Recherche d'Information

Cours « Document et Web Sémantique »

Nicolas Delestre, Nicolas Malandain

INSA ROUEN NORMANDIE

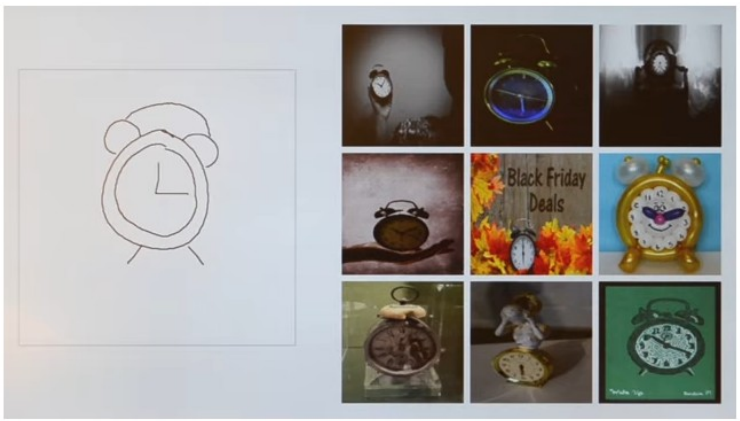
INSA ROUEN NORMANDIE

Contexte 1 / 2

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



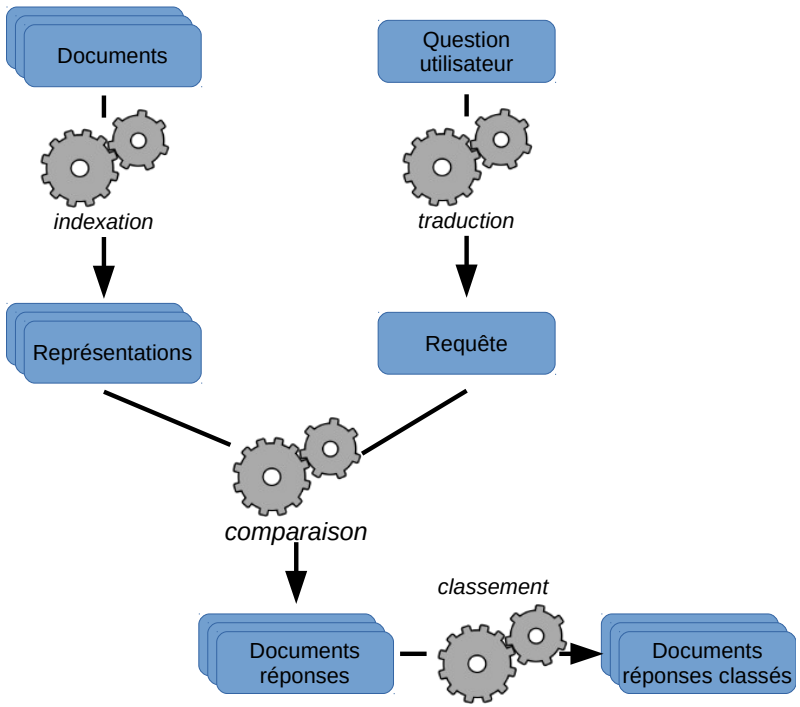
http://sketchy.eye.gatech.edu/

INSA ROUEN NORMANDIE

Contexte 2 / 2

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



INSA ROUEN NORMANDIE

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**

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

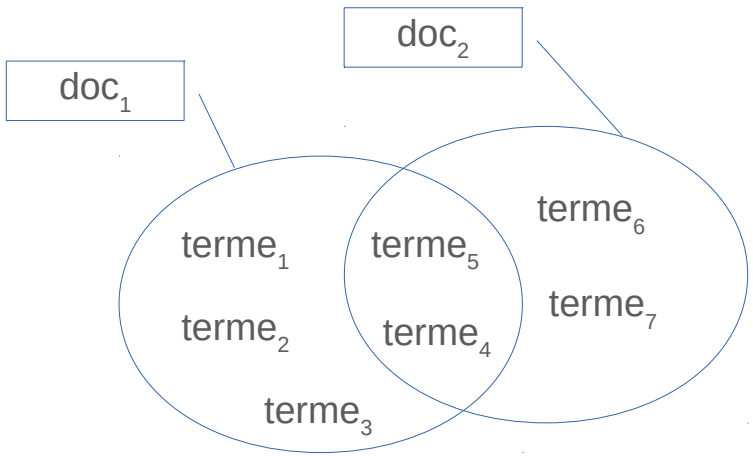
Deux types de représentation

1. Représentation vectorielle : chaque document est représenté par un vecteur mathématique (cf. cours sur le TAL)
2. Représentation ensembliste : chaque mot est représenté par un ensemble de documents

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

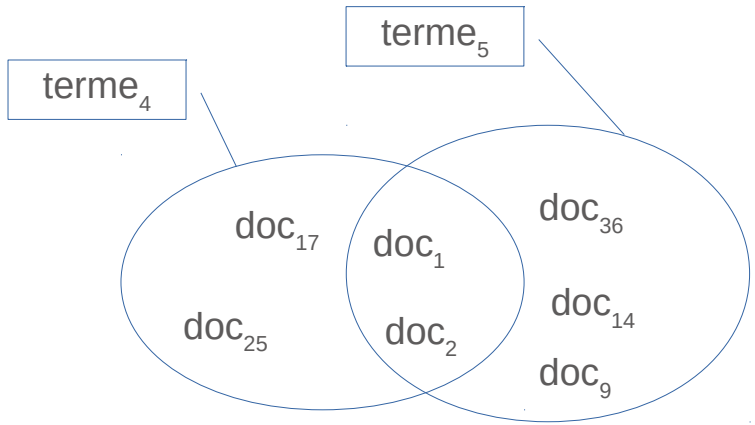


doc$_1$: terme$_1$ (pos$_1$, pos$_2$, ...)
terme$_2$ (...)
terme$_3$ (...)
terme$_4$ (...)
terme$_5$ (...)
doc$_2$: ...

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

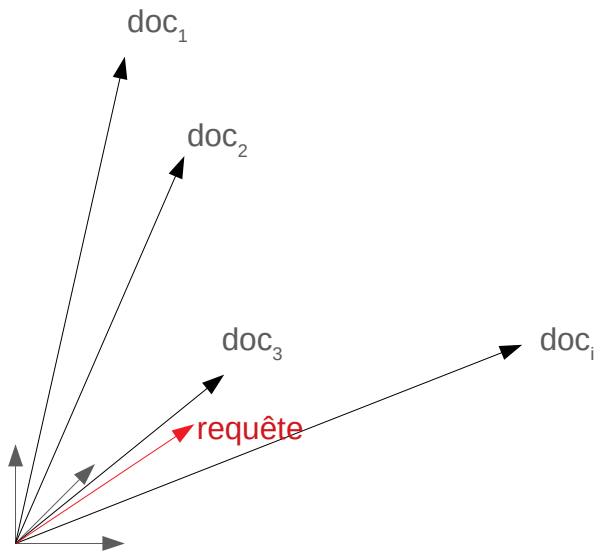
- Documents et requêtes représentés par un vecteur mathématique
- Comparaison et classement : suivant une similarité entre la requête et les documents, souvent similarité cosinus

Avantages

- Modèle homogène
- Classification des documents possibles

Inconvénients

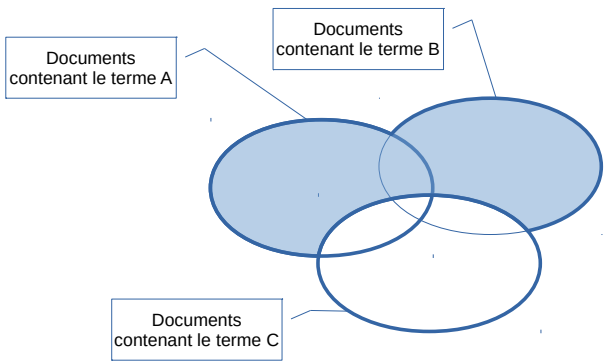
- Nombreux calculs
- Conjonction uniquement



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

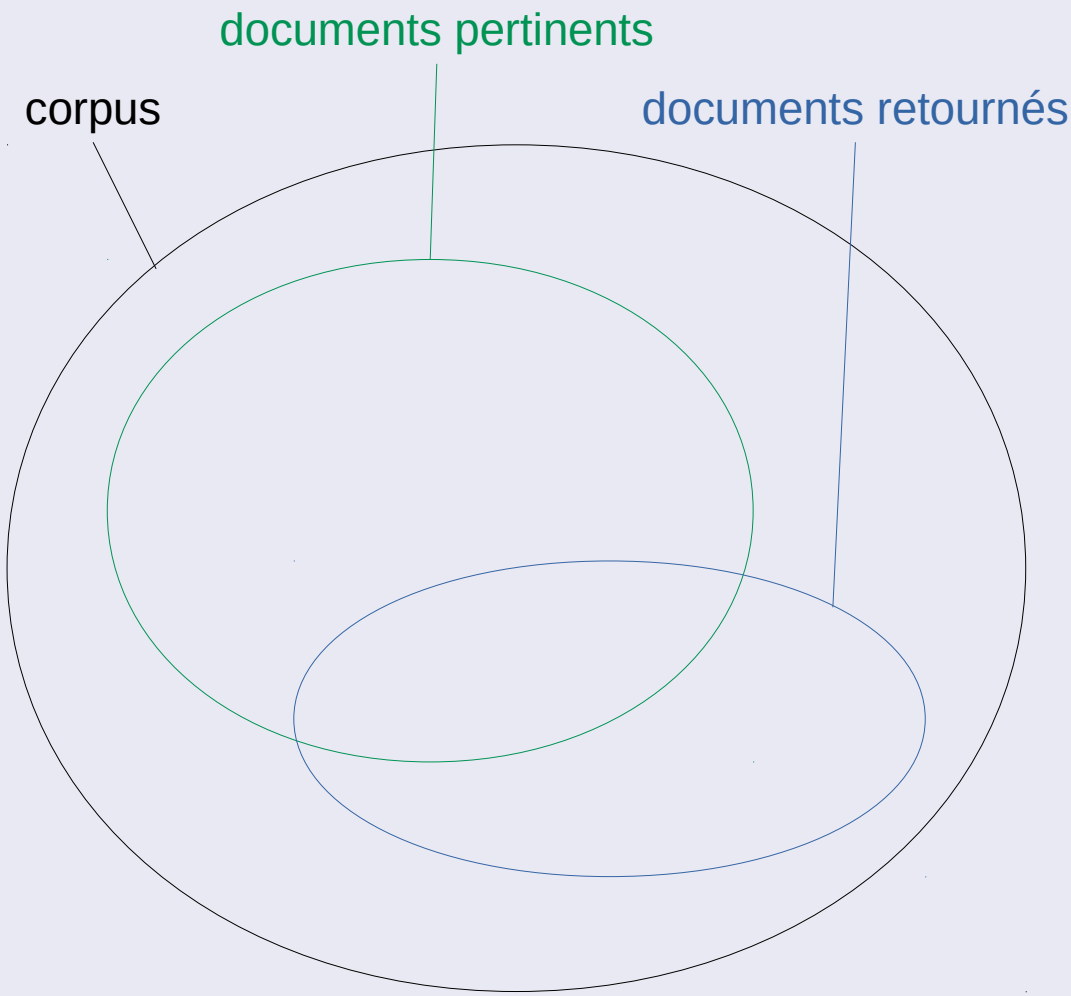
Un classement objectif ?

- 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

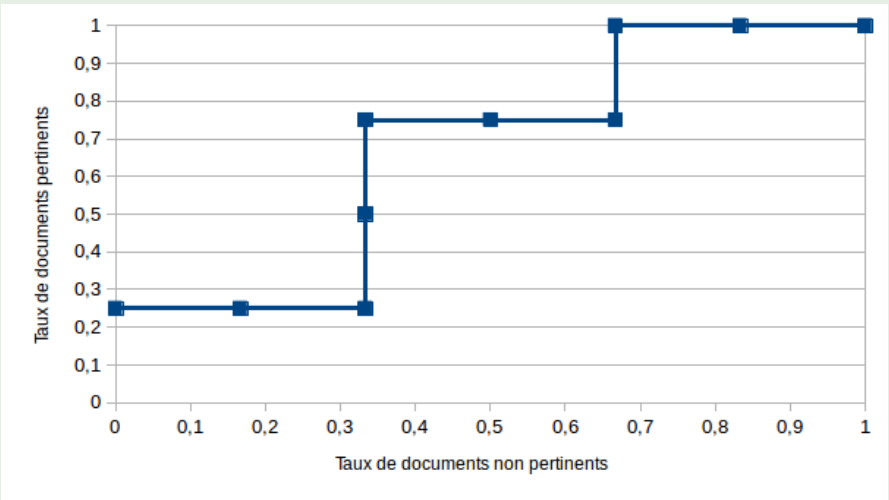
Évaluation de la recherche d'information 4 / 4

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.	Taux doc. pert.
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

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

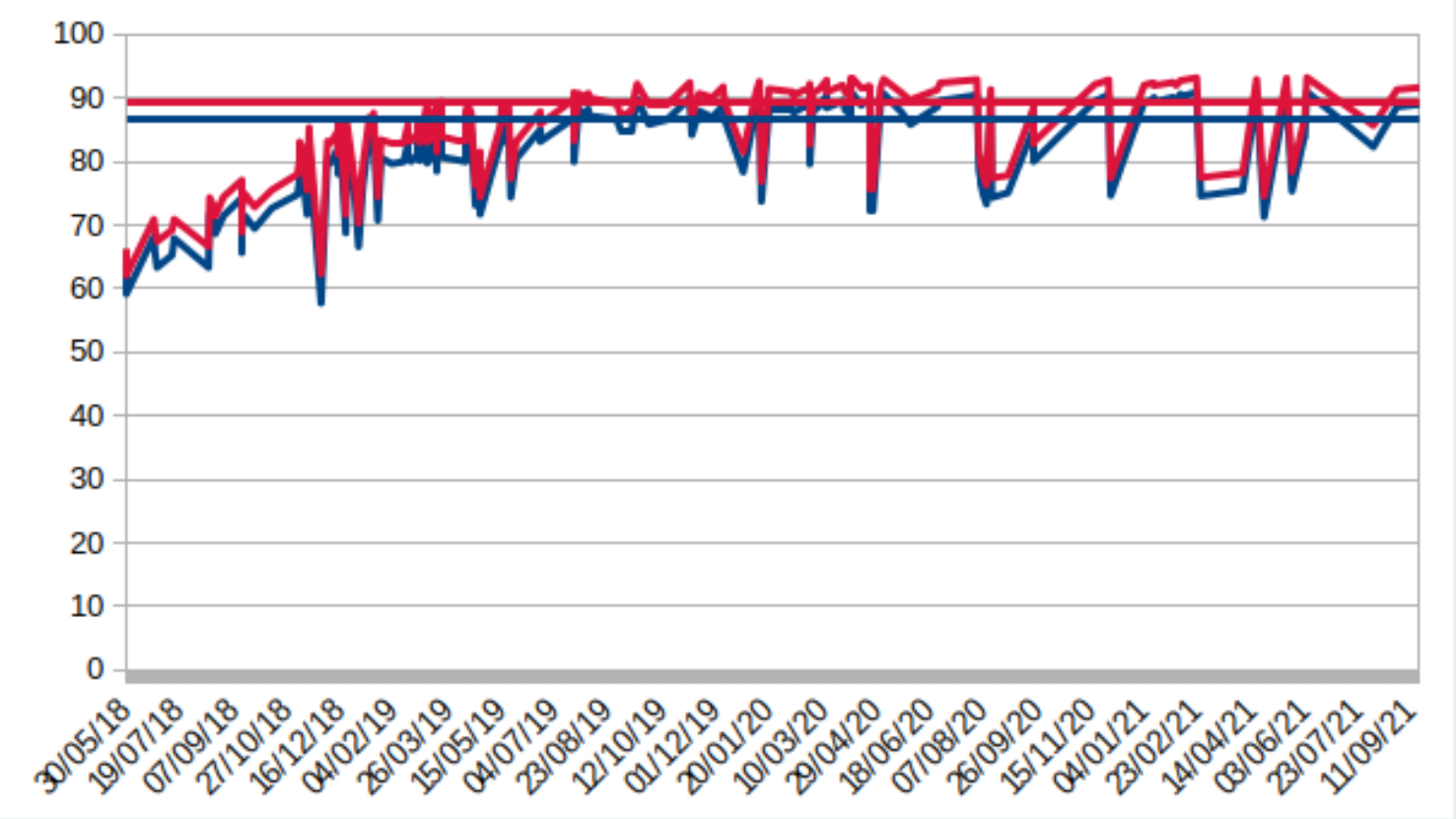
ex : Normans

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 les LLM ?

Précision ChatGTP = Chat + GPT

- GPT (*Generative Pre-trained Transformer*) est un LLM : un perroquet stochastique
- Chat : un programme qui controle ce que produit le LLM
  - Preprompt : contextualise les interactions avec l'utilisateur
  - Gère le contexte de la conversation, l'historique du dialogue
  - Filtre : vérifie que ce que produit le LLM ne pose pas de problème (sociologiquement, sécuritairement, etc.)

Les ChatBot ne sont pas (au départ) des moteurs de RI

- `https://www.youtube.com/watch?v=R2fjRbc9Sa0`
- `https://www.youtube.com/watch?v=JcFRbecX6bk`

INSA ROUEN NORMANDIE

Moteur de RI + LLM 1 / 2

Dernières versions des Chatbot

- Elles peuvent aller chercher les informations sur le Web, mais c'est souvent à l'utilisateur de le demander
- Agentification des ChatGPT : possibilité d'agir sur l'environnement et d'approfondir une recherche pour réaliser des synthèse, par exemple DeepSearch

`https://www.youtube.com/watch?v=c-uMcNa2UXg`

INSA ROUEN NORMANDIE

Moteur de RI + LLM 2 / 2

Retrieval-Augmented Generation

- Synthèse d'un moteur de recherche documentaire vectorielle et d'un LLM, par exemple NotebookLM

  `https://notebooklm.google.com/`

- Phase d'indexation : indexer les représentations vectorielles des phrases d'un corpus
- Phase de recherche :
  - récupérer les phrases les plus similaires à la question de l'utilisateur
  - créer un prompt composé de la concaténation des contextes de ces phrases (par exemple les paragraphes) et de la question initiale
  - fournir ce prompt au LLM pour produire la réponse

INSA ROUEN NORMANDIE

Conclusion sur la recherche d'information

Constats

- Aucune prise en compte de la sémantique (req, doc) dans les moteurs de RI classique
- 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 d'ontologies (Google Knowledge Graph)
- 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