REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE: Mathématique et Informatique
DEPARTEMENT: Informatique
N°:……..

DOMAINE: Mathématique et Informatique
FILIERE: Informatique
OPTION: T.I.C

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par : KHODJA Ala Eddine

Intitulé

Un système d'extraction d' information pour la langue arabe

Soutenu devant le jury composé de :

A.ATIR	Université de M'sila	Président
B.BELKACEM	Université de M'sila	Rapporteur
N.OULD MEHAMEDI	Université de M'sila	Examinateur

Année universitaire : 2016 /2017

Remerciements

Avant tout je désirais remercier mon créateur (DIEU) pour m'avoir donné de la force à accomplir ce travail.

Je voudrais adresser toutes mes gratitudes à directeurs de ce mémoire, M. ***BRAHIMI Belkacem*** pour son patience, *leur disponibilité et surtout leurs judicieux conseils, qui ont contribué à alimenter mes réflexions.*

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance à tout le personnel de département math et informatique, pour cette expérience enrichissante et pleine d'intérêt durant ce projet.

Je voudrai également remercier et exprimer mon profond respect à tous les membres du jury qui ont bien voulu me faire l'honneur d'assister à ma soutenance du Master afin de juger la qualité du travail réalisé.

Je tiens à remercier tout particulièrement M.MOKHETARI Rabah pour son aide et son soutien durant c'est trois mois.

Mes remerciements vont enfin à toute personne qui a contribué de près ou de loin à l'élaboration de ce travail.

Dédicace :

Je dédie ce modeste travail :

A mes très chers parents qui m'ont couvert d'amour et de soutient

A mes frère et *mes sœurs j'espère que la vie réserve le meilleur pour eux*

A mes oncles et mes tantes pour toute l'affection qu'ils m'ont donné et pour leurs précieux encouragements.

A mes cousins et mes cousines

A tous mes chers amis surtout : Salah Eddine, Rahim et Akram

ALA Eddine

Table de matière :

Liste des figures

Liste des tables :

Introduction générale :

La disponibilité croissante de la quantité énorme de l'information et la cause de l'inflation du volume des données, lorsqu'on parle des données massives nos sources du volume qui arrive à des centaines de téraoctets, ou bêta octets.

Le domaine de la fouille de Textes (Fidelia, 2007) réunit et intègre dans ses applications des méthodes d'extraction d'information, de recherche d'information, de questions-réponses, de résumé automatique, de catégorisation de textes, de classification et de routage des documents textuels ainsi que les recours à des techniques de fouille de données. Dans ce mémoire, nous intéressons à une seule tâche : l'extraction d'information (EI).

La fouille de donnée (DM) et la fouille de donnée textuelle (TM) sont des technologies modernes qui sont utilisées dans le système d'information, le Text mining (TM) à savoir de l'extraction de l'information utile à partir de gros volumes de contenus textes.

Les différentes recherches précédentes de l'extraction d'information a partir des textes conçus beaucoup plus sur des textes français et des textes anglaise, les travaux de recherches relatives à l'extraction d'information a partir des textes arabes sont moins nombreux que les autres langues.

Parmi les tâches les plus importantes du la fouille de textes (text mining), nous pouvons citer : la recherche d'information, la classification des textes et l'extraction de l'information textuelle dans les textes non structurés (article, blog, message). L'objectif de l'extraction d'information textuelle est de trouver une information précise (nom, date, nombre, relation…) dans un gros volume de données textuelles, alimenter une base de données, générer un résumé de texte,etc.

L'objectif de notre travail est d'étudier la tâche d'extraction d'information pour la langue arabe, pour cela, nous avons proposé un système d'extraction d'Information à partir des articles de presse arabes dans le domaine des accidents de la circulation en Algérie.

Ce travail contient trois chapitres :

Le premier chapitre intitulé « Text mining » dans ce chapitre, dans lequel nous avons parlé principalement sur les notions de la fouille de donnée textuelle (TM).

Le deuxième chapitre intitulé « Extraction d'information » ce chapitre traite le problème de l'extraction d'information et plus spécifiquement la reconnaissance des entités

nommées. Au cours de ce chapitre, nous examinons les origines, l'évolution de ce concept ainsi que ses objectifs et ses différents besoins.

Le troisième chapitre intitulé « Méthodologie résultats et analyse » concerne la partie pratique dans lequel nous avons appliqué notre approche pour l'extraction des propriétés des accidents (le corpus collecté, approche utilisée, résultats, évaluations et discussion des résultats obtenus).

Nous terminons notre travail par une conclusion générale, les perspectives pour améliorer ce travail sont également fournies.

Chapitre 01
Text Mining

1 Introduction :

La fouille de texte (souvent appelée « text mining ») est l'exploration et l'analyse de grandes quantités de données afin d'y découvrir de l'information implicite. Cette information peut être de différente nature, par exemple on recherchera des règles d'association, une classification ou une segmentation de population.

Dans ce chapitre on va résumer en brève la définition de la fouille de texte, leur processus on expliquant aussi les différant taches de la fouille de texte.

2 Fouille de texte :

2.1 Définition du Fouille de texte:

La fouille de textes (text mining) est l'héritière directe de la fouille de données (data mining), née dans les années 90. À cette époque, les ordinateurs personnels se généralisent, leur capacité de calcul et de mémorisation atteignent des seuils tels qu'ils commencent à pouvoir traiter de grandes quantités d'informations. La fouille de données vise à tirer le meilleur profit possible de cette situation inédite (hors contexte militaire !) pour créer des programmes capables de prendre des décisions pertinentes. Elle naît dans différents environnements qui ont l'habitude de gérer des bases de données conséquentes. C'est notamment le cas des banques et des assurances (pour décider de l'attribution d'un crédit, par exemple), de la médecine (pour effectuer un diagnostic ou évaluer l'efficacité d'un médicament) ou encore de la vente et du marketing (pour cibler les publicités aux clients) : autant de domaines ou l'efficacité est directement monétisable.

C'est le cas du TAL (traitement automatique des langues), qui vise à écrire des programmes capables de comprendre les langues "naturelles", celles que les humains utilisent entre eux. Les outils traditionnels du TAL proviennent de l'informatique théorique et de l'IA classique: automates, grammaires formelles, représentations logiques...

Ils sont malheureusement coûteux en développement et en temps de calcul, et de ce fait peu adaptés au traitement de grandes quantités de textes "tout venant", c'est-à-dire ne respectant pas nécessairement les règles de bonnes constructions syntaxiques. [1]

2.2 Les tâches de la fouille de textes :

Il est temps désormais de passer en revue les tâches de fouille de textes que nous qualifions d''‘élémentaires", parce qu'elles servent de ‘‘briques de base" aux autres tâches plus complexes. Nous avons identifié :

1- La recherche d'information.
2- La classification.
3- L'annotation.
4- L'extraction d'information.

Nous les présenterons dans cet ordre, de la moins spécifique au plus spécifique d'un point de vue linguistique. Pour chacune d'entre elles, nous explicitons ici leur nature et leur intérêt applicatif, les données sur lesquelles elles peuvent s'appliquer, les types de ressources qu'elles requièrent ou qui peuvent aider à les effectuer ainsi que les mesures utilisées pour évaluer les programmes qui s'y confrontent.

On expliquant ces différentes tâches :

2.2.1 La Recherche d'Information :

Le but de cette tâche est de retrouver un ou plusieurs document(s) pertinent(s) dans un corpus, à l'aide d'une requête plus ou moins informelle.

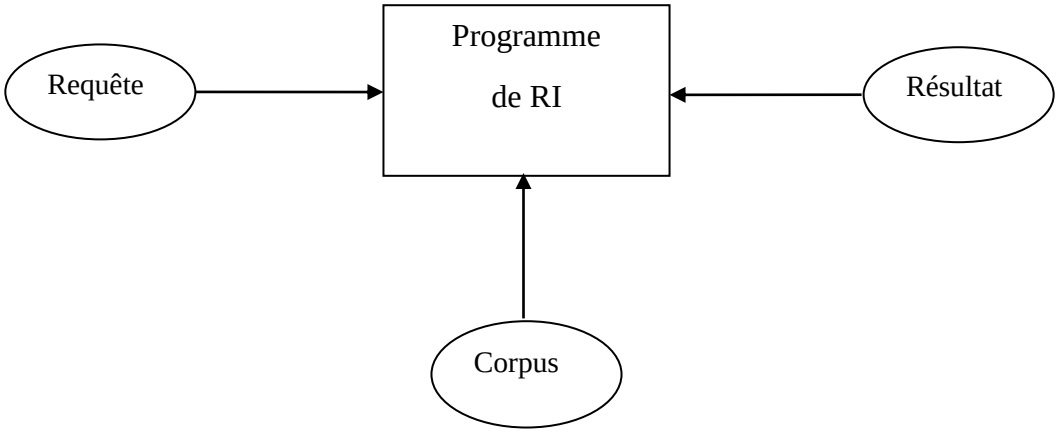

Figure 1.1 - Schéma général de la tâche de Recherche d'Information

2.2.2 La Classification :

La classification est la tâche phare de la fouille de données, pour laquelle une multitude de programmes sont implémentés dans le logiciel Weka. Elle consiste à associer une "classe" à chaque donnée d'entrée. Elle est présente dans la plupart des gestionnaires de courriers électroniques. [1]

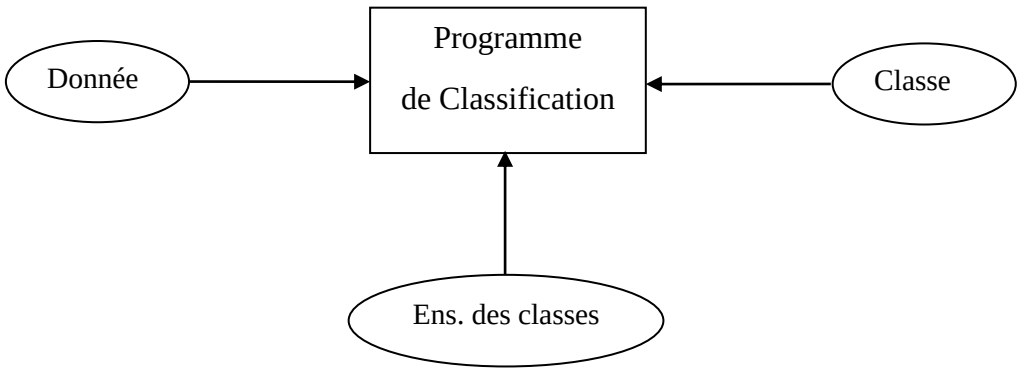

Figure 1.2 - Schéma général de la tâche de Classification

2.2.3 L'Annotation :

L'annotation (ou l'étiquetage), telle qu'elle sera définie ici, est une tâche plus spécifiquement linguistique que les précédentes, au sens où elle ne s'applique pas aux données tabulaires et ne relève donc pas de la fouille de données. La figure 1.3 la présente globalement. [1]

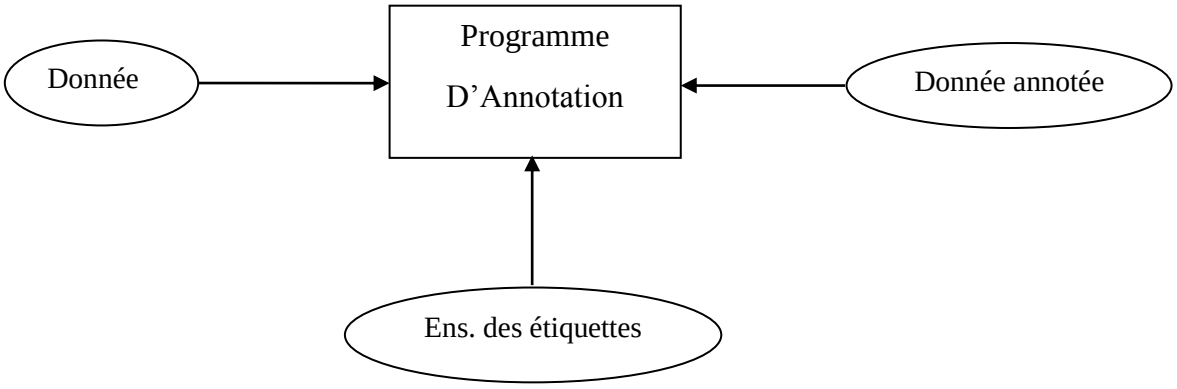

Figure 1.3 - Schéma général de la tâche d'Annotation

2.2.4 L'Extraction d'Information (EI) :

L'Extraction d'Information (EI ou Information Extraction en anglais, abrégé en IE) est décrite par le schéma de la figure 1.4 Le but de cette tache, qui relève de l'ingénierie linguistique, est d'extraire automatiquement de documents textuels des informations factuelles servant à remplir les champs d'un formulaire prédéfini. [1]

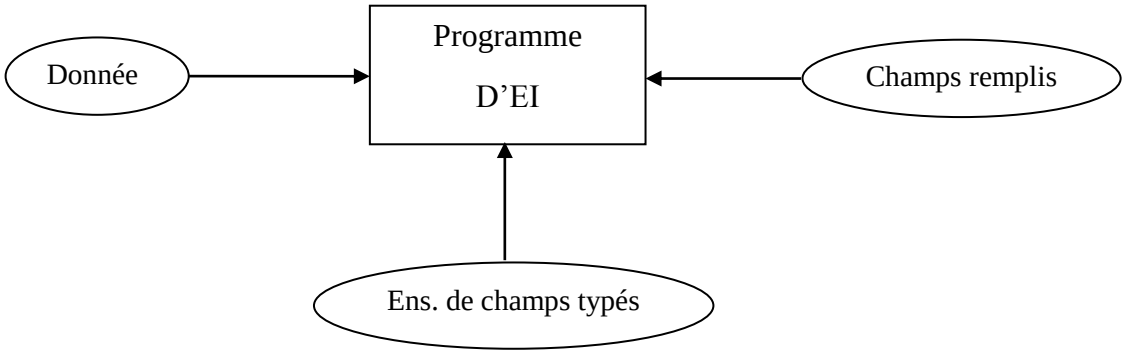

Figure 1.4 - Schéma général de la tâche d'Extraction d'Information

3 Relations entre tâches :

Il est important de distinguer les tâches les unes des autres, parce que les programmes qui seront décrits dans les chapitres suivants sont spécialisés dans la réalisation d'une et une seule d'entre elles. Pour autant, les quatre tâches élémentaires que nous venons de présenter ne sont pas complètement indépendantes les unes des autres. Tout d'abord, il est souvent possible, via une reformulation du problème ou un « codage » astucieux des données, d'en transformer une en une autre, et de permettre par la même occasion d'employer un programme prévu pour accomplir une certaine tâche dans un autre but. C'est ce que nous évoquons dans la première section de cette partie. Ensuite, nous montrons que pour réaliser des traitements moins « élémentaires » que ceux cités jusqu' à présent, il suffit de les décomposer en sous-problèmes correspondant à nos quatre tâches de référence, et d'utiliser des programmes les résolvant les uns après les autres. Jouer avec les entrées/sorties d'une tâche, les reformuler et les enchaîner, font partie des compétences indispensables aux usagers de la fouille de textes. [1]

4 Fouille de données textuelle :

Le Fouille de donnée textuelle, (en anglais appelé Text Mining) est une technique permettant d'automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière statistique, les différents sujets évoqués ainsi découvrir des connaissances et des relations à partir des documents disponibles.

L’outil de Text Mining va générer de l’information sur le contenu du document. Cette information n’était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document. [9]

4.1 Objectifs de la fouille de données textuelles :

La fouille de données textuelle peut être utilisée en particulier dans les cas suivants :

- Pour mieux comprendre le positionnement d'un discours, d'une thèse, d'un communiqué.
- Pour appréhender les thèmes récurrents qui sont associés à une activité, une entreprise ou des concurrents.
- Pour mesurer les points faibles et les points forts dans une revue de presse.
- Pour comparer des textes sur un même thème afin d'en déterminer les points communs ou au contraire de distinguer les différences stylistiques.
- Pour créer automatiquement des répertoires de sites Web ou emails associés à des thématiques. Pour quantifier un texte ou les parties d'un texte pour en extraire les structures signifiantes les plus fortes telles que le résumé automatique et la segmentation thématique.
- Pour établir des liens entre les termes et les documents utilisés dans l’indexation.
- Pour établir des règles de classification automatique de documents (classification supervisée ou non supervisée).

5 Conclusion :

Ce chapitre présent la fouille des textes TM, les différentes tâches et leurs relations, la fouille de données textuelle et ses objectifs.

Chapitre 02
Extraction d’information

Présentation :

L'extraction d'information (en anglais, Information Extraction ou IE) est un sujet de recherche important dans le domaine du Traitement Automatique des Langues naturelles (TAL). Elle connaît ces dernières années un intérêt grandissant, car elle répond à un besoin devenu incontournable dans la société de l'information .

1 Définition et Objectifs :

L'extraction d'information est une technologie récente, mais qui cherche à répondre à un besoin très ancien : acquérir de la connaissance à partir de textes. Cette nécessité s'est accrue ces vingt dernières années avec l'essor considérable de la masse de documents disponibles au format électronique (Internet, courrier et documentation électroniques) qu'il faut gérer afin d'extraire ou de filtrer les informations utiles et pertinentes parmi toutes celles contenues dans ces documents. Comme la recherche d'information, le résumé automatique ou les systèmes de questions-réponses (QA), l'extraction d'information a l'ambition de répondre à ce défi, d'où le développement de nombreuses applications destinées à des institutions, au monde des affaires et/ou de l'industrie.[2]

Ces informations sont destinées à créer ou alimenter un entrepôt de données (appelé aussi banque de données). La tâche d'extraction est réalisée grâce au remplissage des formulaires prédéfinis (template).

Ces formulaires, dits formulaires d'extraction, sont définis dans le but de représenter la connaissance à rechercher par une structure déterminée a priori. Ils décrivent un ensemble d'entités, les relations entre celles-ci et les événements impliquant ces entités . Par exemple, un formulaire concernant des accidents de la route devra spécifier des champs comme « Lieu de l'accident », « Nombre de victimes », « Identité des victimes » ou encore « Cause de l'accident ».

Les informations extraites par un système d'Extraction d'Information peuvent être consultées par des utilisateurs humains (par exemple via la génération de rapports d'événements), alimenter une base de données afin d'être analysées plus tard (interrogation par requêtes ou fouille de données).

Un exemple d'extraction les informations depuis un accident de circulation à partir d'articles de journaux est présentées dans l'exemple 2.1

Texte :

لقي 11 شخصا بورقلة بينهم 03 نساء وأستاذة جامعية وطفلة لم يتجاوز عمرها سنتين، حتفهم في حادث مرور مروع، اقشعرت له الأبدان، بينما أصيب 28 راكبا بينهم عسكري بجروح متفاوتة الخطورة لا يزالون يتقلون العلاج بمستشفى محمد بوضياف وسط المدينة، حالة 06 منهم خطيرة، واثنان في قاعة الإنعاش.

الحادث المأسوي وقع الأحد في الساعة السادسة صباحا بالطريق الوطني رقم 56 الرابط بين عاصمة الولاية ومدينة تقرت بالنقطة الكيلومترية رقم 60، نجم عن احتكاك حافلة لنقل المسافرين من نوع "هيجر" تابعة لمؤسسة خاصة، كانت قادمة من العاصمة يقودها شيخ، بشاحنة ذات مقطورة يقودها شاب، مما استدعى تدخل مصالح الحماية المدنية معززة بـ08 سيارات إسعاف و04 شاحنات إطفاء وتسخير 40 عونا.

Exemple 2.1 - un article de journal

Alimentation d'une base de données :

Table Accident	
Type	**Information de l'accident**
Nombre de victimes	11
Nombre de blessés	28
Lieu	ورقلة
Date	الأحد في الساعة السادسة صباحا
Cause	عن احتكاك حافلة لنقل المسافرين من نوع "هيجر" تابعة لمؤسسة خاصة، كانت قادمة من العاصمة يقودها شيخ، بشاحنة ذات مقطورة يقودها شاب

Tableau 2.1 – résultat de l'extraction

2 Les conférences MUC :

Les recherches actuelles en EI ont été influencées par les conférences MUC. Ces conférences qui se sont déroulées entre 1987 et 1998 faisaient partie du programme TIPSTER financé par DARPA. Ce programme comportait trois tâches: la détection des documents, l'extraction d'information, et le résumé de textes. Les campagnes d'évaluation MUC ont été organisées afin de confronter les systèmes d'extraction d'information réalisés par différentes équipes en comparant leurs performances avec des mesures précises et objectives. Ces mesures, inspirées de celles définies pour le domaine de la recherche d'information, sont devenues un standard pour toute évaluation des résultats de l'EI. Ainsi, la précision mesure la qualité du système, c'est-à-dire le nombre d'informations extraites correctement par rapport au nombre d'informations extraites. Le rappel lui mesure la couverture du système, c'est-à-dire le nombre d'informations correctement extraites par rapport au nombre d'informations correctes présentes dans le corpus. Enfin, le F-mesure permet de disposer d'une évaluation globale du système en combinant précision et rappel. [3]

L'apport des conférences MUC a été considérable ; aussi bien en termes d'identification des problèmes à prendre en compte (linguistique, représentation des connaissances, acquisition de ressources, travail sur corpus…) qu'en termes de méthodes et de techniques pour les résoudre. [3]

Les textes servant de support à l'évaluation provenaient de différents domaines. Les premières conférences ont porté sur l'extraction d'information à partir des messages militaires, par contre ce thème a été développé dans les conférences ultérieures pour couvrir les rapports de presse. Divers systèmes d'extraction d'information ont été testés sur différents types de textes : récits d'attentats (MUC-3 et MUC-4), annonces de produits (MUC-5), annonces financières concernant les prises de participation des entreprises (MUC-6), etc. Les systèmes en compétition devaient remplir un ou plusieurs formulaires fixés à l'avance en fonction du domaine. Par exemple, pour les annonces financières, ils devaient extraire les différentes sociétés (acheteurs, vendeurs, achetés), la date, le lieu et le montant de la transaction financière, etc. [3]

Le Tableau 2.2 résume les différents contenus de textes traités dans chaque conférence.

(MUC 1, 1987) et (MUC-2, 1989)	ont traité et analysé les rapports d'opérations tactiques navales.
(MUC3, 1991) et (MUC4, 1992)	l'objectif était d'analyser des textes journalistiques traitant du terrorisme en Amérique Latine, afin d'extraire, des dépêches d'agence de presse, le maximum d'information sur des actes terroristes.
(MUC5, 1995)	ont traité un corpus de nature économique pour extraire des informations de type fusion, rachat, et création d'entreprises internationales et la fabrication de circuits électroniques.
(MUC 6,1996)	une suite de MUC 5, a traité les changements de dirigeants à la tête des entreprises.
(MUC7, 1998)	s'est intéressée à l'analyse de textes journalistiques rapportant des crashs d'avion et de tirs de missiles.

Tableau 2.2 -Les conférences MUC [3]

3 Les tâches de l'extraction d'information :

Les composants trouvés dans les systèmes d'EI d'aujourd'hui reflètent largement les tâches définies dans ces conférences. Les tâches de la dernière conférence, MUC-7, en 1998 (les plus difficiles dans la série) ont été les suivantes :

- reconnaissance des entités nommées,
- détection de la coréférence,
- reconnaissance des éléments du formulaire,
- reconnaissance des relations,
- reconnaissance des scénarios («scénario template»). [3]

Nous limitons à la première tâche de l'EI : la reconnaissance des entités nommées. Dans la section 3.1.

3.1 Reconnaissance des entités nommées :

Cette tâche consiste à repérer toutes les formes linguistiques bien identifiées, à l'instar des noms propres de personnes, d'organisations, de lieux, etc., mais aussi les expressions temporelles (dates, durées,…), les quantités (monétaires, unités de mesure, pourcentages…) et à leur affecter une étiquette sémantique choisie dans une liste prédéfinie. [3]

3.1.1 Mesures d'évaluation :

Le rappel (R) : est une évaluation de la couverture du système. Il mesure la quantité de réponses pertinentes d'un système par rapport au nombre de réponses idéales. [3]

$$\mathbf{R} = \frac{\mathbf{Nombred'entités\ correctes\ détectées}}{\mathbf{Nombred'entités\ \ manuellement\ identifiées}} \mathbf{(1)}$$

La précision (P) : est une évaluation du bruit du système. Elle mesure la proportion des réponses correctes parmi l'ensemble des réponses fournies par le système. [3]

$$\boldsymbol{P} = \frac{\mathbf{Nombred'entités\ correctes\ détectées}}{\mathbf{Nombred'entités\ détectées}} \mathbf{(2)}$$

Le F-mesure (F) : C'est la moyenne harmonique de la précision et du rappel qui mesure la capacité du système. À donner toutes les solutions pertinentes et à refuser les autres, une mesure populaire qui combine la précision et le rappel est leur pondération. [16]

$$\mathbf{F} = \frac{\mathbf{2\,(P\,.\,R)}}{\mathbf{(P+R)}} \mathbf{(3)}$$

4 Les différentes approches :

Les méthodes d'extraction peuvent être classées en trois catégories : les méthodes linguistiques, les méthodes statistiques et les méthodes mixtes.

4.1 Les méthodes linguistiques :

Ces méthodes sont essentiellement syntaxiques et s'appuient sur une analyse syntaxique des textes, des travaux axés sur une approche sémantique et l'acquisition de terminologie .Nous présentons ici quelques systèmes utilisant les méthodes linguistiques. [4]

4.2 Les méthodes statistiques :

Elles permettent d'extraire des adresses termes sans analyse linguistique préalable. Les méthodes statistiques sont devenues très présentes dans le traitement du langage naturel. La plupart d'entre elles se résument en calculs de valeurs numériques tels que les fréquences. [4]

4.3 Les méthodes mixtes :

La tendance actuelle consiste à combiner des approches linguistiques avec des approches statistiques. Généralement la partie essentielle de la méthode d'extraction est statistique, la partie « linguistique » consistant à filtrer les termes en fonction de leur catégorie syntaxique. [4]

5 Contexte :

5.1 Un besoin ancien et essentiel :

L'Extraction d'Information est désormais un sujet de recherche important dans le domaine du Traitement Automatique des Langues naturelles. Elle connaît ces dernières années un intérêt grandissant, car elle répond à un besoin devenu incontournable dans la société de l'information. Il faut souligner que la collecte d'informations dans des textes est une activité qui remonte à l'Antiquité. Depuis que l'écriture existe, l'humanité s'est penchée sur les textes pour y trouver des réponses à ses questions, a étudié les écrits pour acquérir des connaissances. Cette quête de savoir a connu ces dernières décennies un essor considérable avec le passage à la civilisation de l'information dont une des principales conséquences est la production en masse de documents textuels sous format électronique. [4]

A) Enjeux

Dans la plupart des domaines, qu'il s'agisse de l'économie, de la société ou de la sécurité, obtenir et traiter régulièrement des informations est devenu une nécessité, notamment afin de s'appuyer sur des bases solides lors des prises de décision. [4]

Dans le domaine de l'économie, la collecte d'information est un enjeu essentiel. Les entreprises ont en permanence besoin d'informations fiables et pertinentes sur les marchés ainsi que sur leurs concurrents afin d'élaborer les stratégies leur permettant d'améliorer leurs résultats et de gagner des parts de marché2. Pour répondre à ce besoin, les acteurs économiques se tournent vers les documents issus de la presse, et principalement de la presse économique. Ce processus concerne particulièrement le monde du finance dans lequel il est

nécessaire de connaître au jour le jour les fluctuations au sein des différents secteurs de l'économie. Les prises de décision s'appuient sur des événements particuliers extraits de l'étude de très importantes quantités de textes. Par exemple, en Grande-Bretagne, la banque Lloyds emploie des centaines de personnes pour chercher quotidiennement dans des journaux du monde entier les naufrages de bateaux à travers le globe dans le cadre de son activité d'assureur. [4]

Acquérir de l'information est également un enjeu au niveau sociologique, notamment pour les acteurs politiques. L'analyse de documents traitant d'une société amène à discerner et comprendre les comportements des différentes composantes d'une population, les multiples problèmes de la société et les opinions publiques. Ce type d'analyse permet de définir et de proposer des politiques répondant à ces problèmes3 ou de trouver les moyens de faire comprendre et accepter des mesures à une population. [4]

Dans les secteurs de la défense ou de la sécurité, la collecte d'information a toujours été au cœur des services de renseignements militaires ou policiers. Elle est essentielle dans la lutte contre le terrorisme afin de déceler les prémisses d'actions terroristes. Au niveau militaro-politique, elle est utile en temps de paix pour découvrir les germes des futurs conflits, et en temps de guerre pour déceler certains faits et gestes ennemis afin de prévoir les stratégies militaires à mettre en place. Pour remplir ces objectifs, les services de renseignement se focalisent d'une part sur l'étude de documents traitant de sujets policiers ou militaires (dans la presse par exemple) et d'autre part sur l'analyse de textes relatant des correspondances (courriers papiers ou électroniques, transcriptions d'écoutes téléphoniques) ou de conversations (issues par exemple de l'espionnage d'individus au moyen de microphones).[4]

B) Évolution de la tâche d'extraction

Les textes en langue naturelle véhiculent une grande quantité d'informations. Pour pouvoir analyser et manipuler automatiquement ces informations, chacune d'elles doit être représentée dans une forme structurée qui rend accessible l'ensemble des éléments la constituant.[4]

Jusqu'à récemment, les méthodes utilisées dans la collecte d'information à partir de textes consistaient à confier à un être humain l'étude d'un ensemble de documents afin de recueillir et de structurer les données contenant des informations pertinentes en regard du but fixé .[4]

Une telle tâche est un travail long, coûteux et fastidieux qui s'avère rapidement titanesque tant la quantité de textes à traiter se révèle colossale. La quantité d'information augmente très régulièrement et met en échec la capacité humaine à lire, comprendre et synthétiser une telle masse de documents.[4]

L'évaluation de la tâche d'extraction est également difficile car l'appréciation de la qualité et de la pertinence des informations extraites connaît les mêmes soucis de temps, de coût et de subjectivité que l'exécution de la tâche elle-même.[4]

L'accroissement du nombre de documents électroniques et des capacités de traitement électronique de l'information (augmentation de la taille des mémoires et de la vitesse des systèmes) ont imposé le principe d'automatisation de la tâche d'extraction et ont fait émerger les recherches en Extraction d'Information.[4]

L'Extraction d'Information dispose du potentiel nécessaire pour extraire des informations avec nettement plus de rapidité que la collecte réalisée par des humains. Les travaux de C. A. Will ont montré que la réalisation de la tâche d'extraction par des processus automatiques produit des résultats dont la qualité, mesurée en termes de précision et de taux d'erreurs, est comparable et même parfois supérieure à celle des résultats de travaux menés par des humains, même s'il s'agit d'analystes entraînés spécifiquement.[4]Le domaine de l'Extraction d'Information intègre de plus un grand nombre de sous-problèmes nontriviaux d'analyse de la langue comme la recherche de termes ou l'identification de relations sémantiques ou syntaxiques entre entités. [4]

6Langue arabe :

La langue arabe est la langue des populations arabes qui firent leur entrée dans l'histoire depuis 3 millénaires environ et qui occupaient les zones septentrionales de l'Arabie.

La langue arabe est considérée comme la 5éme langue courante utilisée dans le monde. Elle est parlée par plus de 422 millions de personnes en tant que première langue et de 250 millions en tant que langue secondaire, la langue arabe fait partie de la grande famille des langues sémitiques.[10]

Le système archaïque d'écriture arabe était consonantique. Chaque lettre de l'alphabet arabe représente une consonne unique depuis les temps anciens. Cependant, la fin du VIIe siècle,

Les diacritiques arabes qui sont des symboles graphiques qui discriminent entre la variété des prononciations des consonnes ont été inventés par "Abou Al-Aswad Al-Du'ali".

Néanmoins, ils sont très souvent éliminés du texte écrit d'aujourd'hui. Lecteurs arabes pouvaient discerner les mots avec la même forme d'écriture par l'intermédiaire de son contexte.

Diviser le texte d'entrée en fractions désirées est généralement la phase initiale dans la plupart des tâches de traitement de texte.

Ces fractions pourraient être des phrases, des chiffres, des mots, des caractères ou toute autre fraction utile. Chaque fraction est appelée un **« Token »** et le processus est appelé **« Tokenization ».**

En arabe **Token** peut spécifier toute une phrase grammaticale par exemple« و سنساعدهم » (ce qui signifie : "et nous allons les aider»).

L'un des éléments les plus efficaces dans les phrases distinctives ou limites symboliques est des signes de ponctuation.

Ils ont émergé dans le système d'écriture arabe en 1912. En fait, l'utilisation de la ponctuation ne persiste pas dans la langue arabe.

L'alphabet arabe se compose de vingt-huit (28) lettres fondamentales

(أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي), (vingt-neuf (29) lettres si on n'a pas exclu la hamza (الهمزة), qui se comporte soit comme une lettre à part entière soit comme un diacritique).Il s'écrit de droite à gauche.

Dans la langue arabe, il n'y a pas de majuscules ou minuscules pour les lettres comme les lettres anglaises.

Les signes diacritiques (الحركات) dans la langue arabe permettent d'exprimer les voyelles brèves et d'apporter différentes modulations aux voyelles longues ainsi qu'aux consonnes.

Le but de signes diacritiques pour apprendre à les reconnaitre et à les prononcer correctement en contexte, pour distinguer des lettres ambigües et pour faciliter la lecture.

La majeure partie de l'écriture arabe est écrite sans Harakat. Cependant, ils sont couramment utilisés dans certains textes religieux qui exigent le strict respect des règles de prononciation telles que Qur'an (القرآن). Il est fréquent d'ajouter Harakat à hadiths (الحديث), ainsiune autre utilisation dans la littérature pour enfants pour connaitre le sens des mots arabes. Harakat sont également utilisés dans les textes ordinaires quand une ambiguïté de la prononciation pourrait se poser. [7]

Les Diacritiques arabes comprennent :

Fatha(◌َ) (الفتحة), Kasra(◌ِ) (الكسرة), Damma(◌ُ) (الضمة), Soukoune(◌ْ) (السكون), Shadda (◌ّ) (الشدة)et Tanwin(◌ٌ) (التنوين).

Fatha : permet la réalisation de la voyelle brève [a]. Il se présente sous la forme d'un accent aigu placé juste au-dessus de la lettre. [11]

Damma : permet la réalisation de la voyelle brève [u].Il se présente sous la forme d'un mini waw (و)placé juste au-dessus de la lettre. [11]

Kasra : permet la réalisation de la voyelle brève [i]. Il se présente sous la forme d'un accent aigu placé juste endessous de la lettre.[11]

Soukoune :Les syllabes peuvent être ouvertes ou fermées. C'est-à-dire si la syllabe se termine par une consonne, elle est fermée. Si la syllabe se termine par une voyelle, elle est ouverte. Pour indiquer qu'une syllabe est fermée (à la prononciation), on ajoute simplement un soukoune (petit cercle) au-dessus de la lettre. [11]

Tanouine :

- Tanouinefatha ou fathatan :permet la réalisation du son [an]. Il se présente sous la forme d'un double fatha.
- Tanouinedamma ou dammatan : permet la réalisation du son [on].Il se présente sous la forme d'un double damma.
- Tanouinekasra ou kasratan : permet la réalisation du son [en] ou [an]. Il se présente sous la forme d'un double kasra.[11]

6.1 Complexité de la langue arabe :

L'arabe est une langue difficile pour un certain nombre de raisons :

- L'orthographe avec diacritiques est moins ambiguë et plus phonétique en arabe, certaines combinaisons de caractères peuvent être écrites de différentes manières. [12]
- La langue arabe a des voyelles courtes qui donnent la prononciation différente. Grammaticalement ils sont nécessaires, mais omis dans les textes arabes écrits. [13]
- La langue arabe a une morphologie très complexe par rapport à la langue anglaise.
- Les synonymes sont très répandus. [14]
- La classification automatique du texte dépend du contenu des documents, un grand nombre de fonctionnalités ou des mots-clés peuvent être trouvées dans le texte arabe, tel que les morphèmes qui peuvent être générés à partir d'une racine qui peut conduire à une mauvaise performance en termes de précision et de temps. [14]

6.2 Statut géographique de la langue arabe :

L'arabe est une langue parlée par plus de 200 millions de personnes. Elle est langue officielle d'au moins 22 pays :

1. Péninsule arabique : l'Arabie saoudite, Bahreïn, les Émirats Arabes Unis, Oman, le Qatar, le Yémen.
2. Moyen-Orient : l'Irak, la Jordanie, le Koweït, le Liban, la Palestine, la Syrie.
3. Afrique : l'Algérie, l'Égypte, les Comores, Djibouti, la Libye, le Maroc, la Mauritanie, la Tunisie, la Somalie, le Soudan.

C'est aussi la langue de référence pour plus d'un milliard de musulmans.

6.3 La richesse de la langue arabe :

L'arabe est une langue très riche ; les Arabes se vantent, selon Ernest Renan, d'avoir 80 mots pour désigner le miel, 200 pour le serpent, 500 pour le lion, 1000 pour le chameau et l'épée, et jusqu'à 4400 pour rendre l'idée de malheur. Le vocabulaire comprend 60 000 mots. Les grammairiens arabes prétendent que toutes les racines de leur langue ont été primitivement des verbes, et ils élèvent considérablement le nombre de ces racines. Il est en réalité de 6000. D'après Maurice Gloton le Coran a utilisé environ 5000 termes, ce qui correspond à 1726 racines différentes.[15]

7 Conclusion :

Ce chapitre présente une introduction sur le domaine de l'Extraction de l'Information (EI) dans lequel se situe notre travail de Master. Nous commençons par une petite définition et nous passons en vue le domaine dans son contexte historique vis-à-vis des autres approches informatiques cherchant à collecter de l'information à partir de textes en langue naturelle. Aussi on a focalisé sur la langue arabe et sa complexité.

Chapitre 03
Conception et Réalisation

Introduction :

Ce chapitre est consacré aux étapes fondamentales pour le développement de notre système d'extraction d'information.

Cette partie sera consacrée à la formalisation conceptuelle qui est l'étape la plus importante d'un projet informatique. Elle a pour but de fixer les choix des informations et traitements à manipuler dans le SI. En plus la formalisation organisationnelle consiste à spécifier l'organisation qui régira les données et les traitements étudiés lors de la formalisation conceptuelle. Après avoir effectué la conception de notre système d'extraction d'Information, nous allons à présent entamer sa réalisation.

I. Conception :

1 Construction du corpus :

Le développement d'un système d'extraction d'entités nommées nécessite, au préalable, de rassembler un nombre suffisant de textes qui serviront non seulement de corpus d'observation et d'analyse (pour construire les règles) mais également de corpus de test (extraction des entités).

Pour couvrir le domaine de recherche, il nous a fallu recueillir des articles liés aux accidents de la circulation et constituer un corpus exploitable pour la création de notre système.

L'information et la sensibilisation aux problèmes des accidents de la route constituent une dimension à part entière de la politique de sécurité routière, la méthodologie suivie consistait donc, à analyser un corpus composé de divers éléments : articles, documents . L'échantillon des documents de la presse écrite est tiré des quotidiens les plus lus en langue arabe, notre corpus est constitué de 50 articles ayant abordé le problème des accidents de la route et publiés par les quotidiens. Pour constituer ce corpus, nous avons dépouillé tous les articles publiés par ces journaux durant l'année d'étude.

2 Construction des règles :

Toutes les entités utilisées dans les règles de conception sont les suivantes:

1. Nombre de morts.

2. Nombre de blessés.

3. Adresse (lieu de l'accident).

4. La date (la date de l'accident).

5. Cause (cause de l'accident).

Nous avons établi un ensemble de dictionnaires :

Nombre de morts :

Dictionnaire :{قتل, قتيلا, قتلى, لقي, بحياة, وفاة, ضحيتها,ضحايا, توفي, موت, موتى}

Nombre de blessés :

Dictionnaire :{جرح, أصيب, إصابة, تعرض, جريحا}

Cause d'accident :

Dictionnaire :{نتج, إثر, نجم, تصادم, انقلاب, اصطدام, بفعل, بسبب, ناتج, راجع, عائد}

Lieu d'accident :

Dictionnaire : { أدرار, الشـــلف, الأغـــواط,أم البواقي,باتنة,بجاية,بسكرة,بشار,البليدة,البويرة,تمنراســـت, تبسة,تلمسـان,تيارت,تيزي وزو,الجزائر,الجلفة,جيجل,سطيف,سعيدة,سكيكدة,سـيدي بلعباس,عنابة,قالمـة, قسـنطينة, المديـة, مسـتغانم, المسـيلة, معسـكر, ورقلـة, وهـران, البـيض ,إليزي,بـرج بوعريريج,بومرداس,الطارف,تندوف,تيسمسـيلت, الوادي,خنشلة,سوق أهراس, تيبازة,ميلة,عين الدفلى,النعامة,عين تموشنت,غرداية,غليزان}.

La date d'accident :

Dictionnaire :{يوم, صباح, نهار, ليلة, فجر, مساء, عشية, اليوم}

Nombre :

Dictionnaire :{ واحد, اثنان, ثلاثة, أربعة, خمسة, ستة, سبعة, ثمانية, تسعة, عشرة, اثنين, قتيلان, راكبان, جريحان, مسافران}.

2.1 Règle générale :

On utilisé la fonction Getword() pour sélectionner mot par mot et cherché a quel dictionnaire correspondant a chaque mot,après ça si le mot il excite dans notre dictionnaire on appliqué la fonction GetWord()à une autre fois, pour garder le 2éme terme suivant dans le bdd sur le type correspondant a ce mot, La figure 3.1 la présente globalement.

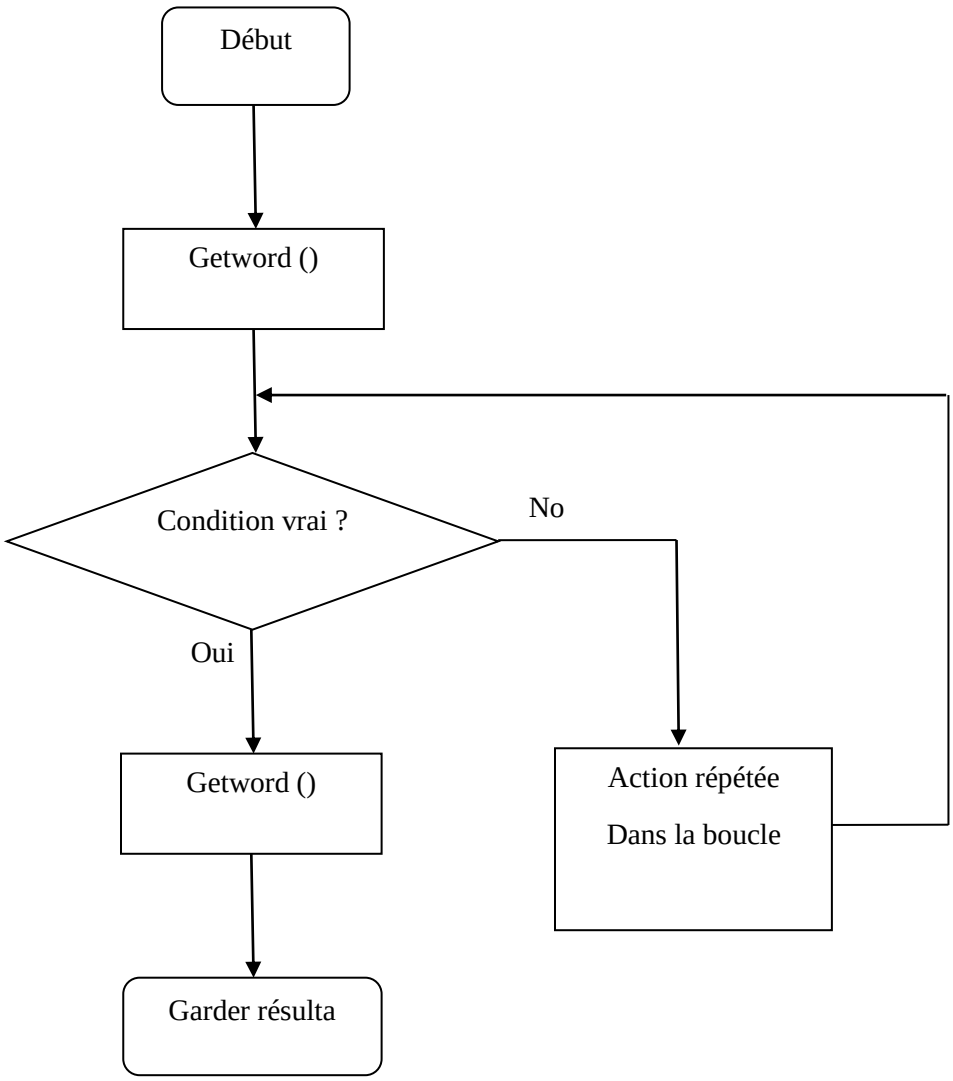

Figure 3.1 - Schéma de la règle générale getword ()

On a expliqué cette fonction sur exemple suivant :

Texte :

لقي خمسة أشخاص مصرعهم وأصيب إثنان آخران بجروح في حادث مرور وقع اليوم الأحد بمنطقة الميلق ولاية الأغواط حسبما علم من مصالح الحماية المدنية , و تمثل الحادث في اصطدام ثلاث سيارات .

Nombre de morts :

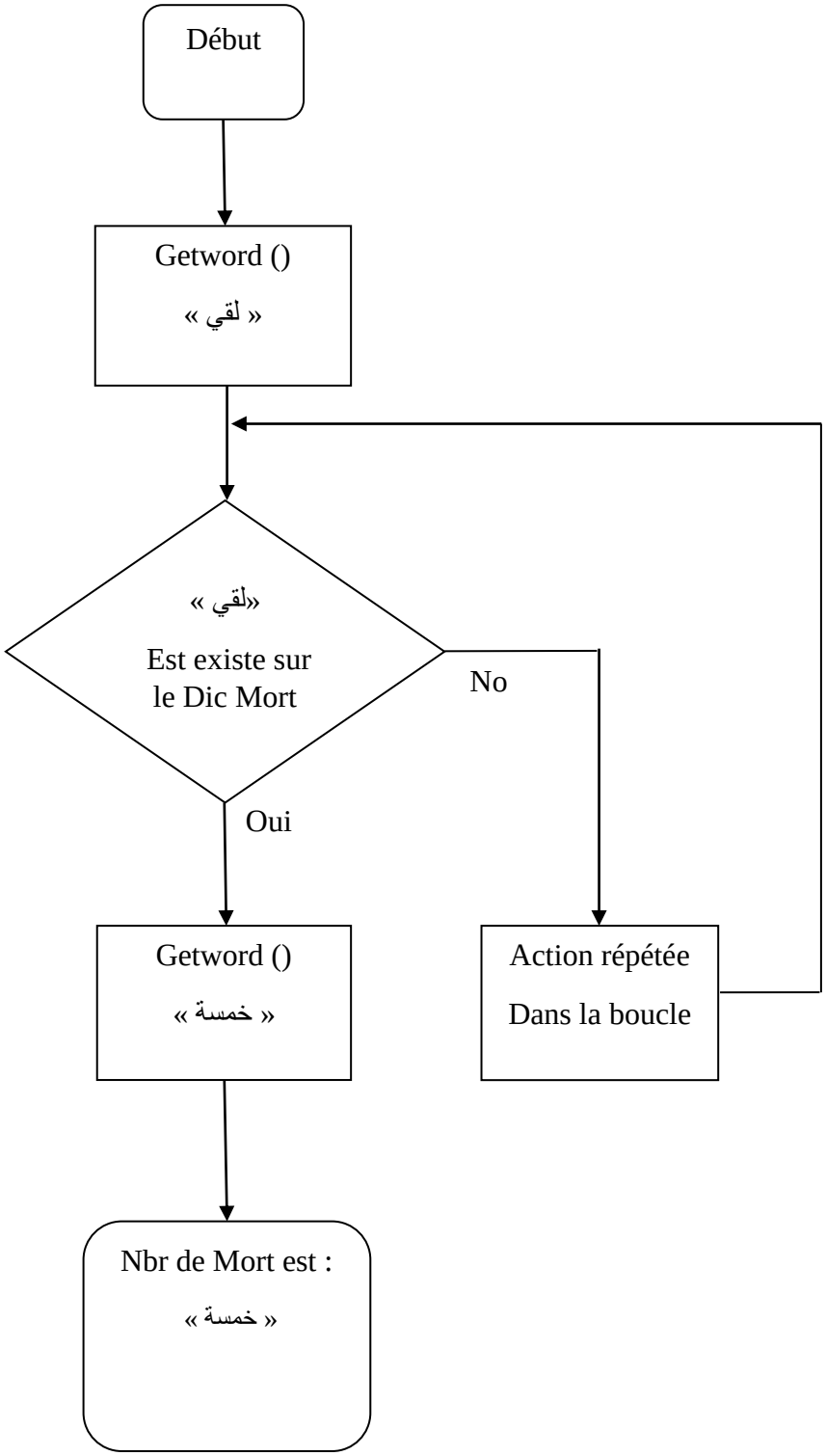

Figure 3.2 - Exemplepour extraire nombre de Morts

Nombre de blessés :

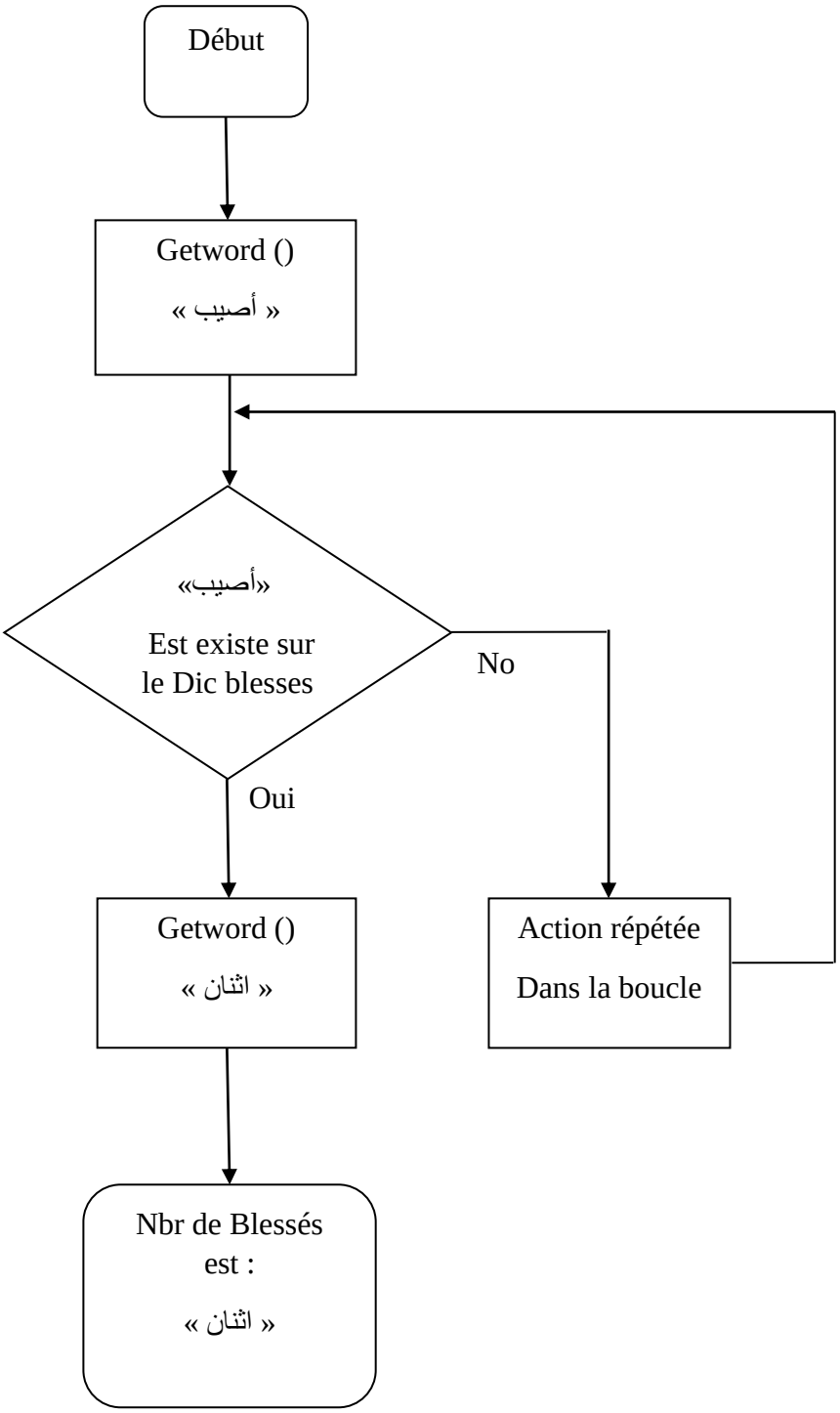

Figure 3.3 - Exemplepour extraire nombre de blessés

Lieu d'accident :

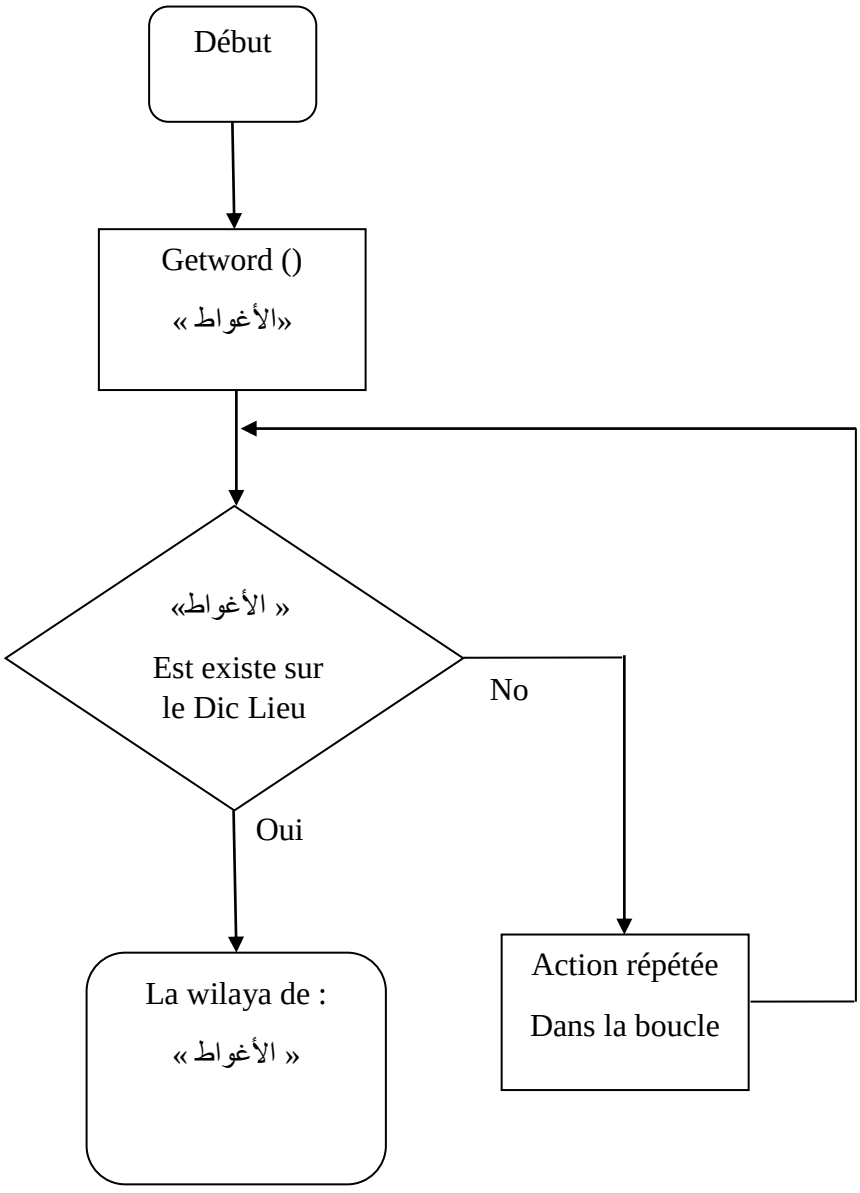

Figure 3.4 - Exemplepour extraire la wilaya d'accident

3 Mesures :

Ce sont des mesures standard pour l'évaluation de la capacité du système à détecter les entités nommées. Trois indicateurs ont été utilisés pour chaque type d'entité afin de mesurer la performance du notre système : la précision, le rappel et le F-mesure. Il est défini dans le chapitre 3 section 3.1.1.

3.1 Résultats expérimentaux :

L'analyse des résultats nous a permis de mieux comprendre les raisons de la baisse de la performance, en particulier le rappel pour certains types d'entités. Ce faible taux du rappel est dû principalement à la couverture insuffisante de notre ensemble de règles. Le système ne reconnaît pas toutes les entités, car il ne dispose pas de règles suffisantes pour les identifier.

Globalement, le système effectue une bonne extraction. Sur 50 entités, il détecte " 35" correct, et identifie "10" entités par erreur. Cela donne une précision globale de "77%" et un rappel global de "70%". Ces résultats sont très intéressants mais doivent être vérifiés dans une collection des articles consistante. Le système repose sur une bibliothèque de règles et un lexique de noms propres pour identifier les entités.

Ressources utilisées :

Nous avons réuni les éléments suivants :

1. une liste de concepts pour les accidents, nous avons réalisé une base de données de concepts (mort,blesse, cause ...).
2. Une liste des wilayas de l'Algérie.
3. Protection civile pour les causes des accidents.

II. Réalisation :

1Environnement du développement :

Avant de commencer l'implémentation de l'architecture conceptuelle de notre système, nous allons tout d'abord spécifier les outils utilisés qui nous ont semblés être un bon choix de par les avantages qu'ils offrent. [8]

1.1 Choix du système d'exploitation :

Notre application a été développée sous le système d'exploitation Windows 7, mais comme elle est développée en langage java, elle peut être intégrée dans n'importe quel autre système d'exploitation supportant la machine virtuelle java (Windows 98/00, Linux …).

1.2 MySQL :

MySQL est un serveur de bases de données relationnelles Open Source. Un serveur de bases de données stocke les données dans des tables séparées plutôt que de tout rassembler dans une seule table. Cela améliore la rapidité et la souplesse de l'ensemble. Les tables sont reliées par des relations définies, qui rendent possible la combinaison de données entre plusieurs tables durant une requête. Le SQL dans "MySQL" signifie "Structured Query Language" : le langage standard pour les traitements de bases de données.

1.3 Le langage Java :

Pour le choix de programmation de notre système, nous avons opté pour le langage JAVA et cela pour de nombreuses raisons :

- JAVA est un langage orienté objet simple, qui réduit le risque des erreurs d'incohérences.
- Il est indépendant de toute plateforme, il est possible d'exécuter des programmes JAVA sur tous les environnements qui possèdent une Java Virtual Machine (JVM).
- Il est doté d'une riche bibliothèque de classes, comprenant la gestion des interfaces graphiques (fenêtres, menus, graphismes, boites de dialogue, contrôles), la programmation multi-threads (multitâches) et la gestion des exceptions.
- Il permet un accès aux bases de données simplifié soit à travers la passerelle JDBC-ODBC ou à travers un pilote JDBC spécifique au SGBD.
- Il est caractérisé aussi par la réutilisation de son code ainsi que la simplicité de sa mise en œuvre.

1.4 NetBeans IDE 8.2 :

Pour le choix de l'environnement de développement, on a opté pour NetBeans, car il possède de nombreux points forts qui sont à l'origine de son énorme succès dont les principaux sont :

- Une plateforme ouverte pour le développement d'applications et extensible grâce à un mécanisme de plugins.
- Support de plusieurs plates-formes d'exécution : Windows, Linux, Mac OSX.
- Malgré son écriture en Java, NetBeans est très rapide à l'exécution grâce à l'utilisation de la bibliothèque SWT.

- La construction incrémentale des projets Java grâce à son propre compilateur qui permet en plus de compiler le code même avec des erreurs, de générer des messages d'erreurs personnalisés, de sélectionner la cible.
- Un historique local des dernières modifications.

2 Interfaces principales de notre système sont :

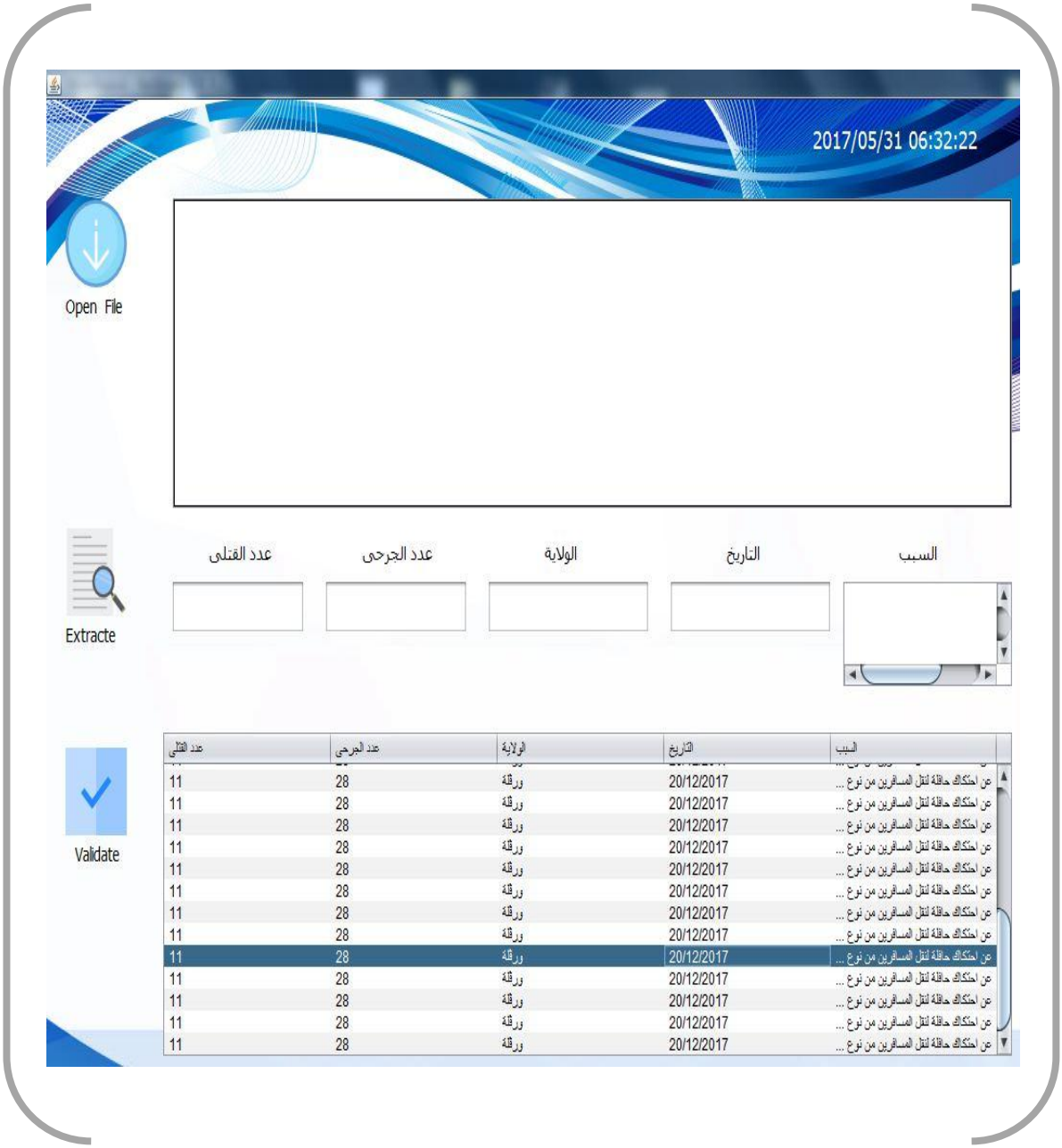

Figure 3.6 - Interface principale

3 Étude des Résultats :

Dans le système l'extraction d'information est faite après analyse linguistique et utilisation de "dictionnaire" pour trouver les informations correspondant, il y a peu de bruit, mais toutes les informations ne sont pas extraites.

Nous pensons que ces résultats peuvent être améliorés :

- Le rajout de termes utilisé dans la BDD.
- L'utilisation du dictionnaire de mots-clés, afin de réduire le bruit.
- Le rajout de règles de levée des ambigüités.

Conclusion :

Dans cette dernière partie, nous avons présenté la conception et la réalisation de notre système d'extraction. Ce dernier a été commencé par les résultats obtenus avec les mesures d'extraction connus (F-mesure moyenne et Accuracy), et on a exploré les outils utilisés, ensuite la description l'interface du système à travers les captures d'écran.

Conclusion Générale :

Dans le cadre de ce mémoire, j'ai présenté le système d'extraction d'Information sur le domaine des accidents de la circulation. Ce système est actuellement en développement ; la majorité des modules peuvent donc être nettement améliorés. Cependant, les résultats préliminaires obtenus démontrent l'avantage de la réutilisation des modules existants et la faisabilité du développement rapide d'un système d'extraction d'information multilingue.

Plusieurs technologies ont été nécessaires pour la réalisation de ce projet, je cite donc le SQL pour l'élaboration des requêtes d'interrogation de la base de données, et j'ai réalisé mon système dans l'environnement NetBeans avec le langage de programmation JAVA et enfin j'ai utilisé la bibliothèque "SAFAR" pour le côté linguistique.

Après le passage par les différentes étapes de développement, l'application a abouti à un logiciel fonctionnel qui répond globalement aux critères imposés dans ce domaine.

Ce travail a permis d'acquérir mes connaissances dans le domaine de la programmation Java, et de conforter mes connaissances en conception logicielle linguistiques.

Pour les perspectives de ce travail. Nous proposons pour l'amélioration de cette étude les points suivants :

- Le développement d'un système d'extraction multilingue à partir d'un système d'extraction monolingue.
- Nous souhaitons de développer une interface pour l'affichage des résultats aux les utilisateurs.

Bibliographie :

[1] I. Tellier, Introduction à la fouille de textes, Université de Paris 3 - Sorbonne

[2]Fabrice Even, Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale, Thèse de Doctorat, Université de Nantes, 05/10/2005.

[3]Barigou Fatiha, Contribution à la catégorisation de textes et à l'extraction d'information, Thèse de Doctorat, Université d'Oran, 2012/2013.

[4]Farida Yamouni Aoughlis, Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales, Thèse de Doctorat, Université Mouloud Mammeri de Tizi Ouzou, 12/10/2010.

[5] http://arabic.emi.ac.ma/safar/ consulté le : 07/05/2017.

[8] M .MEDDAH, F.Z. Laallam « Conception et Réalisation d'un système intelligent en aquaculture », 57 -59 pages, université Ouargla, 2012.

[9]Matallah Hocine, Classification automatique de textes approche orientée agent ; UNIVERSITE université Aboubekr Belkaid-Tlemcen faculté des sciences département d'informatique.

[10]Taïeb.Baccouche L'Information Grammaticale Année 1998 Volume 2 Numéro 1 pp. 49-54 Fait partie d'un numéro thématique : Numéro spécial Tunisie .

[11] https://abjadia.wordpress.com/tag/alif-wasla consulté le : 18/04/2017.

[12]Taghva, K., Elkhoury, R., Coombs, J., "Arabic stemming without a root dictionary", Information Technology: Coding and Computing, ITCC, Vol. 1, pp. 154 , 2005.'

[13]Said D., Wanas N., Darwish N., Hegazy N., "A Study of Arabic Text preprocessing methods for Text Categorization", In the 2nd Int. conf. On Arabic Language Resources and Tools, Cairo, Egypt, 2009.

[14]Kanaan G., Al-Shalabi R., Ghwanmeh S., "A comparaison of Texte-classification techniques applied to arabic text", Journal of the American Society for Information Science and Technology, 60(9), pp. 1837– 1838, 2009.

[15] http://www.agoravox.fr/actualites/religions/article/la-langue-arabe-son-histoire-son-77459, consulté le : 31/05/2017.

[16]Dominik francoeur, Machines A Vecteurs de support une introduction /CaMUS 1 (2010).

الملخص:

في هذه المذكرة نبين كيفية إنشاء و تصميم نظام آلي يقوم استخراج المعلومات من الصحف العربية في مجال حوادث السير في الجزائر. من أجل إنجاز هذا النظام الآلي , استعملنا NetBeans كمحيط تطوير.
اخترنا JAVA كلغة برمجة ومن أجل تسيير قاعدة المعطيات اخترنا MySQL.

الكلمات المفتاحية:استخراج المعلومات , اللغة العربية , جافا ,حوادث السير, NetBeans.

Résumé :

Dans ce mémoire nous avons conçu et réalisé un système d'extraction d'information à partir des articles en arabes dans le domaine des accidents de la circulation en Algérie.

Nous avons réalisé notre système dans l'environnement NetBeans avec le langage de programmation JAVA. Pour la gestion des données, nous avons utilisé le Système de gestion de bases de données MYSQL.

Les mots clés : Extraction d'information, Langue Arab,Accidents de la circulation JAVA, NetBeans.

Abstract :

In this brief we designed and realized a system of extraction of information from the Arab articles in the field of traffic accidents in Algeria.

We realized our system in the NetBeans environment with the JAVA programming language. For data management, we used the MYSQL Database Management System.

Keywords : Extraction of Information, ArabicLanguage, Traffic Accidents, JAVA, NetBeans.