Recherche d'Information

Cours « Document et Web Sémantique »

Nicolas Delestre, Nicolas Malandain

INSA ROUEN NORMANDIE

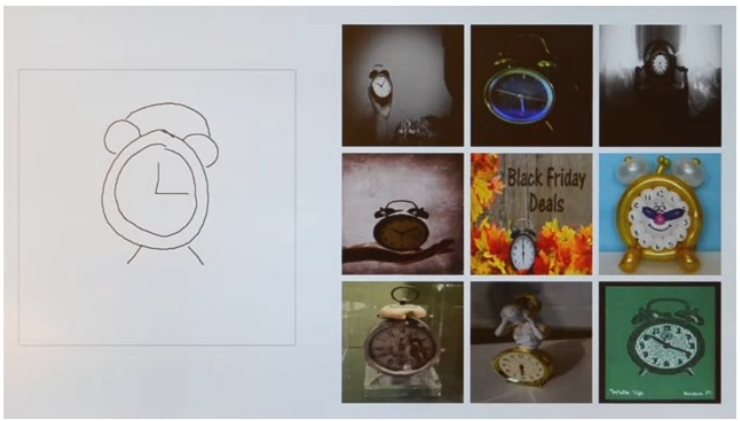
INSA ROUEN NORMANDIE

Contexte 1 / 3

Moteur de recherche

- Type document : texte
  - sonore :
    - exact : Shazam, SoundHound
    - approché : SoundHound, midomi
  - image (dessin ou photo)
- Sur le Web
- Corpus généraliste
- Indexation automatique

midomi



http://sketchy.eye.gatech.edu/

INSA ROUEN NORMANDIE

Contexte 2 / 3

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

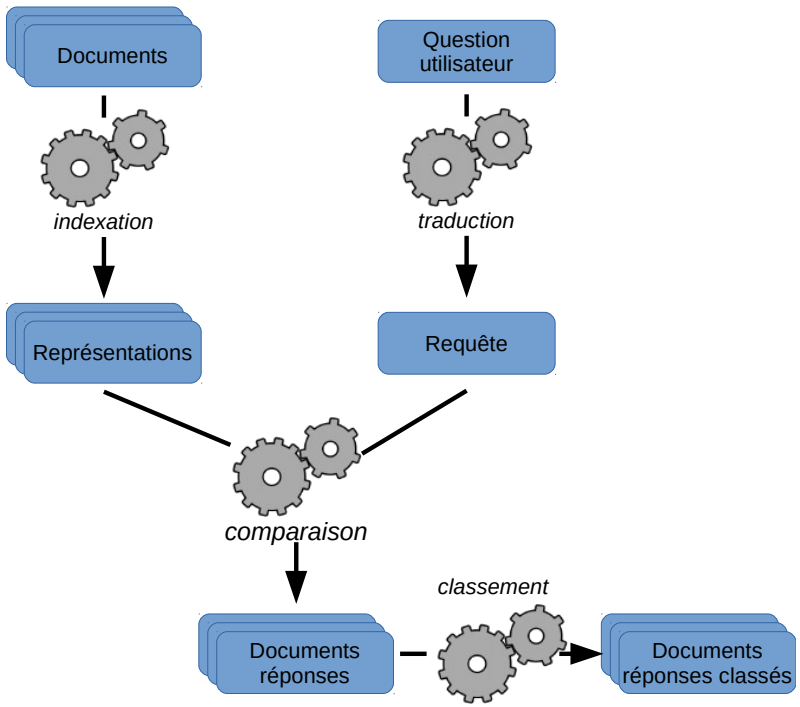
Show and tell : A neural image caption generator
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan
`https://arxiv.org/abs/1411.4555`

INSA ROUEN NORMANDIE

Contexte 3 / 3

Deux phases

- Phase d'indexation
  - Parcours du corpus
  - Création pour chaque document d'une représentation manipulable
- Phase de requêtage
  - Création d'une représentation de la question utilisateur
  - Sélection des documents « pertinents »
  - Classement de ces documents
- Chaque représentation nécessite des traitements du texte à indexer

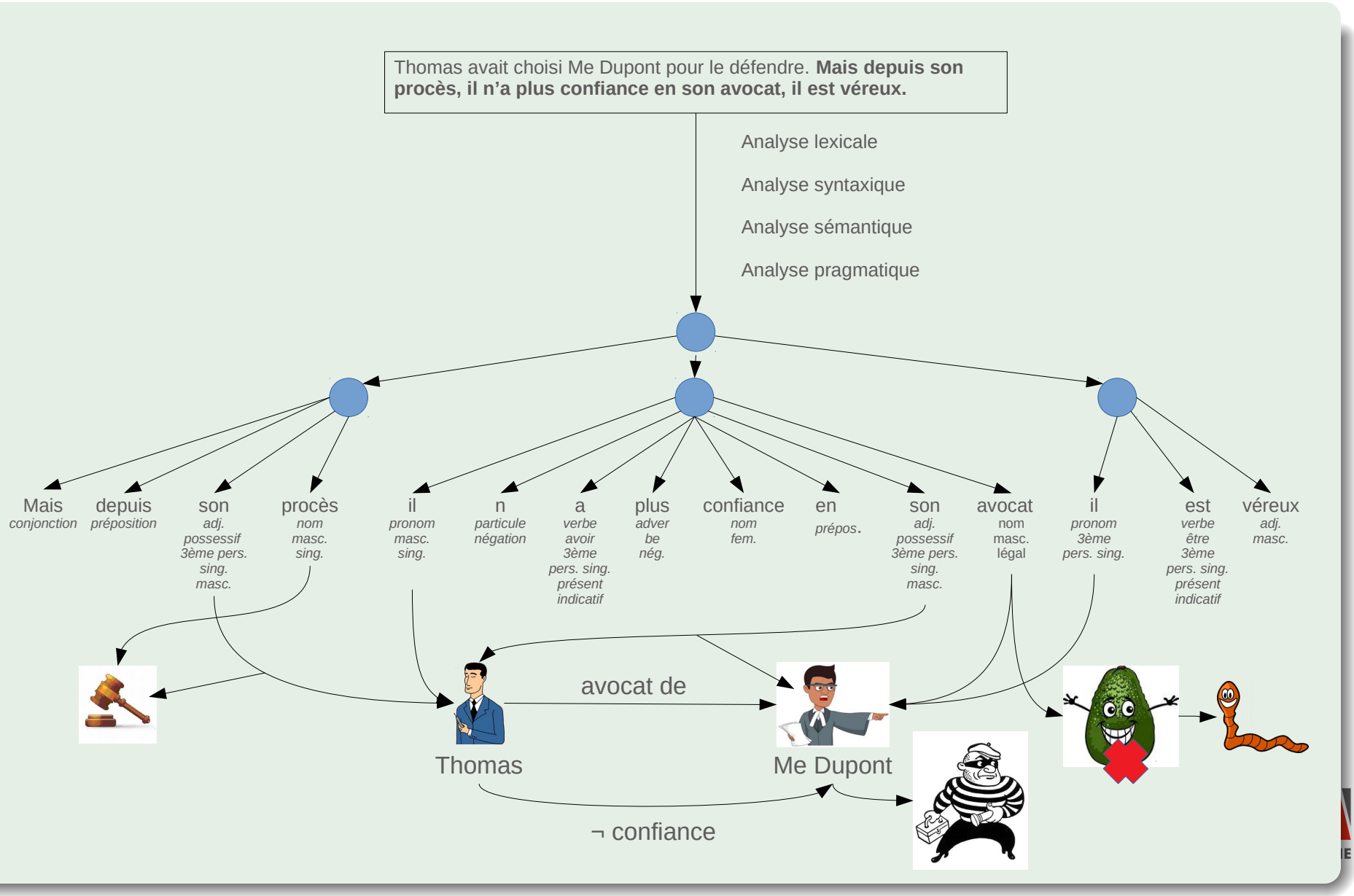


Le TAL

- Le TAL a pour objectif d'interpréter des documents textuels, pour :
  - les classer
  - les traduire
  - les résumer
  - etc.
- Dans le cas des moteurs de recherche, ils permettent de « normaliser » les textes

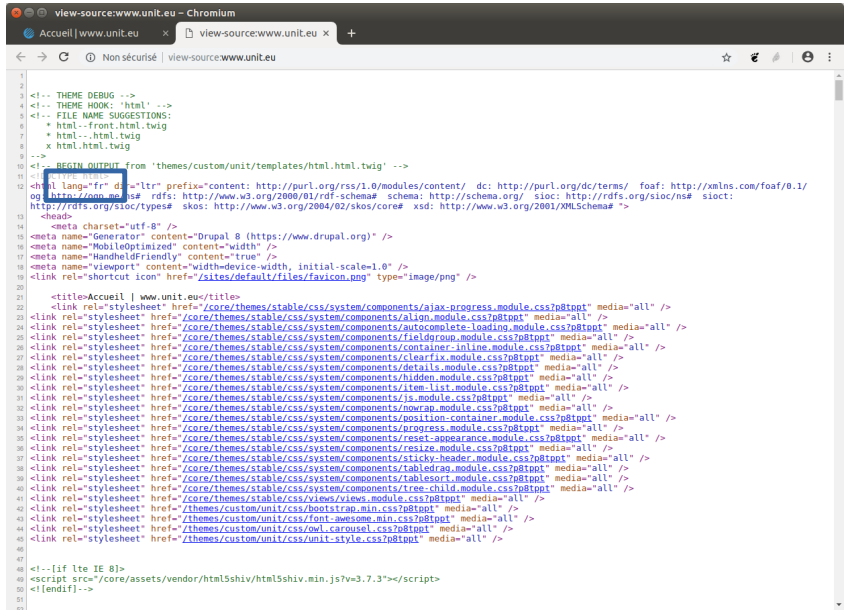
INSA ROUEN NORMANDIE

Une chaîne de traitements



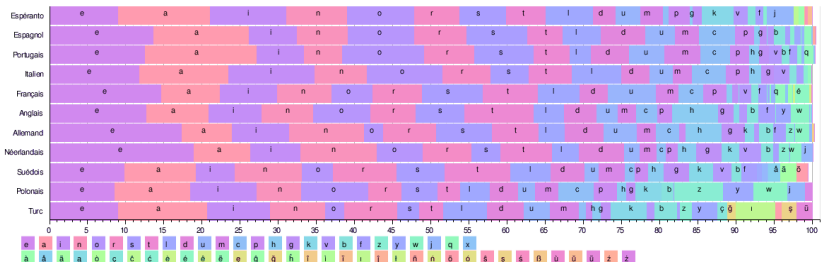
Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue



Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue



https://fr.wikipedia.org/wiki/Fréquence_d'apparition_des_lettres_en_français

Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue

sticef .org
Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation

version pleine page
version à télécharger (pdf)
Volume 14, 2007
Article de recherche

Sommaire
Rechercher
auteur
année
titre
résumé
abstract
rubrique
Contact : infos@sticef.org

Analyse et représentation en deux dimensions de traces pour le suivi de l'apprenant

Nicolas DELESTRE, Nicolas MALANDAIN [LITIS, INSA de Rouen]

RÉSUMÉ : Le suivi d'apprenants lors de la résolution de problèmes est difficile, surtout lorsque le nombre d'apprenants est important ou lorsque la résolution de problèmes se fait à distance. Nous proposons ici une représentation graphique en deux dimensions des traces de ces apprenants qui pourrait être utilisée dans un logiciel de « monitoring ». Pour arriver à ce résultat nous avons adapté et combiné des algorithmes d'analyse numérique (principalement des algorithmes de réduction de dimensions : carte de Kohonen et SNE). Nous avons aussi abordé la problématique de distance entre ensembles en proposant une nouvelle mesure de similarité lorsque leurs éléments sont sémantiquement proches. Enfin nous avons validé et amélioré notre approche à l'aide tout d'abord de données simulées, puis de données réelles issues d'une expérimentation.

MOTS CLÉS : Visualisation de traces, projection 2D de données symboliques, distance/similarité entre ensembles, cartes conceptuelles, carte de Kohonen, algorithme du SNE.

ABSTRACT : The learner follow-up in problem solving is a hard issue. It is more difficult when there are a lot of learners or when those learners use distance learning. We propose in this paper a two-dimensional graphic representation of student's traces. To achieve this goal, we use and modify numerical analysis algorithms (automatic dimensionality reduction algorithms like Self Organizing Map and Stochastic Neighbour Embedding). We also propose a new distance between sets whose elements have semantic similarity. Finally, we validate and improve our algorithm with simulated data and experimental data.

KEYWORDS : Display of student traces, symbolic data 2D

INSA ROUEN NORMANDIE

Identification de la langue

- Information présente sous forme de métadonnées
- Inféré à partir de statistique
  - Taux compression
  - Calculer la fréquence d'apparition de n-gram (succession de n caractères) : méthode tolérante aux fautes d'orthographe
- Difficultés lorsque le document
  - est court
  - est multilingue
  - contient du texte qui n'est pas de la langue

Dès lors, l'algorithme d'apprentissage d'une carte de Kohonen devient un algorithme d'initialisation (Cf. figure 12).

Données en entrée :

X : les cartes conceptuelles d'apprentissage (de l'enseignant)

N : les neurones

Données en sortie :

W : les prototypes de la carte de Kohonen

début

Initialiser les W_i avec $\emptyset$

$W_{N(X_i)} \leftarrow \{X_i\}, i \in [1..|X|]$

Pour chaque neurone $n \notin N$ **faire**

Calculer les cartes influentes $C_i \subset X$ avec $i > 0$

Calculer les attributs att de l'ensemble des cartes C_i

Calculer le nombre d'attributs nb_{att} des cartes pour W_n

$W_n \leftarrow$ l'ensemble des cartes générées à partir de att et nb_{att}

fin

Figure 12 • Phase d'initialisation d'une carte de Kohonen de cartes conceptuelles

Précisons quelques points de cet algorithme :

- pour un neurone *n* de coordonnées *(i,j)*, les cartes d'influence sont les cartes se trouvant dans le cercle de centre *(i,j)* et de rayon *r*. Ce rayon est par défaut fixé au nombre maximal d'attributs que possèdent les cartes d'apprentissage,

INSA ROUEN NORMANDIE

Segmentation

Découper le texte en mot

- utiliser les séparateurs de mots
  - espaces, ponctuations, apostrophe
    - l'enfant
    - aujourd'hui
  - trait d'union
    - Mont-Saint-Michel
    - qu'en est-il ?
- Cela peut être difficile avec des langues acceptant des mots composés
  - ≪ Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz ≫
  - ≪ loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine ≫

Thomas avait choisi Me Dupont pour le défendre. Mais depuis son procès, il n'a plus confiance dans son avocat, il est véreux.

⇓

Thomas avait choisi Me Dupont pour le défendre Mais depuis son procès il n a plus confiance dans son avocat il est véreux

INSA ROUEN NORMANDIE

Pré-étiquetage

Hypothèse quant au rôle des mots

- mots connus (dictionnaire)
- mots inconnus :
  - les terminaisons : noms (-eur), adjectifs (-able), adverbes (-ment), verbes (-er, -ir, -aient), ...
  - genre/nombre fonction des terminaisons
  - statistiques
- problèmes avec les ≪ homographes ≫

Thomas avait choisi Me Dupont pour le défendre Mais depuis son procès il n a plus confiance dans son avocat il est véreux

⇓

- Thomas : *EN*
- avait : *avoir 3ème pers imparfait*
- choisi : *choisir p. passé* ou *adj.*
- Me Dupont : *EN*
- pour : *prep.*
- le : *art. def.*
- défendre : *défendre inf.*
- Mais : *conj.* ou *nom* ou *adv.*
- depuis : *prep.* ou *adv.*
- son : *adj. pos.* ou *nom*
- procès : *nom*
- il : *pron.*
- n : *particule*
- a : *pron. pers.* ou *avoir, 3ème pers. sing. présent*
- plus : adv. ou conj.
- confiance : *nom* ou *confiancer, 3ème 3ème pers. sing. présent indic.* ou *subj.* ou *imper.*
- dans : *prép.*
- son : *adj. pos.* ou *nom*
- avocat : nom
- il : *pron.*
- est : *être 3ème sing. présent* ou *nom*
- véreux : *adj.*

Treetagger

Logiciel d'analyse lexicale gratuit (non opensource) multi-plateformes, multi-lingues

- Site Web : `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
- Tagset français : `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html`

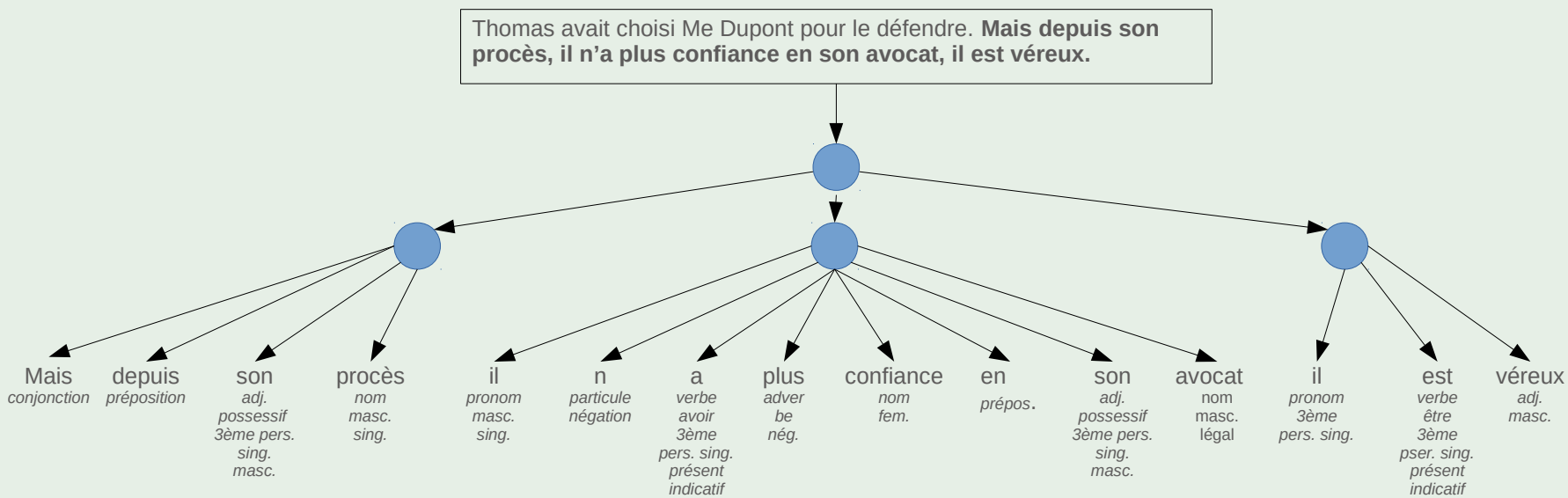
```
$echo "Thomas avait choisi Me Dupont pour le défendre Mais depuis son procès il n a plus confiance
dans son avocat il est véreux" | cmd/tree-tagger-french
reading parameters ...
tagging ...
Thomas NAM Thomas
avait VER:impf avoir
choisi VER:pper choisir
Me ABR Me
Dupont NAM Dupont
pour PRP pour
le PRO:PER le
défendre VER:infi défendre
```

Treetagger peut proposer plusieurs étiquetages (ex : ≪ Je **suis** allé au cinéma ≫)

Analyse syntaxique

Objectifs

- Créer un arbre syntaxique
- Préciser l'étiquetage des mots
  - Trois méthodes : automatique à partir d'un corpus étiqueté et avec un algorithme d'étiquetage, à partir d'une grammaire formelle, méthode mixte

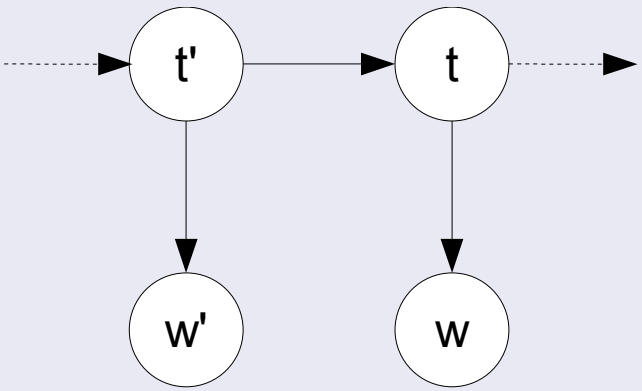


Exemple d'étiquetage automatique 1 / 3

Hidden Markov Model / Chaînes de Markov Cachées

On cherche l'étiquette t du mot w :

Apprentissage :



$$p(w|t) = \frac{|(w,t)|}{\sum_x (x,t)} = \frac{|(w,t)|}{|t|}$$

(w,t) signifie que w est étiqueté par le tag t

$$p(t|t') = \frac{|t't|}{\sum_y |t'y|} = \frac{|t't|}{|t'|}$$

$t't$ représente la séquence des deux tags

Décision : Choix de la séquence de tags de plus forte probabilité pour une séquence w_1^n de n mots :

$$\hat{t}_1^n = \underset{t_1^n}{argmax} \prod_{i=1}^{n} \underbrace{p(w_i|t_i)}_{\text{émission}} \times \underbrace{p(t_i|t_{i-1})}_{\text{transition}} \qquad \textit{(taggeur bigramme)}$$

Exemple d'étiquetage automatique 2 / 3

Tagger bigramme pour « les beaux avions »

- « les » : article ou pronom ?
- « beaux » : nom, adjectif ou un adverbe (« il a beau partir tôt, il arrivera en retard ») ?
- « avions » : verbe ou nom ?

« les »

- $P(les|article) \times P(article|debutDePhrase)$
- $P(les|pronom) \times P(pronom|debutDePhrase)$

$\Rightarrow$ *les* est un *article*

« beaux »

- $P(beaux|nom) \times P(nom|article)$
- $P(beaux|adverbe) \times P(adverbe|article)$
- $P(beaux|adjectif) \times P(adjectif|article)$

$\Rightarrow$ *beaux* est un *adjectif*

Exemple d'étiquetage automatique 3 / 3

≪ avions ≫

- $P(avions|nom) \times P(nom|adjectif)$
- $P(avions|verbe) \times P(verbe|adjectif)$

$\Rightarrow$ *avions* est un *nom*

Attention

- Certains cas peuvent nécessiter des taggeur trigramme (voire plus)

≪ il la fatigue ≫ , ≪ de la fatigue ≫ [Éric Laporte]

INSA ROUEN NORMANDIE

Analyse sémantique

Objectif : ≪ extraire ≫ le sens d'un texte

Analyse du sens des mots → de la phrase → du texte

Réalité d'une machine

Analyse ≪ automatique ≫ limitée à un domaine

- création d'un modèle du domaine
  - logique
  - réseaux sémantiques
  - graphes conceptuels (représentation des connaissances basée sur la logique)
- filtrage du ≪ sens ≫ des mots liés au domaine
  (réduit les problèmes comme celui de la polysémie)

INSA ROUEN NORMANDIE

Analyse pragmatique

Objectif

Re-situer l'analyse sémantique dans le contexte d'énonciation
« interprétation »

- de l'implicite (ex : « Je vais à la poste », laquelle ? certainement la plus proche)
- des présupposés (culture commune, contexte commun, etc.)
- des actes de langage (la façon de dire des choses, d'insister sur certains, figure de style)

Par exemple savoir que le procès de Thomas se déroule en France, indique que la justice « est inquisitoire. C'est l'enquête qui est au centre de la procédure, et non l'accusation. » (cf. `http://www.politique.net/2011060602-differences-justice-francaise-justice-americaine.htm`), alors qu'aux états unis elle est accusatoire.

Le TAL pour les moteurs de recherche

- Les moteurs de recherche se limitent à l'analyse lexicale, à laquelle ils ajoutent deux étapes :
  - identification de la langue
  - segmentation
  - **normalisation**
  - **filtrage**

INSA ROUEN NORMANDIE

Normalisation

Objectifs

- Simplifier l'indexation
- Représenter tous les mots dans des formes plus simples
- Deux types de normalisation :
  - textuelle
  - linguistique
- Deux types de normalisation linguistique :
  - racinisation : mot $\rightarrow$ racine. Par exemple les mots *motoriser*, *motrice*, *motoriste*, *motricité* ont pour racine *moteur*
  - lemmatisation : mot $\rightarrow$ représentation encyclopédique

Normalisation textuelle + lemmatisation

Thomas avait choisi Me Dupont pour le défendre Mais depuis son procès il n a plus confiance dans son avocat il est véreux

$\Downarrow$

Thomas avoir choisir Me Dupont pour le defendre mais depuis son proces il ne avoir plus confiance dans son avocat il etre véreux

INSA ROUEN NORMANDIE

Filtrage

Objectifs

- Simplifier l'indexation
- Supprimer les mots les plus fréquents de la langue, car supposer peu significatifs
- Mots les plus fréquents de la langue française :

1-10	de, la, le, et, les, des, en, un, du, une
11-20	que, est, pour, qui dans, a, par, plus, pas, au
21-30	...

Thomas avoir choisir Me Dupont pour le defendre mais depuis son proces il ne avoir plus confiance dans son avocat il etre véreux

⇓

Thomas choisir Me Dupont defendre proces confiance avocat véreux

INSA ROUEN NORMANDIE

Indexation

Rappel, deux étapes

1. Parcours du corpus
2. Représentation du document après l'application des algo. de TAL

Parcours

- Un programme nommé « Robot » part de pages référencées
- Il analyse le code HTML pour en extraire
  - Le contenu textuel du document (+ métadonnées) → représentation du document
  - Les liens vers les autres documents → continuer le parcours

Deux types de représentation

1. Représentation vectorielle : chaque document est représenté par un vecteur mathématique
2. Représentation ensembliste : chaque mot est représenté par un ensemble de documents

Représentation simple

Définition

- n documents d et m termes t
- $v_{i,j}$ = nombre d'occurences du terme t_j dans un document d_i

Inconvénients

- matrice sparce
- non prise en compte du nombre d'occurrences d'un terme au sein du corpus
- non prise en compte des synonymes
- ajout d'un nouveau document

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_j \\ \vdots \\ t_m \end{array}
\begin{array}{c}
\begin{array}{cccccc} d_1 & d_2 & \dots & d_i & \dots & d_n \end{array} \\
\left[\begin{array}{cccccc}
2 & 0 & \dots & & \dots & 5 \\
0 & 4 & \dots & \dots & \dots & 1 \\
\vdots & \vdots & & & & \vdots \\
 & & & v_{i,j} & & \\
\vdots & \vdots & & & & \vdots \\
0 & 0 & \dots & & \dots & 1
\end{array}\right]
\end{array}
$$

Représentation *tf .idf*

Définition

- $tf_{i,j}$: nombre d'occurrences du terme t_j dans dans le document d_i
- idf_j : $log(\frac{|D|}{\{d_i : t_j \in d_i\}})$
- $v_{i,j} = tf_{i,j} \times idf_j$

Inconvénients

- matrice sparce
- non prise en compte des synonymes
- ajout d'un nouveau document

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_j \\ \vdots \\ t_m \end{array}
\begin{array}{c}
\begin{array}{ccccccc} d_1 & d_2 & \dots & d_i & \dots & d_n \end{array} \\
\left[\begin{array}{ccccccc}
0.3 & 0 & \dots & & \dots & 0.3 \\
0 & 0.9 & \dots & \dots & \dots & 0.1 \\
\vdots & \vdots & & & & \vdots \\
 & & & v_{i,j} & & \\
\vdots & \vdots & & & & \vdots \\
0 & 0 & \dots & & \dots & 0.8
\end{array}\right]
\end{array}
$$

Représentation LSA

Définition

- *Latent Semantic Analysis* : réunir les termes qui sont corrélés (concept c_j)
- Reprendre la matrice *tf.idf* V, calculer les valeurs et vecteurs singuliers : $V = U\Sigma W^t$
- Retenir uniquement les vecteurs ayant les valeurs singulières supérieures α : $V = U_k\Sigma_k W_k^t$ avec $k << m$

$$
\begin{array}{c}
 \\
c_1 \\
c_2 \\
\vdots \\
c_j \\
\vdots \\
c_k
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} d_1 & d_2 & \dots & d_i & \dots & d_n & \end{array} \\
\begin{bmatrix}
0.3 & 0.3 & \dots & & \dots & 0.3 & \\
0.5 & 0.6 & \dots & \dots & \dots & 0.1 & \\
\vdots & \vdots & & & & \vdots & \\
 & & & w_{i,j} & & & \\
\vdots & \vdots & & & & \vdots & \\
0.1 & & 0.7 & \dots & & \dots & 0.8
\end{bmatrix}
\end{array}
$$

Inconvénients

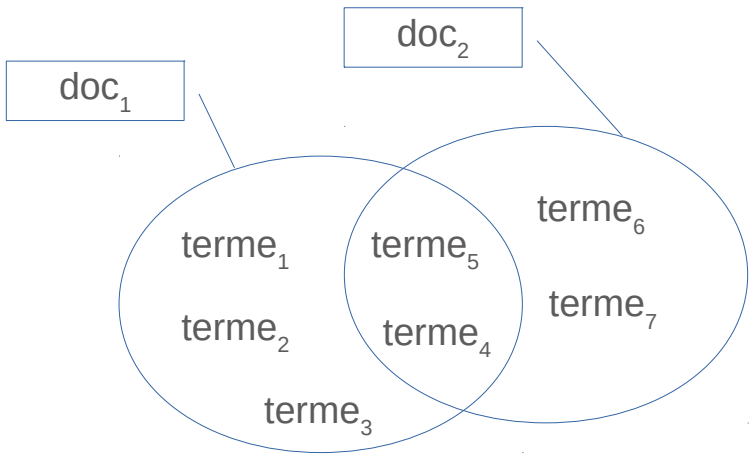
- ajout d'un nouveau document

INSA ROUEN NORMANDIE

Représentation ensemblistes 1 / 2

Constat

- Document = Suite de termes normalisés
- Création d'un index
  - Termes sont associés aux documents
  - Possibilité d'ajouter des informations aux termes
- Problèmes
  - On ne recherche pas les termes qui appartiennent aux documents
  - On cherche des documents qui contiennent certains termes

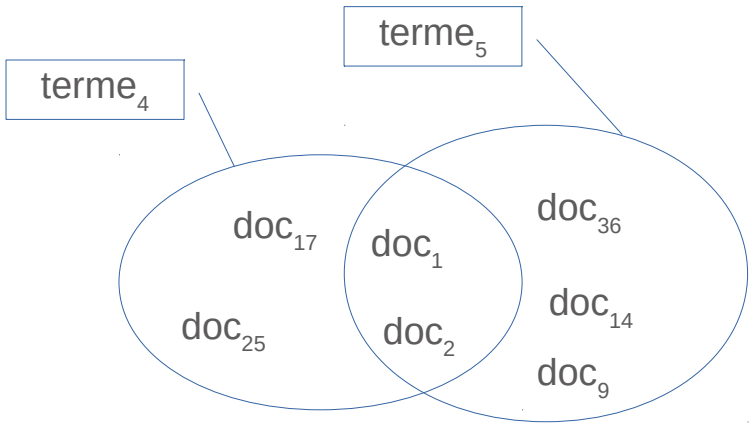


doc_1 : $terme_1$ (pos_1, pos_2, ...)
$terme_2$ (...)
$terme_3$ (...)
$terme_4$ (...)
$terme_5$ (...)
doc_2 : ...

Représentation ensemblistes 2 / 2

Index inversé

- Les documents sont associés aux termes
- Opérations ensemblistes :
  - Intersection : documents qui partagent certains termes
  - Union : documents qui contiennent certains termes
  - Soustraction : documents qui ne contiennent pas certains termes
- Avantage
  - L'ajout d'un nouveau document ne pose pas de problème



$terme_4$: doc_1 (pos_1, pos_2, ...)
doc_2 (...)
doc_{17} (...)
doc_{25} (...)
$terme_5$: ...

Phase de requétage

Rappels

- Création d'une représentation de la question utilisateur
- Sélection des documents « pertinents »
- Classement de ces documents

Question → requête

- Requête :
  - Représentation de la question de l'utilisateur
- La question peut être :
  - Une phrase en langue naturelle
  - Une suite de mots clés
  - Une expression booléenne de mots clés
- La requête est une expression booléenne de termes normalisés
  - Par défaut utilisation de l'opérateur *et*

ROUEN NORMANDIE

Sélection des documents pertinents 1 / 2

Représentation vectorielle

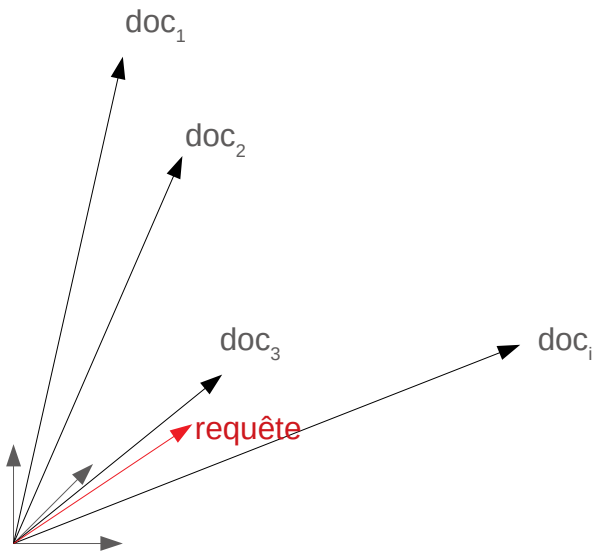
- Documents et requêtes représentés par un vecteur mathématique
- Comparaison et classement : suivant une similarité entre la requête et les documents, souvent similarité cosinus

Avantages

- Modèle homogène
- Classification des documents possibles

Inconvénients

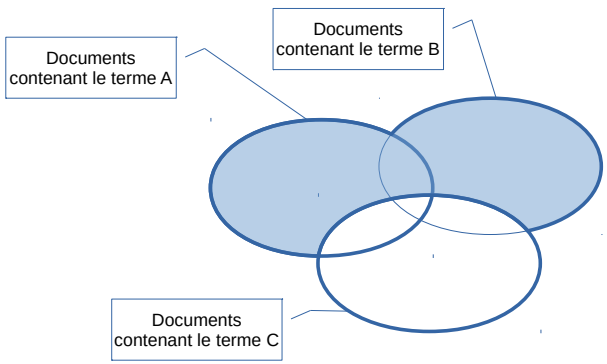
- Nombreux calculs
- Conjonction uniquement



Sélection des documents pertinents 2 / 2

Représentation ensembliste

- Comparaison :
  - Traduction des opérateurs booléens de la requête par des opérations ensemblistes :
    - Et → intersection
    - Ou → union
    - Non → Soustraction



Avantages

- Opérations très efficaces
- Permet d'ajouter facilement de nouveau documents

Inconvénients

- Pas de classement

Classement dans la représentation ensembliste 1 / 3

Historiquement

- Classement fonction du calcul d'un score qui prend en compte les termes de la requête dans les documents :
  - nombre d'occurrences
  - positions (titre, sous-titre, etc.)

Page Rank

- Classement indépendant de la requête
  - Peut être calculé en « tâche de fond » et régulièrement
- Tire parti de la topologie du graphe du Web
- Fonction de la popularité des documents du corpus
  - Plus il y a des liens qui pointent sur un document d plus d est populaire
  - Plus les documents d_i qui référencent d sont populaires, plus d est populaire

Classement dans la représentation ensembliste 2 / 3

Principe de l'algorithme du *page rank*

- Imaginer un ≪ surfeur ≫ éternelle qui parcourt le Web. La notoriété d'une page p est le nombre de fois que ce surfeur a parcouru la page p (cf. [Abi12])
- Dans les faits, les n documents d_i et les liens forment un graphe orienté :
  1. Calcul de la matrice M de transferts (matrice d'adjacence tel que les poids d'un lien allant de d_a à d_b est $1/|arcsSortant(d_a)|$. La matrice est de dimension $n \times n$
  2. Initialisation d'un vecteur N_0 de taille n représentant la notoriété de chaque document (initialisation aléatoire ou équiprobable)
  3. Mise à jour du vecteur N ($N_{i+1} \leftarrow N_i M$) jusqu'à un point fixe ($N_{i+1} - N_i < \epsilon$)

INSA ROUEN NORMANDIE

Classement dans la représentation ensembliste 3 / 3

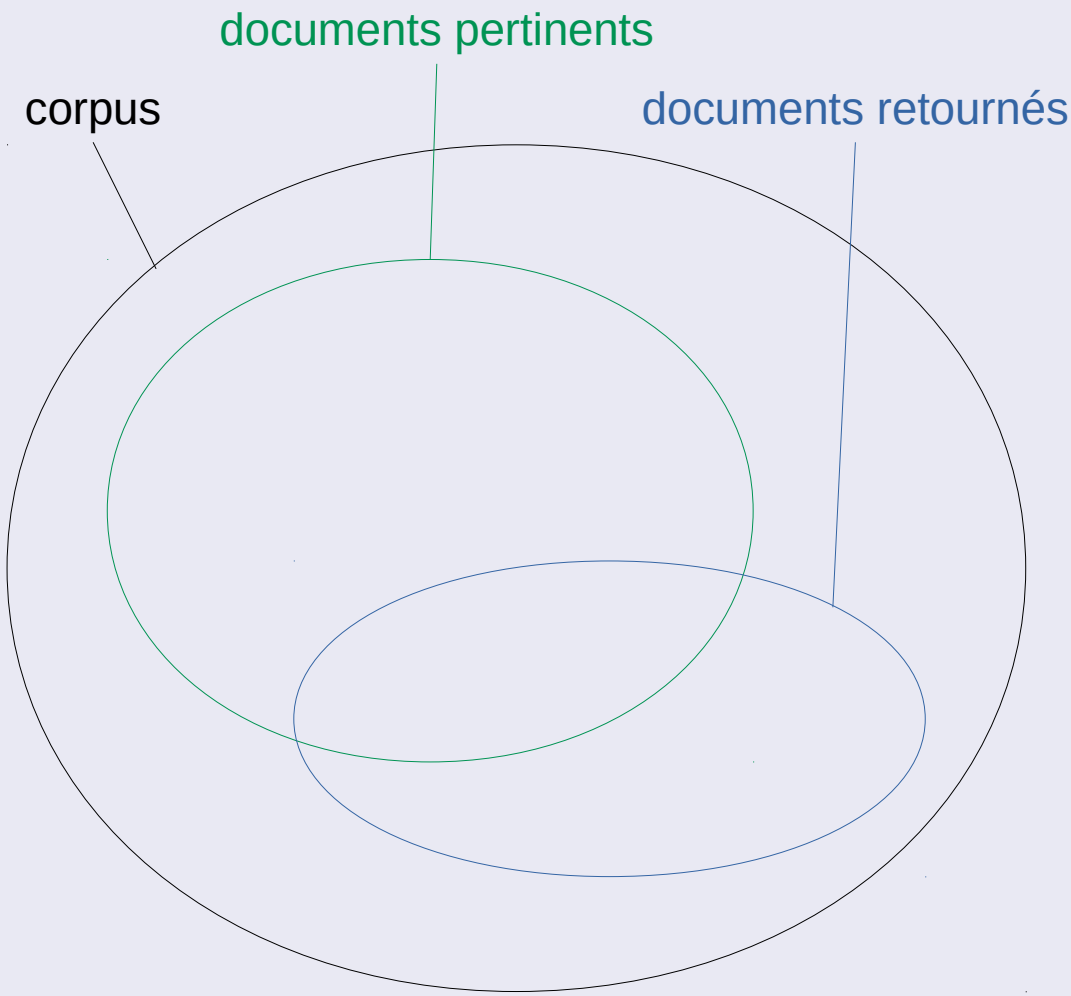
Un classement objectif ?

- Le classement effectué par certains moteurs de recherche est aussi fonction :
  - du contenu des documents (en dehors du texte)
  - de votre profil
  - des enjeux commerciaux

INSA ROUEN NORMANDIE

Évaluation de la recherche d'information 1 / 4

Ensembles pour la mesure de la qualité



Évaluation de la recherche d'information 2 / 4

Résultats non ordonnés : deux couples d'indicateurs

1. Précision, rappel :

$$précision = \frac{|Nb\ documents\ pertinents\ retournés|}{|Nb\ documents\ retournés|}$$

$$rappel = \frac{|Nb\ documents\ pertinents\ retournés|}{|Nb\ documents\ pertinents|}$$

2. Bruit, Silence :

$$bruit = \frac{|Nb\ documents\ non\ pertinents\ retournés|}{|Nb\ documents\ retournés|}$$

$$silence = \frac{|Nb\ documents\ pertinents\ non\ retournés|}{|Nb\ documents\ pertinents|}$$

Évaluation de la recherche d'information 3 / 4

Résultats non ordonnés : F-Mesure

- Très souvent l'amélioration d'un indicateur est au dépend de l'autre :
  - Lorsque l'on essaye de réduire le bruit, le silence grandit
  - Lorsque l'on essaye d'augmenter la précision, le rappel diminue
- F-mesure
  - Critère de qualité utilisant le couple rappel-précision Moyenne harmonique :

$$F_1 = 2 \times \frac{precision \times rappel}{precision + rappel}$$

INSA ROUEN NORMANDIE

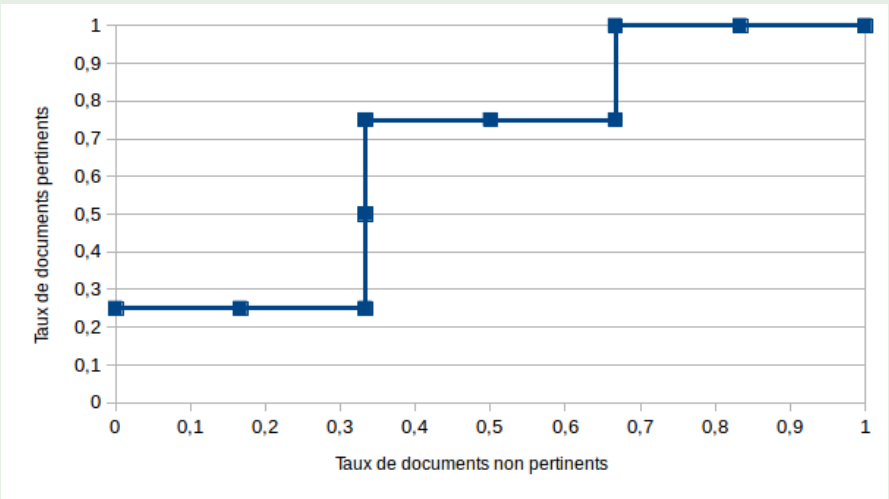
Évaluation de la recherche d'information 4 / 4

Résultats ordonnées

1. La courbe ROC (*Receiver Operating Characteristics*) :
   - Taux de documents non pertinents en abscisse (bruit), Taux de documents pertinents en ordonnée (rappel)
   - « Cette courbe se construit en estimant le taux de documents pertinents à chaque rang, en fonction du taux de documents non pertinents [à ce même rang] » [AG13]
   - Non sensible au ratio des deux classes [Ouf08]

Exemple avec un corpus de 10 documents [AG13]

Rang	Pertinence	Taux doc. non pert.	Rappel
1	1	0	1/4
2	0	1/6	1/4
3	0	1/3	1/4
4	1	1/3	1/2
5	1	1/3	3/4
6	0	1/2	3/4
7	0	2/3	3/4
8	1	2/3	3/4
9	0	5/6	1
10	0	1	1



De la recherche documentaire à la recherche d'information

Constat

- Que contiennent les documents sur le Web ? des informations
- Mais pourquoi utilisons nous un moteur de recherche ?
  - Pour avoir une réponse à une question, pour obtenir une information
  - Plutôt que de retourner des documents qui contiennent « peut-être » la réponse à la question posée, il faudrait que les moteurs de recherche retournent l'information désirée
  - C'est ce qu'essaye de faire :
    - Certains moteurs de recherche pour des questions simples
    - Les assistants personnels (Google Home, Siri, Alexa, etc.)
- Difficulté supplémentaire : les informations ne sont pas seulement dans le texte des documents, mais aussi dans des tableaux, figures, etc.

INSA ROUEN NORMANDIE

Vers la compréhension du texte

Concernant la compréhension du texte...

- Les moteurs de recherche ne comprennent pas les documents qu'ils indexent
  - Seule l'analyse lexicale est réellement réalisée
  - Document = ensemble de termes normalisés
  - « Il ne faut pas 5 minutes pour caraméliser des oignons » → (5, minute, caraméliser, oignon)
- Il faudrait soit :
  - dépasser l'analyse lexicale et donner « sens » aux mots, phrases et documents
  - produire des informations que les moteurs de recherche puissent comprendre (cf. le prochain cours)

INSA ROUEN NORMANDIE

Une meilleur représentation des documents du WEB

Cela nécessite

- Ne plus représenter les documents par un ensemble de termes normalisés
- Donner « sens » aux mots, aux phrases, aux documents
  - Qu'est ce que donner du sens ?
    - Relier les mots, phrases et documents entre eux par des liens sémantiques
    - Permettre des traitements sur ces mots, phrases, ou documents
  - exemple : word2vec (Google 2013)
    - Apprentissage automatique à partir d'un corpus de documents
    - Chaque terme est représenté par un vecteur, tels que : deux vecteurs proches sont sémantiquement proches des opérations sont possibles sur ces vecteurs, per ex. $roi - homme + femme \approx reine$

INSA ROUEN NORMANDIE

word2vec et doc2vec 1 / 2

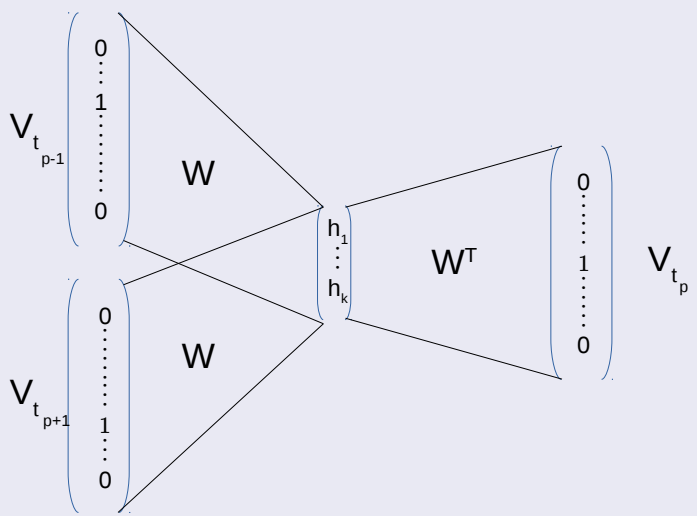
word2vec : principe

cf. https://www.youtube.com/watch?v=_YYQNpjvvLE

- Construction d'une matrice carrée symétrique M (de taille t) du nombre de co occurrences de termes t_i du corpus
- Calcul d'une matrice W (de taille t, k) tel que $W.W^t \simeq M$, W est la représentation des termes t_i

word2vec dans la pratique : utilisation d'un réseau de neurones

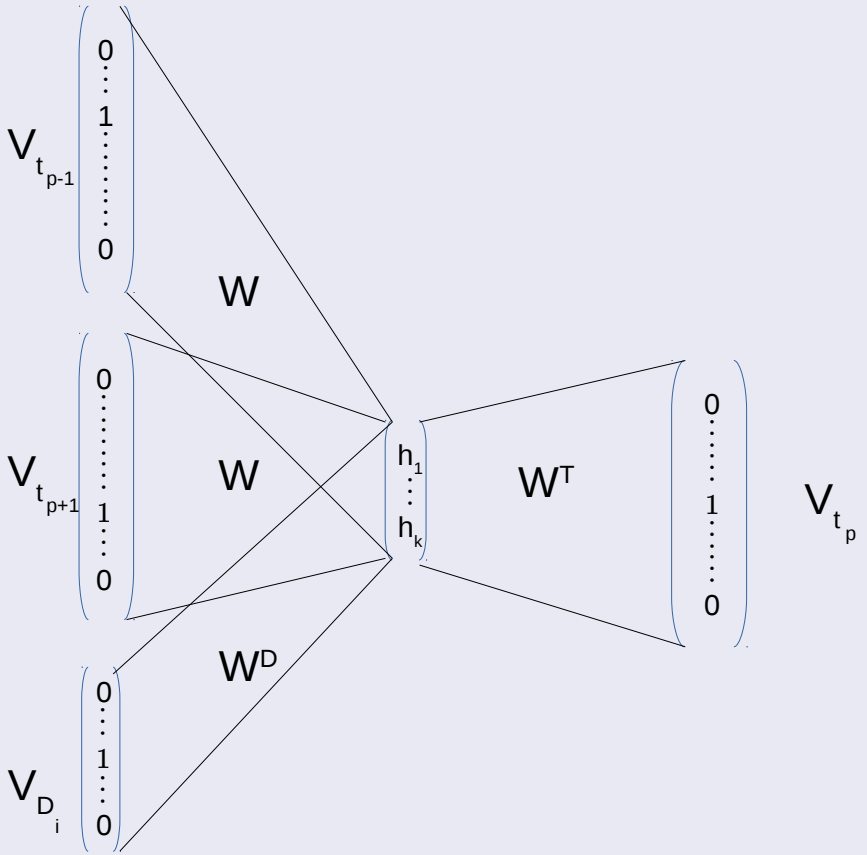
- Chaque terme t_i est représenté par un vecteur v_{t_i} de taille t (un seul 1, le reste à 0)
- Pour chaque document, on présente au réseau les représentations vectorielles de termes entourant le terme t_p à reconnaître (ici fenêtre de 3)
- La représentation finale du terme t_p est $W^T.V_{t_p}$



word2vec et doc2vec 2 / 2

doc2vec : une extension de word2vec

- Lorsque l'on apprend le réseau, on ajoute en entrée le document pour lequel on apprend la représentation du terme t_p
- La représentation du document D_i est $W^{D^T}.V_{D_i}$
- Plus adapté à la thématisation des documents plutôt qu'à la recherche d'information



Compétition SQuAD 1 / 2

The Standford Question Answering Dataset

- `https://rajpurkar.github.io/SQuAD-explorer/`
- La version 2.0 : 150000 questions/réponses sur plus de 500 extrait d'articles Wikipédia anglophone
- Corpus
  - Public : texte + questions/réponses + un script d'évaluation
  - Privé : questions/réponses pour évaluer les contributions
  - Réponse = terme(s) d'un article

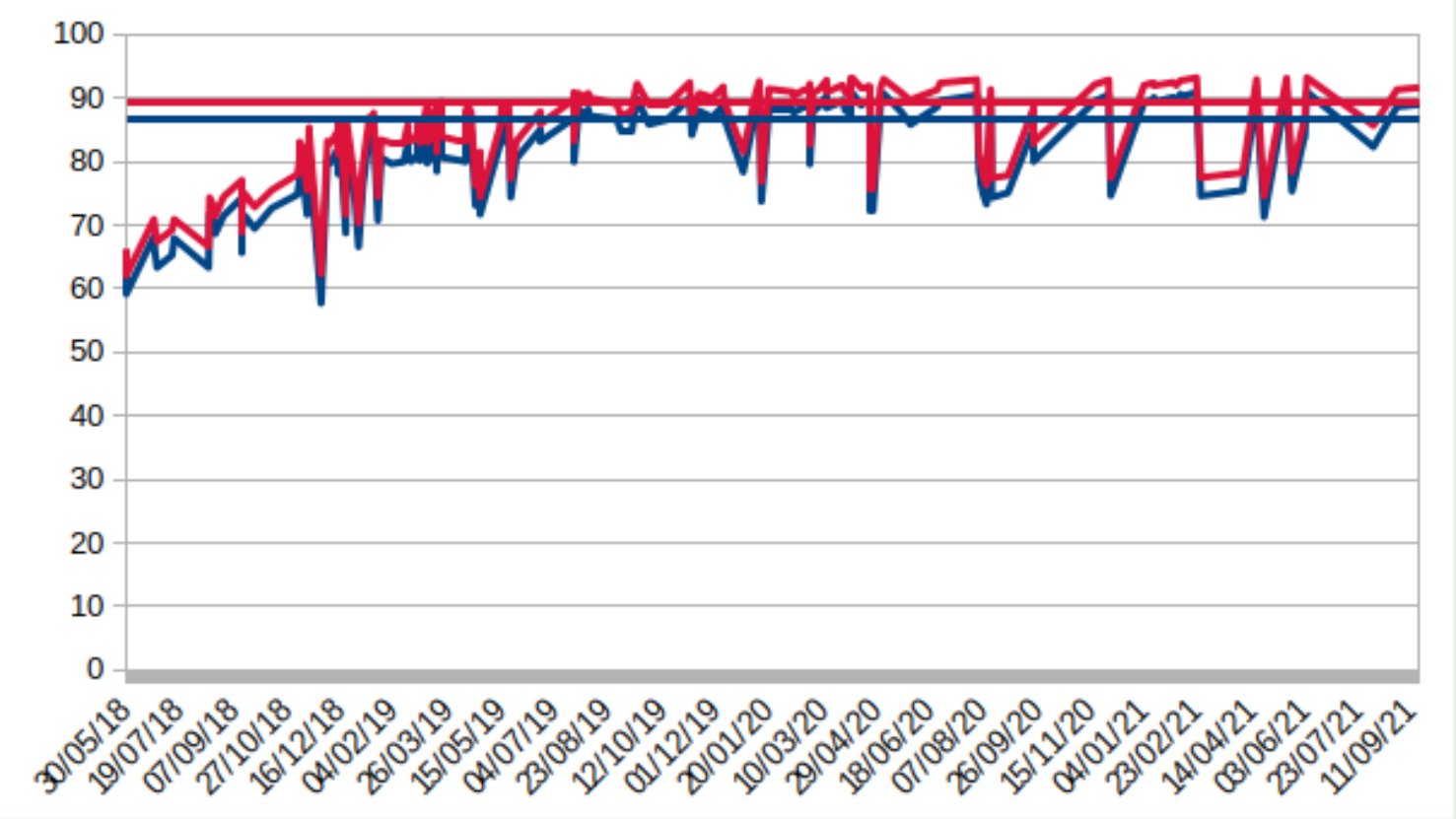
ex : Normans

The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

- In what country is Normandy located ? France
- What century did the Normans first gain their separate identity ? the first half of the 10th century, the 10th century, the 10th

Compétition SQuAD 2 / 2

Résultats en mars 2022



- En rouge : *Exact Match*
- En bleu : F_1 Mesure

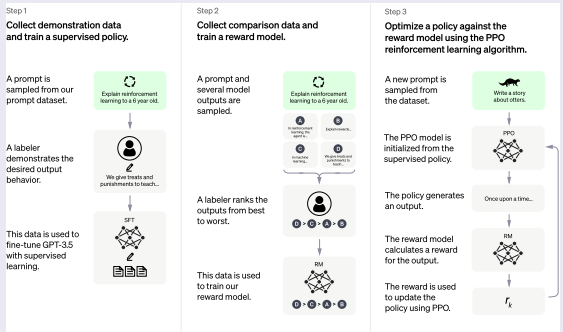
Et ChatGPT ?

GPT

- Un modèle de langage large (LLM)
- Capable de produire du texte contextualisé
- Version 3.5 :
  - 175 milliards de paramètres
  - Contexte de 2048 tokens

ChatGPT

- Une configuration et utilisation de GPT :



- Ce n'est pas un moteur de recherche d'information :
  - `https://www.youtube.com/watch?v=R2fjRbc9Sa0`
  - `https://www.youtube.com/watch?v=JcFRbecX6bk`

Conclusion sur la recherche d'information

Constats

- Aucune prise en compte de la sémantique (req, doc)
- Domaine en perpétuelle évolution
- À la croisée de différentes compétences : experts en mathématiques, linguistique, informatique, science des données, apprentissage artificiel (TAL)

Conclusion

- Recherche très active
- Utilisation de graphes conceptuels représentants les documents et requêtes
- Utilisation d'ontologies
- Vers un dialogue H/M pour la recherche d'informations : requête $\rightarrow$ proposition d'interprétation $\rightarrow$ affinage $\rightarrow$ acceptation $\rightarrow$ recherche $\rightarrow$ résultats

Références I

[Abi12] Serge Abiteboul.
Moteur de recherche de la toile.
http://www.college-de-france.fr/site/serge-abiteboul/course-2012-05-02-10h00.htm, mai 2012.

[AG13] Massih-Reza Amini and Éric Gaussier.
Recherche d'Information - applications, modèles et algorithmes.
Eyrolles, 2013.

[Gen00] David Genest.
Extension du modèle des graphes conceptuels pour la recherche d'informations.
PhD thesis, Montpellier 2, Grenoble, 2000.
Th. : informatique.

[JT01] Radwan Jalam and Olivier Teytaud.
Identification de la langue et catégorisation de textes basées sur les n-grammes.
In Henri Briand and Fabrice Guillet, editors, *EGC*, volume 1 of *Extraction des Connaissances et Apprentissage*, pages 227–238. Hermes Science Publications, 2001.

[Lar14] Hugo Larochelle.
Cours de traitements automatiques de la langue.
https://www.youtube.com/channel/UCiDouKcxRmAdc5OeZdiRwAg, 2014.

INSA ROUEN NORMANDIE

Références II

[Ouf08] Yannick Oufella.
Évolution du concept de front roc et combinaison de classifieur.
Master's thesis, Université de Rouen, 2008.

INSA ROUEN NORMANDIE