



INNOSCAPE

## Projet fil rouge: Product Matching sur base de textes et images

Innoscape est une startup *RetailTech* d'intelligence de marché digitale. Nous analysons les canaux de distribution digitalisés et fournissons des outils d'aide à la gestion marketing et commerciale. Notre plateforme de *BigData as a Service* fournit à nos clients, grandes marques B2C, la visibilité opérationnelle d'un seul clic afin de contrôler leur présence en magasin, optimiser les actions de leurs forces de vente, traquer le référencement des produits concurrents et leurs prix de vente, analyser les avis consommateurs... pour gagner des parts de marché.

Notre technologie et business model ont été validés et nous préparons l'expansion suivante.

### Contexte :

- Nous réalisons de façon continue et itérative des POCs autour des problématiques de la donnée et une des problématiques actuelles concerne le **“Product Matching”** (détection de produits à travers différentes **descriptions littérales / attributs** et à travers leurs images collectés sur des nombreux sites marchands).
- Le projet s'inscrit dans le cadre d'un programme de développement à long terme avec un objectif d'amélioration continu et d'enrichissement de nouvelles sources et fonctionnalités.

### Contenu du projet :

- Réaliser une analyse sémantique (lexicale / syntaxique ) du corpus technique caractéristique des produits de nos clients, de faire du Feature engineering grâce aux différentes techniques et approches **d'apprentissage Statistique / Machine Learning supervisées, semi ou non supervisées (NLP, NLTK, CRF, TF-IDF, Word Embedding)** afin de construire des modèles prédictifs via Classification / Regroupement / Clustering:
  - Analyse/évaluation de la similarité par des techniques classiques ( Similarity & metrics learning).
  - Prédiction de référence d'un produit suivant des descriptions / images.
  - Améliorer les modèles prédictifs grâce à la prise en compte des images et les approches de Deep Learning.
- Réaliser une analyse sémantique afin de construire une base de connaissance lexicale (verbatim) des produits ainsi que leurs liens grâce aux techniques d'extraction d'information (Information retrieval) : NER, CRF etc...
- **Outils / techniques** : Python / Pyspark, scikit-learn, Numpy, Pandas, Gensim, Machine learning, NER, NLTK, NLP, Word Embedding, TF-IDF, Tensorflow, Keras, Deep Learning, expérimenter et enrichir les modèles pré-entraînés de Google, Ms, AWS.

