



Analytics Design Lab

ビッグデータ活用展2019 spring
株式会社NTTデータ数理システム ブース内セミナー

ビッグデータのテキストマイニングに欠かせないAI技術

トピックモデルの適用と新たな応用技術

株式会社アナリティクスデザインラボ
代表取締役 野守耕爾

2019年5月8日、10日

人工知能技術を応用したデータ分析の研究開発とビジネスコンサルティングの経験を活かし、2017年6月にデータ活用コンサルティングの新会社を設立しました

株式会社アナリティクスデザインラボ

企業におけるデータ活用を支援するコンサルティング会社です。



データというスタートから課題の解決というゴールまでをいかにつなげばよいのか、どのようなデータ処理、分析手法、考察、アクションを検討していけばよいのか、というデータ活用するプロセスを企業の抱える課題や思惑・事情などに応じてしっかりとデザインし、それを実行することで企業の課題解決を支援します。

設立	2017年6月1日
事業内容	● 企業におけるデータ活用のコンサルティング ● データ分析技術の研究開発
資本金	5,000,000円
所在地	東京都中野区東中野1-58-8-204

野守 耕爾



- 2012年3月
早稲田大学大学院 創造理工学研究科
経営システム工学専攻 博士課程修了
博士(工学)
 - 人間行動の計算モデルの開発を研究
- 2012年4月～(技術研修生としては2008年～)
独立行政法人産業技術総合研究所
デジタルヒューマン工学研究センター 入所
 - センシング技術を応用した子どもの行動計測と人工知能技術を応用した行動の確率モデルの開発を研究
- 2012年12月～
デロイトトーマツグループ 有限責任監査法人トーマツ
デロイトアナリティクス 入所
 - データサイエンティストとしてビッグデータを活用したビジネスコンサルティング及び分析技術の研究開発に従事
- 2017年6月～
株式会社アナリティクスデザインラボ 設立

ビッグデータのテキストマイニングに効果を発揮するAI技術

トピックモデルとは

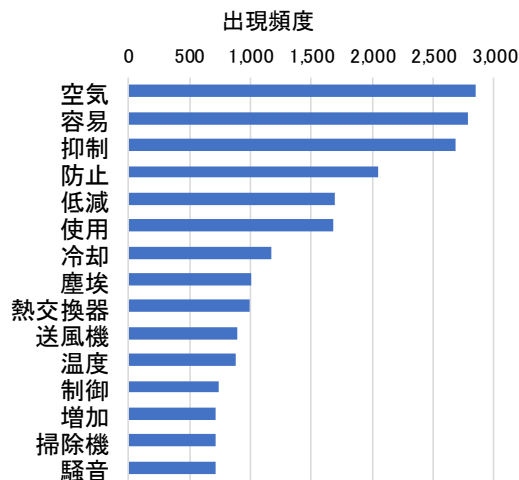
従来のテキストマイニングと課題

従来のテキストマイニングは単語をベースに全体像を把握しますが、膨大な単語で可視化される結果は複雑で解釈が難しいため、単語をグルーピングすることが有効となります

従来のテキストマイニングのアウトプットの例

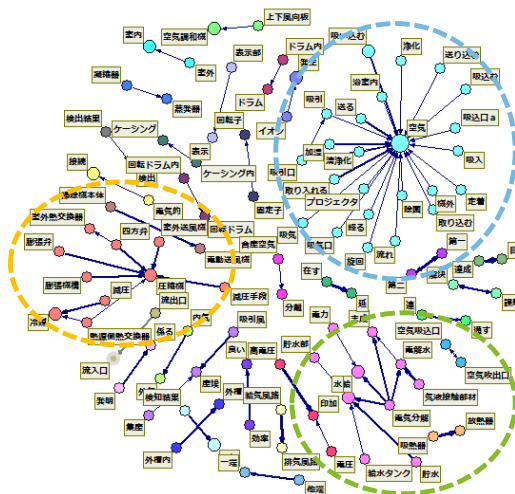
頻度集計

単語や係り受け表現の出現頻度を集計して、どのような記述が多いのか、おおまかな全体像を把握する



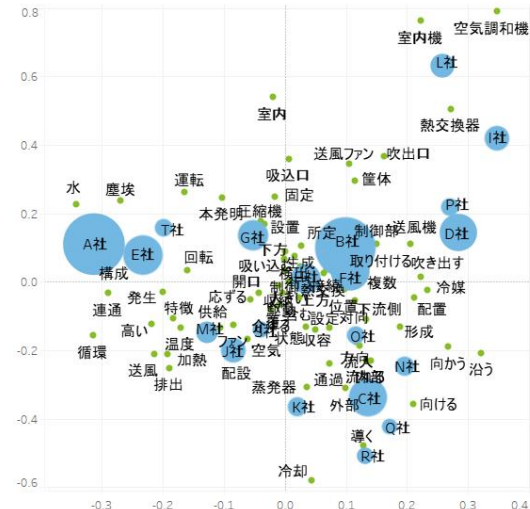
共起ネットワーク

同時に出現しやすい単語同士をネットワークでつなぎ、そのかたまりからどのような話題があるか考察する



コレスポネンス分析

属性情報と出現単語との対応関係を同じ平面上にマッピングし、その位置関係から属性の傾向を把握する



従来のテキストマイニングの課題

結果が複雑で解釈がしづらい

- 単語をベースにした結果は複雑で、解釈が難しい
- 特に読み込むテキストデータがビッグデータになると、マイニングで抽出される単語も膨大となる

解決方法

単語をグルーピングする

- 単語を人がグルーピングすることもあるが、その作業は主観的で作業負荷があまりにも大きい
- 機械的にグルーピングを実現することが望まれる

列数の多い複雑なデータのクラスタリングにはトピックモデルと呼ばれるAI技術が有効です

階層型 クラスタ分析

- Ward法など
- 要素間の距離を計算し、距離の近い要素同士を結合してクラスタを構成していく
- 結合の過程が樹形図で表され、結果を見てからクラスタ数を決められる(ボトムアップ的なクラスタ分析)
- データ数が多くなると計算が膨大となる

非階層型 クラスタ分析

- k-means法など
- あらかじめクラスタ数を決め、そのクラスタ数に全要素を一回でグループ化する
- 各クラスタ(の重心)に対して要素の距離を計算し、距離の近い要素で集められたクラスタとなるように分類結果を調整する
- 階層型クラスタ分析よりも計算量が抑えられる

LSA (Latent Semantic Analysis)

- 特異値分解と呼ばれる
- $(m \times n)$ の行列を、 $(m \times k), (k \times k), (k \times n)$ に分解する
- m 個のデータと n 個の変数を、 k 個の潜在クラスで表現する(クラス数はあらかじめ設定する)
- 大きな値をとりやすいクラスが残る傾向にあるため、各要素は事前に重み付けする必要がある

PLSA (Probabilistic Latent Semantic Analysis)

- LSAを確率的に処理
- LSAのような事前の重み付けは必要がない
- $P(x,y)$ の確率を、 $P(x|z), P(y|z), P(z)$ に分解する
- 行要素 x と列要素 y を、潜在クラス z で表現する(クラス数はあらかじめ設定する)
- 結果は観測データのみから定義され、新規データはクラスで表現できない(過学習)

LDA (Latent Dirichlet Allocation)

- PLSAの拡張手法
- PLSA(他左3つの手法も含め)の過学習の問題に対して、LDAではディレクレ分布を仮定し新規データのクラスを推定できる
- 新規データに対応するため、抽出されるクラスは観測データを忠実に再現するものではなく、クラスの抽象度が高い傾向がある

従来のクラスタ分析

- 一つの要素は必ず一つのクラスタに所属し、重複所属を許さないハードクラスタリングとなる
- 基本的に要素間の距離に基づいて分類を行う
- 列要素の距離に基づいて行要素を分類するか、行要素の距離に基づいて列要素を分類し、行と列どちらか一方を分類する
- 要素数が多くなると要素間の距離が離れていき妥当な結果が得られにくい(次元の呪い)

トピックモデル

- 一つの要素は全てのクラスに所属するソフトクラスタリングで、その所属の重みを計算するため、データが複数の特徴をまたがる場合でも表現できる
- 行の要素と列の要素の背後にある共通する特徴をクラスとして抽出するため、行と列の両方をクラスタリングでき、クラスの持つ情報が多い
- 要素間の距離の近さで分類するのではなく、高次元データの情報をできるだけ保存した形で低次元に変換する次元圧縮手法であるため、要素数が多い複雑なデータにも対応できる

PLSAは、トピックモデルと呼ばれる人工知能技術で、複雑なデータをいくつかの潜在変数で説明するクラスタリング手法として用いられ、複雑な観測情報でも分かりやすく理解できます

PLSAの概要

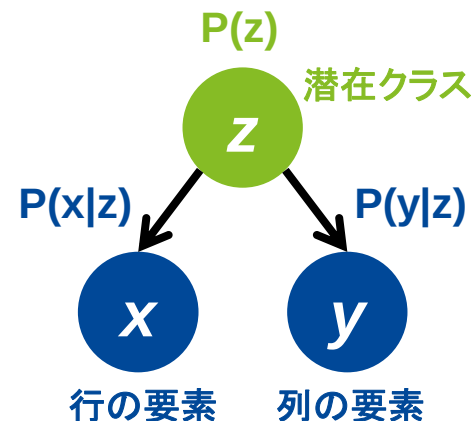
- 行列データの行の要素xと列の要素yの背後にある共通特徴となる潜在クラスzを抽出する手法である
- 元々は文書分類のための手法として開発されている (Hofman, 1999)
- 各文書の出現単語を記録した文書(行) × 単語(列) という高次元(列数の多い)共起行列データに適用することで複数の潜在トピックを抽出し、文書(行) × トピック(列) という低次元データに変換して文書を分類する

文書ID	単語 1	単語 2	単語 3	...	単語 5,014	単語 5,015
1	0	0	1		1	0
2	1	0	1		0	1
...						

文書ID	トピック 1	トピック 2	...	トピック 11
1	0.09%	0.03%		0.04%
2	0.01%	0.12%		0.06%
...				

例えば数千列ある高次元のデータでも十数個の潜在トピックで説明することができる

PLSAのグラフィカルモデル



- $P(z)$, $P(x|z)$, $P(y|z)$ の3つの確率が計算される
- 潜在クラスzの数はあらかじめ設定する

※条件付確率 $P(A|B)$
事象Bが起こる条件下で事象Aの起こる確率

xとyの共起確率を潜在クラスzを使って表現する

$$P(x, y) = \sum_z P(z)P(x|z)P(y|z)$$

PLSAのメリット

行の要素と列の要素を同時にクラスタリングできる

潜在クラスは行の要素と列の要素の2つの軸の変動量に基づいて抽出され、結果も2つの軸の情報から潜在クラスの意味を解釈することができる

ソフトクラスタリングできる

全ての変数が全てのクラスに所属し、その各所属度合いが確率で計算されるため、複数の意味を持つ変数がある場合でも自然と表現できる

PLSAの適用事例

電気自動車関連の特許文書データの分析

「車」「電気」を含む10年分の特許データ26,419件の要約文を対象に、テキストマイニングとPLSAでトピックを抽出し、各トピックの特徴を可視化します

データの抽出条件と分析対象

- 対象
 - 公開特許公報
- キーワード
 - 要約と請求項に「車」と「電気」を含む
- 出願日
 - 2007年1月1日～2016年12月31日
- 抽出方法
 - Patent Integrationを使用
- 抽出件数
 - 26,419件
- 分析対象
 - 要約文のテキスト情報



分析プロセス

テキストマイニング

Text Mining Studio

要約文から単語や係り受けを抽出

トピック化 (PLSA)

Visual Mining Studio

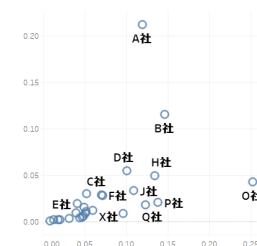
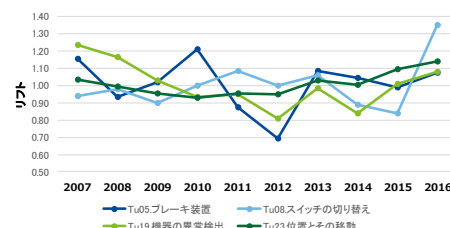
単語や係り受けを複数のトピックに集約

トピックのスコアリング

全データに対する各トピックのスコアを計算

特徴の可視化

トピックのスコアを属性を軸に集計・可視化



トピック抽出のアプローチ

テキストマイニングで単語と係り受け表現を抽出し、単語×係り受けで構成される共起行列にPLSAを適用することで単語と係り受けの出現の背後にある潜在トピックを抽出します

テキストマイニングの実行

要約文に含まれる「単語(名詞)」と「係り受け」を抽出する

単語	頻度
構成	4,997
制御	4,360
配置	3,895
モータ	3,486
形成	3,459
供給	3,309
検出	3,215
電気自動車	3,181
...	...

係り受け表現	頻度
電力⇒供給	1,208
否⇒判定	517
モータ⇒駆動	460
バッテリー⇒充電	440
効率⇒良い	419
供給⇒電力	332
電気自動車⇒提供	285
充電⇒行う	273
...	...

共起行列の作成

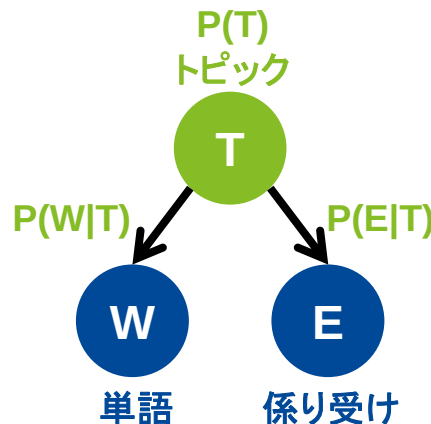
「単語×係り受け」の共起行列(文章単位で同時に出現する頻度のクロス集計表)を作成する

係り受け表現						
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	…	
単語	構成	118	33	36	33	
	制御	268	73	108	85	
	配置	69	2	29	8	
	モータ	239	61	494	58	
	…					

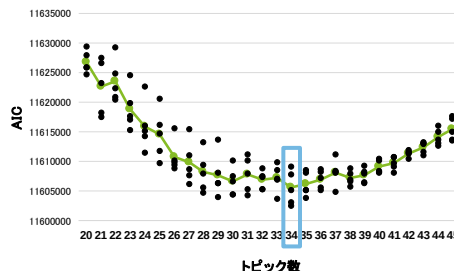
単語(名詞): 3,020語
係り受け: 2,128表現
※頻度20件以上を対象

PLSAの実行

共起行列にPLSAを適用する



トピック数を幅を持たせて設定し、各トピック数に対してPLSAを初期値を変えて5回ずつ実行して情報量基準AICを計算し、AIC最小の解を採用する



トピックの抽出

各トピックについて以下の3つの確率が計算される

- ① $P(T)$
トピックの存在確率
- ② $P(W|T)$
トピックにおける単語の所属確率
- ③ $P(E|T)$
トピックにおける係り受けの所属確率

トピックにおける $P(W|T)$ と $P(E|T)$ からトピックの意味を解釈する

トピック Tu13			
$P(T) = 5.0\%$			
$P(W T)$	単語	$P(E T)$	係り受け
12.6%	充電	5.1%	バッテリー充電
8.9%	電気自動車	4.0%	充電-行う
6.5%	蓄電装置	3.9%	電気自動車-充電
3.0%	バッテリー	1.9%	蓄電池-充電
2.0%	充電システム	1.6%	蓄電装置-充電
2.0%	蓄電池	1.6%	電力-供給
1.9%	電力	1.3%	充電-開始
1.7%	制御	1.2%	電気自動車-接続
1.5%	充電スタンド	1.2%	充電-蓄電装置
1.5%	放電	1.1%	用いる-充電
...	...	...	...

確率の高い構成要素から、Tu13のトピックは「電気自動車の蓄電池充電」に関するトピックと解釈できる

特許トピック34個の一覧①

26,419件の特許は、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、小型化、安全性などの34個のトピックに集約されました

Tu01.エンジンの始動と停止

検出 充電 駆動
運転装置 充電状態 要求 動作 駆動
制御手段 クラッチ 発電機 モータ 走行 制御 判定 内燃機関
停止 ECU ハイブリッド車両 エンジン
動力 制御装置 開始 モータジェネレータ バッテリ
成立 再始動 駆動力 充電

Tu02.動力の伝達

動力
制御 制御装置 ハイブリッド車両 車輪 構成 トルク 駆動部
モータ クラッチ 駆動輪 エンジン
内燃機関 変換 トランスミッション 伝達 連結
入力 入力軸 駆動装置 モータジェネレータ ギア 駆動 出力軸
電気機械 駆動輪 発電機 駆動力 配置

Tu03.モータ駆動

検出 駆動 制御
配速 走行 制御装置 回転 電気自動車 トルク
制御 駆動 構成 駆動力
車輪 バッテリ モータ 駆動
駆動輪 電力 充電機
油圧ポンプ 供給 電気エネルギー 車体 電源
インバータ ハイブリッド車両

Tu04.ロータ・ステータなど回転部品の構成

一体 同軸 対向 回転体 中心 ロータ
ブラッシュ軸方向 配置 固定 周方向 外周 周囲
回転軸 支持 回転 軸 シャフト 回転+できる
固定子巻線 形成 永久磁石 構成 ステータ 収容 モータ
コイル ハウジング 磁石 反対側

Tu05.ブレーキ装置

演算 液圧付与 制御
駆動 マスタシリンダ 操作 操作者 摩耗部材
電気ブレーキ 入力 装置 電気信号
制動力 車輪 ブレーキ 制動トルク 入力 作動
伝達 モータブレーキ 液圧 操作量 検出 運転者
モータシリンダ装置 ブレーキ 操作 固定 調整

Tu06.動作制御

停止 アクチュエータ
駆動調整 回転速度 供給 トルク 給電 充電 充電
入力 制御 制御回路 ECU 速度 制御手段
動作 モータ 制御 制御装置
インバータ 検出 演算
生成 回転 エンジン 電流 温度 制御部 電力

Tu07.動力伝達の制御

変速 駆動 制御 構成 変速時 変更
差動 燃費 変速部 駆動輪 制御 電気式 差動部
エンジン 車両 用動力 伝達装置 制御装置
動力 伝達経路 モータ 車両 用駆動装置 運転状態
変化 回転部材 ハイブリッド車両 差動機構 差動部 変速 ショック
変速比 回転速度 連結

Tu08.スイッチの切り替え

給電 インバータ 遮断 負荷 電力 直流 直列
制御装置 電流 電気負荷 リレー 検出 車両 用電源装置
並列 バッテリ 電圧 スイッチ コンデンサ
モータ DCインバータ オン 制御 オフ 放電 蓄電装置
スイッチング素子 充電 通電 印加

Tu09.交流・直流の変換

モータ 制御 電圧 バッテリ 検出 負荷 電気自動車 制御装置
電気自動車 直流 電力 変換装置 制御装置
直流電力 インバータ 交流 変換
動作 直流電圧 供給 交流電圧 駆動 入力 変換
コンバータ 蓄電装置 制御部 交流電力
生成 電力変換部

Tu10.エネルギーの変換

生成 風車 水車 水車 供給 走行中
変換 回転 蓄電 回収 蓄電池
エネルギー 電気エネルギー 発電機
蓄電 発電システム 熱エネルギー エンジン
バッテリ 電力 回転 電力 充電 走行 発電 運動エネルギー

Tu11.電池モジュールの提供

一対 並列 隣接
結合 電気接続 隣接 ユニット 構成 冷却 連結
細電池 電極 バッテリ 電池パック パスバー
装置 バッテリケース 位置 配置
電源 電気自動車 形成 相互 収容 電池モジュール

Tu12.二次電池の構成

活物質 収容 積層 充電 形成 層構造 電解質
構成 電解質 電池特性 対向表面 形成 負極 活物質
正極 活物質 二次電池 正極 負極
セパレータ 電極 非水電解質 リチウムイオン電池
非水電解質電池 集電体 配置 含有

Tu13.電気自動車の蓄電池充電

蓄電池
放電 取得 情報 構成 充電システム
電力 充電ケーブル 充電+できる 充電 充電 充電
バッテリー 電気自動車 充電 開始 制御部
充電スタンド 外部電源 制御 ユーザ 蓄電装置
充電制御装置 充電制御 供給

Tu14.非接触受電など給電装置

変換 外部 駐車 充電プラグ 給電 送電 受電部
駐車装置 電源 駐車スペース 電気自動車 給電+できる
電力供給システム 制御 停止 受電 給電装置
非接触 充電装置 送電コイル バレット 電力
受電コイル 給電制御部 供給

Tu15.外部への電力供給

蓄電 蓄電池 動作 消費
外部電源装置 外部電源 負荷 放電 モータ 制御装置
発電機 供給+できる 駆動 電源
電気負荷 制御 検出 電力
蓄電 高電圧 バッテリ 電気機器 制御部
エンジン 充電 電圧 バッテリ 蓄電装置

Tu16.空調などの冷却・加熱

暖房 構成 熱 電力 冷却水
車室内 燃料冷却装置 空気 冷媒 通電 排出 制御 冷却システム
燃料電池システム 車両 用空調装置 燃料電池 温度
放熱 加熱 電気ヒータ 温度センサ 流通 圧縮機 冷却
排気ガス 供給 熱交換 循環 ヒータ

Tu17.情報通信

処理 記憶 特定 表示装置 データ 判定
制御 電気自動車 生成 取得 制御装置 検出 表示
信号 制御部 電気信号 情報 受信
入力 表示部 制御信号 記憶部 通信 送信
演算 ユーザ 車載機器 利用者 位置 変換

Tu18.演算・推定

走行 情報 モータ 電圧 方法
予測 閾値 推定 判定 電気自動車 演算
充電状態 充電 検出 時間
修正 取得 値 プログラム 記憶 差 温度 バッテリ
測定 ECU

Tu19.機器の異常検出

信号 異常 検出 停止
異常 電圧 演算 制御部 検出 信号 比較
構成 電流 センサ 印加 温度 電流 センサ
変化 制御装置 異常 電流 判定
検出部 有無 検出 信号 検出 結果
検出手段 検出+できる 故障 制御
ECU 回転角度

Tu20.操作スイッチ

操作 安定 破壊 形成 空気 掃除機 ストップボタン 操作
制御部 スイッチ 検出 検出 検出 検出 検出 検出
車室内 安価 スイッチ 電子機器
自動車 操作 検出 検出 検出 検出 検出 検出
検出 検出 検出 検出 検出 検出

26,419件の特許は、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、小型化、安全性などの34個のトピックに集約されました

Tu21.筐体

一端カバー 取出 取容部 保持 モータ外部開口部
筐体ケース内 形成 電子制御ユニット
電気部品 開口 ハウジングケース 収容
露出 基板 制御回路配置 電気接続部 一体 挿入
装着 コネクタ貫通孔 構成 固定

Tu22.表面の形成

一対外周面 被覆対向凹部 光 面
周囲 外側一体 形状先端発光素子
導電性 突出 形成 接触 位置 表面
構成 基板 本体 端部 部分
配置 反対側 貫通孔 挿入電極
絶縁

Tu23.位置とその移動

直交 許容形成 ロック 支持 係合 方向
解除 保持 係 位置 移動
アクチュエータ 回転 力 連結 駆動
回転 反対側 規制 固定 接触 移動+できる ハンドル
電気信号

Tu24.配置・位置・方向

供給 近傍領域 軸方向 モータ 配置
取容 ハウジング形成 対向隣接 配置
対向隣接 配置 長手方向 方向
流路 反対側 構成 移動+できる 周囲端部 下方
一対

Tu25.構成の方位

方向 車両前後 下面 配設
電気自動車空間位置側面構成連結突出 配置
車輪 車輪 後方下方 開口 上方 一対 支持
先端 形成前方 収容上面 バッテリー 電気掃除機
固定 開口部ケース

Tu26.構成

モータ制御+できる センサ 外部長さ
自動車 生成 設置 構成
電気自動車 充電 変換 制御装置 プロセッサ
電圧 電源 供給 制御部配置
コイル

Tu27.接続

電源ケーブル 車体 固定 検出 保持
位置 嵌合電源プラグ 配線構成 基板 供給 接続部
端子他端 コネクタ 一端 接続+できる
配置端部ケーブルバスバー 外部 形成 ワイヤハーネス
接地 電線 収容 回路 一対

Tu28.方法の提供

配置監視 段階 自動車 工程生成
製造 エネルギー 電気機械調整 方法
内燃機関 動作 存在 電気エネルギー 制御モータ 分離
センサ

Tu29.損傷や浸水など不具合の防止

耐久性 発明 衝撃 電気機器 不具合 静電気 構造 振動
電気接続箱 製造コスト外部 確保+できる 影響
損傷 侵入電動パワーステアリング装置 未然
外力 電気自動車ノイズ 水 温度変化 信頼性
起因衝突破損変動 異音 安全 浸入

Tu30.小型化・簡素化・低コスト化など付加価値

リレー 大型化 自動車 リードフレーム 安価 小型 端子材 実現
確保 低コスト 必要+ないコンパクト 製造コスト車両用灯具
部品点数コスト 小型化 信頼性 構造
簡素化 電気接続箱 耐久性 軽量化 製造方法
コネクタ ワイヤハーネス 作業性削減
強度

Tu31.効率性・安全性の向上

燃費 温度上昇 電源システム 安定 実現
構成 精度 確保 劣化 ハイブリッド車両 バッテリー
安全 正確 電気自動車 モータ 効率
短縮 車両制御装置 消費電力 問題 充電+できる 検出+できる
制御手段 エネルギー効率走行中 運転者
必要+ない技術

Tu32.既存エンジンへの警鐘・樹脂組成物の提供

含有成形品 耐熱性成分 組成物
電気部品 自動車部品 成形品外観 金型離型性電気特性
発電量 理論最良エンジン 電気部品用途
ポリマーレソルフィド 耐衝撃性 後追いエンジン発明阻止
高校大学 機械的強度 既存エンジン
化合物 溶融流動性

Tu33.重力発電の活用による地球温暖化防止

船舶 電気駆動 既存火力原子力発電全廃圧縮空気加速 船舶
落差燃料費ゼロ 垂直下方 全面電化住宅全盛
工場電化全盛 駆動二酸化炭素排気ゼロ 全地球
人間絶滅 大気圧同速度同容積仕事率 既存世界
海水温度上昇ゼロ 先送り 既存蒸気タービン発電
安価 重力加速度加速 重力発電運用 水発電量増大
タービン 重力発電蓄電池駆動 地球温暖化 自動車 発電量

Tu34.タービン発電の出力向上・燃費低減

反転 最大速度水 発電原価 静置 永遠運用改善
横軸自動車 軽量蒸気速度 マッハ狙い 容積圧縮仕事率
安価電気駆動 燃料費ゼロ 太陽光加熱器熱製造
容積 電気+液体空気+過熱蒸気温熱供給設備
宇宙到達費用 空気圧縮液体酸素圧縮駆動 発電量
軽量物発電日帰り旅行 飛行機 製造物全部
燃費 既存蒸気タービン発電 自動車 既存 全動置 船舶
蒸気速度 出力発電

Tu32, Tu33, Tu34は
特定の出願人による
重複した要約内容の
特許から抽出された

※文字の大きさはトピックに対する関係の強さを表現している（上位5つの単語を赤色で表示している）

全特許データに対して各トピックのスコア(該当有無)を計算することで、トピックをベースとした様々な集計・分析を実行することができます

トピックのスコア(該当有無)を紐づけた特許データ

特許ID	出願番号	要約	出願年	出願人	トピック Tu01	トピック Tu02	...	トピック Tu34
1	特願2007-XXXX	【課題】電気式変速操作装...	2007	A社	1	1		0
2	特願2009-XXXX	【課題】従来の電気自動車...	2009	B社	0	1		1
3	特願2012-XXXX	エンジンのための方法及び...	2012	C社	0	1		1
4	特願2014-XXXX	【課題】駐車場に設置された...	2014	D社	1	0		0
...	...		...	...	...	...		...
26,419	特願2016-XXXX	充電ステーションが電気エネ...	2016	X社	1	0		1

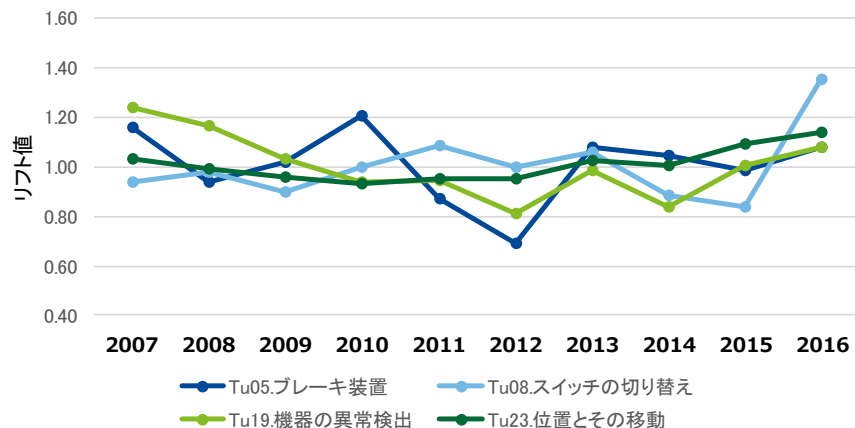
トピック×属性の様々な集計・分析が可能に

トピックをベースとした特徴分析

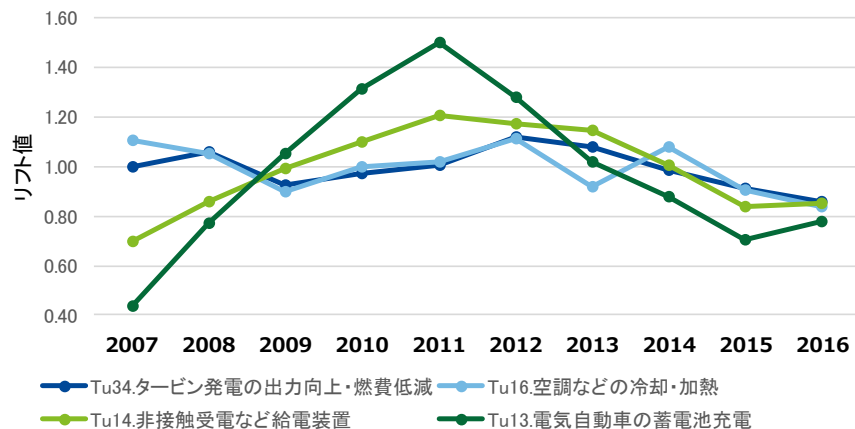
従来の単語ベースではなく集約されたトピックをベースにした分析を実行することで、膨大な特許情報に潜む特徴を分かりやすく理解することができます

トピック×出願年のトレンド分析

近年上昇傾向にあるトピック



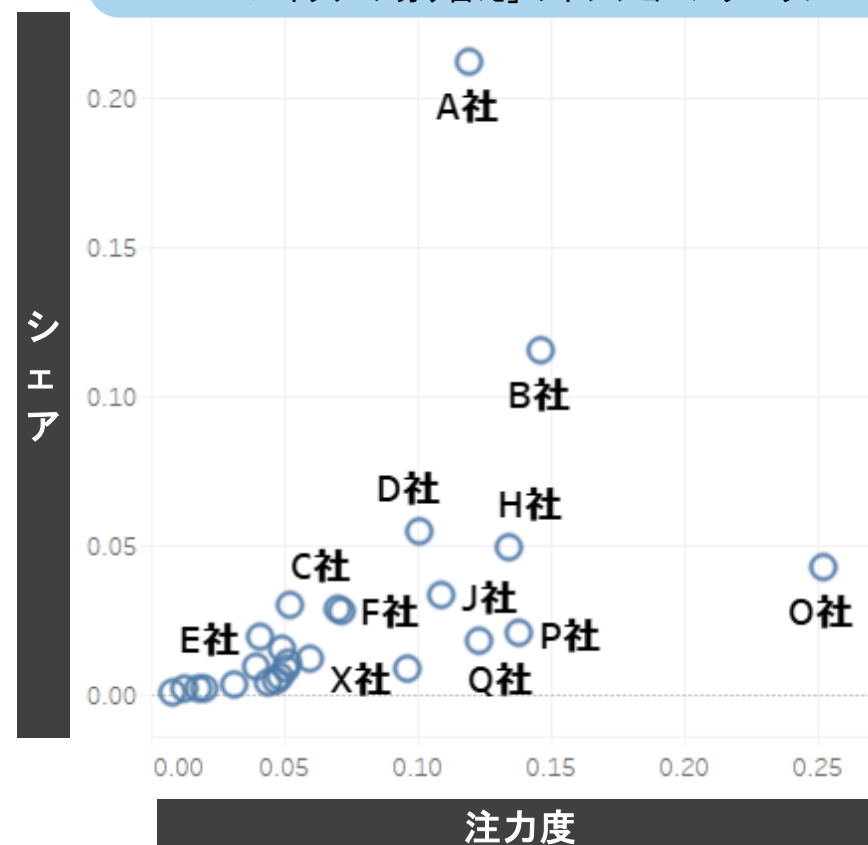
近年下降傾向にあるトピック



トピック×出願人の競合分析

各トピックに対する出願シェアと注力度を軸とした出願人のポジショニングマップ

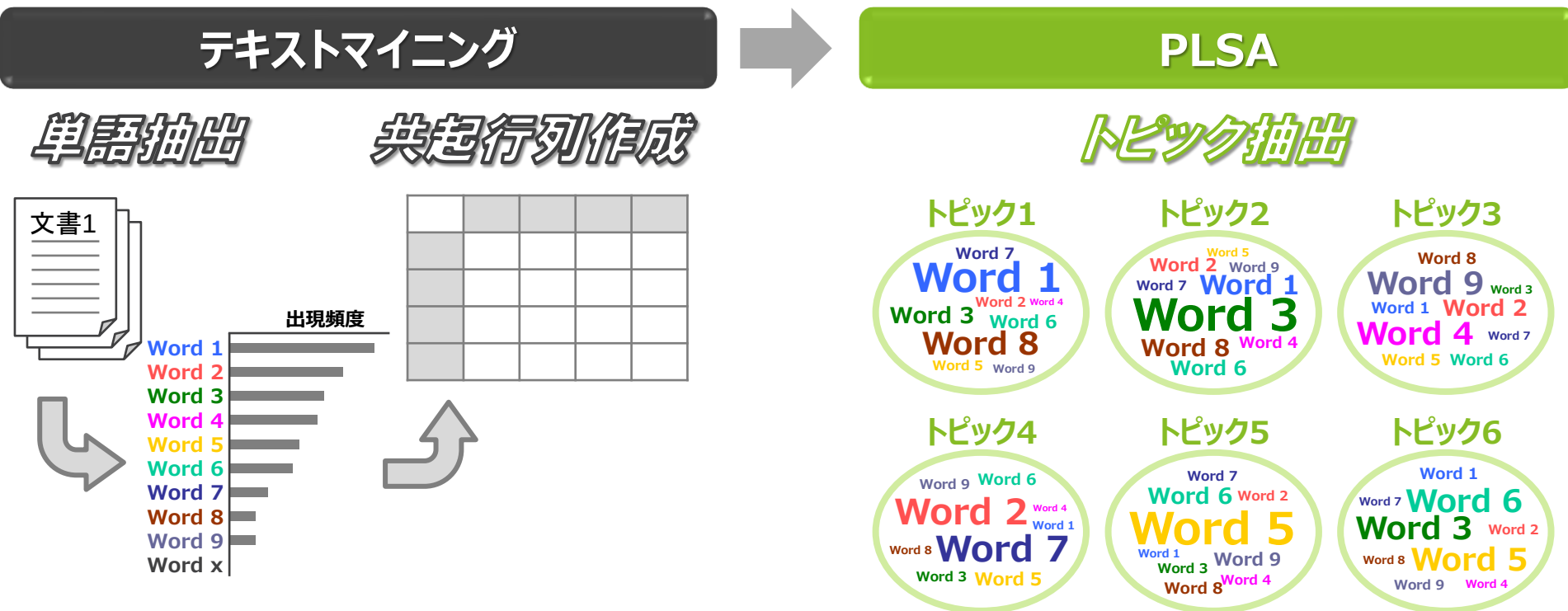
「Tu08.スイッチの切り替え」のポジショニングマップ



PLSAを応用した新技術

PCSA(確率的因果意味解析)

PLSAでは、テキストマイニングで全体のデータから抽出された単語に基づいた共起行列をインプットにすることで、データ全体を表現する平均的なトピックを抽出します



テキストデータにテキストマイニングを実行して単語を抽出し、その単語の共起頻度を集計した共起行列を作成する

作成した共起行列にPLSAを適用し、単語をトピックに集約する(使われ方の似ている単語をその重みと共にまとめる)

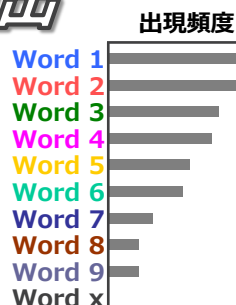
ターゲット事象の該当データと非該当データからそれぞれ構築した2つの共起行列の差分にPLSAを適用することで、そのターゲット事象に影響を与えるトピックを優先して抽出します

PCSA[®] (Probabilistic Causal Semantic Analysis : 確率的因果意味解析)

テキストマイニング

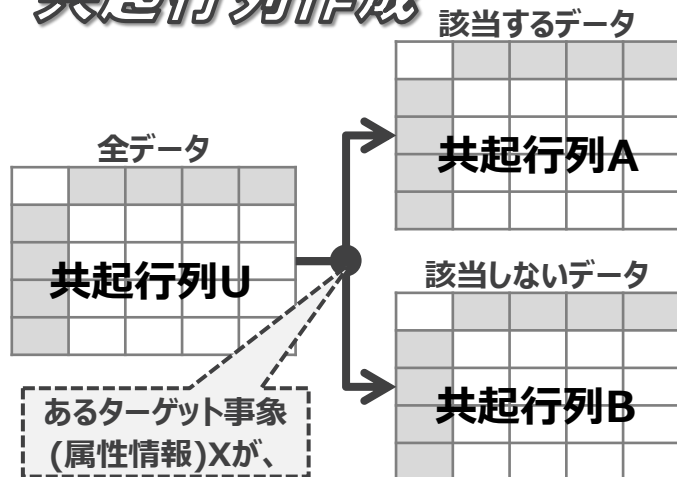
PLSA

単語抽出



全データから構築した共起行列Uを、あるターゲット事象(属性情報)Xが該当するデータから構築した共起行列Aと、該当しないデータから構築した共起行列Bに分割し、その2つの共起行列の差分を取った共起行列(A-B)に対してPLSAを適用する

共起行列作成

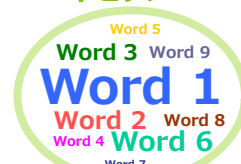


PLSAの適用

2つの共起行列の差分
共起行列A-B

トピック抽出

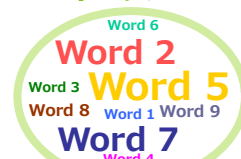
トピック1



トピック2



トピック3



ターゲット事象Xの該当有無に影響を与える潜在トピックを優先的にテキスト情報から抽出できる

ターゲット事象 X

【人工知能学会 2018年度全国大会優秀賞 受賞】

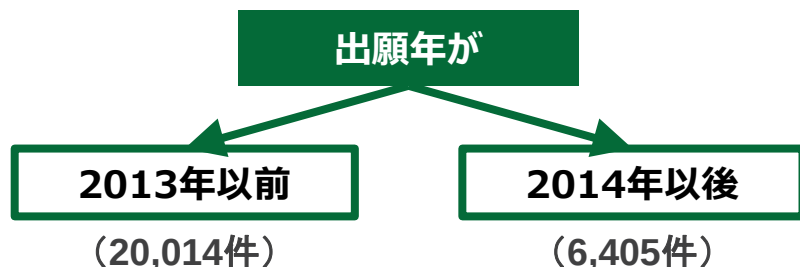
先ほどと同様の電気自動車関連の特許データを対象に、「出願年」をターゲットにPCSAを適用し、近年上昇傾向あるいは下降傾向にある技術トピックを抽出します

データの抽出条件と分析対象(先ほどと同様)

- 対象
 - 公開特許公報
- キーワード
 - 要約と請求項に「車」と「電気」を含む
- 出願日
 - 2007年1月1日～2016年12月31日
- 抽出方法
 - Patent Integrationを使用
- 抽出件数
 - 26,419件
- 分析対象
 - 要約文のテキスト情報



PCSAのターゲット



分析プロセス

テキストマイニング

Text Mining Studio

要約文から単語や係り受けを抽出

差分共起行列の作成

出願年が2013年以前の共起行列と2014年以後の共起行列の差分を取った共起行列を作成

トピック化 (PLSA)

Visual Mining Studio

差分共起行列にPLSAを適用して2014年前後の出願に特徴がある技術トピックを抽出

トピックのスコアリング

全データに対する各トピックのスコアを計算

特徴の可視化

トピックのスコアを出願年を軸に集計・可視化

①2013年以前データ、②2014年以後データで作成した2つの共起行列の差分にPLSAを適用することで、2014年前後の出願に特徴がある技術トピックを優先的に抽出します



①出願年が**2013年以前**のデータ
20,014件(文章数:33,113件)

係り受け表現					
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	⋮
単語	構成	97	30	27	25
	制御	208	49	80	66
	配置	56	1	23	8
	モータ	192	42	356	46
	...				

②出願年が**2014年以後**のデータ
6,405件(文章数:10,723件)

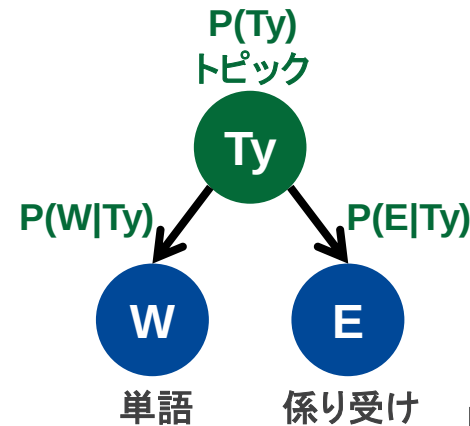
係り受け表現					
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	⋮
単語	構成	21	3	9	8
	制御	60	24	28	19
	配置	13	1	6	0
	モータ	47	19	138	12
	...				

①2013年以前の共起行列と
②2014年以後の共起行列の
差の絶対値を計算した共起行列
を作成する

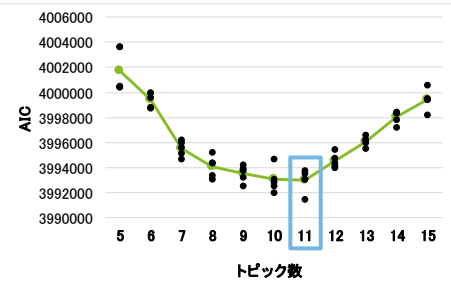
係り受け表現					
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	⋮
単語	構成	10.4	6.7	0.3	0.1
	制御	7.4	8.1	2.1	2.4
	配置	5.1	0.7	1.4	2.6
	モータ	15.2	5.4	22.7	2.9
	...				

ただし、2つの共起行列は異なる
文章数のデータから作成されて
いるので、①2013年以前の共起
行列の頻度を2つの文章数の比
率(10,723/33,113)で重み調整
してから、②2014年以後の共起
行列との差を計算する

差分の共起行列にPLSAを適用
する



トピック数を幅を持たせて設定し、
各トピック数に対してPLSAを初
期値を変えて5回ずつ実行して
情報量基準AICを計算し、AIC最
小の解を採用する



各トピックについて以下の3つの
確率が計算される

- ① $P(Ty)$
トピックの存在確率
- ② $P(W|Ty)$
トピックにおける単語の所属確率
- ③ $P(E|Ty)$
トピックにおける係り受けの所属確率

トピックにおける $P(W|Ty)$ と $P(E|Ty)$
からトピックの意味を解釈する

トピック Ty05			
P(T) = 2.5%			
P(W T)	単語	P(E T)	係り受け
3.1%	製造	4.5%	水素-製造
2.2%	二次電池	3.0%	発電-電気
1.5%	負極	2.6%	製造方法-提供
1.5%	正極	2.6%	製造-方法
1.5%	製造方法	1.6%	含む-リチウムイオン電池
1.2%	エネルギー	1.3%	成形品-提供
1.1%	リチウムイオン電池	1.3%	強度-有する
1.1%	水素	1.3%	表面-形成
1.1%	セパレータ	1.2%	含む-組成物
1.0%	電解液	1.2%	リレー-スイッチ
...	...	...	...

確率の高い構成要素から、Ty05の
トピックは「二次電池の製造方法」
に関するトピックと解釈できる

出願年の特徴トピック11個の一覧

出願年(2014年前後)で特徴を示すトピックは、エンジン駆動、モータ構成、電力変換、充電、二次電池の製造方法、冷却・加熱、検出判定、小型化・低コスト化など11個抽出されました

Ty01.エンジン駆動・動力伝達の制御

動力 演算 発電機
駆動輪伝達 クラッチ 出力軸 開始構成
トランスミッション制御手段駆動 運転者 **モータ**
エンジン **制御装置** **ハイブリッド車両**
判定 **制御** ブレーキ モータジェネレータ トルク
連結 駆動力バッテリー停止 車輪 内燃機関
作動 検出

Ty02.モータの回転構成

位置 伝達 方向形成
中心固定ハウジング自動車 **回転**
収容移動+できる 電気機械 駆動部 **回転**
モータ 支持回転+できる **配置** **回転軸**
軸方向 ステータ連結駆動周囲 移動口
構成 車輪 発電機 シャフト 結合

Ty03.交流・直流の変換

放電直列直流 スイッチ 電力 並列 蓄電池
負荷 **供給** **変換** バッテリー検出 直流電力
交流電力充電電気自動車コンバータ **インバータ**
制御 制御部駆動 電力変換装置 蓄電装置
コンデンサ 制御装置 **電圧** 電流オフ電源モータ

Ty04.電気自動車への充電、給電装置

検出判定走行 構成 外部
充電ケーブル **電力** 給電装置 コネクタ 取得電源
収容 蓄電池 **電気自動車** 蓄電装置
充電システム **充電** **バッテリー**
充電スタンド外部電源 **充電** **バッテリー**
給電開始供給情報 設置 送信

Ty05.二次電池の製造方法

由来 負極活物質 電解液 成膜品
セパレータエネルギー **正極** 電子機器正極活物質
積層エネルギーシステム
製造 **負極** **二次電池** リチウムイオン電池
再生可能エネルギー電気部品 **製造方法** スイッチ
方法 自動車部品電池特性 水素 形成安全 活物質
電極含有表面発電

Ty06.空調などの冷却・加熱

制御
変換電気ヒータ 燃料 配置 冷媒 排出 空気
モータ内燃機関 **駆動** **加熱** 構成 冷却水 **発電**
エンジン 電気エネルギー **供給** **発電機**
バッテリー 燃料電池 制御装置 温度 **電力**
排気ガス冷却 車室内 作動 熱 循環

Ty07.情報通信、検出判定システム

電流送信自動車 調整 比較 情報
信号受信 **構成** センサ 取得 制御装置
システム **制御** **検出** 速度 **方法** 制御部
演算 **判定** 生成 測定電圧 電気信号
動作 記憶 入力 変化 モータ 位置

Ty08.形成・配置

開口
他端外部バッテリー 方向 下方
開口部 ケース 位置端子一端上方 **収容**
コネクタ **固定** **形成** 面ハウジング **配置**
電気部品 基板 端部 接触 突出 **構成**
保持 支持 対向 筐体 挿入

Ty09.小型化・低コスト化・簡素化・操作性向上

精度 低コスト 効率 バッテリー安定 **操作** 起因 安価
コスト自動車 小型化車室内 損傷 ワイヤハーネス
操作+できる **電気自動車** 確保ハイブリッド車両
振動 **構成** 信頼性 **構造** 電子機器 **スイッチ**
部品点数必要+ない モータ 影響
実現安全

Ty10.重力発電の活用による地球温暖化防止

駆動 タービン既存火力発電 発電量 安価
重力発電運用燃料費ゼロ 圧縮空気加速 発電量増大
垂直下方海水温度上昇ゼロ 人類絶滅
海草 重力加速度加速二酸化炭素排気ゼロ水
落差 大気圧同速度同容積仕事率先送り **船舶**
工場電化全盛 **自動車** 既存蒸気タービン発電
電気駆動 既存世界 **全面電化住宅** **全盛**
重力発電蓄電池駆動既存火力原子力発電全盛 全盛地球
地球温暖化人類

Ty11.既存エンジンへの警鐘、タービン発電・重力発電

自動車 軽量物発電 運用
安価電気駆動 軽量蒸気速度 横軸h車両 反転永遠
全動翼 太陽光加熱器熱製造 飛行機液体酸素圧縮駆動
既存エンジン 宇宙到達費用 **理論最良エンジン**
船舶 電気+液体空気+過熱蒸気熱供給設備 **大学**
発電量 既存既存蒸気タービン発電
容積圧縮仕事率 発電原価 後追いエンジン発明阻止
製造物全部 日帰り旅行 燃料費ゼロ空気圧縮
高校 静翼 燃費

Ty10,Ty11は特定の
出願人による重複した
要約内容の特許から
抽出された

※文字の大きさはトピックに対する関係の強さを表現している（上位5つの単語を赤色で表示している）

全特許データに対して各トピックの該当有無を計算し、各トピックが該当するデータのうち、 ②2014年以後となる割合について、PLSAの結果とPCSAの結果を比較します

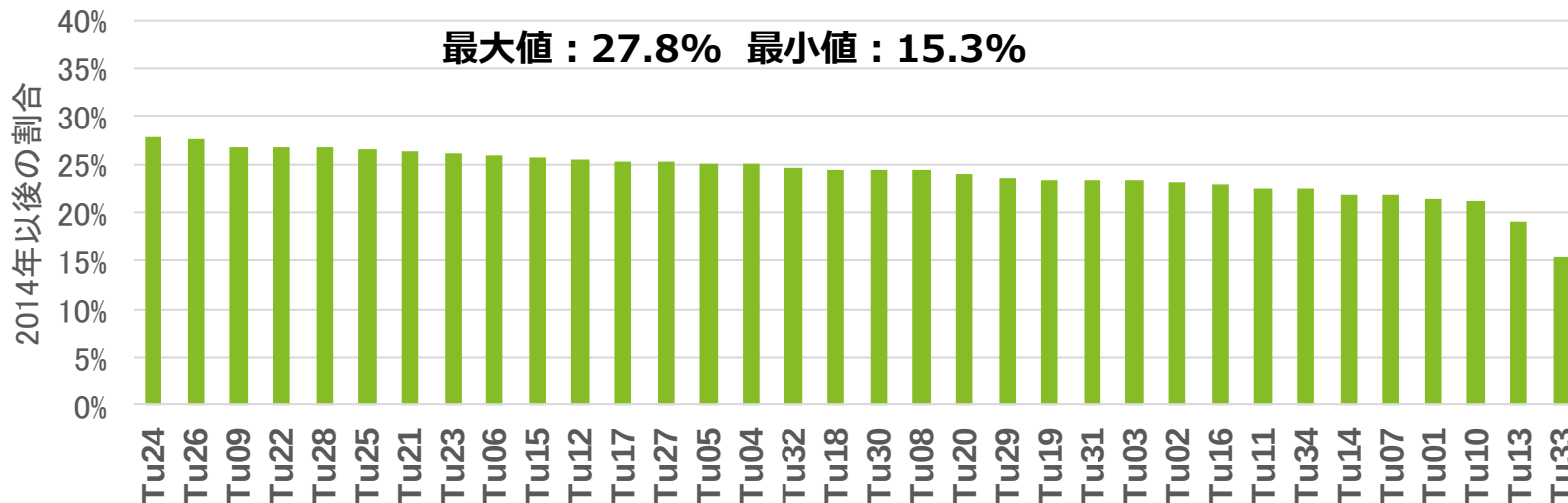
PLSAによるトピックとPCSAによるトピックのスコア(該当有無)を紐づけた特許データ

特許ID	出願番号	出願年	出願年 グループ	PLSAによる34個の全体トピック				PCSAによる11個の要因トピック			
				トピック Tu01	トピック Tu02	...	トピック Tu34	トピック Ty01	トピック Ty02	...	トピック Ty11
1	特願2007-XXXX	2007	①2013年以前	1	1		0	0	1		1
2	特願2009-XXXX	2009	①2013年以前	0	1		1	0	0		1
3	特願2012-XXXX	2012	①2013年以前	0	1		1	0	1		0
4	特願2014-XXXX	2014	②2014年以後	1	0		0	1	0		0
...	...	...	...	...	...		...	...	...		...
26,419	特願2016-XXXX	2016	②2014年以後	1	0		1	1	0		0

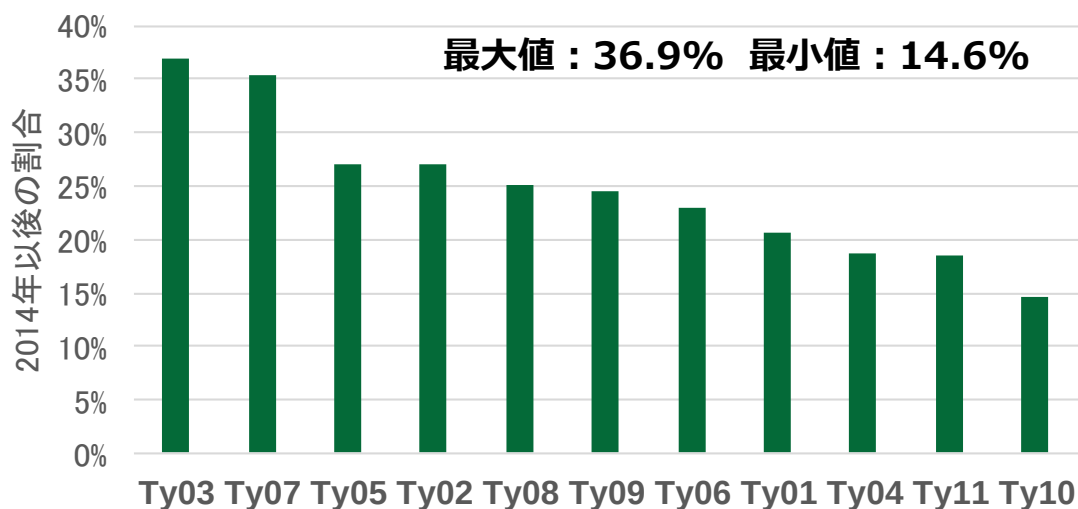
各トピックが該当するデータのうち、②2014年以後の割合を比較し、2014年前後においてPCSAではどれくらい偏ったトピックを抽出できているか確認する

各トピックの2014年以後の割合は、PLSAではおおむね25%前後ですが、PCSAでは割合が高いものと低いものに偏っており、2014年前後に対して特徴的なトピックとなっています

PLSA による 全体 トピック



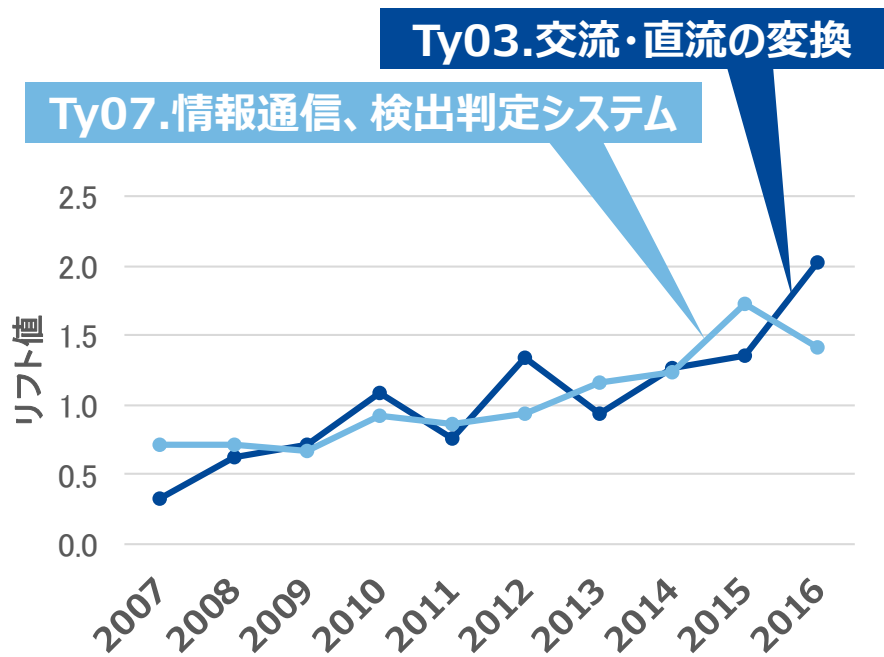
PCSA による 要因 トピック



- それぞれ2014年以後の該当割合の高い順にトピックを並べている
- 全体のトピックではおおむね25%前後のトピックが抽出されているが、PCSAのトピックでは割合が高いものから低いものまで抽出されており、最大値も高い
- 元々の2014年以後データの割合は24.2%

2014年以後の割合が高いトピックとして、電力変換に関する制御技術(Ty03)、データに基づいた運転者のアシスト技術(Ty07)が近年上昇傾向にあることが分かります

2014年以後の割合がトップ2のトレンド



※リフト値 =
$$\frac{P(\text{出願年} \mid \text{トピックTy}=1)}{P(\text{出願年})}$$

- リフト値は出願年とトピックの関係を示す指標
- トピック毎の各出願年の出願割合に対して、その出願年の出願割合で正規化した値であり、全体における各出願年の出願割合がそのトピックを条件にすることで何倍に変化するかを示す

該当特許の要約の例

Ty03.交流・直流の変換

発明の名称	出願年
電気自動車	2016

要約文（抜粋）

車両の衝突時に平滑化コンデンサを放電する確実性を向上させる。衝突検知装置が、衝突を検知したときに第1信号と第1信号に続く第2信号を送信する。第1信号を受信したときに、インバータ制御回路がスイッチング素子駆動回路への電力の供給を停止する。第2信号を受信したときに、インバータ制御回路が平滑化コンデンサを放電する。

Ty07.情報通信、検出判定システム

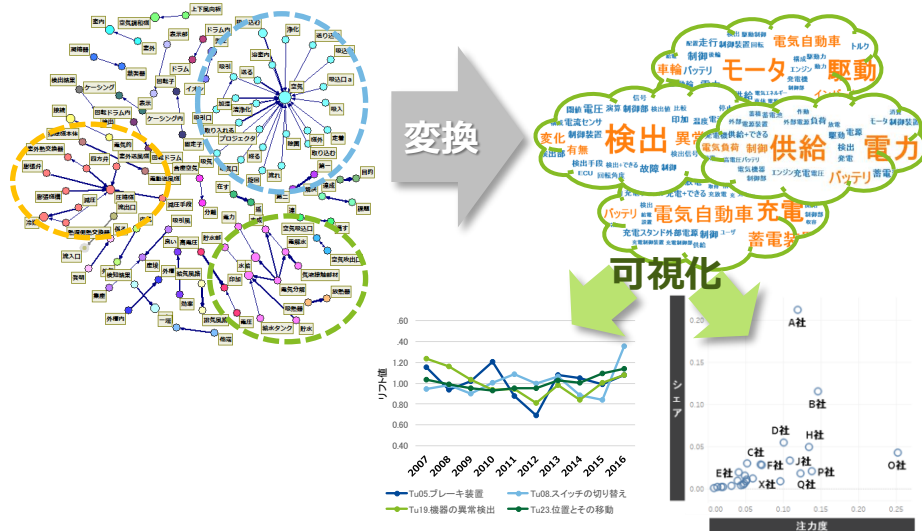
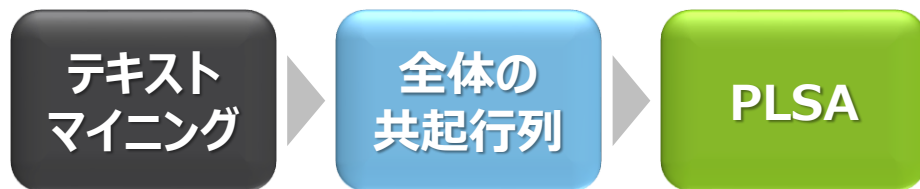
発明の名称	出願年
車両速度の制御方法	2016

要約文（抜粋）

経路及び交通状況に関する情報の応答として、自動車の速度を制御する方法に関するものである。本方法は、計画経路データ及び／または、繰返行程ロガーデータにより特定された想定経路に基づいて、最適な制動または加速点を決定し、最適な制動または加速点に基づいて、運転者に速度プロファイルを調整するためのサインを送る。

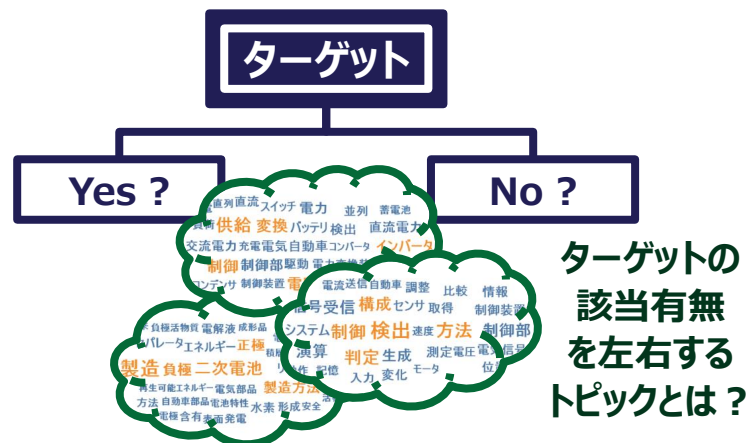
まとめ

テキストデータ全体を表す平均的なトピックを理解し、
そうしたトピックの特徴を様々な分析軸で探索する



膨大な単語をトピックに変換して特徴を把握できる

探索したい特徴に特化したトピックを優先的に抽出し、より顕著な要因を深く分析してインサイトを得る



確認したい対象の要因となるトピックを抽出できる

PCSAの技術は様々な業務のテキストデータに適用でき、効果的なビジネスアクションの検討を支援します



口コミ

評価得点に影響を与える
口コミのトピックとは？



アンケート

顧客満足度に影響を与える
自由記述トピックとは？



コールセンター履歴

解約・退会に影響を与える
問い合わせトピックとは？



特許文書

近年出願が伸びている
技術トピックとは？



営業日報

契約成立に影響を与える
営業活動トピックとは？



有価証券報告書

業績指標に影響を与える
定性的なトピックとは？



エントリーシート

選考通過に影響を与える
自己PRトピックとは？



診療記録

検査指標に影響を与える
定性的なトピックとは？



問題発生レポート

問題の発生に影響を与える
報告記録トピックとは？

ビッグデータからインサイト獲得のためのダブルアプローチ

ビッグデータはそのままでは複雑で理解不能なので、まず理解できる形に抽象化して特徴を発見しますが、今度は抽象度が高くて業務に活用できないので、その特徴の個別のデータに着目して再度具体化します

特徴を発見しやすくするために抽象化する（量的分析に軍配を上げ平均的な存在の間にある普遍的な特徴を得る）

マクロ探索アプローチ

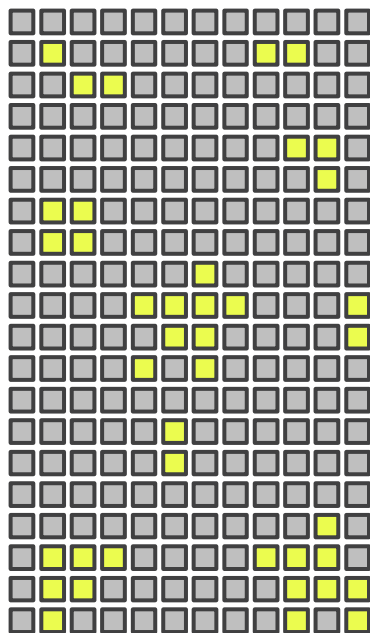
業務課題



確認したい
データの抽出

価値ある
個別データの
詳細調査と業務
活用への
昇華

膨大で複雑な ビッグデータ

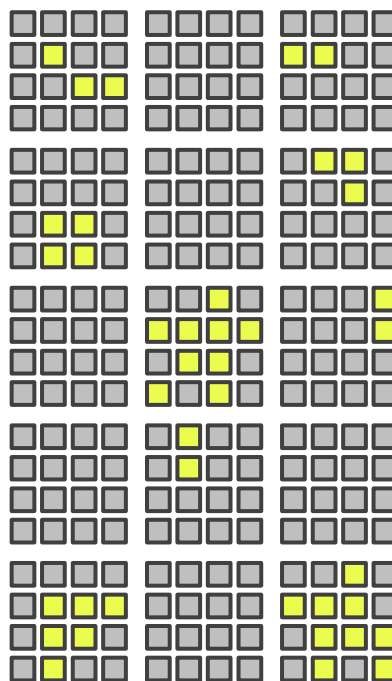


■ 価値あるデータ

特徴の似
ているデー
タの分類

特徴ピック
から個別
のデータの
絞り込み

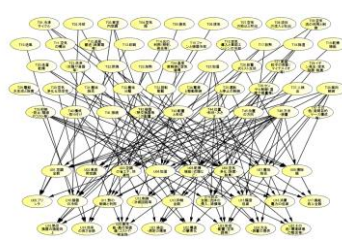
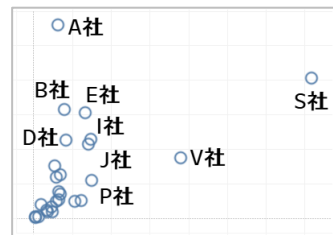
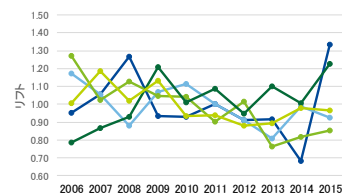
類型化



各トピック
の特徴の
可視化

注目すべ
きトピック
や要因条
件の発見

可視化・モデル化



ミクロ探索アプローチ

価値ある個別のデータを発見するために具体化する（質的分析に軍配を上げ平均の中の個別の特性を確認する）

資料に関するお問い合わせやコンサルティングの
ご相談は以下までお願いします。

analytics.office@analyticsdlab.co.jp

会社ホームページもご参考にしてください。
過去の講演・論文資料や技術解説も掲載しています。

<http://www.analyticsdlab.co.jp/>

株式会社アナリティクスデザインラボ

