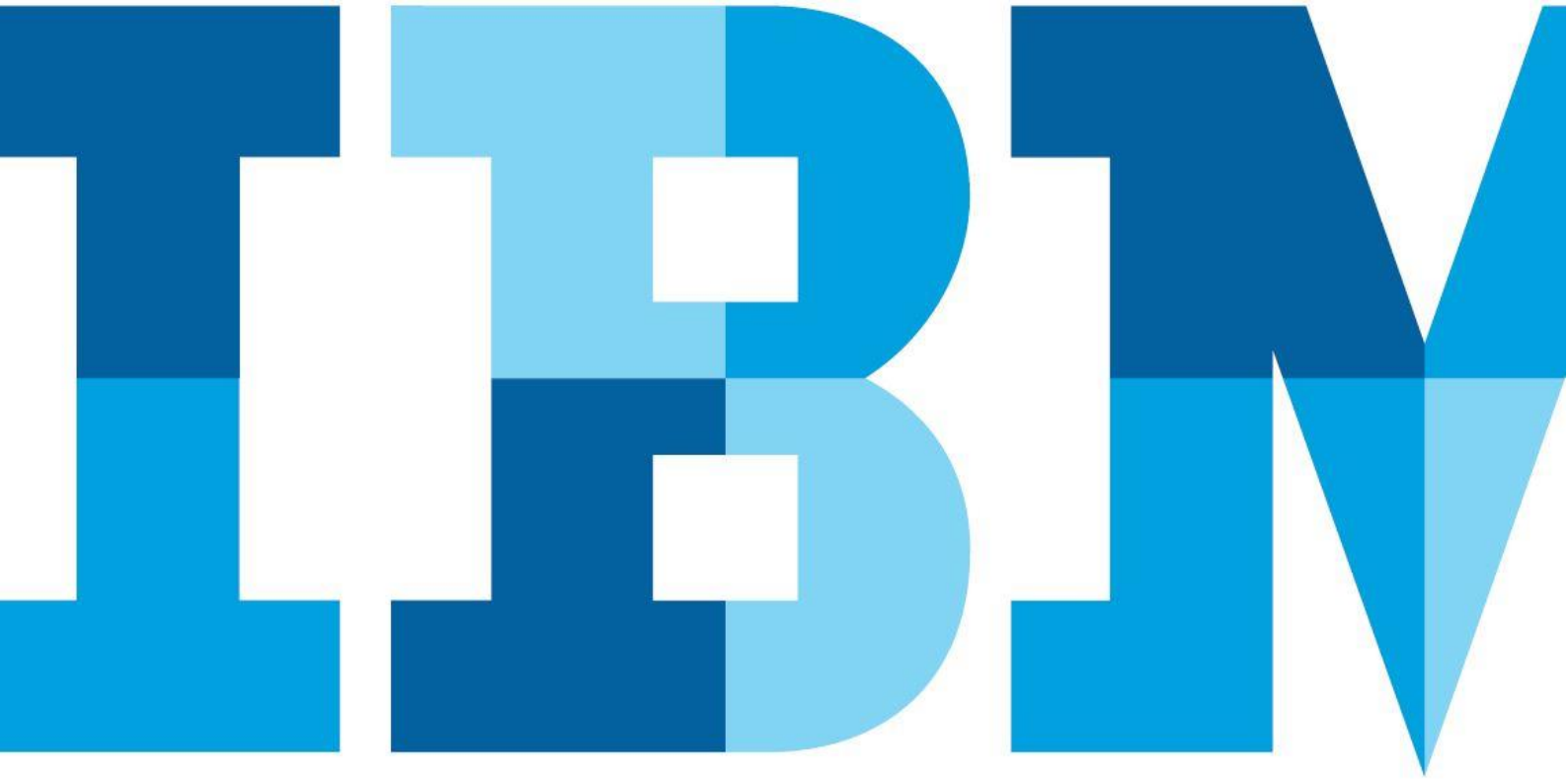
IBM Watson
ソート・リーダーシップ・ホワイト・ペーパー

テキスト分析：
エンタープライズ・データ・イニシアチブの重要な役割

非構造化データに隠された価値をテキスト分析サービスによって引き出す



IBM

エグゼクティブ・サマリー

- 専門家によると、フォーチュン 500 の企業は、テキストなどの非構造化データを活用していないために毎年 120 億ドルの価値を失っていると推定されています。
- 賢い組織は、競合企業の多くが見逃している機会を活用しています。ビッグデータへの取り組みの一環として、テキスト分析ソフトウェアを利用して非構造化データを処理することで、データを「再利用」しています。
- テキスト分析ソフトウェアは、非構造化データに隠された価値を引き出し、企業の意思決定者がブランドの資産価値や収益を高め、運用コストを削減できるよう支援します。
- テキスト分析に価値があることに気付き、テキスト分析ソフトウェアの使用を検討する組織は、それを購入するか独自に構築するかのジレンマに陥ります。
- IBM Watson などの完成された堅固なテキスト分析ソフトウェアの購入には、低コストなうえ迅速な市場投入を実現できるといった多数の利点があります。

テキスト分析とは

テキスト分析は、テキストに構造を追加するプロセスです。データはいったん構造化されると、抽出してビジネス・インテリジェンスに変換できます。一般的にテキスト分析という場合は自動テキスト分析を指しますが、それは手動で行うコストと時間のかかる方法とは対照的です。手動による方法は小規模の場合に有益ですが、エンタープライズ規模のデータ・セットには応用できません。

例えば、iPhone 5 へのメディアの反応を把握したいメディア分析企業を考えてみましょう。「iPhone 5」というキーワードを含むブログ投稿が 24 時間ごとに何百件も書き込まれます。この企業は業務の一環として以下を把握する必要があります。

- これらのブログ投稿のうち、iPhone 5 に関連があり、実際にそれについての意見を述べているのはどれくらいか。
- これらのブログ投稿のうち、単に「iPhone 5」という語が含まれているだけの投稿はどれくらいか。
- 関連するブログ投稿で否定的なものと肯定的なものはそれぞれどれくらいか。
- iPhone 5 の機能固有の外観や特徴で評判が良い点または悪いと思われる点は何か (価格や画面サイズなど)。
- 上記すべてに関する過去 6 カ月間の傾向はどのようなものか。

このメディア分析企業が過去 6 カ月間の一日あたり何百件というブログ投稿を人手によって把握するとしたら、多大なコストがかかります。テキスト分析では、その作業をコンピューターが行います。IBM Watson は、言語、統計、ニューラル・ネットワークの複雑なアルゴリズムを活用して超高速でテキストを読み、理解します。IBM Watson を利用することで、このメディア分析企業は上記の質問に答えるレポートを容易に作成できます。

図 1 は未加工の入力形式であるニュース記事の例を示し、表 1 はテキスト分析後の構造化された出力を示します。


The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

図 1: 未加工の入力形式であるニュース記事の例 [1]

記者	テキスト属性	値	評判
Walter S. Mossberg	エンティティー	iPhone	肯定的
Walter S. Mossberg	エンティティー	Apple	肯定的
Walter S. Mossberg	エンティティー	Google Maps	肯定的
Walter S. Mossberg	キーワード	price	否定的
Walter S. Mossberg	キーワード	new maps	肯定的
Walter S. Mossberg	キーワード	screen	肯定的

表 1: テキスト分析後の構造化された出力

「我々の手に情報は溢れていますが、知識は不足しています。」
- John Naisbitt、The New York Times ベストセラー作家[2]

テキスト分析を行う前は、元の未加工の入力形式は、上記の質問に容易に答えられるようにはフォーマット化されていません。テキスト分析を行った後は、エンティティー固有の感情スコアなど、データベースに適した構造化データが生成されます。

この新たな構造化データは、既存のレポート作成ソフトウェアやビジネス・インテリジェンス・ソフトウェアでの利用がより容易になります。そして最終レポートを基にした洞察を、PR 企業とその顧客の意思決定に利用できるのです。

テキスト分析のアプリケーション

「ブランドに関する発言、不満、関心に耳を傾けることが、信頼できるソーシャル・エンゲージメント・プログラムの第一の要素です。耳を傾ける企業こそが、販売機会を見いだし、満足度を測定し、マーケティング・キャンペーンやメッセージ・テーマへの反応を評価するとともに、問題の背後にある根本原因を明らかにし、評判や競合企業の脅威を特定してそれに対応できるのです。」[3]

テキスト分析は、ブランド評価管理、市場調査、対競合インテリジェンス、顧客サービスおよびサポートなど、企業内のさまざまなアプリケーションに適用できます。このセクションでは、エンタープライズ・アプリケーションの概要を見ていきましょう。

お客様の声は、お客様のコミュニケーションを分析することで知ることができます。お客様のコミュニケーションには、コール・センターの記録、サポート・サイトへの E メール、アンケート・フォームへの回答などの社内文書も含まれます。

外部サイト上のお客様のコメントを参照することにより、以前では知り得なかったお客様との関係の問題点を明らかにし、傾向を把握し、問題が多発する箇所を判別できます。

顧客の声は、オンライン・レビューから拾うこともできます。外部サイト上のお客様のコメントを参照することにより、以前は知らなかった顧客との関係の問題点を明らかにし、傾向を把握し、問題が多発する箇所を判別できます。市場調査と同様に、顧客が関心を持つ新しいトピックを特定することもできます。

市場調査では、新しい傾向を見極め、新たな市場を発掘することのできるテキスト分析が有益です。これには、組織にとって戦略上興味深いイベントを監視することも含まれます。SemanticWeb.com[4] では、以下の質問をすることが提案されています。

- 競合企業が獲得した新しい取引は何か。
- 業界は安定しているか。
- 会社の幹部は尊敬されているか。また、ニュースでどのように扱われているか (肯定的/否定的)。
- 市場参加者は貴社の競合企業をどのように見ているか。
- 市場の傾向はどのようなものか。
- どの企業どうしが取引しているか。
- 主力社員の離職率はどれくらいか。

組織が販売機会を見いだすことも可能です。例えば、何らかの問題について不満を述べている人たちは、それが貴社の製品で解決できる問題である場合、未開拓のお客様となります。

メディア監視では、お客様かジャーナリストかに関わらず、発言を特定して分析します。発言には、自社、競合企業、自社製品、競合製品に関するものがあります。「X について Twitter、Facebook、ブログ投稿、ブログ・コメントで何と言われているか。どんな言葉が使用されているか。」ソーシャル・メディアの投稿を見れば、特に「自社の車は速いと評価されているか。自社の製品は素晴らしいと思われているか。」といった細かい問題に関するお客様の感情の全体像をつかむことができます。

新製品の発売では、同時に、消費者の反応を把握するためのメディア監視を行う必要があります。特に、新たな問題を迅速に特定し、コストがかさむ前にそれを阻止することが重要なためです。

感情分析は、お客様の声、市場調査、メディア監視などの多くのテキスト分析アプリケーションの中で主要なコンポーネントです。感情分析は以下の特性を備えている必要があります。

- 正確である
- 拡張が容易
- きめ細かい

推奨システムはテキストを分析し、それを関連する記事、ビデオ、広告につなげます。周知のとおり、Google は Web ページ・コンテンツのテキスト分析を利用することで、関連する広告を表示する AdSense ビジネス・モデルを開発しました。The Huffington Post のようなメディアでは、通常、このような推奨システムを利用して関連する記事を表示しています。このようにしてユーザーをサイトに留め、さらにコンテンツがクリックされるようにします。YouTube でも関連動画機能の一部でテキスト分析が使用されています。

履歴的な視点は、履歴データを調査し傾向を明らかにすることで得ることができます。傾向分析はお客様の声、市場調査、メディア監視、感情など、さまざまなテキスト分析アプリケーションで実行できます。「自社のブランド認知は時間と共にどのように進展しているか。肯定的な関心や否定的な関心が急増するのはどのような時か。さらにその理由は。」

主要な PR イベントの後の数分から数時間といった短時間の変動を評価する場合にも、傾向分析を利用できます。IBM Watson など、拡張可能なテキスト分析ソリューションを利用する大きな利点の 1 つに、履歴全体の分析を実行する場合のコスト効果が高く、数秒で作成できるということがあります。

テキスト分析の実行方法

これまで、賢い企業組織で行われる可能性のあるテキスト分析について示し、企業でのテキスト分析の実際の使用例を紹介してきました。テキスト分析はどのように実行されているのでしょうか。以下のセクションでは、テキスト分析の一般的な落とし穴と最良事例を示します。

落とし穴: 単語のままの分析

一般的な落とし穴の 1 つに、単語そのままの情報を使用してテキスト分析を実行しようとする点があります。この方法は単純さゆえに魅力的ですが、コンテキスト情報を取得できず、コンポーネントの信頼性が低くなります。

- 例えば、ある記事が「apple」について言及しているとします。それはテクノロジー企業の Apple の意味でしょうか。それとも食べ物の apple (りんご) の意味でしょうか。この区別は、ブランド認知に関連するアプリケーションでは特に重要です。Vision Critical は次のように説明しています。「多くの企業は、非常に限定的な曖昧性除去の課題に直面しています。それは、一般的な英語 (またはその他の言語) を使用した企業名やブランド名が多く、言葉の指す意味を区別するのが難しくなっているということです。例えば、Avon Cosmetics と Avon Theatre、Avon Indiana、さらにテレビ・キャラクターの Avon Barksdale の Avon の区別などです。[5]
- 単語そのままの情報を基に判別した感情は信頼できません。例えば、否定的な表現 (「この映画は良くなかった」) は単語そのままの情報に対しては検出できません。

信頼できるテキスト分析には、コンテキストを認識する言語および統計アルゴリズムを含む高度な技術が必要です。

信頼できるコンポーネントを構築するのは困難です。隠れた多額のコストを伴い、総所有コストが予測不能なほど高くなる可能性があります。

落とし穴: 自社開発のテキスト分析

テキスト分析コンポーネントを自社で開発する場合、問題が 2 つあります。

1. リスク: 信頼できるコンポーネントを構築するのは困難です。隠れた多額のコストを伴い、総所有コストが予測不能なほど高くなる可能性があります。予測不能なコストに加え、プロジェクトが失敗するリスクもあります。

2. 成果が出るまでの時間: 信頼できる高品質なテキスト分析コンポーネントを構築するには、最低でも数カ月から数年程度かかる場合があります。

テキスト分析コンポーネントを構築する上でのリスクと成果が出るまでの時間に関する課題は、以下のように分類できます。

- **技術的な専門知識:** 多くの場合、自然言語処理 (NLP) の専門家の役割を果たすのは容易ではありません。NLP の専門家はトレーニング・データを手に入れ、モデルを構築し、ソリューションを拡張して最新状態に維持できなければなりません。それには統計、言語、ソフトウェア・エンジニアリングを組み合わせた専門知識が必要となります。マルチリンガル・システムの構築に関する知識が必要な場合に、その知識やスキルがないこともあります。豊富な資金を持つ組織であっても、必要な NLP の人材を雇用するのは非常に困難です。
- **構築時間:** NLP システムの構築には、最低でも数ヶ月から数年もかかることがあります。例えば Twitter 上のテキストは、カジュアルな表現、つづりの誤り、スラング、正しい文法でない場合には特に困難です。コーディング時間に加え、コンポーネントをトレーニングする時間も必要になります。最高水準の正確さに到達するまでには、多くのモデルごとに数週間か数カ月のトレーニング期間が必要となります。
- **スケーラビリティー:** 高速で大量のテキストを処理するために、拡張性の高いシステムである必要があります。使用要件が予測できない場合もあります。大きなイベントではソーシャル・メディアの活動が急増する要因となります。要求に応じてリソースを拡張し、その後分析要求が低下したら縮小するといった対応は適切な運用が困難になります。
- **メンテナンス:** システムは稼働し続ける必要があります。理想的には、システムを定期的に再トレーニングするべきです。なぜなら言語は進化しており、私たちの語彙には新しい単語や用語が加わるからです。現在稼働中の信頼できるシステムでも、1 年後にはその成果の信頼性が下がっている可能性があります。

上記の問題によって、テキスト分析コンポーネントの社内開発には潜在的リスクが加わります。複合的なリスクによりプロジェクトの予算超過や期間延長が発生したり、失敗に終わってしまう可能性があります。

多くの組織にとって、NLP コンポーネントの開発と維持はコア・コンピテンシーでもなければ、必須事項でもありません。独自のテキスト分析コンポーネントを構築し維持しようとする場合には、必要な労力と総所有コストを徹底的に評価してください。

最良事例: 最善の既存コンポーネントの利用

テキスト分析を実装する最も良い方法は、定評のある業界リーダーの最善のコンポーネントを取り入れることです。ゼロから構築することに比べ、既存コンポーネントを購入することの総合的な利点は、一般的にコストの透明性が高く、成果が出るまでの期間が短縮できることです。

API を利用すると、テキスト分析を取り入れるプロセスを簡素化し、より少ない初期投資でさまざまなソフトウェアへのアクセス権を獲得することができます。

既存のテキスト分析コンポーネントを利用する特に魅力的な方法の 1 つに、使用量に応じて課金されるアプリケーション・プログラミング・インターフェース (API) の利用があります。API を利用すると、テキスト分析を取り入れるプロセスを簡素化し、より少ない初期投資でさまざまなソフトウェアへのアクセス権を獲得することができます。

ソフトウェア・パッケージを購入するのではなく、API を使用する最善のソリューションを購入する利点は以下のとおりです。

- **使い易さ:** API は接続が比較的簡単で、即実行可能です。
- **クリアな価格構造:** 使用量に基づいて費用が課金され、あらかじめ費用が分かります。初期費用もメンテナンス費用もありません。API を利用すると、資本支出を運用支出に移行し、大規模な先行投資に必要なキャッシュ・フローを回避できます。これに比べ、ソフトウェアの購入には稼動するハードウェア、インストール、メンテナンスといった隠れたコストが含まれることがあります。

API は柔軟性が高いうえ、比較的使いやすく初期コストも低いため、最もリスクの低い選択肢といえます。また API を利用すると、企業は大規模投資を行うことなく迅速に価値を実現できます。

タイミングと実行速度が重要です。API を利用すると、企業は API を利用しなければ不可能だった、短期のプロジェクトに取り組むことができます。例えば、お客様がドイツ語のテキスト分析を必要とする場合、数カ月後にまた来るよう伝えなければならないということもありえますが、API を使用すればすぐに支援することができます。

ある組織が逸した機会が別の組織の競争優位となることがあります。テキスト分析ソフトウェアを利用して非構造化データの可能性を引き出すことで、スマートな組織は競合企業が実現できない価値を実現できます。

IBM Watson の利点

規模上の利点

IBM Watson トレーニング・セットの規模は Wikipedia の 250 倍です。IBM Watson は、何億ものツイートを含む何百億もの Web ページをクロールします。最新のクロールでは、そのデータをすべて収集して分析するために千台を超えるコンピューターが使用されています。この規模は膨大で、これが IBM Watson 製品の汎用性につながっています。これは多くの企業の通常の運用能力を超えています。

言語に対して継続的に生じる変化に対応できる唯一の方法は、ソーシャル・メディアと Web の常時分析です。言語は進化しており、新しい単語、イディオム、製品名が日々生まれます。モデルを常に最新に維持するには、Web を繰り返しクロールし続ける必要があります。IBM Watson は毎月何百億もの Web ページを再クロールするため、Web 上のテキスト・データの量だけでなく、増加量についても規模上の利点となっています。

クロスドメインの利点

多くのテキスト分析 API は、非常に狭い分野に重点を置くことを選択しています。例えば、主にインテリジェンス・コミュニティーや反テロリズムに重点を置くテキスト分析 API があります。その一方で、お客様の声のみに重点を置くテキスト分析 API もあります。これに比べ、IBM Watson はさまざまなタイプや分野のテキストを対象とします。お客様基盤は広く、10 以上の異なる業界に及びます。これは、コンテキスト広告からビジネス・インテリジェンスや金融サービス・アプリケーションに至るほぼすべてになります。さらに、IBM Watson のシステムには自己改善機能が組み込まれており、それにより狭義のテキスト分析システムを超えた汎用化が可能となっています。

人の利点

テキスト分析ソフトウェアの作成には膨大な工数が必要となりますが、それは始めにラベル付けされたトレーニング・データにアノテーション(注釈)を付けるために利用されます。このトレーニング・データは、ソフトウェアが一定の正確さを達成できるようにするために使用されます。多くの場合、これらの人間のアノテーターには言語学の学位や相当量のトレーニングなど、非常に高度な素養が必要となります。トレーニング・データを作成するのに人材を雇用すると高いコストがかかりますが、テキスト分析システムを独自に実現させるためには必須です。IBM Watson では、アノテーションの作成と改良に多数の人年を費やしてきました。このことは、企業が独自のテキスト分析ソフトウェアを構築したいと考える場合に克服すべき高額な障害の可能性と、既存システムを利用するもう一つの利点を示しています。

展望

多くのエンタープライズ・データ・イニシアチブでは、既に構造化されているデータの活用に重点が置かれていますが、ビッグデータの可能性の多くは、構造化されていない 80% の中にあります。ただし、非構造化データは使いにくいという点があります。ある組織が逸した機会が別の組織の競争優位となりえます。テキスト分析ソフトウェアを利用して非構造化データの可能性を引き出すことで、スマートな組織は、競合企業が実現できない価値を実現できます。そのような組織は組織内外にある非構造化データを活用しています。

非構造化データの隠された価値を引き出す第一歩は、テキスト分析を利用した構造の追加です。テキスト分析により非構造化データの情報が明らかになり、データベース・ツールやビジネス・インテリジェンス・ツールといったより一般的なエンタープライズ・データ・ソフトウェアからアクセス可能となります。テキスト分析は、ブランド評価管理、市場調査、対競合インテリジェンス、顧客サービスおよびサポートなど、企業内のさまざまなアプリケーションに適用できます。

テキスト分析を実装する最も良い方法の 1 つに、定評ある企業の最善のコンポーネントを購入することがあります。特に IBM Watson テキスト分析ソフトウェアは柔軟性が高く、比較的使いやすく、低コストです。IBM Watson は規模上の利点、クロスドメインの利点、人の利点を備え、これらの 3 点により多くの競合企業に勝る高い精度を実現しています。 未加工のテキストを有用な情報に変換することで、IBM Watson は意思決定者が業界最先端の洞察を手に入れ、競争上の優位を実現できるよう支援します。

Joseph Turian について

Joseph Turian, Ph.D. は、データ・サイエンス、NLP、機械学習のコンサルティングを行う MetaOptimize LLC の代表取締役です。また、機械学習や自然言語処理の専門家が知識を共有する MetaOptimize Q&A サイトも運営しています。専門は大規模データ・セットです。

Joseph Turian は科学者として、14 を超える査読付論文を NLP + ML のトップ・コンファレンスで発表しています。彼のチームが出品したパーサーは、EVALITA 2009 Main+Pilot タスクで最優秀パーサーとなりました。オープン・ノートブック・サイエンスの支持者で、自身の GitHub でリサーチ・コードを公開し、インターネットを通した科学面での幅広いコラボレーションに取り組んでいます。

IBM Watson について

2014 年 1 月、IBM は、クラウドで提供するコグニティブ・コンピューティング・テクノロジーの開発と商用化に特化したビジネス、IBM Watson ユニットを立ち上げました。この動きは、学習によって向上し、膨大なビッグデータから洞察を引き出す新しいクラスのソフトウェア、サービス、アプリケーションの提供に向けた IBM の戦略的変更を示すものです。IBM Watson の詳細については、ibm.co/watsonecosystem をご覧ください。

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
December 2015

[1]「The iPhone Takes to the Big Screen」Mossberg, W.S.、Wallstreet Journal、2012 年 9 月 20 日、
http://www.wsj.com/articles/SB10000872396390444450004578004370248427736

[2]「Megatrends: Ten New Directions Transforming Our Lives」Naisbitt, J. 著、1982 年 10 月 27 日

[3]「Six Types of Sentiment Analysis, and a Look Ahead; Breakthrough Analysis;」Grimes, S. 著、2012 年 9 月 10 日、Alta Plana Corporation、
http://breakthroughanalysis.com/tag/

[4] SemanticWeb.com:
http://www.dataversity.net/category/data-topics/smart-data-data-topics/

[5]「5 Ways Text Analytics Can Enrich Your Research」(ブログ) Shulman, S、Vision Critical Communications, Inc.、2013 年 1 月 21 日、
https://www.visioncritical.com/5-ways-text-analytics-can-enrich-your-research/

Please Recycle

YTW03423-JPJA-00