CorpLink–AI
データ処理 WEB APP
ガイドライン

作者：楊　天楽
協力：李　宗昊
参考資料：李　佳璇さんの「ネットワーク分析手順書(PHR)」

yotenra.com

はじめに

CorpLink–AI とは、従来手作業で行っていた膨大なビジネス文書やニュース記事（Factiva や LexisNexis 等）からのデータ抽出作業や名寄せ作業を、AIで全自動化するアプリです。
自然言語処理（NLP）と大規模言語モデル（LLM）を組み合わせることで、テキスト内の「組織名」を正確に抽出し、表記揺れを自動でクレンジングして、組織間のネットワークをデータベース化・可視化するまでを一貫して行います。
名寄せの基準：ある擬似組織名が、データベースで累積した正規組織名との関連があったら、そのまま正規組織名に転換。擬似組織名がデータベースで既存してないならば、APIで接続した大規模言語モデルで正規組織名を確認し、転換し、データベースに保存するという仕組みです。

簡単！

複数なニュースデータのファイル（.rtf / .docx）

↓↓↓↓↓クッリクだけで…↓↓↓↓↓

組織名の抽出結果・隣接行列・ピボット表

コアコードについて

企業等の組織間関係は経営意思決定において非常に重要
企業間関係の特徴を研究・分析し、共通点を探り、論文を執筆し、方法論を提供する必要
その分析には大量のペア関係データが必要
商用ニュースデータベース内の英語記事には大量の関係情報が暗黙的に含まれている
直接の全文分析：USD 25 / 1,000,000 英語文字ニュースデータ
コストが高すぎるため、安価なソリューションが必要
docxまたはrtfファイルをインポート
ローカルで全文検索を行い、cooperatやbindなど関係を示すキーワードを含む文を特定
文に対してspaCy、en_core_web_smプラグインを使用し、初期の疑似組織名を抽出
初期の疑似組織名に対し、Sentence-TransformersやRapidFuzzを使用してフィルタリングとあいまいマッチングを行う
さらに、MySQLデータベースの既存のbanデータテーブル、疑似組織名–正規組織名データテーブルと比較・マッチングを行う

- banデータテーブル：すでに非組織名としてマークされている疑似組織名を除外するために使用
- 疑似組織名–正規組織名データテーブル：既存の疑似組織名について、直接正規組織名に置換

banデータテーブル、疑似組織名–正規組織名データテーブルに存在しない疑似組織名について：

- OpenAI APIの 4o–mini モデルを呼び出し、正規組織名を取得する。費用について：
    - データベースに類似の業界・シナリオの履歴データが存在しない場合：USD 0.006 / 1,000,000 英語文字ニュースデータ
        - かかる費用は全文認識の4,000分の1
    - データベースに類似の業界・シナリオの履歴データが存在する場合：USD 0.0015 / 1,000,000 英語文字ニュースデータ
        - かかる費用は全文認識の17,000分の1
    - データベースに完全一致する履歴データが存在する場合：費用 0

今回の処理結果をMySQLデータベースに保存し、以後の処理で既存データを優先的に使用できるように
以降のOpenAI API呼び出し時のトークン数をさらに削減し、それによりさらなるコスト削減
まとめ：使えば使うほどスマートになり、低コストになるローカライズされた「ニュースデータ→組織関係」の自動処理ツールを構築

付属の Web APP について

ローカルで python ファイルを実行するには、一定の操作ハードルがある（環境構築、コードファイルなど）
ワンストップ、ワンクリックの APP を提供し、使用を簡素化してハードルを下げる
ユーザーは .zip に圧縮されたニュースデータファイルをアップロードし、処理パラメータを選択し、API key を入力するだけでよい
「処理開始」をクリックし、数分から数時間待って、結果をダウンロードするだけで完了

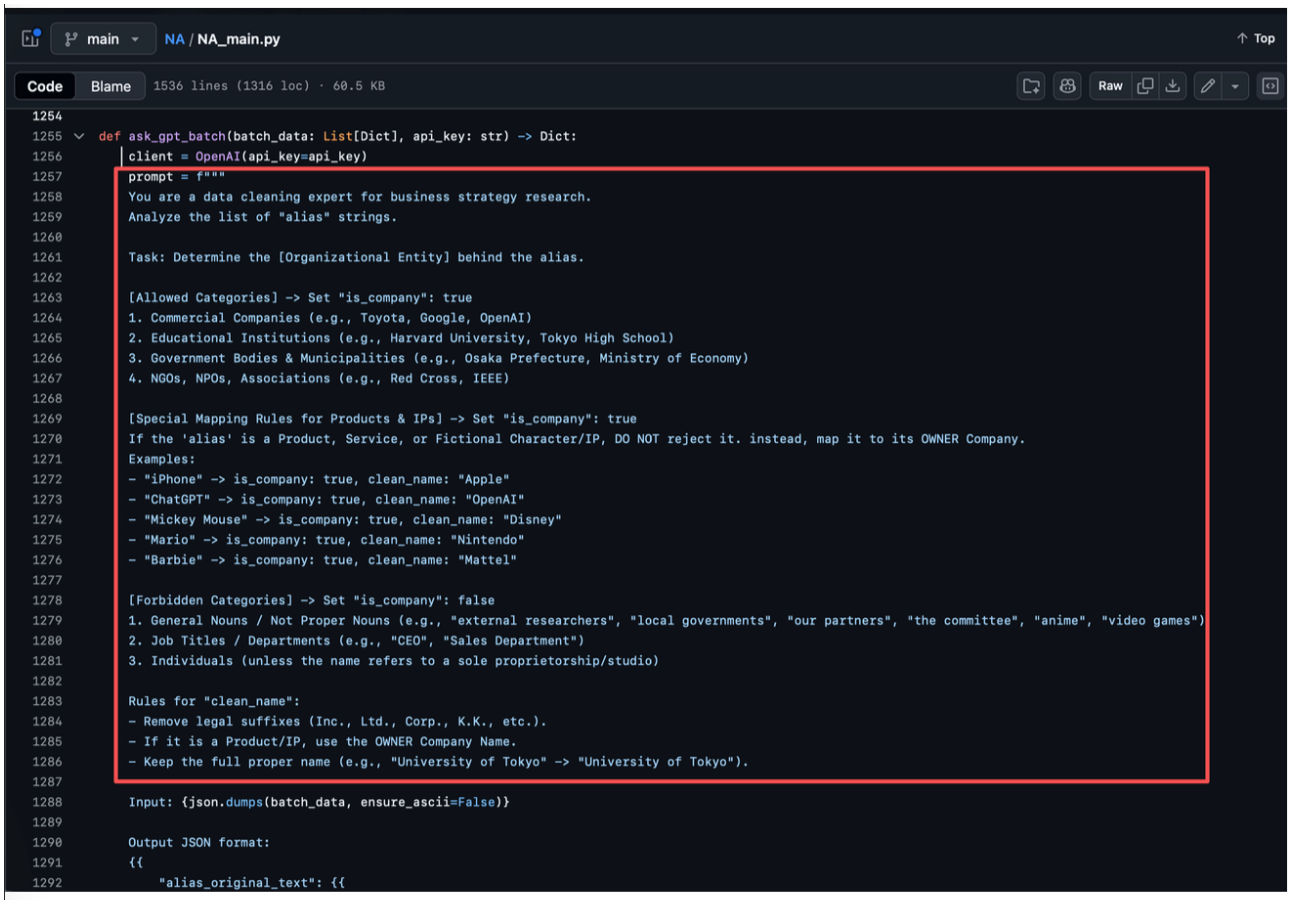
Q&A

Q: AIって、本当に信用できるの?

A: 結論から申し上げますと、本アプリが対象とするデータ処理業務においては「人間による手作業よりもはるかに信頼性が高い」と言えます。AIを業務に組み込むの理由は、圧倒的な「標準化」と「規模化」を実現するためです。

1. 「対話型AI」とは異なる、APIによる独立した事実検証
本アプリはChatGPTのような「チャット形式」ではなく、アプリ裏側で「API」としてGPTを制御しています。 対話型AIのように過去の無駄な会話履歴に引きずられて混乱することはなく、1件1件のデータに対して毎回クリーンな状態で独立した検証を行います。必要に応じて最新のウェブ検索も実行するため、極めて事実に基づいた客観的な処理が可能です。

2. 属人化の排除と、決して「疲れない」安定した集中力人間が膨大なテキストから組織名を抽出し、表記揺れを統一しようとすると、どうしても「作業者による基準のブレ」「その日の気分」「長時間の作業による集中力の低下や見落とし」が避けられません。
本APPでは、AIに対する「プロンプト（判定基準の指示書）」をシステムレベルで厳密に固定・統一しています。そのため、最初の1件目であろうと、1万件目であろうと、常に全く同じ高い基準で、ブレのないデータクレンジングを瞬時に実行します。



↑採用したプロンプト↑

3. AIに丸投げしない、ハイブリッドな設計（CorpLink–AIの強み）
本APPは、ただAIにデータを丸投げしているわけではありません。 事前に自然言語処理（NLP）やベクトルマッチングを用いて徹底的にノイズ（不要な単語や無関係な情報）を弾き出し、本当に高度な判断が必要な「組織名の同一性判定」の工程にのみ、LLM（大規模言語モデル）の推論能力を投入しています。AIの強みを最も安全かつ確実な形でコントロールする設計になっています。

別記：
今使っているプロンプトの論理の説明：
1. AIの役割設定 「あなたはビジネス戦略研究のためのデータ整理の専門家です。リストにある『擬似組織名』を見て、その対応している『組織・団体』を特定してください。」

2. 抽出する対象（これらは「組織」として扱います） 以下の4つのカテゴリーに当てはまるものは、有効なデータとします。

民間組織（例：トヨタ、Google、OpenAIなど）
教育機関（例：ハーバード大学、○○高校など）
政府・自治体（例：大阪府、ニューヨーク市、経済産業省など）
非営利団体・協会（例：赤十字、IEEEなどの学会・協会）

3. 製品・キャラクターの特別ルール 「iPhone」や「ミッキーマウス」のような製品名

やキャラクター名が来た場合、これまでは除外していましたが、今回は除外しません。 代わりに、「その持ち主（親会社）」に変換して抽出します。
例：「iPhone」というデータが来たら → 「Apple」という組織として登録
例：「ChatGPT」 → 「OpenAI」
例：「ミッキーマウス」 → 「Disney」
例：「マリオ」 → 「Nintendo」
例：「バービー人形」 → 「Mattel」

4. 除外する対象
一般名詞（特定の名前ではないもの。例：「外部研究者」「地方自治体」「我々のパートナー」「委員会」「アニメ」「ビデオゲーム」など）
役職・部署名（例：「CEO」「営業部」など）
単なる個人名（個人事務所やスタジオ名として機能していない場合）

5. 名称の整形ルール（Clean Nameの作り方） 抽出した名称は、以下のようにきれいに整えます。
法人格は削除する（Inc., Ltd., Corp., 株式会社 などを取る）。
製品/IPの場合は、親会社名に置き換える。
大学や公的機関は省略しすぎず、正式な固有名詞を残す（例：University of Tokyo はそのまま）。

Q:　API料金が高いですか?
A:　トークン消費を極限まで節約する設計になっているため、非常に低コストで運用可能です。
目安：５万件＝１米ドル＝１６０円
一般的なAIの使い方のように、ニュース記事などの「全文」をそのままAPIに投げて意味解析をさせると、文章が長すぎるためトークン消費量が膨大になり、コスト的に全く実用的ではありません。

1. ローカルでの事前スクリーニング
全文をAIに読ませるのではなく、まずはアプリ内部のキーワード処理等を用いて、無関係な情報を徹底的に除外します。

2. 「名寄せ」工程にのみAPIを限定使用
OpenAIのAPIを呼び出すのは、本当に高度な推論能力が必要とされる「擬似組織名の正規化（名寄せ）」のタイミングのみです。

3. 送信データの最小化とキャッシュ活用
APIに送信するのは短い「擬似組織名」の文字列のみです。さらに、過去に名寄せしたことがある組織名はローカルのデータベースを参照するため、未知の擬似組織名にしかAPIを使用しません。

4. 使えば使うほど「賢く・安く」なるコスト低減できる、成長できるアプリ
本アプリは、過去の処理結果をローカルデータベースに蓄積していく仕組みを持っています。そのため、過去に処理したことのある組織や関連トピックが再び出現した場合は、APIを一切呼び出さずにローカルで瞬時に解決します。 つまり、「使えば使うほどアプリが賢くなり、それに比例してAPIコストが劇的に下がっていく」という、長期的・大規模な運用に非常に適した設計になっています。

どう使えば、

まず…
http://yotenra.com/
を開ける

クリック

クリック

CorpLink-AI データ処理 WEB APP

作者：楊　天楽 (YO Tenra)

📂 ファイルアップロード & 設定

アップロードするファイルを選択 (.zip, .docx, .rtf):

选取文件 未选择文件　ここで、ファイルをアップロード

※1つのファイルを処理する場合は、.docxまたは.rtfの拡張子のファイルをアップロードしてください。
※複数のファイル又はフォルダを処理する場合は、zip形式に圧縮してアップロードしてください。なお、すべてのファイルの拡張子は同一（すべて.docx、またはすべて.rtf）である必要があります。

OpenAI API Key: *　OpenAIでKeyをもらって、入力。

sk-...　又は、楊天楽から貸してもらえます。

OpenAIの関連ガイドを参照して、APIキーを取得してください。
https://help.openai.com/en/articles/4936850-where-do-i-find-my-openai-api-key

データソース形式: *

最近は主にFactiva

- (●) Factiva (.rtf)
- ( ) LexisNexis (.docx)

キーワードモード: *

エンティティ間にリレーションが存在すると判定するためのキーワードを指定してください。

- (●) プリセット (2025 AI x Healthcare)
- ( ) カスタムキーワード

普段はディフォルトでオッケーです。

データベース接続先: *

特にない限り、本アプリのデフォルトの内蔵データベースをそのまま使用してください。

- (●) 既存のデータベース
- ( ) カスタムURL

自分のデータベースがない限り、
本アプリのデータベースを使用ことを推奨

▶ 処理を開始

そして、処理を開始

⚙️ 処理ステータス

現在の状態: 待機中 (新しいタスクを開始できます)

システム準備完了...

ここで、状態や進捗を確認できます。

⚙️ 処理ステータス

現在の状態: 処理中 (システムロック中...)

システム準備完了...
[0:17:07.619] アップロードと処理を開始します...

これはちょいちょい処理中でございます。
数分から一時間程度の時間をかかる場合があります。そのまま待ってください。🥲

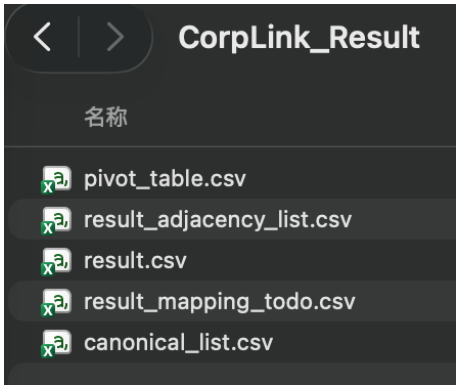
⚙️ 処理ステータス

現在の状態: 完了 (結果をダウンロードし、クリーンアップしてください)

⬇️ 結果をダウンロード　🗑️ 環境をクリーンアップ

システム準備完了...
[0:17:07.619] アップロードと処理を開始します...
[0:22:06.616] 完了: 処理が完了しました！

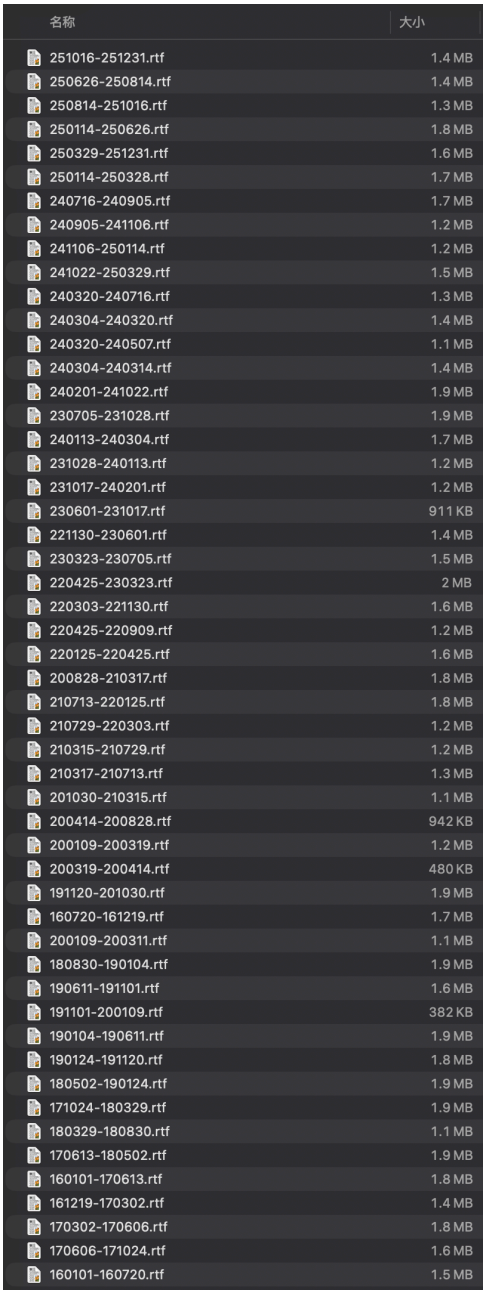
完成！ダウンロードする！



Pivot_table	ピボット表
result	抽出結果
result_adjacency_list	隣接行列

利用例：
同時に一つタスクしか実行できないので、必ず、事後にクリーンアップすること!
李伊迪爾さんのプロジェクトに、
3時間で、
3786万文字・60959擬似組織名・76.23MB

の新聞記事データを処理した実例です。
しかも、APIコストは USD 0.25でした。



	time	item
.RTF File Loading	0:08:22	52
Name Recognizing	1:03:00	60,595
GPT Cleaning	1:53:00	328
Adjacency Editing	0:00:21	60,595
Time Cost	3:04:43	
Characters Amount	37,866,000	
Name Amount	60,595	
Cost for OpenAI API	US$ 0.25	

業界フィルタリング・ノイズ除去機能について

1. この機能は何のためにあるのか?
ニュース記事（LexisNexisやFactiva）から企業関係を抽出すると、「PR Newswire」や「Reuters」といったニュース配信社やメディア名が頻繁に登場します。これらがデータに含まれると、社会ネットワーク分析（SNA）において以下のような問題が発生します:

- 構造的ノイズ: 配信社がハブノードとなり、本来無関係な企業同士が「ニュースソース」を介して繋がって見える。
- 分析の歪み: 中心性（Centrality）指標を計算した際、実体のないメディア企業が上位を独占してしまう。

本機能は、AI（GPT-4o-mini）を活用して組織の業界を自動判定し、分析に不要な業界を一括で排除（行列から削除）するために開発されました。

2. 具体的な使い方 (操作手順)
ステップ 1: ファイルの準備
「メイン（関係抽出）」タブで生成された pivot_table.csv（隣接行列ファイル）を手元に用意します。

ステップ 2: 業界フィルタ・タブへ切り替え
画面上部の 「業界フィルタ (ノイズ除去)」 タブをクリックします。

ステップ 3: 情報の入力とアップロード

1. 対象行列ファイル: 先ほどの pivot_table.csv を選択します。
2. OpenAI API Key: 業界判定に使用するAPIキーを入力します。
3. 除外する業界を選択: * デフォルトで 「3. ニュース・PRメディア」 がチェックされています。
    - 研究目的に応じて、「政府機関」や「教育機関」なども自由に除外対象に設定可能です。

ステップ 4: 処理の実行
「▶ フィルタ処理開始」 ボタンをクリックします。
※組織数が多い場合、AIの判定に30秒〜1分ほどかかります。サーバーのログ（PM2 logs）で進捗を確認できます。

ステップ 5: 結果のダウンロード
処理が完了すると 「⬇️ 結果ダウンロード」 ボタンが表示されます。 ダウンロードされる ZIP ファイルには以下の 2 点が含まれます:

- pivot_table_filtered.csv: 指定した業界が削除された、SNA用の「綺麗な」隣接行列。
- organization_list.csv: 全組織の業界判定結果と、今回「排除」されたか「保留」されたかのリスト。

3. 推奨される活用例

- 純粋な競合分析: 「金融」「メディア」「政府」をすべて除外。
- 産学連携の研究: 「教育・研究機関」と「製造業」のみを保留し、他をすべて除外。