Sujet 7 — Trading à partir d'information textuelle

Objectif

L'idée de ce sujet est d'exploiter des informations textuelles récentes (actualités économiques, géopolitiques, décisions de banques centrales, etc.) pour générer un signal de trading, en pariant sur le fait que le marché met un certain temps à intégrer pleinement une information nouvelle dans les prix.

Sources de données suggérées

`yfinance` peut être utilisé pour récupérer les cours historiques, et l'API de *The Guardian* pour récupérer un flux d'articles de presse. **Attention :** la densité d'articles disponibles a fortement changé au cours des dernières années (volume de publication, couverture des sujets financiers) ; il faut impérativement contrôler l'effet de ce changement de densité sur les modèles construits (une variation du nombre de signaux détectés dans le temps peut refléter un changement dans le volume de publication plutôt qu'un réel changement du flux d'information pertinente).

Le rôle des LLM

Faire trader directement un LLM à partir du flux d'informations brut (c'est-à-dire lui demander de décider directement une position ou une quantité à trader) est une mauvaise stratégie : les LLM généralistes ne sont pas conçus pour raisonner quantitativement sur des séries temporelles ni pour intégrer une notion de risque, et donnent en pratique de mauvais résultats en trading direct. Il est en revanche pertinent de les utiliser pour des tâches de traitement de l'information en amont de la décision de trading, par exemple :

- filtrer et trier les articles pertinents parmi un flux important (ne garder que ceux ayant un impact potentiel sur le marché ou l'actif considéré) ;
- extraire des informations structurées (entités concernées, type d'événement, date d'effet) ;
- aider à interpréter la direction probable de l'impact d'une information macroéconomique ou géopolitique (variation de taux directeur, conflit, crise) sur tel ou tel marché.

Pour des besoins d'inférence LLM gratuits et exécutables localement, on peut se tourner vers Ollama (permettant de faire tourner des modèles en local), avec par exemple Mistral 7B si la machine disponible est suffisamment performante.

Pour l'analyse de sentiment à proprement parler, il est préférable d'utiliser un modèle spécialisé et entraîné pour ce cas d'usage financier précis, tel que FinBERT, plutôt qu'un LLM généraliste : ces modèles spécialisés sont généralement plus fiables et moins coûteux à faire tourner pour cette tâche ciblée.

De la direction à la quantité

Un point de vigilance central : déterminer la *direction* probable du marché suite à une information (le marché devrait-il monter ou baisser ?) est une tâche bien plus simple que de déterminer *combien* trader. La direction seule ne suffit pas à construire une stratégie : il faut ensuite un cadre de sizing des positions (par exemple un critère de Kelly, un ciblage de volatilité du portefeuille, ou un modèle de régression permettant d'estimer l'amplitude attendue du mouvement plutôt que son seul signe) pour transformer le signal directionnel en une décision de trading complète.

Pistes supplémentaires

Ce sujet est réputé délicat à bien mettre en œuvre. Quelques points supplémentaires à considérer :

- **Alignement temporel précis** entre l'heure de publication de l'information et l'heure des cours utilisés : un décalage, même de quelques minutes, peut introduire un biais de *look-ahead* qui rendrait la stratégie artificiellement performante en backtest sans être exploitable en réalité.
- **Vitesse de réaction du marché** : une grande partie de l'information publique est déjà intégrée dans les prix en quelques secondes à quelques minutes pour les actifs liquides ; il faut donc s'interroger sur l'horizon de trading réaliste sur lequel un edge informationnel peut effectivement subsister (et si la stratégie a une chance d'être exécutable compte tenu de la latence disponible).
- **Étude d'événements (event study)** : mesurer statistiquement, sur un historique, le rendement anormal moyen (et sa significativité) suite à différents types d'événements, avant de construire une stratégie dessus.
- **Validation rigoureuse** : découpage train/validation/test strictement chronologique (jamais de mélange temporel), et vigilance particulière sur le risque de tests multiples (tester un grand nombre de variantes de signaux augmente la probabilité d'en trouver un significatif par pur hasard).
- **Sources complémentaires** : au-delà de la presse généraliste, on peut envisager des transcriptions d'earnings calls, des documents réglementaires (SEC filings), ou des réseaux sociaux financiers, chacun ayant ses propres biais et limites qu'il faudra étudier.
- **Coûts de transaction et slippage** réalistes, particulièrement importants ici car les mouvements de marché suite à une information peuvent être rapides et donc coûteux à capturer en pratique.

Remarque générale

Ce qui précède n'est qu'une trame indicative. Ce sujet demande un travail de recherche approfondi, à la fois sur le plan du traitement de données textuelles (NLP financier), sur le plan statistique (significativité, validation), et sur le plan de la construction d'une stratégie de trading exploitable en pratique.