

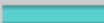

# FOUILLE DE TEXTES SUPERVISÉE

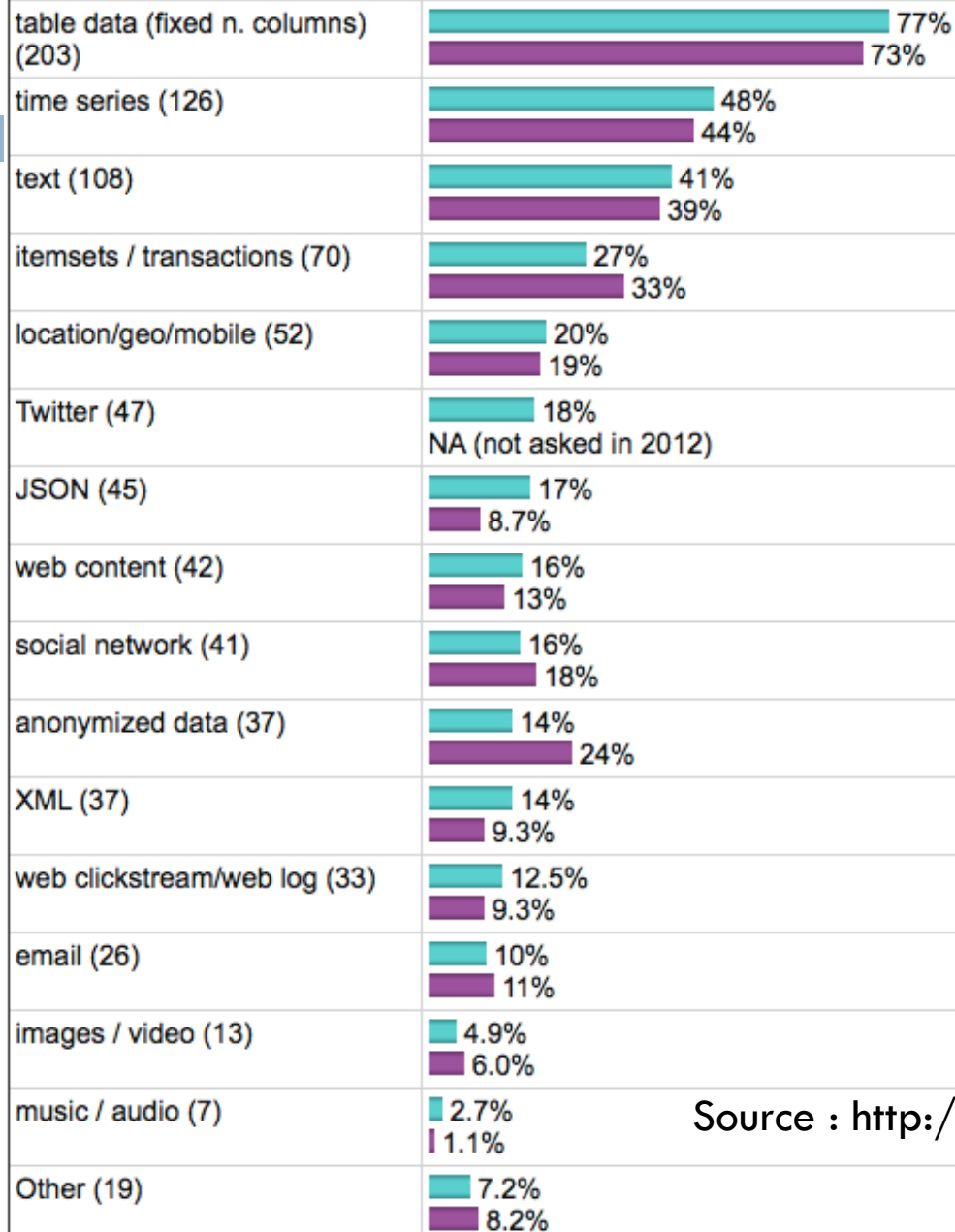
# Text Mining

2

- Ensemble des techniques et méthodes destinées au traitement **automatique** de documents **non structurés** contenant des données textuelles **en langage naturel**
  - ▣ Tous types de formats :
    - articles de presse, documents word, email, pdf, powerpoint, blogs, reviews, ...
  - ▣ But?
    - Dégager du sens
    - et/ou structurer le contenu

What data types/sources you analyzed in the past 12 months? [264 votes total]

 % users in 2014  % users in 2012



Source : <http://www.kdnuggets.com/>

# Problématique

4

- Association automatique entre
  - ▣ Documents textuels (page web, tweet, document xml...)
  - ▣ Classes prédéfinies (catégories, étiquettes, opinions...)
- Exemple : DEFT 2007, corpus jeux vidéos

```
<DOCUMENT id="2:18">
<EVALUATION nombre="1">
<NOTE valeur="1" confiance="1.00" />
</EVALUATION>
<TEXTE>
<![CDATA[
Lego Star Wars II : La Trilogie Originale
LEGO Star Wars, premier du nom, ayant surpris bon nombre
de joueurs par sa mise en scène pleine d'humour, une vraie
fidélité à l'oeuvre de papa Lucas, et la variété de ses
phases de jeu, le second opus était pour le moins attendu.
        [... suppression de 1358 mots ...]
Quoi qu'il en soit, ne vous faites pas prier si vous aimez
l'oeuvre de George Lucas et l'humour en règle générale.]]>
</TEXTE>
</DOCUMENT>
```

- ▣ **valeur** dans la balise **NOTE** est la variable à prédire
- ▣ Le contenu de la balise **CDATA** est l'ensemble des variables prédictives

→ Classification par le contenu

# Catégorisation de documents

5

- Approches statistiques
  - ▣ Représentation numérique du document par un ensemble de descripteurs
  - ▣ Utilisation d'algorithmes d'apprentissage statistique supervisé
- Combinaison entre algorithme et données d'apprentissage

# Exemples concrets de tâches

6

- Catégorisation
  - ▣ SNCF : Classification de documents techniques
  - ▣ LS2N : Dossier de candidatures à la fac
- Détection d'opinions
  - ▣ Orange : Satisfaction client à partir de plateforme service client
  - ▣ MMA : Analyse besoins clients sur retours
  - ▣ IPPON : Prospection commerciale
- Extraction d'entités nommées
  - ▣ Ina : Indexation des fonds audiovisuels

# Extraction de connaissances

7

- Un sous-ensemble de  
*Knowledge Discovery in Databases (KDD)*
- ▣ En français : Extraction de Connaissances à partir de Données (ECD)

# Une étape primordiale du KDD

8

- ❑ Étapes du processus de KDD
  1. Définition du problème et ses objectifs
  2. Inventaire/Intégration des données
  3. Sélection/Préparation des données
  4. Fouille de données
  5. Evaluation des performances
  6. Représentation des connaissances pour prise de décisions
  7. Déploiement, enrichissement des modèles
- ❑ Souvent : confusion entre Data mining et le KDD



# Exemple de charges en temps

9

Etape	Charge (en jours)	
	Projet Léger	Projet moyen
Définition de la cible et des objectifs	4j	8j
Inventaire des données	7j	10j
Collecte et préparation des données	15j	28j
Elaboration et validation des modèles	15j	25j
Analyse complémentaire, restitution des résultats	9j	12j
Documentation - Présentation	5j	7j
Analyse des premiers test	5j	10j
Total	60j	100j

# Catégorisation de texte

10

- Résumé en trois grands temps
  - ▣ À partir d'un ensemble de documents
    - Choix d'une description *pertinente* du document
  - ▣ Mise en œuvre *efficace* d'un algorithme
    - Fonction du type de problème à résoudre
  - ▣ Evaluation et/ou analyse des résultats obtenus
    - Fonction de l'application visée
- Tâche transversale
  - ▣ Cible et objectif

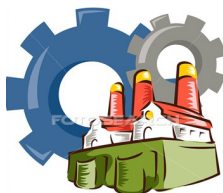
# Les données : les documents

# Qu'est-ce qu'une donnée-document ?

12

- Instance du corpus caractérisée par un ensemble de *descripteurs (variables)*
- Représentation plus formelle
  - ▣  $x$  un document de l'ensemble du corpus  $C$
  - ▣ chaque document  $x$  est défini par  $p$  descripteurs
  - ▣ chaque descripteur  $d$  prend sa valeur dans  $V_d$
  - ▣ Tout document appartient alors à un espace euclidien à  $p$  dimensions

Document  $x$



d1	d2	d3	d4	d5

# Types de descripteurs

13

## □ Descripteurs qualitatifs

### ▣ Variable discrète

- Ensemble de valeurs prédéfinies

### ▣ Pas d'application d'opérations arithmétiques habituels

### ▣ Exemples :

- une couleur, une marque, une ville, ...

### ▣ Nature de la valeur :

- nominale

- Ensemble de valeurs arbitraires, incomparables *a priori*

- Ex couleur : rose et orange

# Types de descripteurs

14

- Descripteurs quantitatifs
  - ▣ Type : entier, réel, date
    - $\neq$  numérique et réciproquement
  - ▣ Possibilité d'appliquer des opérateurs arithmétiques habituels
  - ▣ Nature de la valeur :
    - Ordinale
      - Ensemble de valeurs arbitraires mais comparables SELON une unité de mesure
    - Absolue
      - Ensemble de valeurs non arbitraires

# Types de descripteurs

15

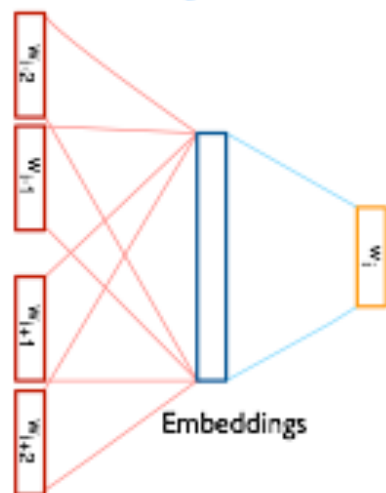
- Descripteur textuel (particularités...)
  - ▣ Représentation vectorielle
    - Cas usuel
      - **Nombre** de composantes = taille du vocabulaire
      - **Sélectionner** les composantes pertinentes
      - Définir les **valeurs** de chaque composante
    - Prise en compte de la sémantique (LSA, LDA,...)
      - Moins de composantes
      - Valeurs continues obtenues lors d'un apprentissage
    - Word embeddings
      - Nombre de composantes à définir
      - Valeurs continues obtenues lors d'un apprentissage

# Word embeddings

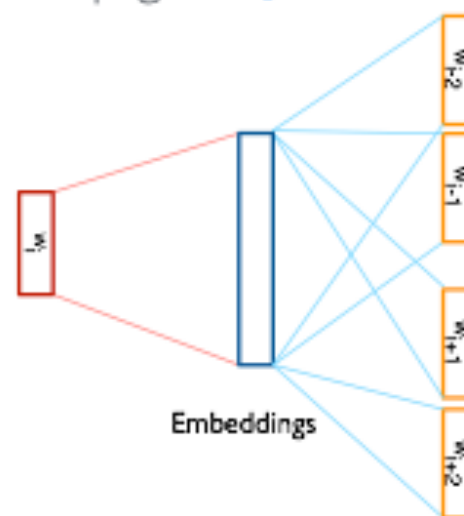
16

Embeddings Linguistiques

CBOw [T. Mikolov et al. 2013]



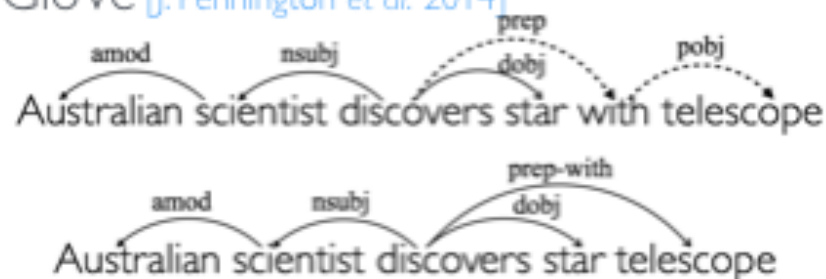
Skip-gram [T. Mikolov et al. 2013]



w2vf-deps [O. Levy et al. 2014]

- Calcul d'une matrice de co-occurrence  $X$
- Factorisation de  $X$  pour obtenir les word embeddings

GloVe [J. Pennington et al. 2014]





# Représentation... concrètement!

17

Non séquentialité / Séquentialité

Approches traditionnelles

Vecteur?

Matrice?

Approches neuronales

Entrée de taille fixe ?  
Padding/truncating

Combinaison avec descripteurs globaux ?

Doc2Vec?

# Préparation des données

18

- Préparation des documents
  - ▣ Inventaire, collecte et intégration
  - ▣ Sélection (détails slide suivant)
    - Choix de la représentation des mots
    - Choix et intégration d'autres descripteurs?
  - ▣ Regroupement en corpus (détails slide d'après)
    - Ensemble des données disponibles

# Sélection des descripteurs

19

- Pertinence
  - ▣ Importance de la sélection des descripteurs en fonction de **l'application visée**
    - Définir le problème et les objectifs
- Fiabilité / Bruit
  - ▣ Représentation complète?
  - ▣ Validité des valeurs des descripteurs?
- Quantité
  - ▣ Peu : apprentissage simplifié... performance?
  - ▣ Beaucoup : apprentissage complexe... performance?

# Les corpus

20

- Ensemble des données disponibles
- Corpus d'apprentissage (APP)
  - ▣ Entraînement du modèle
- Corpus de développement (DEV) (facultatif)
  - ▣ Optimisation des paramètres d'ajustement du modèle (si nécessaire)
- Corpus de test (TEST)
  - ▣ Évaluation des performances du modèle en généralisation

!! La taille est critique ...

# Au final

21

- Dictionnaires de descripteurs et de classes



- Corpus APP
- Corpus DEV
- Corpus TEST

d1				...			dp	Y

d1				...			dp	Y

d1				...			dp	Y

- Nouvelle donnée

d1				...				dp

# Domaine de définition des classes

22

- $Y$  est un ensemble fini : Problème de *classement*
  - ▣ Associer une donnée à une valeur discrète parmi plusieurs classes prédéfinies
  - ▣ **Classement binaire** :  $Y = \{0,1\}$
  - ▣ **Classement multi-classes** :  $Y = \{0,1,\dots,I\}$
- $Y$  est un ensemble infini : Problème de *régression*
  - ▣ Associer une donnée à une valeur continue
  - ▣ **Régression** :  $Y \subset \mathbb{R}$

# Classification multi-classes :

## Cas particulier

23

- Possibilité d'associer plusieurs classes à une seule donnée
  - ▣ *Ensemble de classes discrètes non exclusives*
$$Y = \{a, b, c, d, \dots\}$$
  - ▣ *Si une donnée n'est associée qu'à une seule classe*
    - **Classement uni-label**
  - ▣ *Si une donnée peut être associée à plusieurs classes*
    - **Classement multi-labels**
- Cas proposé par peu d'algorithmes
- Correspond souvent à plusieurs classifications binaires.

# L'algorithme – La construction du modèle



# Apprentissage supervisé

25

- Création automatique d'un *modèle* à partir d'un corpus de données d'apprentissage *annotées*
  - ▣ **Prédire** une classe par donnée « connue »
  - ▣ **Généraliser** : Prédiction sur une donnée non connue
- Modèle permettant d'associer à toute donnée correctement décrite une valeur définie
- Objectif de l'apprentissage :
  - ▣ Identifier une liaison fonctionnelle qui soit la plus « efficace » possible entre le document et la classe

# Classification supervisée

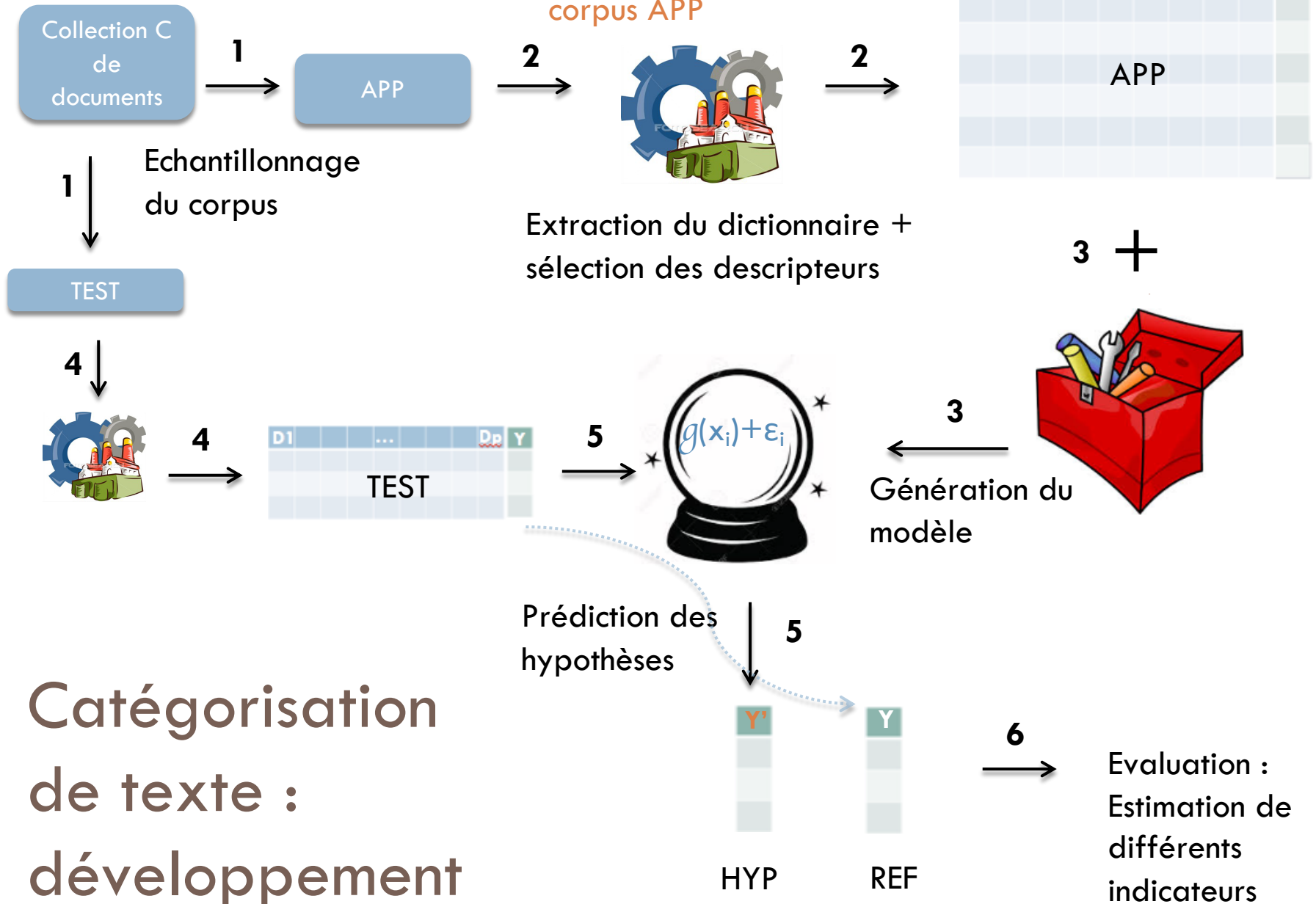
26

Plus formellement :

- Ensemble de couples document/classe :  $(x_i, y_i)$ 
  - ▣ Avec  $x_i \in \mathcal{C}$  , l'ensemble des données d'apprentissage
  - ▣ Avec  $y_i \in \mathcal{Y}$  , l'ensemble des classes à prédire
  - ▣ Tel que :  $y_i = f(x_i) + w_i$  ( $w_i$  bruit de mesure)
- Construction d'un modèle
  - ▣ Déterminer la représentation compacte de  $f$  par  $g$  appelée fonction de prédiction.
  - ▣ Tel que :  $y_i = g(x_i) + \varepsilon_i$  ,  $\varepsilon_i$  erreur de prédiction

Corpus

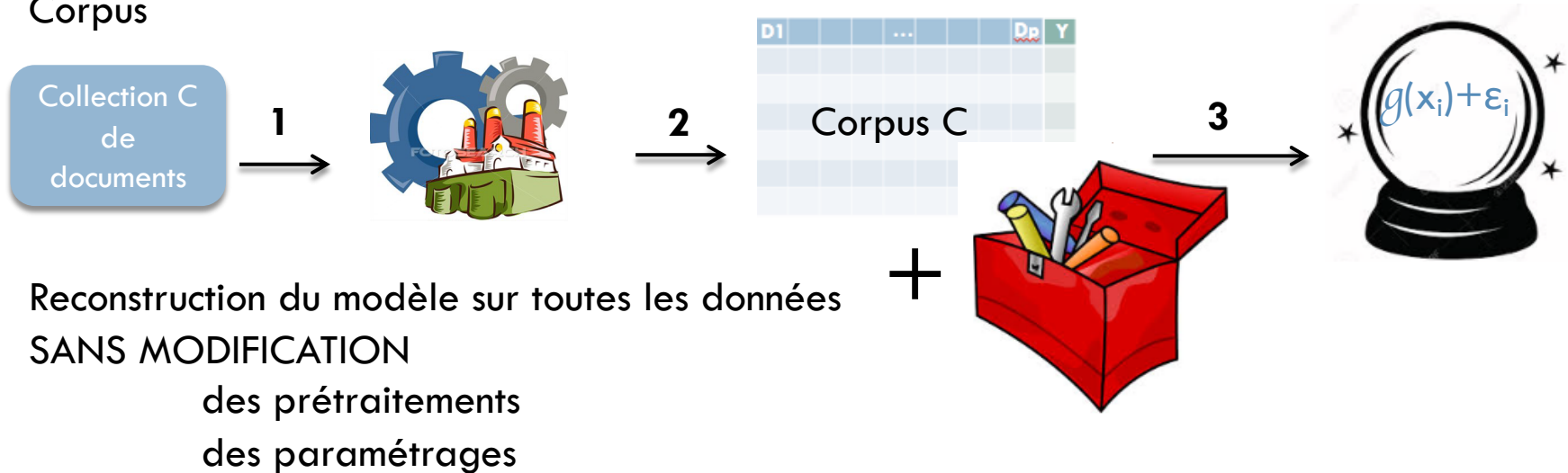
!! La construction de  
l'usine se fait sur le  
corpus APP



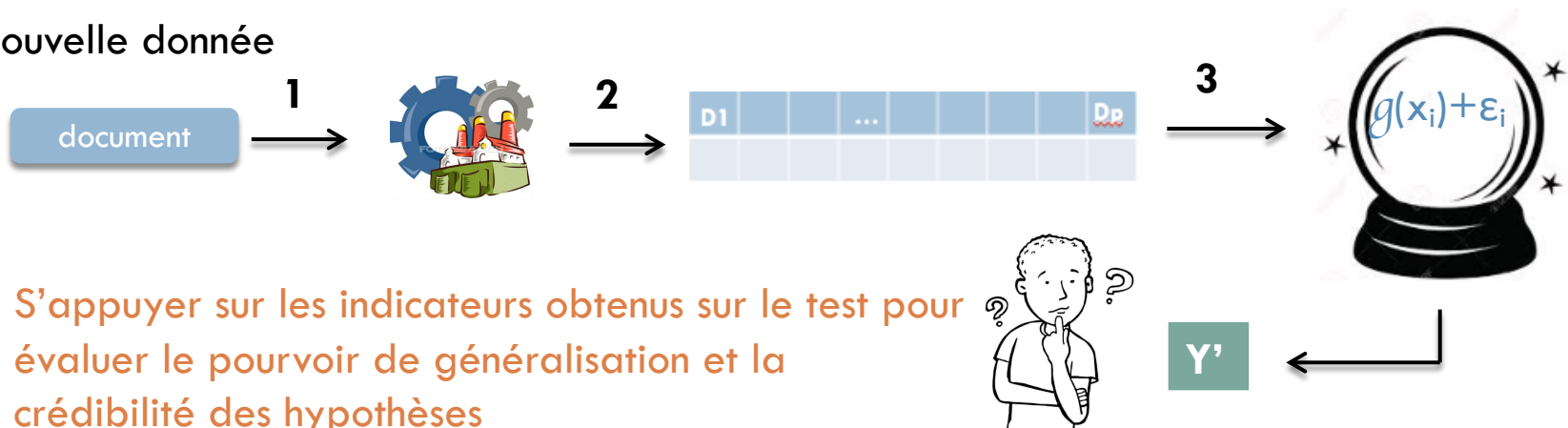
Catégorisation  
de texte :  
développement

# Catégorisation de texte : mise en production

Corpus



Nouvelle donnée



# À propos du modèle

29

- Peut-être considéré comme une boîte noire
    - ▣ Simple utilisateur... mais
  - Selon l'algorithme choisi :
    - ▣ Différentes représentations possibles
    - ▣ Différents paramètres à ajuster
- Meilleur choix et optimisation de l'apprentissage si on connaît l'algorithme

# L'évaluation

# Validation classique des résultats

31

- Cas classique : Assez de données annotées
  - ▣ Ex : 1 APP (70%) et 1 TEST (30%)
  - ▣ Estimation de l'erreur de prédiction
    - Évaluation du modèle sur l'APP
    - Taux de mauvaise classification sur l'APP
  - ▣ Estimation de l'erreur de généralisation
    - Evaluation du modèle sur le TEST
    - Taux de mauvaise classification sur le TEST
- Mise en production
  - ▣ Ré-apprentissage du modèle sur TOUT le corpus annoté

# Mesures de performance du modèle

32

- Évaluation de l'erreur
  - ▣ Soit le couple  $(x_i, y_i)$ ,  $y_i$  classe de **référence**
  - ▣ Soit le modèle  $g$
  - ▣ Soit l'**hypothèse**  $y'_i = g(x_i)$   
Est-ce que  $y'_i = y_i$  ?
- Comment évaluer l'erreur?
  - ▣ Dépend de l'objectif de l'application visée



# Terminologies des mesures d'évaluation en classification binaire

33

		Predicted condition			
		Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	
True condition	condition positive	<i>tp</i> True positive	<i>fn</i> False Negative (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$
	condition negative	<i>fp</i> False Positive (Type I error)	<i>tn</i> True negative	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$
Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$		Positive predictive value (PPV), Precision $= \frac{\Sigma \text{True positive}}{\Sigma \text{Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{False negative}}{\Sigma \text{Test outcome negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{False positive}}{\Sigma \text{Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{True negative}}{\Sigma \text{Test outcome negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

# Tableau de contingence

34

- Aussi appelée Matrice de confusion
- Représentation des documents en fonction de leur dépendance à deux critères:
  - ▣ Hypothèse : HYP
  - ▣ Référence : REF
- Alignement des classes REF et HYP
- Remplir par comptage un tableau
  - ▣ Chaque donnée doit appartenir à l'effectif d'une case

Y'	Y
HYP	REF

# Classification binaire - Mesures classiques

35

## □ Matrice de confusion

REF	HYP			
	+'	-'	Σ	
	+	A	B	A+B
	-	C	D	C+D
	Σ	A+C	B+D	N

## □ Taux d'erreur (error rate)

$$CER = \frac{B + C}{N}$$

## □ Précision (precision)

- ▣ pourcentage de documents pertinents
- ▣ Précision élevée, moins de bruit

$$prec = \frac{A}{A + C}$$

## □ Rappel (recall)

- ▣ pourcentage de documents pertinents retrouvés
- ▣ Rappel élevé, moins de silence

$$rapp = \frac{A}{A + B}$$

Choix!!

# Confiance dans l'estimation de l'erreur?

36

- Erreur = variable aléatoire
  - ▣ Après classification, 2 valeurs possibles pour la donnée : bien ou mal classé
    - Erreur = probabilité de l'événement « mal classé »
  - ▣ En déterminer la moyenne? Un intervalle?
- Calcul de l'Erreur sur \*1\* corpus de test par le CER
  - ▣ Sur 100 exemples de test, 15 sont faux
    - Le taux d'erreur du système est de 15%?

# Intervalle de confiance

37

- Estimation du *taux d'erreur réel* du système à partir du taux d'erreur observé sur un ensemble de test T
  - ▣ Approximation de la loi binomiale par la loi normale  
Intervalle de confiance à 95%
  - ▣ On estime l'erreur par l'intervalle de confiance :

$$CER \pm 1.96 \sqrt{\frac{CER(1 - CER)}{N}}$$

!! Nombre d'exemples du jeu de test suffisant

# Courbe précision/rappel

38

Classer les données  
selon un score décroissant

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Positifs = 6  
Négatifs = 14

Seuil = 1

	^positif	^négatif	Total
positf	1	5	6
négatif	0	14	14
Total	1	19	20

recall =  $1/6 = 0.2$  ; precision =  $1/1 = 1$

Seuil = 0.95

	^positif	^négatif	Total
positf	2	4	6
négatif	0	14	14
Total	2	18	20

recall =  $2/6 = 0.33$  ; precision =  $2/2 = 1$

Seuil = 0.9

	^positif	^négatif	Total
positf	3	3	6
négatif	0	14	14
Total	3	17	20

recall =  $3/6 = 0.5$  ; precision =  $3/3 = 1$

Seuil = 0.85

	^positif	^négatif	Total
positf	3	3	6
négatif	1	13	14
Total	4	16	20

recall =  $3/6 = 0.5$  ; precision =  $3/4 = 0.75$

Seuil = 0

	^positif	^négatif	Total
positf	6	0	6
négatif	14	0	14
Total	20	0	20

recall =  $6/6 = 1$  ; precision =  $6/20 = 0.3$

# F-mesure ou F-score

39

- F-mesure :
  - ▣ combinaison de la précision et du rappel
  - ▣ mesure unique pour accélérer les comparaisons

$$fmes = \frac{(1 + \beta^2) * prec * rapp}{\beta^2 * prec + rapp}$$

- Moyenne harmonique pondérée
  - ▣ généralement  $\beta=1$  (F1-mesure)
  - ▣  $\beta=2$  : poids du rappel double p/r précision
  - ▣  $\beta=0,5$  : poids de la précision double p/r rappel

# Classification multiclass

40

- Uni-label ou multi-label, mesures par classes
- Précision de la classe  $i$  :

$$\text{prec}_i = \frac{\# \text{ instances correctement classées } i}{\# \text{ instances classées } i}$$

- Rappel de la classe  $i$  :

$$\text{rapp}_i = \frac{\# \text{ instances correctement classées } i}{\# \text{ instances réellement } i}$$

- Comment combiner ces mesures?



# Classification multiclass

41

## □ Macro-mesures

### ▣ Même poids pour toutes les classes

- + Ne pas masquer les classes rares
- Classes rares et très présentes ont même importance

$$\text{mes}_{macro} = \frac{\sum_i \text{mes}_i}{\# \text{ classes}}$$

$$\text{mes} \in \{prec, rapp\}$$

## □ Micro-mesures

- ▣ Même poids pour tous les documents
- ▣ Classe très présentes masque les résultats sur la classe rare

**HYP**

	+	-
+	TP = $\sum TP_i$	FN = $\sum FN_i$
-	FP = $\sum FP_i$	

**REF**

$$\text{prec}_{micro} = \frac{TP}{TP + FP}$$

$$\text{rapp}_{micro} = \frac{TP}{TP + FN}$$

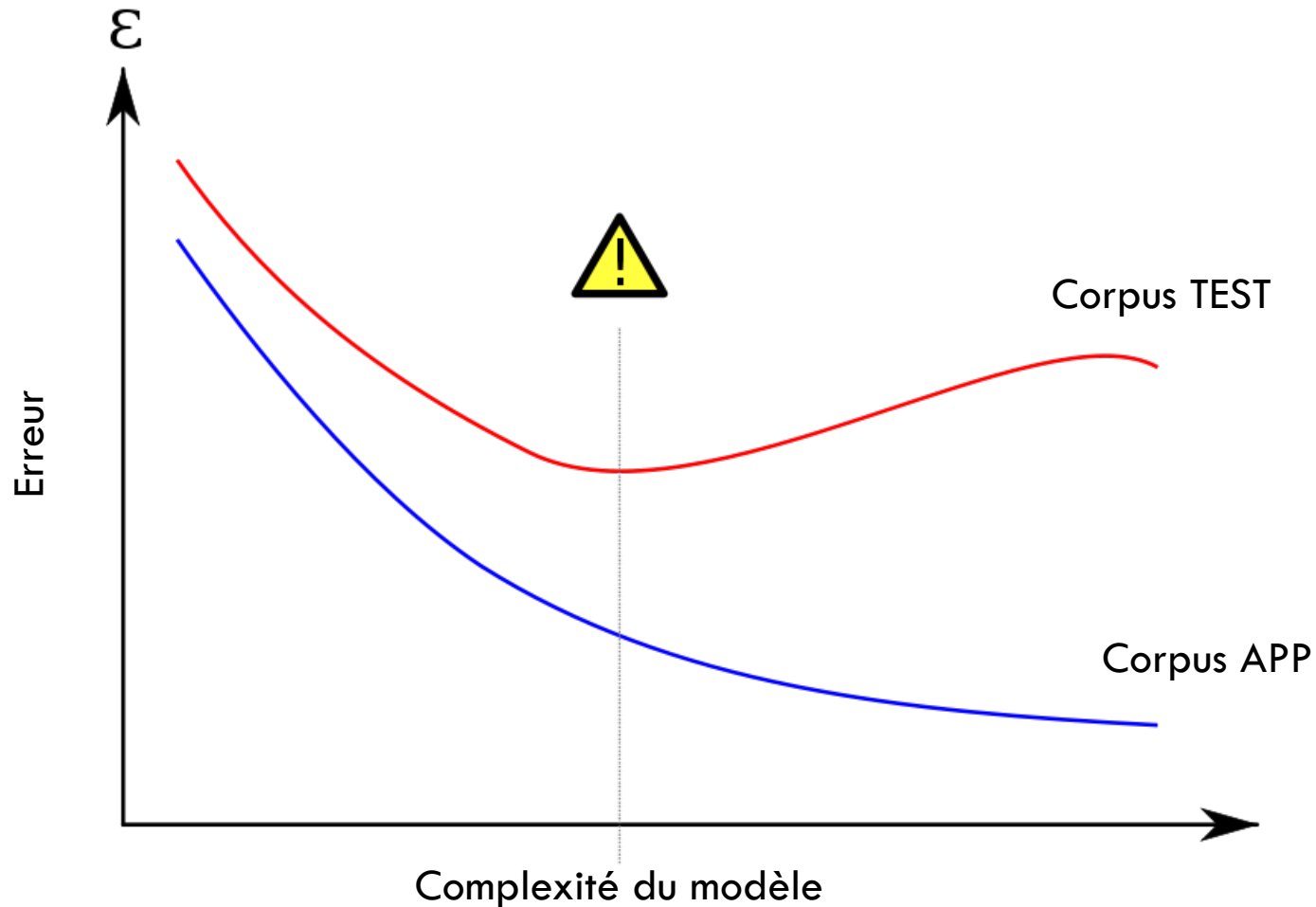
# Problème de sur-apprentissage

42

- Deux critères à considérer
  - ▣ Erreur de prédiction
  - ▣ Erreur de généralisation
- Quand « arrêter » d'apprendre?
  - ▣ Erreur de prédiction diminue ET l'erreur de généralisation augmente
- Comment faire?
  - ▣ Taille du corpus d'apprentissage
  - ▣ **Paramètres d'ajustement** du modèle

# Problème de sur-apprentissage

43



# Validation croisée

44

- Problème : manque de données annotées
- Approche par « leave one out »
  - ▣ Soit un ensemble de  $P$  données annotées
  - ▣ Construction de  $P$  modèles différents sur (APP-1 donnée)
  - ▣ Test de chacun des modèles sur la donnée mise de côté
- Généralisation au « N-fold »
  - ▣ Découpage de l'APP en  $N$  sous-ensembles distincts
  - ▣ Apprentissage sur  $N-1$  fold et test sur le fold restant
- Erreur : moyenne des erreurs de chaque fold

# Quelques réflexions

# Données d'apprentissage

46

- Le point sensible de la classification supervisée
  - ▣ Nécessité *suffisamment* de données annotées
  - ▣ *Suffisamment*? Dépend de :
    - La difficulté de la tâche
    - La complexité de représentation des données
- Problème:
  - ▣ L'annotation du corpus d'apprentissage/test est humaine
    - Coût très élevé
- Mais les méthodes ont fait leurs preuves!

# Facteurs de succès d'un projet

47

- Objectifs précis, stratégiques et réalistes
  - ▣ !! Données existantes et disponibles
- Qualité et richesse des informations collectées
  - ▣ !! Collecte des données, contrôle qualité
- Maîtrise des techniques de data mining utilisées
  - ▣ !! En utiliser plusieurs
- Bonne restitution des résultats

# Les Data miners

48

- De nombreuses compétences :
  - ▣ Maîtrise des outils d'exploitation performants
  - ▣ Expertise mathématique pour analyse des résultats
  - ▣ Bonne connaissance métier
- Besoin croissant de data miners...

<http://archives.lesechos.fr/archives/2012/lesechos.fr/07/15/0202173368914.htm>

<http://news.efinancialcareers.com/fr-fr/139910/data-miner-un-job-davenir-en-it-finance-mais-qui-reste-ultra-selectif/>

« Data scientist : The Sexiest Job of the 21st Century », T.H. Davenport et D.J Patil, Harvard Business Review, 2012



# Ressources en ligne

67

- <http://chirouble.univ-lyon2.fr/~ricco/data-mining>
  - ▣ Un portail pour la documentation : liens, supports de cours en ligne, logiciels, données
- <http://www.kdnuggets.com>
  - ▣ « Le » portail du DATA MINING, avec toute l'actualité du domaine
- <http://data.mining.free.fr>
  - ▣ Le site de Stéphane Tufféry
- Wikipédia

# Quelques sources

68

## Livres :

- « Apprentissage artificiel, concepts et algorithmes », A.Cornéjuols et L.Miclet

## Cours sur le web :

- <http://www.grappa.univ-lille3.fr/~ppreux/fouille/>
- <http://www.dsi.unive.it/~marek/files/06%20-%20datamining.pdf>
- <http://www.public.asu.edu/~jye02/>
- <https://eric.univ-lyon2.fr/~ricco/cours/slides/TM.C%20-%20categorisation%20de%20textes.pdf>
- <http://freedownloadb.com/ppt/data-mining-data-warehousing-lecture-notes>