Preuve du Concept (POC)

> ☝️ Le proof of concept est l'étude de faisabilité d'un projet à réaliser avant de vous engager, afin de démontrer la viabilité du projet à vos clients ou à l'équipe produit. Le POC figure souvent parmi les étapes essentielles de la méthodologie de gestion de projet et des méthodes agiles en général.

Définition de l'idée

Le système MedVir fonctionne selon le principe de la recherche diagnostique par logique floue. Cela signifie que jusqu'à preuve certaine, laquelle est fournie par les examens complémentaires (biologie, imagerie, anatomopathologie, génétique, exploration…), un diagnostic est plus ou moins probable.

Le système Medvir utilise un réseau de neurones qui fonctionnent à l'aide de "poids". Ainsi, Medvir attribue un « poids » à chaque symptôme ou à chaque caractéristique du symptôme, et ceci pour chaque diagnostic évoqué. C'est le rapport entre le déclaratif du patient et le poids de la totalité des symptômes du diagnostic qui fournit la probabilité du diagnostic.

Actuellement, MedVir ne dispose pour la recherche par mots que d'un champ qui ne comprend qu'un mot ou une expression à la fois (**pas de saisie multicritères**). Chaque mot tapé donne lieu à une liste comprenant le mot tapé et les diverses occurrences possibles correspondantes. On en sélectionne un et le symptôme correspondant apparaît dans le cartouche à droite.

L'objectif du projet est de remplacer cette barre de recherche monocritère par un champ de saisie d'environ 250 caractères tel que : « En sortant de chez moi, je me suis **cassé la gueule** dans l'escalier, et j'ai **peur** de m'être **pété la cheville**. Depuis, j'ai **mal à la tête** et j'ai **vomi** partout. ».

Le nouveau système devra être en capacité d'identifier les symptômes (mots en gras) et de les cocher dans la cartouche droite tout comme le fait le système actuel.

Ainsi, sur demande de Medvir, nous devons utiliser une solution de NLP (Natural Language Processing) afin d'identifier et de traiter le contenu du champ de saisie.

Proof Of Concept

Aujourd'hui il existe pleins de bibliothèques NLP open sources qu'il est possible d'utiliser pour analyser des champs de saisie.

Comparaison des bibliothèques NLP existantes

Nom	Open source	Languages	Tokenization	Part Of Speech tagging (POS)	Named Entity Recognition (NER)	Classification	Sentiment Analysis	Packages of chatbots	Dependency Parsing	Word Vectors	Matrix Factorization	TF-IDF	Parsing	Noun phrase Extraction
NLTK	Yes	English Russian (POS)	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No
spaCy	Yes	English French + 22	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	No
Gensim	Yes	English French and more	No	No	No	No	Yes (latent)	No	No	No	Yes (Non-negative)	Yes	No	No
Pattern	Yes	English	Yes	Yes	Yes	No	Yes	No	No	No	No	No	Yes	No
TextBlob	Yes	Language translation	No	Yes	No	Yes	Yes	No	No	No	No	No	Yes	Yes

Nom	Wordnet Integration	Use-cases	Pros	Cons
NLTK	No	Recommendation systems Building chatbots Sentiment analysis	Most well known and full NLP library	Difficult to lear and use Context of word is ignored Slow No neural network model Generally used as an education and research tool
spaCy	No	Summarization Autocomplete and autocorrect Analyzing reviews	Production usage Fast Esay to learn and use Use neural networks for training	Les flexible
Gensim	No	Converting documents to vectors Finding text similarity Text summarization	Intuitive interface Scalable Implemented algorithms	Designed for unsupervised text models Should be used with ohter libraries
Pattern	No	Spelling correction Search engine optimization Sentiment analysis	Data mining Network analysis and visualization	Not optimized with specific NLP tasks
TextBlob	Yes	Sentiment Analysis Spelling Correction Translation and Language detection	Easy to use Intuitive interface to NLTK Provides language translation and detection	Slow No neural network model No integrated word vectors

Tests des bibliothèques

Afin d'identifier ce que nous sommes réellement capables de faire à l'aide des bibliothèques. Nous avons effectué une série de tests sur les bibliothèques les plus prometteuses.

(*cf. Annexe 3 : Notebook test NLTK, Annexe 4 : Notebook test Spacy*)

Axes de décision

Afin d’être en mesure de choisir la bibliothèques la plus pertinente, nous avons besoin de définir et de prioriser les axes de décision :

1. Open Source
2. Disponible pour la langue française
3. Facile à comprendre et à utiliser
4. Rapide
5. Dispose d’une solution d’apprentissage

Résultats

Grâce à notre POC, nous avons constaté que NLTK est applicable à l'anglais et au russe, tandis que Spacy offre une prise en charge dans une vingtaine de langues, y compris le français. Nous avons donc réalisé des tests avec NLTK en anglais, en effectuant une traduction préalable du texte français vers l'anglais. La langue anglaise étant moins complexe que le français, nous pensons qu’elle peut être plus facile à analyser. Toutefois, il est important de noter que la traduction introduit des incertitudes et des erreurs potentielles. De plus, une étape finale de traduction vers le français est nécessaire pour obtenir le résultat dans la langue souhaité.

Au vu de toutes ces observations, nous avons pris la décision de nous concentrer sur l'utilisation d'une bibliothèque en français pour notre projet. Cette orientation vise à contourner les étapes de traduction, à minimiser les risques d'erreurs liées à celles-ci, et à simplifier le flux de traitement en travaillant directement dans la langue cible. Ce choix contribuera à optimiser notre approche NLP et à garantir une cohérence linguistique tout au long du processus d'analyse des symptômes dans les textes en français.