2017年6月30日

株式会社リクルートホールディングス

リクルート人工知能研究所、データ統合および準備のオープンソースエコシステム「BigGorilla」を提供開始

株式会社リクルートホールディングス（本社：東京都千代田区、代表取締役社長　兼　CEO：峰岸真澄、以下リクルート）は、当社の人工知能研究所であるRecruit Institute of Technology(以下RIT)より、データサイエンティストの分析に必要なデータ準備作業を削減できるPythonベースのデータ統合および準備のオープンソースエコシステム"BigGorilla"をリリースいたしました。

1．本件の背景と目的

ビジネスシーンにおいて人工知能を活用して価値あるインサイト(洞察)を導き出すには、質の高いデータが必要です。今日、データサイエンティストは、質の高いデータを準備するにあたり、様々な情報源からデータを収集し、非構造化データから構造化データを取り出す作業を行っています。また、データから価値あるインサイトを取り出すために、彼らが普段使うアルゴリズムで利用できるようにデータを整備したり、種々雑多なデータを統合しています。ある専門家によれば「データサイエンティストは有用な知恵を見つけ出す前に、彼らの時間の50-80%は手のかかるデータ収集や準備に費やしている」とも言われています。

BigGorillaは、データサイエンティストが手のかかるデータ準備に費やす時間を減らし、データ分析の最も本質的な部分に注力できるよう支援することを目的に開発されました。

2．BigGorillaの概要

BigGorillaは、 RITにてDirector of Researchを務めるWang-Chiew TanとWisconsin大学のAnHai Doan教授とともに2016年9月にプロジェクトが開始されました。RITとDoan教授は、データ統合と準備にまつわる様々なタスクに注力したオープンソースエコシステムの構築を目標にしています。BigGorillaのプロジェクト名は、データ統合と準備作業が膨大で厄介で扱いにくい問題であり、それがゴリラの毛質に似ていることに由来しています。データサイエンティストによるデータ統合と準備作業の様々なプロセスに対して、BigGorillaのサイトでは既存の技術や今後コミュニティによって開発されうるべき技術を紹介しています。

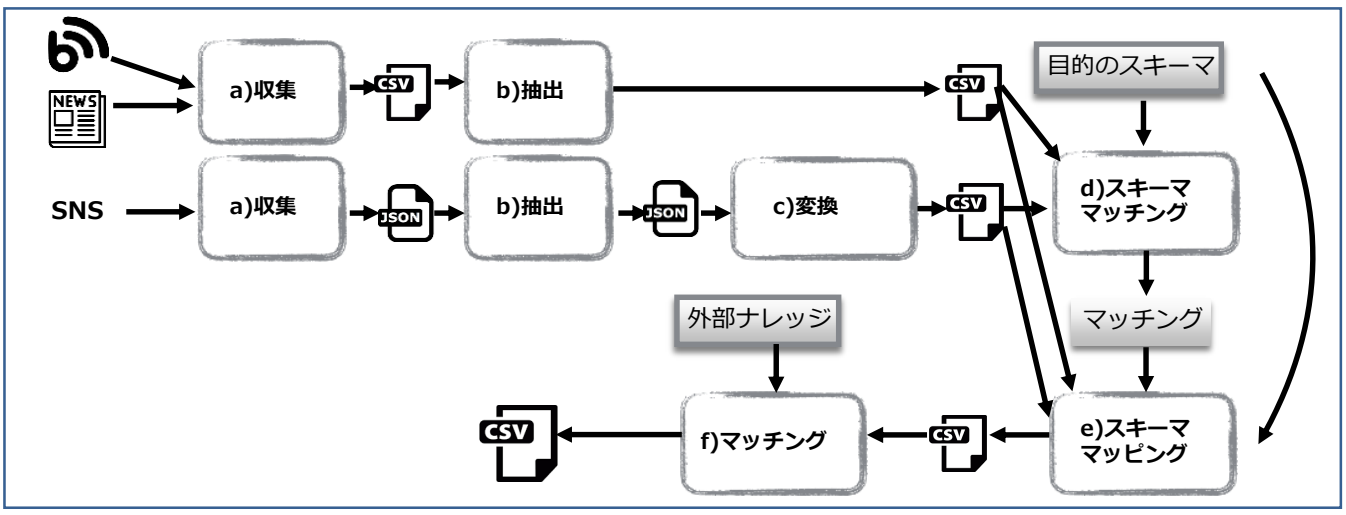
▼BigGorillaのコンポーネント

a) データ収集：外部ソース(Webページ、SNSのつぶやきなど)からデータを収集またはスクレイピング
b) データ抽出：非構造化テキストや自然言語のフリーテキストから人名や組織名などの属性・値のペアを抽出
c) データ変換：データ形式の変換(例. CSVからJSON)や特定の形に再構築または整理
d) スキーママッチング・マージング：2つのスキーマ間で異なる属性をマッチングし2つのスキーマを1つにマージ
e) スキーママッピング：(スキーママッチングから)特定のデータフォーマットに紐付けるコードを生成
f) データマッチングとマージング：2つのエンティティを同一のエンティティと特定、および衝突時の解消

※現在、RITはKOKOおよびFlexMatcherというパッケージにて上記b)およびd)を開発しており、Doan教授のチームではMagellanというパッケージにてf)を開発しています。
※各コンポーネントの詳細は以下公式サイトにて解説

BigGorilla公式サイト(日本語)：http://www.biggorilla.org/ja/

▼ BigGorillaのワークフロー



3． BigGorillaのユースケースとその効果

現在BigGorillaは、リクルートグループ内の8社12グループにて既に利用開始ならびに利用検討が行われています。

実証実験中のユースケースとして、当社が持つ多岐に渡る事業領域において、店舗名の表記揺れの防止、社名や物件名の名寄せ、レセプトデータの変換、複数データソースからのリスト統合、フリーテキストからの店舗名や人名・場所情報の取り出し、Webページからの項目抽出といった様々な業務における有効性が確認されています。例えば、検証結果として、約10,000件の店舗名の名寄せタスクにおいてカバー率98.9%という高い精度を実現しており、これは既存の類似サービスと比較してもトップクラスを誇ります。その他にも、10万店舗分の名寄せ作業を30分程度で完了したり、Webサイトのクローリングによって特定領域の店舗情報を抽出するタスクに関しても、約7割の成功率を実現しています。特に、このWebサイトから特定の情報を探し出す作業においては、人手で作業していた時と比較して工数を最大98%削減することに成功しています。

これらの結果を受けて、BigGorillaは、本来は時間を要する質の高いデータを得るための統合や準備作業を、わずかな労力と時間で実現できることが期待できます。企業のコスト削減に貢献するだけでなく、データサイエンティストが本来解くべき企業の本質的な課題に注力してデータから有用なインサイトを見つけ出すことで、企業における重要な意思決定を加速させることもできます。

RITは今後も、データ統合や準備のプロセス全般に対してツールを開発提供していきます。現場主導でのデータ分析を加速させ、営業員やビジネス企画者がより多くの仮説検証をスピーディーに実施できる環境作りに寄与していくことで、企業の成長に寄与していきます。

4. 今後の展望

BigGorillaはオープンソースコミュニティへ価値貢献し、様々な大学や組織との協業を行いながら発展していきます。現在開発中のツールに関しては、2017年度内を目処に順次オープンソースで公開していく予定です。当オープンソースプロジェクトへの参画希望の方は、下記問い合わせ先までご連絡ください。

▼Big Gorillaのご利用に関するお問い合わせ窓口
thebiggorilla.team@gmail.com

リクルートホールディングスではこれからも、働く、学ぶ、住む、結婚、育児、旅、車、趣味や暮らし情報など、
さまざまな場面でユーザーが新しい発見・機会創出できるサービスを提供し、
一人ひとりにあった「まだ、ここにない、出会い。」を届けることを目指していきます。

【本件に関するお問い合わせ先】
https://www.recruit.jp/support/form/