**Proposition commerciale**

**Projet Youth Talk Edition 1 - Intelligence Artificielle**

**Cyril Vincent**

Il s’agit d’identifier automatiquement les topics pertinents d’empathie par rapport à la base de données de 40K questions ouvertes et fermées

Voici les phases de la modélisation :

**Phase 1 : NLTK + Random Forest + SQL**

* Pour le thème d’empathie PD:
  + Calculer la moyenne des 4 notes
  + Discriminer les notes 3 catégories : négatif, neutre, positif
* Utilisation des modèles NLTK, USE et Random Forest
  + NLTK est un framework de langage naturel basé sur plus d’une centaine de modèle de machine learning et de statistiques. Les modèles Lemma, Gestalt et Leveinstein seront utilisés. Voir <https://www.nltk.org>.
  + USE (Universal Sentence Encoder) est un modèle de Google qui transforme une phrase en un vecteur de 512 flottants. En calculant une distance entre 2 vecteurs il est possible de déterminer la similarité syntaxique des 2 phrases. Voir <https://static.googleusercontent.com/media/research.google.com/fr//pubs/archive/46808.pdf>.
  + Random Forest est un modèle de Machine Learning très utile pour la classification et les corrélations par l’algorithme des features importances. Voir <https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
  + Pearson est un algorithme de corrélation. Voir : <https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.pearsonr.html>.
* Pour la question 1 et ses 3 contributions rechercher les topics principaux par les modèles NLTK
  + Les compter
  + Les regrouper en évitant les synonymes
  + Les associer aux 3 catégories du thème
* Créer la base de données PostgreSql
* Analyser les 50 000 contributions et les stocker en base de données
* Créer les métrics
  + NbTopic : nombre de topic total
  + NbTopicOccurence : nombre d’occurrence d’un topic
  + NbOccurence : nombre d’occurence de tous les topics
  + NbTopicNegative : nombre d’occurrence négative d’un topic, idem pour positive et neutral
  + RatioTopicNegative = NbTopicNegative / NbTopicOccurence, idem pour positive
  + Si RatioTopicNegative ~= RatioTopicPositive, le topic n’est pas pertinent
* Filtrage
  + Enlever les topics très faiblement représentés
  + Enlever les topics non ou peu pertinents
  + Regrouper les topics synonymes
* Retrouver les topics pertinents
  + Si RatioTopicNegative >> RatioTopicPositive le topic est pertinent, l’inverse est vrai également
* Utilisation des modèles Random Forest et USE
  + Trouver les features importances via Random Forest
  + Corrélations de Pearson
  + Calculer les matrices de confusion
  + Calculer les scores : accuracy, recall, precision, f1 score
  + Calculer les corrélations par rapport aux features importance de Random Forest et des ratios
  + Retrouver les similarités avec USE
* Export CSV
* L’ensemble de la phase 1 est estimé à 5 000 €
  + Les modèles pourront être facilement réutilisés pour les 10 autres questions et les 3 autres thèmes d’empathie

**Phase 2 : LLM**

Utilisation du LLM ChatGPT 4o car les données sont libres et anonymes

* Concaténations des 3 contributions de la question 1
* Prompt engineering pour extraction automatiques des topics
* Envoie des 50 000 prompts (facturé par OpenAI)
  + 40h à 50h de calcul
* Comme pour la phase 1, stockage, comptage, regroupement des topics
* Calculs des scores
* Retrouver grâce aux calculs de la phase 1 les faux négatifs et faux positifs évidents
  + Analyse des résultats
* Léger fine tuning sur les questions 1 et EC afin de corriger les faux négatifs et positifs évidents
  + 20h de calcul
  + Réenvoie des 50 000 prompts et calcul des scores (facturé par OpenAI)
* Calcul des corrélations
* Export CSV
* L’ensemble de la phase 2 est estimée à 10 000 €

**Phase 3 : Recherche des signaux faibles**

* Un signal faible est un topic faiblement utilisé mais avec une forte corrélation négative (ou positive)
  + Si RatioTopicNegative est élevé ET NbTopicOccurence / NbTopic est faible ET le score de corrélation est élevé, il s’agit d’un topic peu utilisé mais avec une corrélation très négative (ou positive), il s’agit donc d’un signal faible
  + Le score des features importances de Random Forest sera un indicateur supplémentaire pour les retrouver
* Il faut ensuite refaire une mini-modélisation pour retrouver les topics des questions ouvertes et fermées
  + En effet les gens répètent souvent le contenu des questions
  + Si un topic fait partie de la question, il ne s’agit pas d’un signal faible
* Puis il faut retrancher les topics des questions aux signaux faibles pour retrouver les vrais signaux faibles
* Stocker les résultats
* Exporter les résultats
* L’ensemble de la phase 3 est estimée à 3 000 €

**Phase 4 : Extension aux 3 autres thèmes d’empathie**

* Normalement les modèles pourront être utilisé tel quel
* La base de données sera légèrement modifiée
* J’estime cette phase à 2 000 €, mais je préfère attendre les résultats des phases 1 et 2 avant de faire un devis

**Phase 5 : Extension aux 10 autres questions**

* Il y aura 180h de calcul
* Normalement les modèles pourront être utilisé tel quel
* La base de données sera légèrement modifiée
* J’estime cette phase à 4 000 €, mais je préfère attendre les résultats des phases 1 et 2 avant de faire un devis

**Phase 6 : Fine Tuning**

* Fine tuning sur les questions et les thèmes d’empathie sur tous les faux positifs et faux négatifs en rajoutant aléatoirement des vrai positifs et négatifs pour avoir des données équilibrées
  + Coût estimé par OpenAI de 3 750€
  + 100h à 200h de calcul
  + Puis envoie des 50 000 prompts et calcul des scores (facturé par OpenAI)
* En général cette phase coûte 5 000 € à 10 000 € + 3750 €, à confirmer

La base de données ainsi que la conception des modèles pourront être réutiliser pour la seconde étude YouthTalk