アンケートデータ活用のための
テキストマイニング環境の実装

2019年05月

イノベーション推進部

志賀野　芳宏

PLANET, INC.

１．はじめに

「大量のテキストデータ全体から、関連のある情報を探索する」というタスクが、ビックデータ活用のシーンでは重要となっている。Fromプラネット編集のために収集した大量のアンケートデータと機械学習技術を用いたテキスト分析手法を活用し、分析手法と表現手法について、アプリケーションの実装を通した、効果の検証を行う。

２．提案手法

２．１プロセスの流れ

「大量のテキストデータから関連のある情報を探索する」タスクを、効率的に実行するためのプロセスを下記のように想定した。

（１）アンケートデータ（いわゆる生データ）をデータベースにインポートする。

・質問、回答の相互検索性と、集計処理の容易性を確保する。
・アプリケーション側とのシステム的なつながりを確保する。
・この処理は可能な限り自動で実行されることが望ましい。

（２）情報の関連性をデータに対して付加する。

・情報の属性や活用法が多岐にわたることが想定される。さまざまなテキストマイニング手法を並列でき、任意のタイミングで追加できることが望ましい。

（３）フロント側から必要な検索クエリを作成し、情報を取得する

・フロントとDBとの接続は可能な限り、疎結合（マイクロサービス）の状態であることが望ましい。（可用性の確保）
・フロント開発においては、アンケートデータベース以外のソースとの接続性も考慮する。

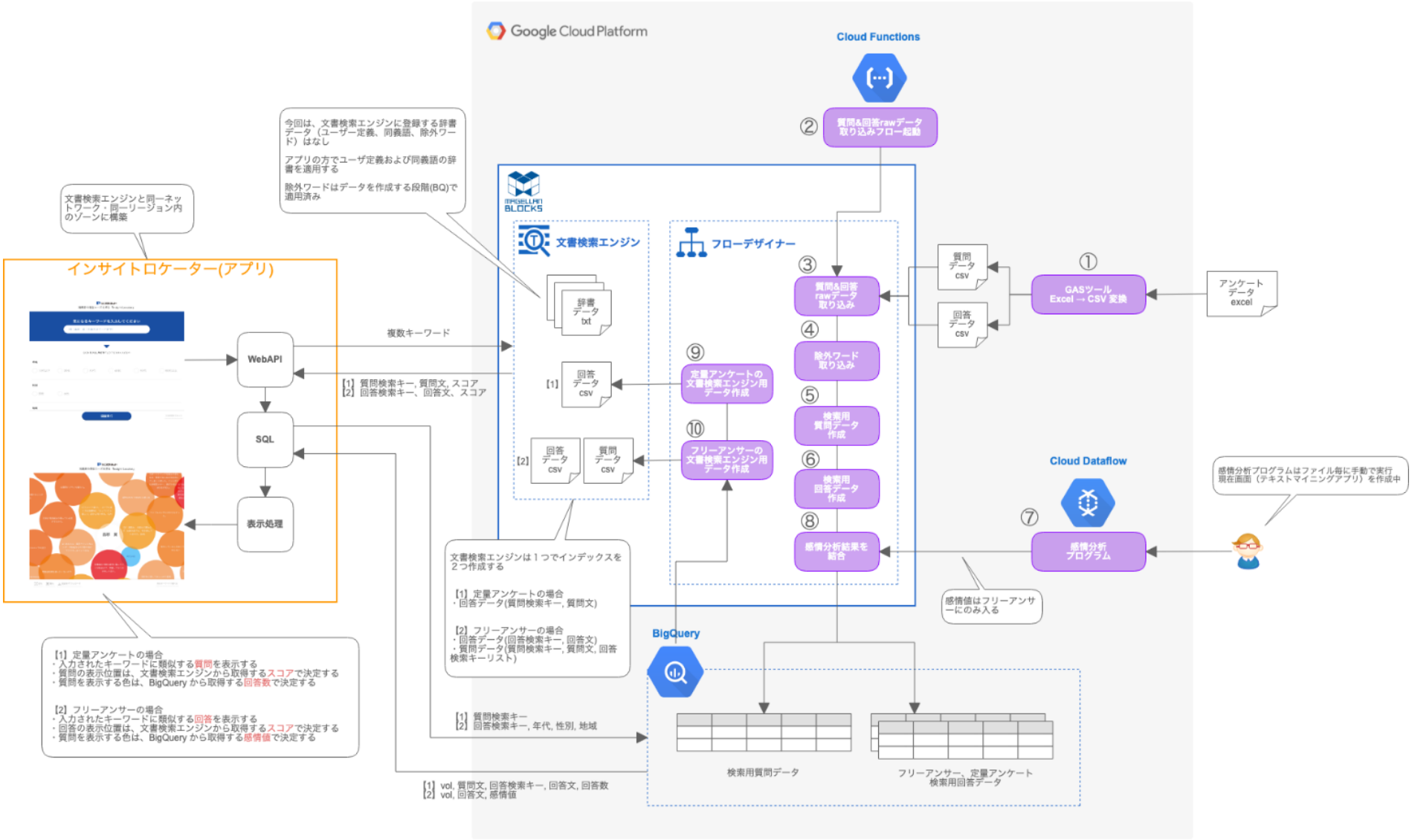
２．２本実装における技術的特徴

今回の実験システムは、大きく2点のシステム的特徴を持って開発した。

・クラウドコンピューティングを活用し、機能間を疎結合（マイクロサービス）とすること。
・可能な限りサーバーレスアーキテクチャ（独自サーバを持たない構成）を採用すること。

テキストデータを情報学（数学）的に解析する手法は、現在も研究が進められており、コンピュータの進化に伴い、より高度化することが期待できる。機密性、可用性への配慮は必要だが、これらの特徴により、適切な拡張性を確保しながら、データ解析を進める環境が実現できる。

３．１システム構成



PLANET, INC.

３．２主なコンポーネントの提供する機能と役割

機能名称	提供する役割
GASツール（Google App Script）	アンケートデータをExcel形式からCSV形式へ変換し、質問データと回答データへ切り分ける。
データ取り込みと検索用データ作成 （Magellan Blocksフローデザイナー）	GASが処理したCSVデータをBigQueryにインポートする。 今回は、付加機能として、無効な回答（フリーアンサーが「nashi」などの意味のない文字列の場合）をインポートしない機能も実装した。 文字列の意味についても、機械学習を用いて判別する。
感情分析プログラム （Google Natural Language API）	テキストのブロック内で示されている全体的な意見、感想、態度の感情を数値化し、回答テーブルに追加する。
文書検索エンジン （Magellan Blocks 文書検索エンジン）	フロントからの検索キーワードを受け取り、BQ検索に必要なキーを返す。
フロント側Webアプリケーション （Google App Engine）	web画面と文書検索エンジンへのキーワード提供を行う。レスポンシブ対応を行う。

３．３本環境が提供する価値

本環境で実現した機能と価値は下記である。

（１）Fromプラネットのアンケート「Vol.1～80」までのアンケート結果の中から、ユーザーが入力したキーワードに類似する回答を表示する（属性選択も可能）

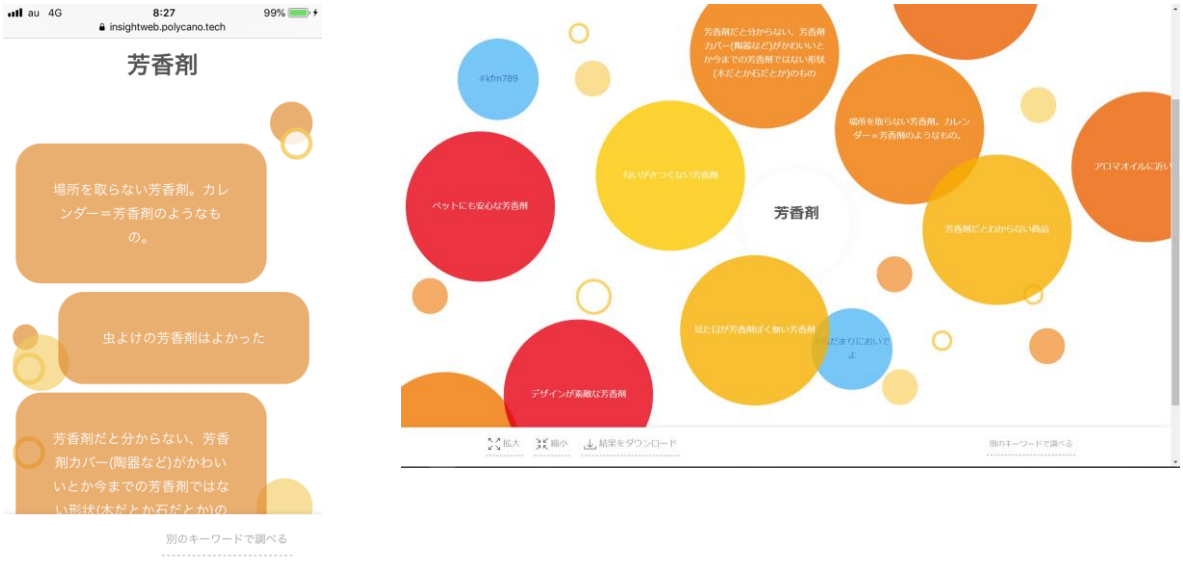
（２）フリーアンサーの感情値の大きさに合わせて、表示位置を調整する

・パソコン版の場合：感情値が高い方が、中央に近い

・スマートフォン版の場合：感情値が高い方が上位に表示される。

（３）検索結果をExcel形式でダウンロード可能。（PC版のみ）

（４）汎用的なフロントエンドとテキストマニング手法への展開可能性



図：「芳香剤」で検索した結果。左がスマートフォン版、右がPC版

４．おわりに

本環境の実証を通じて、下記のような考察とまとめを得た。

４．１考察

（１）利用するツール特性に合わせた追加パラメータの重要性

感情値を設定するために使用した「Google Natural Language API」の特徴として、長文については、感情値が高めに出力される傾向があることがわかった。
フリーアンサーように、文章の長さにばらつきがある場合は、追加のパラメータを採用するなどを検討し、短い文章の価値を高める工夫が必要である。

（２）文章間の類似度におけるパラメータの追加

例えば、「洗剤」というキーワードにおいて、最も近い位置に、
「しゃぶしゃぶの〆はきしめんで昆布だしで食べる事が多く、残った具材は灰汁など取り除いてから豚汁に姿を変えて何度も美味しく戴きます。」
が表示される。

今回アプリケーション側で、「洗剤」というキーワードを複数の検索キーワードに「拡げる」処理を追加している。（検索結果のバラエティ性を向上させるため）
→　洗剤, クリーナー, クレンザー, 中性洗剤, 洗浄剤, 漂白剤, 洗剤, 灰汁

灰汁というキーワードが発生し、結果が戻った後に、検索キーワードとの「関連性」を比較するプロセスが必要である。

（３）検索キーワードのコンテキスト（文脈）軸での表示

文書検索エンジンの背景技術として使われている、文書解析技術を単体で活用した場合、形容詞は単語として認識される。結果、検索キーワードに形容詞を投入した場合、「優しい○○」や「××は柔らかい」等が検索結果として表示される。
一方で、人間が「優しい」や「柔らかい」といったキーワードで検索をしたい場合は、表現的に「優しい」意見の集合を求めている場合がある。
ユーザー側の文脈に寄り添った、分析軸の追加が必要である。

４．２まとめ

本環境で利用対象とした自由回答数は「394,933」であった。

これだけの量の文章を一つ一つ確認していくことは非効率であり、アンケートのカテゴリや属性（性別、年代、職業など）を選択し、回答の方向性を絞ることで、「定量アンケート結果の補足資料」としてこれまでは活用されてきた。

一方で、文章要約やテキストマイニング技術の発達によって、単語の「距離」や「演算」など、テキストを数学的に扱うことも可能になってきた。

本環境では、形態素解析（分かち書き）をベースとした検索技術を中心に実証実験したが、解析ツールの拡張によって、「このアンケートに回答している集団の意識」についても、より理解ができると考えている。

以上