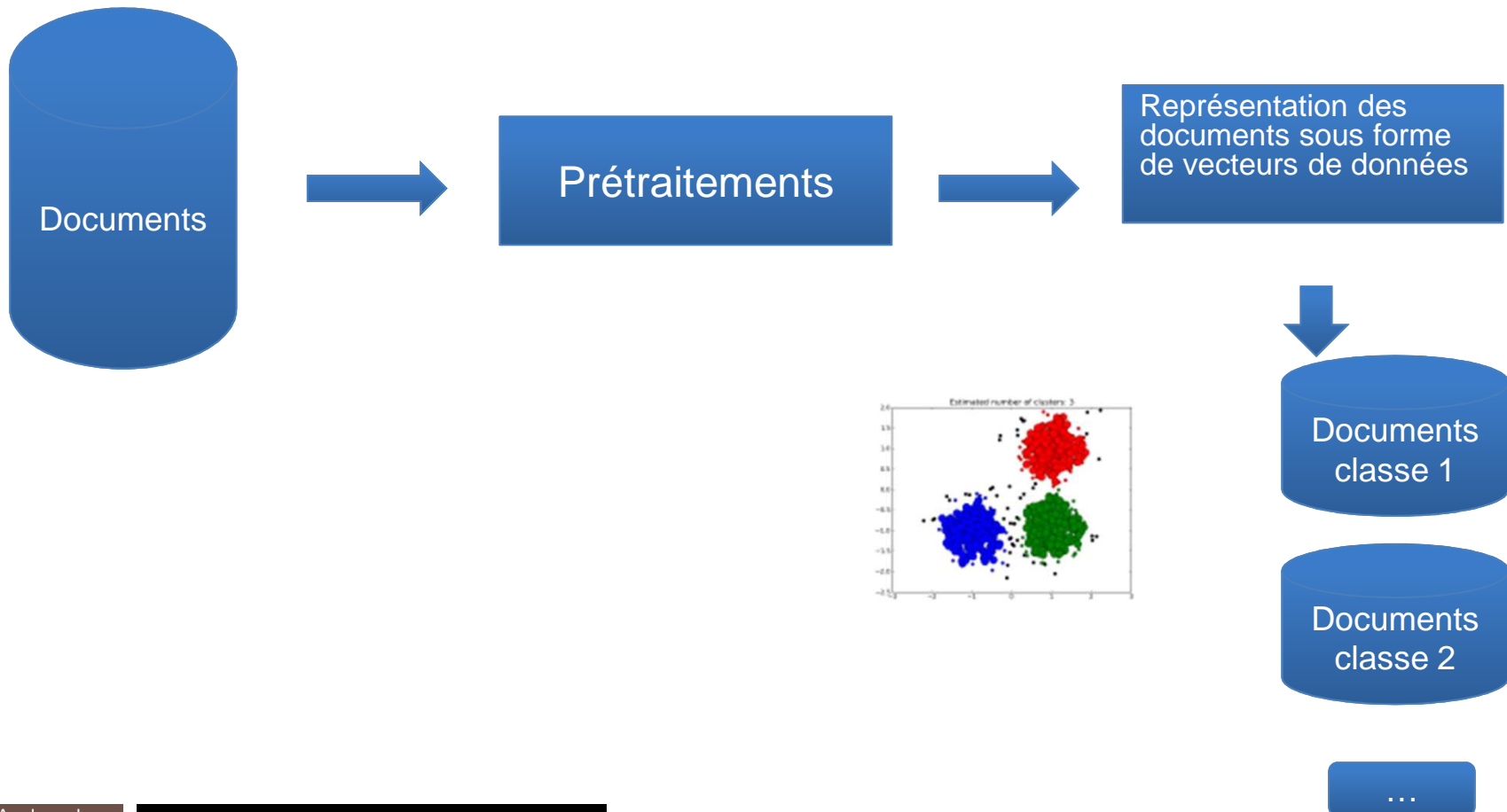


Machine learning et classification supervisée

- **Avantage :**
 - Plus forte interopérabilité des modèles
- **Inconvénient :**
 - Nécessite d'avoir des bases de données annotées (tâche d'annotation fastidieuse)
 - Difficulté d'interpréter les modèles appris
 - Généricité du modèle dépend des données du corpus d'apprentissage

Clustering de documents

■ Classification non supervisée



Classification non supervisée

■ Exemples de méthodes

- K-moyennes

- Principe général

- documents = points d'un espace multi-dimensionnel, muni d'une distance d.
 - Initialisation: Les documents sont dans un premier temps aléatoirement affectés à chaque classe 1...K. + Calcul du centroïde de chaque classe comme barycentre des individus du groupe:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i, \quad \forall k \in \{1, \dots, K\}$$

- Itération: calcul de l'inertie, critère d'arrêt = convergence de l'inertie

$$\sum_{k \in \{1, \dots, K\}} \sum_{i \in C_k} \|x_i - \mu_k\|_2^2$$

Choix de la distance?



Classification non supervisée

- **Choix de la distance/mesure de similarité pour les k-means**

- Métrique la plus courante en texte: similarité cosinus
 - Similarité entre 2 vecteurs de doc A et B en fonction du cosinus de l'angle

$$\cos \theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

- Autre mesure de similarité, l'indice de Jaccard

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- La distance associée

$$J_\delta(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$



Classification non supervisée

■ Exemples de méthodes:

- Analyse sémantique latente, analyse en composantes principales, analyse des correspondances

— Principe:

- décomposition de matrices selon leurs directions propres (ou singulières) pour conserver un maximum d'information sur un nombre minimum de dimensions.
- La décomposition en valeurs singulières de la matrice terme/document permet d'obtenir des thèmes dominants dans le corpus, chacun étant associé à un sous-espace singulier.

— Outil pour l'analyse sémantique latente :

<http://lsa.colorado.edu/>

Méthodes statistiques et machine learning

■ Méthodes récentes

- Reformuler le problème de la détection d'opinions comme un problème d'annotation séquentielle :
 - Ex pour détecter les expressions explicites (DSE) et implicites des opinions (ESE) [Irsoi and Cardie]

The committee , as usual , has
O O O B_ESE I_ESE O B_DSE

refused to make any statements .
I_DSE I_DSE I_DSE I_DSE I_DSE O

Exemple d'annotation en entrée des CRF : le modèle BIO (Beginning , Inside, Out)



Les méthodes d'étiquetage séquentiel pour des tâches d'extraction d'opinion

- même méthodes que celles qui sont utilisées pour l'étiquetage morpho-syntaxique ou la détection d'entités nommées: CRF et HMM

[Localité d'Ukraine] menace les livraisons de gaz à l' UE
. affaire Madoff contient encore de nombreuses zones d
de l' UE sous l'il de **Paris** [Communes de France] . La
tionnisme de **Nicolas Sarkozy** [Chef d'État] . Avec l'
iment culturel . La **Russie** [Pays] a cessé de fournir
ent] n' a pas à craindre pour ses approvisionnements .
le de l' occupation américaine en **Irak** [Pays] . Le
ourées entre jeunes et policiers . Des engins incendiaires

Détection d'entités nommées

- Les données annotées selon le modèle BIO

```
Wolff B-PER  
, O  
currently O  
a O  
journalist O  
in O  
Argentina B-LOC  
, O  
played O  
with O  
Del B-PER  
Bosque I-PER  
in O  
the O  
final O  
years O  
of O  
the O  
seventies O  
in O  
Real B-ORG  
Madrid I-ORG  
. O
```



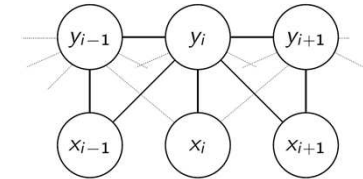
Les méthodes d'étiquetage séquentiel pour des tâches d'extraction d'opinion

■ outils sur étagère (apprentissage à base de CRF):

- pour le français
 - LIA_NE <http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html> (appris sur des données issues de l'oral)
 - SEM <http://www.lattice.cnrs.fr/sites/itellier/SEM.html> (appris sur des données écrites, des phrases tirées du journal Le Monde)
- pour l'anglais:
 - l'étiqueteur d'entités nommées de Stanford appris sur des données variées (CoNLL, MUC-6, MUC-7 and ACE) <http://nlp.stanford.edu/software/CRF-NER.shtml>

Etiqueteur probabiliste : les CRF – les Champs Aléatoires Conditionnels

- généralisation des modèles de Markov cachés



- permettent d'intégrer via leurs fonctions caractéristiques des connaissances de nature très diverse.

$$F_j(\underline{x}, \underline{y}) = \sum_{i=1}^n f_j(y_{i-1}, y_i, \underline{x})$$

- Le modèle appris par un CRF présente également l'avantage d'être relativement propice à l'interprétation :

- l'importance d'une fonction caractéristique dans le modèle est caractérisée par son poids θ

$$p(\underline{y}|\underline{x}; \theta) = \frac{1}{Z(\underline{x}, \theta)} \exp \sum_{j=1}^D \theta_j F_j(\underline{x}, \underline{y})$$

- permet d'identifier les connaissances qui jouent un rôle dans la tâche d'étiquetage



Quelques pointeurs

- **Outils de classification :**
 - NLTK :
 - modules python open source pour le TAL et scikitlearn
<http://nltk.org/> et <http://scikit-learn.org/>