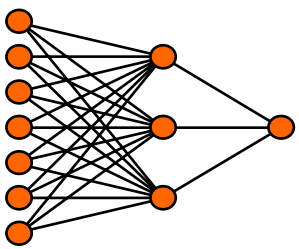
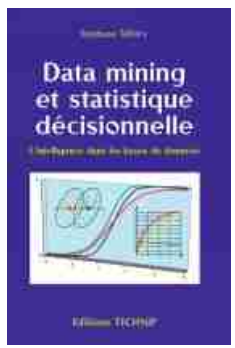
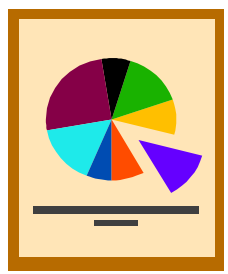
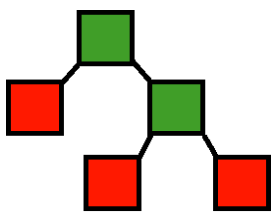
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs à éviter
- L'analyse et la préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- *Le web mining*
- *Le text mining*

Le Web Mining

Définition du Web Mining

- Web Mining = Data Mining appliqué aux données de navigation sur le web
- Objectifs du Web Mining (Web Usage Mining) :
  1) Optimiser la navigation dans un site, afin de maximiser le confort des internautes, d'augmenter le nb de pages consultées et l'impact des liens et des bannières publicitaires ⇒ Analyses globales
  2) Déceler les centres d'intérêt, et donc les attentes, des internautes venant sur le site de l'entreprise ⇒ Analyses individuelles
  3) Mieux connaître les clients qui se connectent nominativement à un site, en croisant leurs données de navigation avec leurs données personnelles détenues par l'entreprise ⇒ Analyses nominatives
- Web Content Mining : Recherche d'informations sur le web et « crawling » des pages web par les moteurs de recherche

1) Analyses globales

- Statistique descriptive
  - « 70 % des internautes ont consulté 3 pages ou moins »
  - « 40 % des internautes accèdent au site sans passer par la page d'accueil »
- Détection des règles d'associations
  - « 20 % des internautes visitant la page A visitent la page B dans la même session »
  - établir la matrice de transition entre les pages du site
  - on tient compte de l'ordre des items (≠ tickets de caisse)
- Typologies d'internautes
  - selon les sites de provenance, les pages d'entrée, le nombre de pages consultées, le temps passé sur les pages, les fichiers téléchargés, les pages de sortie, etc.

Le fichier « log »

- Source de données pour les analyses globales : le fichier « log »
  - est un fichier texte enregistré sur le serveur du site web
  - dans lequel une ligne est écrite à chaque demande de l'internaute (changement de page, téléchargement d'un fichier...)

Format du fichier log

- Common Log Format (CLF)
  - adresse IP de l'internaute, date et heure (avec décalage GMT) de la requête, type de requête, URL demandée, protocole HTTP, code retour du serveur, taille (en bits) de l'envoi
  - ex : 130.5.48.74 [22/May/2002:12:16:57 -0100] "GET /content/index.htm HTTP/1.1" 200 1243
- Extended Log Format (XLF)
  - contient en plus la page d'origine (« referrer »), le navigateur et le système d'exploitation (« user agent », ici : Internet Explorer 6.0 installé sur Windows XP SP2)
  - ex : 130.5.48.74 [22/May/2002:12:16:57 -0100] "GET /content/news.htm HTTP/1.1" 200 4504 "/content/index.htm" "Mozilla/4.0"

Explications sur le fichier log

- Type de requêtes
  - get : télécharger un objet
  - put / delete : stocker / détruire un élément sur le serveur
  - head : variante de get (parfois utilisée par les robots)
- Code retour
  - 200 / 2xx : requête satisfaite totalement/partiellement
  - 3xx : redirection
  - 401 / 404 : accès refusé / URL non trouvée
  - 4xx / 5xx : autres erreurs / erreurs du serveur
- Adresse IP
  - NB : souvent non permanente – attribuée dynamiquement par le fournisseur d'accès au moment de la connexion
  - Difficulté quand l'internaute passe par un réseau d'entreprise

Mise en forme du fichier log

- Les fichiers log sont très gros (> plusieurs centaines de Mo / jour) $\Rightarrow$ il faut les nettoyer
- Suppression des lignes correspondant à des :
  - pages visitées par moins de 5 adresses IP
  - fichiers d'images (gif, jpeg...) ou de scripts, n'apportant rien à l'analyse
  - accès de robots, d'agents ou de testeurs de liens
  - adresses IP aberrantes
- Une visite = un ensemble de requêtes provenant de la même adresse IP, du même « user agent », séparées les unes des autres par un laps de temps maximum (généralement fixé à 30 minutes, ce qui signifie que si une requête suit la précédente de plus de 30 minutes, elle débute une nouvelle visite)

Données extraites du fichier log 1/2

- Identifiant (adresse IP)
- Date de la visite
- Heure de début et de fin de la visite
  - heures de travail, soirée et nuit, week-end et jours fériés
- Type de navigateur (IE, Firefox, Netscape, Opera...)
- Système d'exploitation (Windows, Linux, Mac...)
- Pays du visiteur (voir les sites www.ip2location.com et www.dnsstuff.com/info/geolocation.htm)
- Pages visitées
- Nombre de pages visitées
- Temps moyen passé sur chaque page
- Nombre de clics moyen

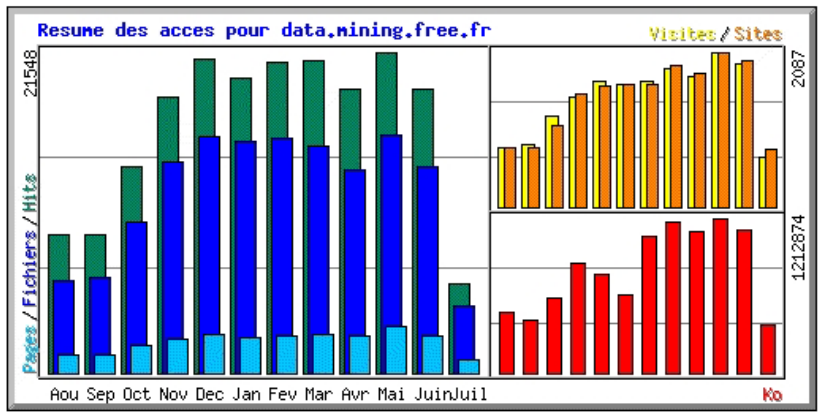
Données extraites du fichier log 2/2

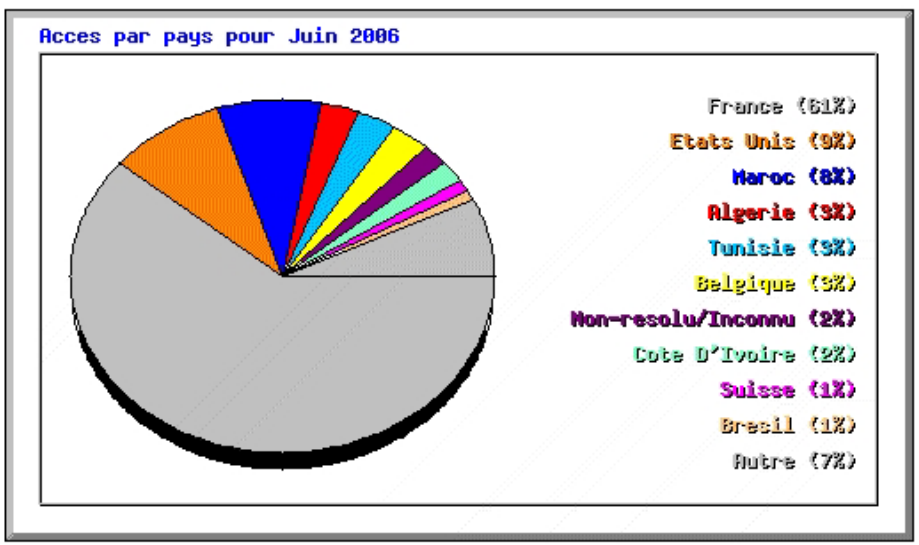
- Pour une adresse IP, on peut agréger les données :
  - dates de première et dernière visite
  - nombre de visites
  - durées totale et moyenne des visites
- Ces données permettent d'en déduire une typologie de visiteurs

- Logiciels d'analyse des fichiers log (pour faire du reporting et produire des tableaux de bord) :
  - commerciaux : Webtrends
  - gratuits : Analog, Awstats et Webanalyzer

Exemple avec Webanalyzer

- Noter la différence entre le nombre de :
  - requêtes (« hits »)
  - fichiers (« files ») = requêtes satisfaites (code retour 200)
  - pages = fichiers HTML (en excluant images, javascript...)





Résumé par mois

Mois	Moyenne journalière				Totaux mensuels					
	Hits	Fichiers	Pages	Visites	Sites	Ko	Visites	Pages	Fichiers	Hits
Juil 2006	373	284	55	41	768	373480	665	881	4547	5976
Juin 2006	635	461	84	64	1960	1117485	1920	2532	13853	19055
Mai 2006	695	515	100	67	2076	1212874	2087	3101	15969	21548
Avr 2006	635	453	84	58	1794	1102017	1767	2536	13619	19065
Mar 2006	676	491	85	59	1909	1178555	1854	2636	15248	20974
Fev 2006	746	563	88	60	1656	1074349	1685	2477	15765	20895
Jan 2006	639	501	75	52	1660	609027	1640	2342	15542	19809
Dec 2005	679	512	82	54	1630	777335	1693	2571	15896	21053
Nov 2005	617	472	74	49	1530	862621	1487	2235	14161	18534
Oct 2005	447	327	59	39	1089	588192	1228	1833	10162	13863
Sep 2005	319	220	42	28	796	418569	834	1242	6390	9265
Aou 2005	320	214	42	27	799	479622	790	1239	6227	9291
Totaux						9794126	17650	25625	147379	199328

2) Analyses individuelles

- Pour passer des analyses globales aux analyses 1:1
- Ex : 35 % des internautes qui consultent la fiche d'un roman de Boileau-Narcejac consultent la fiche d'un film de Hitchcock dans les 2 mois
- Utilisation des cookies :
  - fichiers textes créés sur le disque dur de l'internaute lors de la connexion sur le site Web
  - contiennent un identifiant propre à l'ordinateur connecté, le nb de pages consultées, les pages d'entrée, de sortie, les sites de provenance, les fichiers téléchargés, des informations nominatives demandées par le site...
  - en temps réel ou à la prochaine connexion : transmission du cookie au site Web qui peut proposer des pages personnalisées à l'internaute en fonction de ses centres d'intérêts

Avantages et inconvénients des cookies

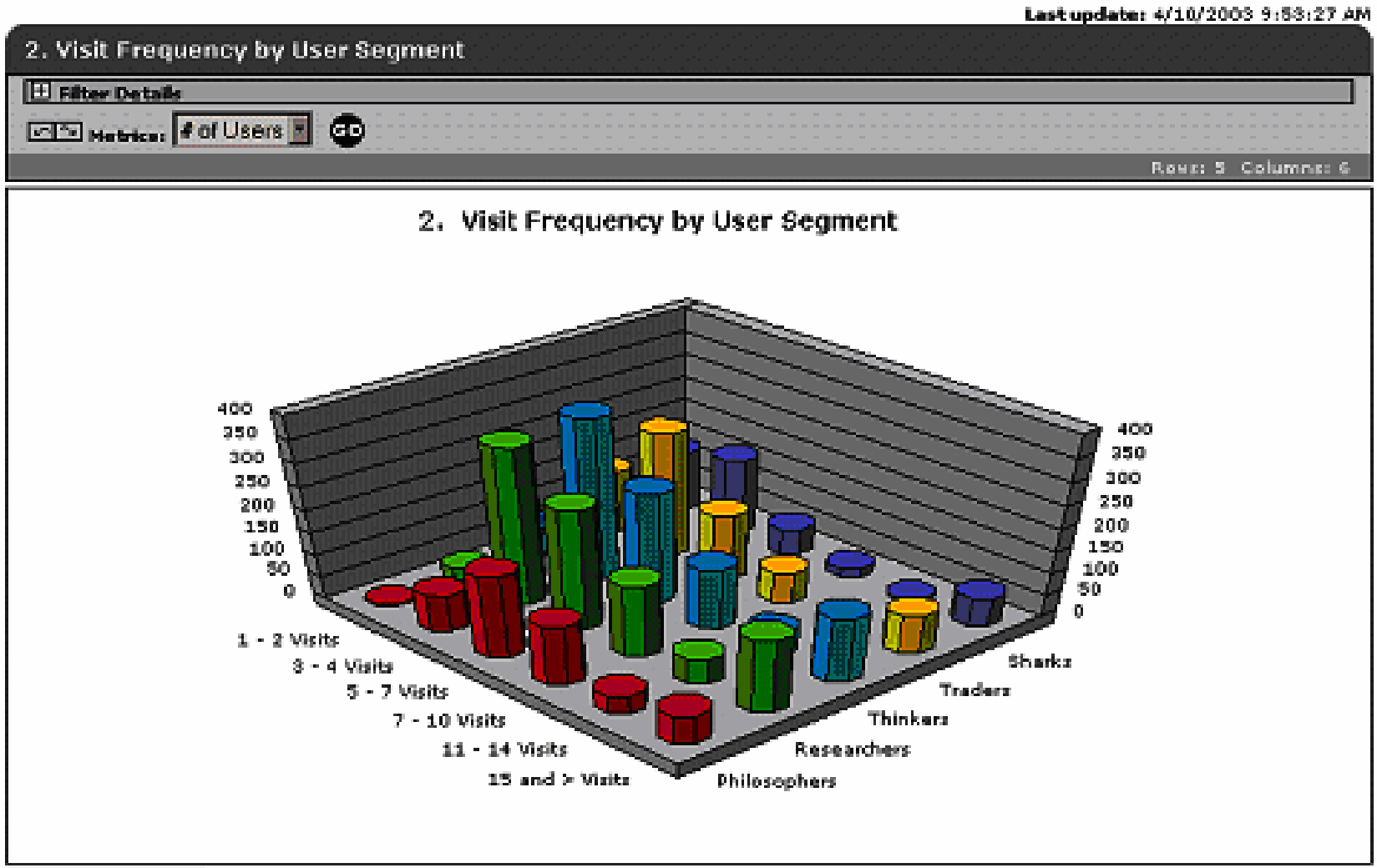
- Avantages
  - mise à jour automatique
  - mise à jour instantanée

- Inconvénients
  - refus ou suppression possible du cookie par l'internaute
  - blocage possible par un pare-feu
  - identification d'un ordinateur et non d'une personne

3) Analyses nominatives

- L'internaute est un client connu de l'entreprise
- Le site web requiert une identification personnelle
  - ex : sites bancaires en ligne
  - indexation non possible par les moteurs de recherche
- Intégration possible dans les bases de données marketing des informations sur la navigation du client
- Possibilité de construire une typologie des clients
- Les pages consultées et les demandes de simulation effectuées fournissent des indices probants sur l'intérêt du client pour tel ou tel produit
  - informations utiles dans des scores d'appétence

Croisement d'une typologie avec un indice de fréquentation



Source : SPSS

Le Text Mining

Définition du Text Mining

- Le **text mining** est l'ensemble des :
  - techniques et méthodes
  - ... destinées au **traitement automatique**
  - ... de **données textuelles en langage naturel**
  - ... disponibles sous forme informatique, en assez grande quantité
  - ... en vue d'en **dégager et structurer le contenu, les thèmes**
  - ... dans une perspective d'analyse rapide (non littéraire !), de découverte d'informations cachées, ou de prise automatique de décision.

Définition du Text Mining

- Text Mining = Lexicométrie + Data Mining

- Comme en Data Mining, on trouve en Text Mining :
  - des algorithmes descriptifs
    - recherche des thèmes abordés dans un ensemble (corpus) de documents, sans connaître à l'avance ces thèmes
  - des algorithmes prédictifs
    - recherche des règles permettant d'affecter automatiquement un document à un thème, parmi plusieurs thèmes prédéfinis

Conditions sur les textes analysés

- Format informatique
  - c'est une autre problématique que la lecture automatique de l'écriture manuscrite
- Nombre minimum de textes
- Compréhensibilité et cohérence minimale
- Pas trop de thèmes différents abordés dans un même texte
- Le moins possible de sous-entendus, d'ironie...

Sources de textes utilisées

- Enquêtes d'opinion
- Baromètres de satisfaction clientèle
- Lettres de réclamation
- Transcriptions des entretiens téléphoniques
- Messageries électroniques
- Comptes-rendus d'entretiens commerciaux
- Revues de presse - Dépêches AFP, Reuters...
- Documentation - Rapports d'experts
- Veille technologique (exemple : brevets déposes)
- Veille stratégique et économique
- Internet - Bases de données en ligne
- CV

Utilisateurs du text mining

- Analystes financiers
- Économistes
- Professionnels du marketing
- Services de satisfaction clientèle
- Recruteurs
- Décideurs

Utilisation du text mining

- Analyse rapide
  - rapports sur l'image de l'entreprise, l'état de la concurrence
  - génération automatique de baromètres de satisfaction
  - indexation automatique de documents
- Découverte d'informations cachées (« techniques descriptives »)
  - nouveaux domaines de recherche (brevets déposés)
  - ajout des informations aux bases de données marketing
  - adaptation du discours marketing à chaque type de client
- Prise de décision (« techniques prédictives »)
  - routage automatique de courriers, d'information
  - filtrage de courriels : spams – non spams
  - filtrage de « news »

Recherche et extraction d'information

- Les découvertes cachées d'informations cachées et la prise de décision appartiennent surtout à la « recherche d'information »
- L'analyse rapide appartient surtout à « l'extraction d'information »

Différences entre EI et RI 1/2

- La RI s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent, pour comparer les documents entre-eux et détecter des typologies.

- L'EI recherche des informations précises dans les documents, sans les comparer entre-eux, en tenant compte de l'ordre et de la proximité des mots pour discriminer des énoncés ≠ ayant des mots clés =

> + grande complexité de l'EI
  - qui doit effectuer une analyse lexicale et morpho-syntaxique pour reconnaître les constituants du texte (phrases, mots), leur nature et leurs relations

Différences entre EI et RI 2/2

- L'EI consiste en l'alimentation d'une base de données structurée à partir de données exprimées en langage naturel.

- Il s'agit de détecter dans le texte en langage naturel les mots ou syntagmes correspondant à chaque champ de la base de données.

> La RI cherche à détecter tous les thèmes présents
> L'EI ne s'intéresse qu'aux thèmes en rapport avec la base de données « cible »

Le Text Mining

Recherche d'information

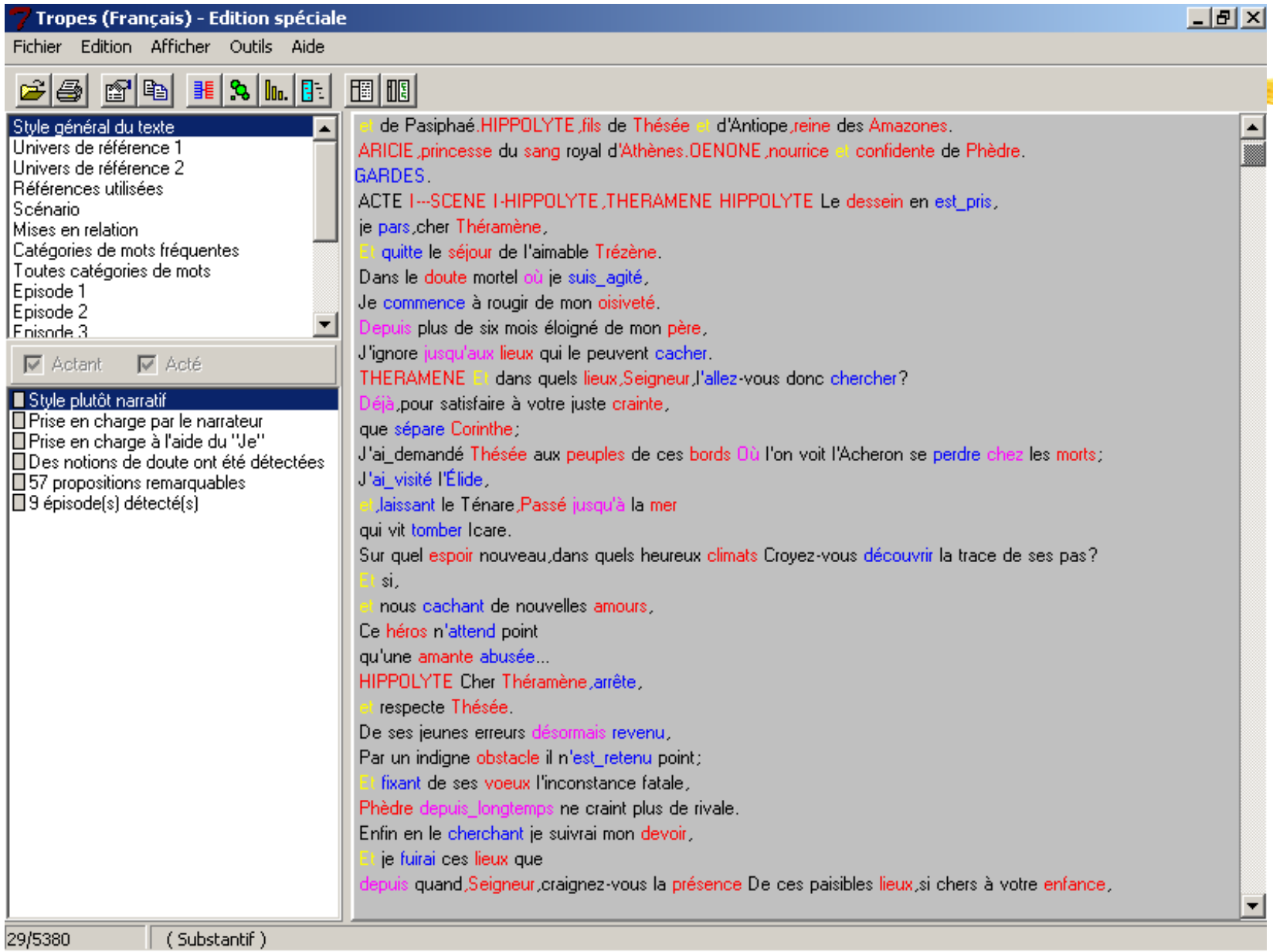
Analyse linguistique

- Identification de la langue
  - le web oblige à gérer le multilinguisme

- Exemples de phrases polyglottes
  - Pendant l'affaire du Watergate :
    - « Nixon put dire comment on tape »
    - = Nixon a enregistré un commentaire désastreux
  - Vu dans les quartiers anglophones de Montréal :
    - « Garage sale »
    - = vente dans un garage (« vide-garage »)

- Identification des catégories grammaticales
  - noms / verbes / adjectifs / adverbes
  - parfois difficile : « les poules du couvent couvent »

Catégories grammaticales de *Phèdre*



Analyse linguistique (suite)

- Désambiguïsation
  - ambiguïtés dues à la typographie pauvre
    - ELEVE -> élève (n), élève (v), élevé (adj), élevé (pp)
    - « Ce bureau ferme à cause des émeutes » ≠ « Ce bureau fermé a causé des émeutes »
  - ambiguïtés dues aux fautes d'orthographe
  - ambiguïtés dues à la polysémie des mots
  - ambiguïtés dues aux ellipses (style « télégraphique »)
  - ambiguïtés dues aux abréviations personnelles
  - ambiguïtés dues aux anaphores (il, elle, lui, celui-ci...)
  - ambiguïtés dues aux homographes
    - « nous portions des portions d'avocats aux avocats »

Analyse linguistique (suite)

- Désambiguïsation (fin)
  - ambiguïté entre le chiffre 0 et la lettre O
  - ambiguïtés dues aux retours à la ligne sans trait d'union

- Reconnaissance des mots composés
  - Expressions comme : France Telecom, le 21 février 2002, le gouverneur de la Banque Centrale Européenne
  - Prise en compte éventuelle d'un lexique spécialisé
    - data mining, text mining, entrepôt de données
    - carte bleue, compte chèques, compte courant …
  - Élaboration d'un lexique propre à l'entreprise
    - en repérant les suites de formes graphiques (souvent 2 ou 3) se répétant plusieurs fois dans le corpus

Analyse linguistique (suite)

- Lemmatisation (mots ramenés à leur forme canonique)
  - substantifs ramenés au singulier
  - adjectifs ramenés au masculin
  - flexions d'un verbe ramenées à l'infinitif
- Un dictionnaire général contient 60.000 entrées qui correspondent à 700.000 formes fléchies
- Le français, l'espagnol et l'allemand ont de nombreuses formes fléchies (conjugaisons ou déclinaisons).

Analyse linguistique (suite)

- Regroupement des variantes
  - variantes graphiques
    - clef = clé
  - variantes syntaxiques
    - complément de nom = complément nominal
  - variantes sémantiques
    - X achète Y à Z = Z vend Y à X
  - synonymes
    - US = USA = États-Unis = Oncle Sam
  - parasynonymes (mots de sens voisins)
    - mécontentement, colère, insatisfaction
  - développement des sigles
    - € = EUR = euro
    - E.D.F. = EDF = Électricité de France

Analyse linguistique (suite)

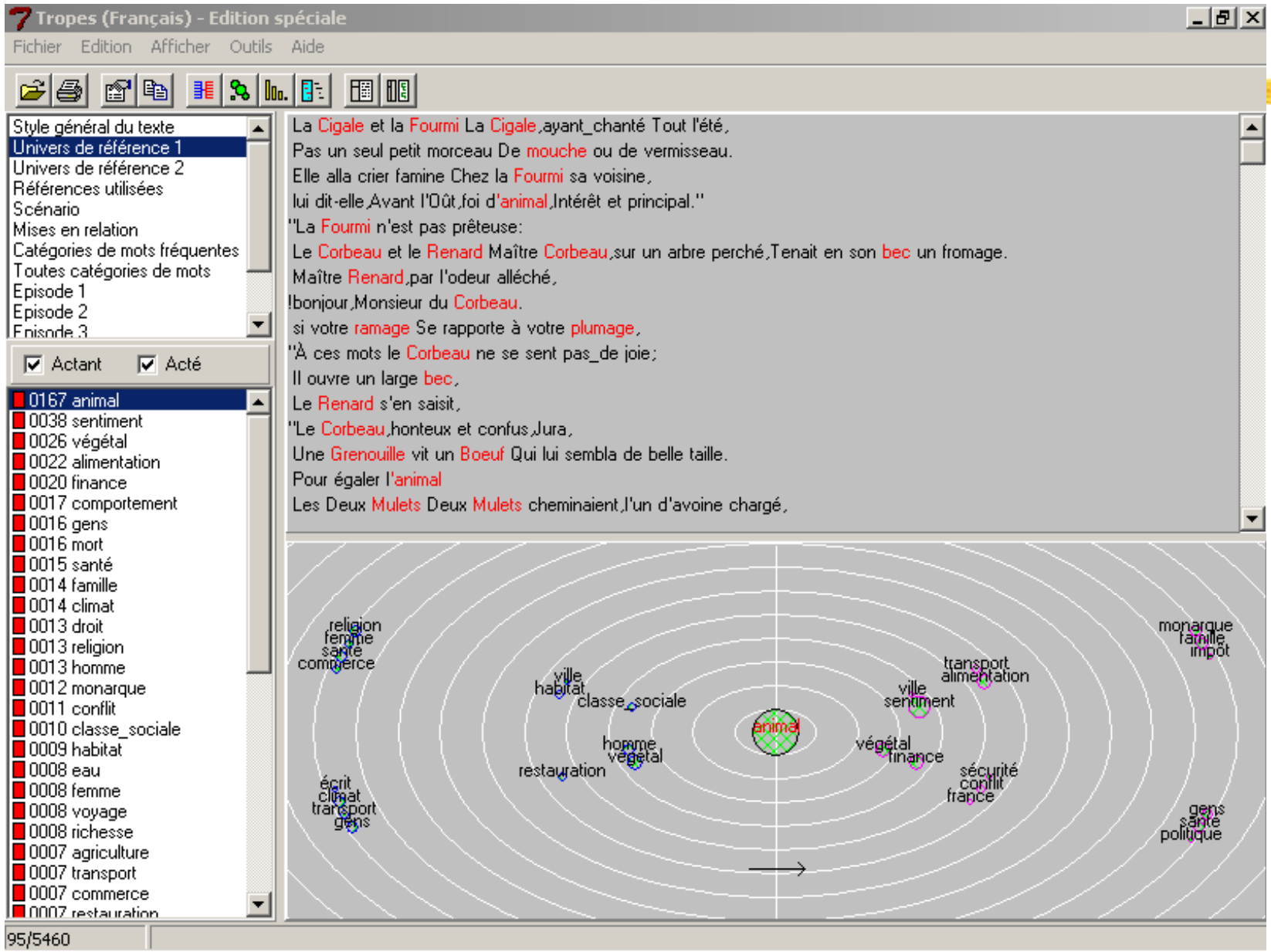
- Regroupement des variantes (fin)
  - métaphores
    - Empire du Soleil Levant, Quai d ’Orsay...
- Regroupement des analogies
  - familles de mots-dérivés
    - crédit / prêt / engagement / dette / emprunter / emprunteur / débiteur
  - marqueurs d’intensité
    - peu / moins / très peu / -
    - beaucoup / plus / très / +

Analyse linguistique (fin)

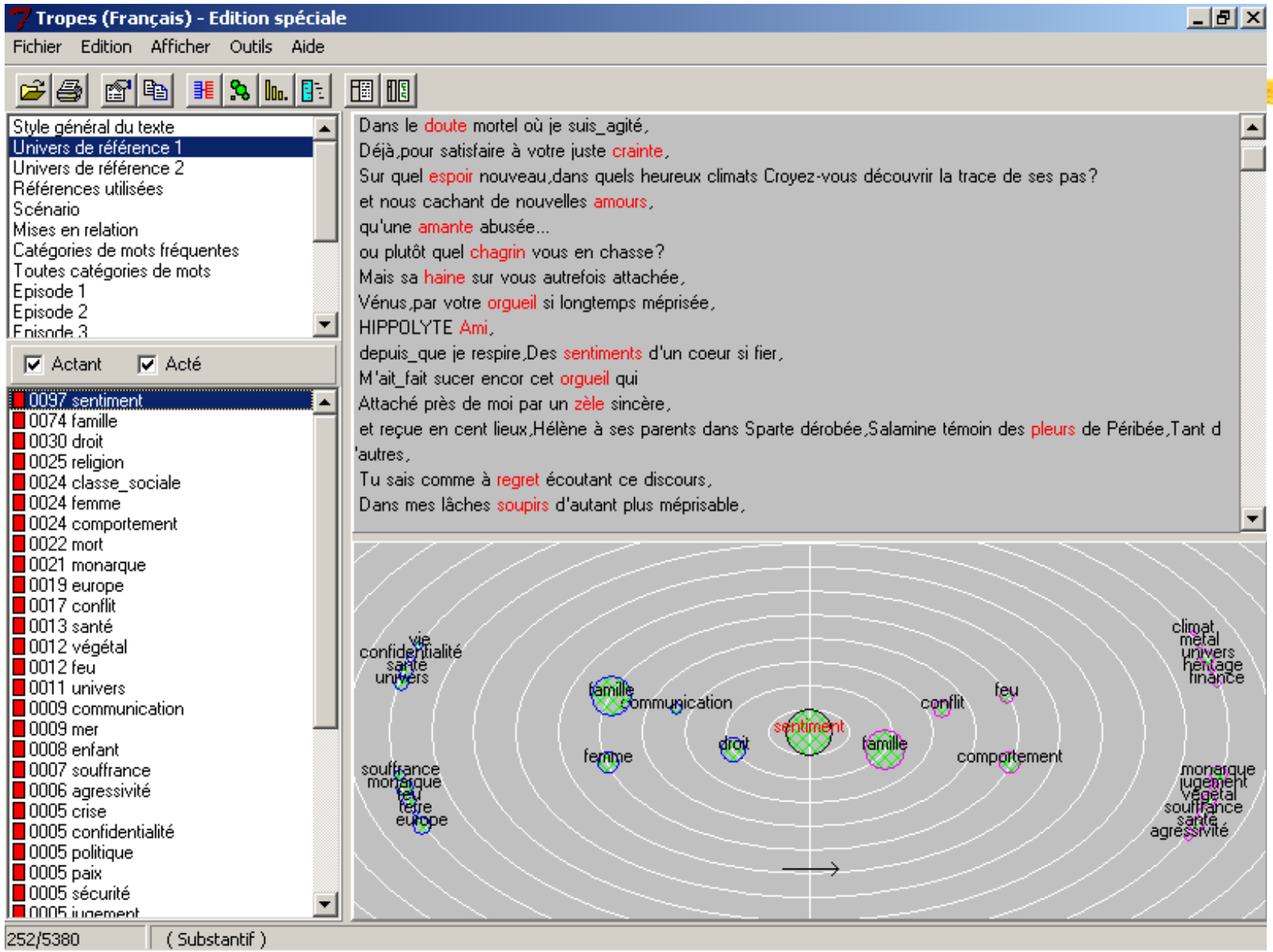
- Identification des thèmes
  - des termes aux thèmes de niveau 1:
    - chéquier / carte bleue / TIP / devises / ... <=> moyen de paiement
  - des concepts de niveau 1 aux thèmes de niveau 2 :
    - moyen de paiement / monnaie / argent / ... <=> banque

- Sélection éventuelle des termes / thèmes
  - d'après un critère statistique : élimination des termes/thèmes fréquents
  - d'après un critère sémantique : sur un sujet donné
  - d'après un corpus : repérage des mots à éviter et de leurs dérivations (expurgation du document)

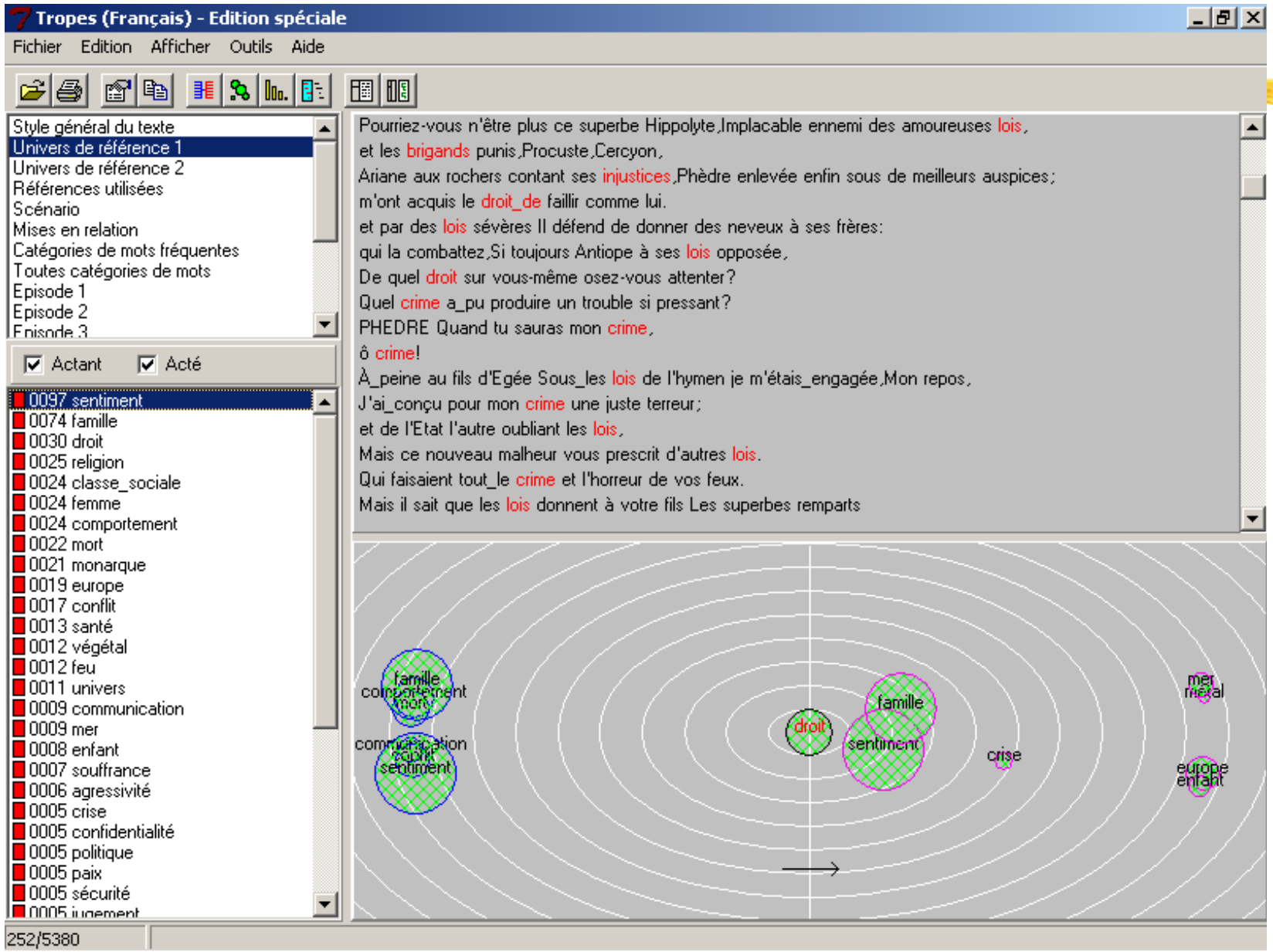
Thèmes des *Fables* de La Fontaine



Thèmes de *Phèdre* de Racine



Thèmes de *Phèdre* de Racine



Application des statistiques et du DM

- On applique ensuite les techniques de data mining :
  - individus = documents (par ex : des courriels)
  - caractères des individus = thèmes/termes des documents

- **Remarque**
- Les thèmes peuvent être très nombreux (plusieurs milliers) si le nombre de documents est important

> On aboutit à des problèmes de data mining avec un grand nombre de variables

> Intérêt de :
  - techniques puissantes de DM
  - réduire le nb de thèmes grâce à l'analyse linguistique

DM : + d'individus que de variables

	variable 1	variable 2	variable 3	...
individu 1				
individu 2				
individu 3				
...				
individu n				

TM : + de variables que d'individus

	thème 1	thème 2	thème 3	...					thème n										
texte 1																			
texte 2																			
texte 3																			
...																			

Techniques descriptives applicables

- Classification des documents
  - selon des thèmes non prédéfinis
    - découverts dans les documents
  - suivie d'une extraction automatique des mots clés
    - thèmes/termes fréquents dans le segment et rares dans l'ensemble des documents

- Analyse factorielle
  - Analyse des Correspondances Multiples
  - en croisant les données textuelles avec les autres données

Techniques prédictives applicables

- Classement des documents
  - selon des thèmes prédéfinis (nomenclature)
  - utilisé pour du routage ou du filtrage de documents
  - emploi des :
    - arbres de décision (CART, C5.0)
    - réseaux de neurones (perceptron multicouches)
- Utilisation des chaînes de Markov pour les requêtes ouvertes (libres)

Représentation graphique

- On peut dresser une **cartographie des documents** et repérer :
  - les thèmes isolés
  - les thèmes formant des ensembles homogènes
  - l'intensité des liens entre thèmes d'un même ensemble (vocabulaire et problématique commune aux thèmes)
  - le nombre de documents pour chaque thème.

Le Text Mining

Extraction d'information

Principaux exemples d'EI 1/2

- Remplissage automatique de formulaires prédéfinis à partir de textes libres
- Constitution automatique de bases de données bibliographiques à partir d'articles de recherche
  - champs à extraire : titre, auteur, revue, date de publication, organisme de recherche...
- Dépouillement automatique de la presse économique : chapitre « people » sur les changements d'emploi des cadres dirigeants

Principaux exemples d'EI 2/2

- Dépouillement automatique de milliers de dépêches Reuters traitant d'achat d'une entreprise par une autre
  - champs à extraire : acquéreur, vendeur, prix, secteur d'activité, chiffre d'affaire, cours de Bourse...

- Détection automatique des projets financiers des clients d'une banque à partir des notes des commerciaux
  - champs à extraire : nom du client, type de produit bancaire proposé, type de projet du client, montant, délai du client, réponse du client (souscription-refus), motif de la réponse du client, autre(s) banque(s) du client...
  - > utilisation dans un score d'appétence

Principe de l'EI

- Systèmes d'EI composés :
  - de mots déclencheurs (verbe ou nom)
  - de formes linguistiques
  - et de contraintes limitant l'application du déclencheur
- Ces systèmes nécessitent :
  - des dictionnaires sémantiques spécifiques du domaine ou de l'entreprise
  - des analyseurs syntaxiques sachant reconnaître les formes linguistiques générales (sujet , verbe, COD…)
- A partir d'une cible à extraire, ils :
  - détectent les phrases la contenant
  - génèrent les résultats

Exemple d'application bancaire 1/3

- Transcription d'entretiens commerciaux
- Les commerciaux détectent chez leurs clients des projets finançables (achat maison, changement voiture...)
- Les commerciaux font une proposition de crédit à leur client et notent leur réaction dans un compte-rendu
- Si la réaction est positive, le C-R est moins important, car on verra bien que le produit a été souscrit
- Si la réaction est négative, l'existence du C-R est plus importante, car sinon on ne saura pas qu'un produit avait été proposé au client.

Exemple d'application bancaire 2/3

- Les comptes-rendus ne sont pas normalisés :
  - écrits au fil de l'eau
    - fautes d'orthographe
    - ellipses (style « télégraphique »)
    - abréviations personnelles
    - ordre pas toujours logique des phrases (des mots liés se trouvent parfois séparés par une certaine distance)
  - les négations ne sont pas toujours explicites
    - « construction Le Vésinet - financement Crédit Lyonnais »

> Difficulté de normalisation automatique des C-R

> Nécessité d'outils puissants de text mining
  - et pas seulement de recherche de mots-clés

Exemple d'application bancaire 3/3

- Résultats de l'analyse des comptes-rendus par text mining
  - détection des clients réfractaires à certains types de crédit
    - utilisation de cette information pour élaborer un **score d'appétence**
  - détection automatique de certains motifs de refus du crédit
    - client « anti-crédit »
    - proposition + intéressante de la concurrence
    - pas de besoin du crédit
  - détection des clients ayant des projets à venir dans un certain délai
    - déclenchement d'une action commerciale à ce moment

Data mining multitype 1/2

- Prise en compte simultanée :
  - des données textuelles
    - issues des traitements de text mining
  - des données paratextuelles
    - date et objet du document
    - type du document (courrier, transcription d'entretien...)
    - service destinataire du document dans l'entreprise
    - ...
  - des données contextuelles
    - sur son auteur (sexe, âge, PCS...)
    - sur ses relations avec l'entreprise (produits achetés, services utilisés...)
    - ...

Data mining multitype 2/2

- Les données textuelles sont :
  - converties en données codées
  - stockées avec les autres données dans les bases de données marketing.

- Le croisement de toutes les données (textuelles et non textuelles) fait du **data mining multitype** un outil très puissant.
  - Exemple : une étude d'attrition gagne en précision à prendre en compte les lettres de réclamation et autres échanges entre l'entreprise et le client.