テキストデータ解析のワークフローとその応用例

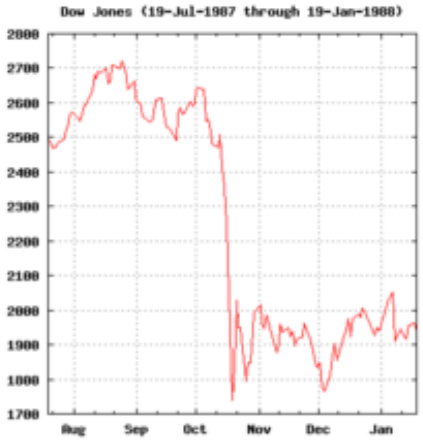
MathWorks® Japan
アプリケーションエンジニアリング部
アプリケーションエンジニア
井原 瑞希

Agenda

- テキスト解析について
- テキストデータ解析に特化した難点と解決策
- テキスト解析のワークフロー
- 文書からのトピックの抽出や類似度の算出

テキスト解析の主な適用分野

- アプリケーション例
  - 故障メンテナンス
  - マーケット分析、株価予測
  - 医療診断
  - 不正検知
  - ソーシャルメディア分析
  - 著者推定



- ドキュメントの…
  - 情報抽出
  - 感情分析
  - カテゴリ分類

テキスト解析をする際の課題

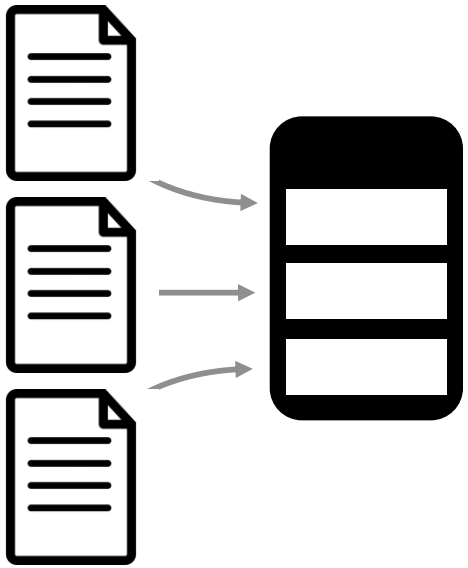
1. 数値型ではない
   - 日本語？英語？○○語？
2. 出現する単語の数が膨大なことが多い
   - 機械学習でいう、「特徴量の数が膨大」と同じ
3. 単語が曖昧性を持つ
   - 同じ単語でも周辺の単語によってコンテキストが変化

テキストに特化した
前処理や解析の工夫が必要

テキスト解析のワークフロー

データアクセス → データの前処理 → 予測モデルの構築

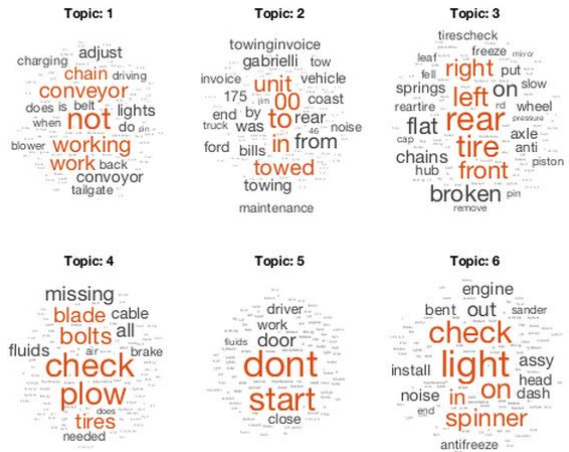
データの前処理 = テキストの整形 → テキストの数値化



GAS IN FUEL TANKHYD
LINES UNDER
CABSTROB LIGHTS

Gas fuel tankhyd lines
cabstrob lights

	gas	fuel	tank	hyd
doc1	1	0	1	0
doc2	1	1	0	1



Text Analytics Toolbox™

ワークフローごとの機能

データアクセス → データの前処理 → 予測モデルの構築

テキストの整形 → テキストの数値化

テキストファイル
スプレッドシート
Web

削除/抽出
正規表現

Word ドキュメント
PDF
R2017b

高頻度の単語抽出
ステミング (マッチング)
トークン化 (分かち書き)

単語カウント (Bag-of-words)
TF-IDF
単語の分散表現

潜在意味解析 (LSA)
潜在的ディリクレ配分法 (LDA)

R2017b, R2018a では英語対応のみ
(外部の形態素解析器を利用)

可視化

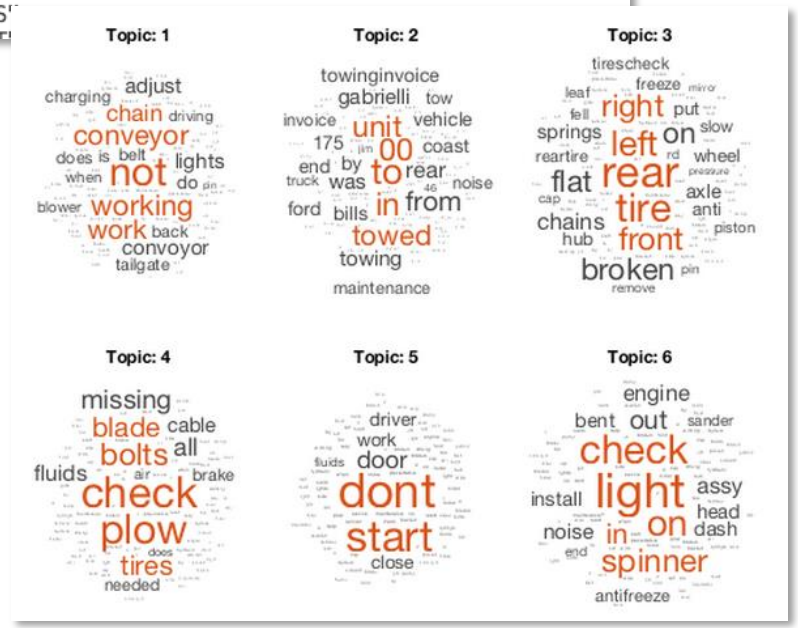
Word Cloud
Text Scatter

例1: 車の故障原因の抽出

英語テキストを対象とした解析

repairNotes = 617×1 string array

```
"PM SERVICE, CHECK TURN SIGNAL, CLUNKING NOISE WHEN DRIVING"
"SERVICEROB,EXT,5604"
"NEED 4 PLOW PINS"
"INSTALL SPINNER ASSY"
"DONT START"
"DOG BONE PIN BROKEN"
"NEED SERVICE, CHECK BRAKES"
"HYD CAP CHECK ENGINE LIGHT ON"
"TARP VALVE STICKINGRIGHT SIDE MIRROR BRACKET BROKEN"
"HANDLES IN CAB LOOSE"
"NO PLOW LIGHTS
```



目的

- 車の修理ログから、メンテナンスを実施した理由を分析

アプローチ

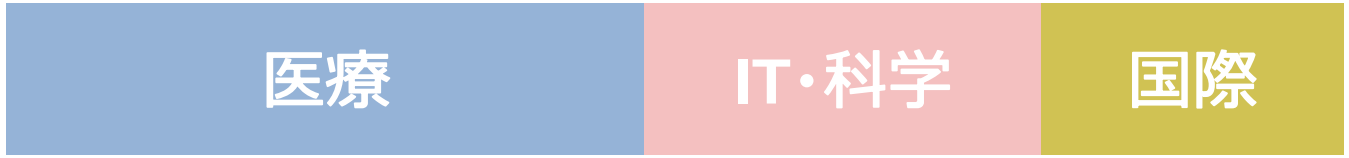
- 車両修理ログの読み込み
- テキストを分析しやすくするための整形
- **分野に特化した用語**に対する前処理
- 教師なし学習の手法を使用してメンテナンス実施の**主な理由を特定**

```
>> vehicleRepairAnalysis_jp
```

潜在的ディリクレ配分法 (Latent Dirichlet allocation; LDA) とは

- 文章をトピックの混合と仮定した文章の生成モデル
- 各文章の**トピック分布**と、各トピックの**単語分布**を求める
  - トピック分布

文章 A



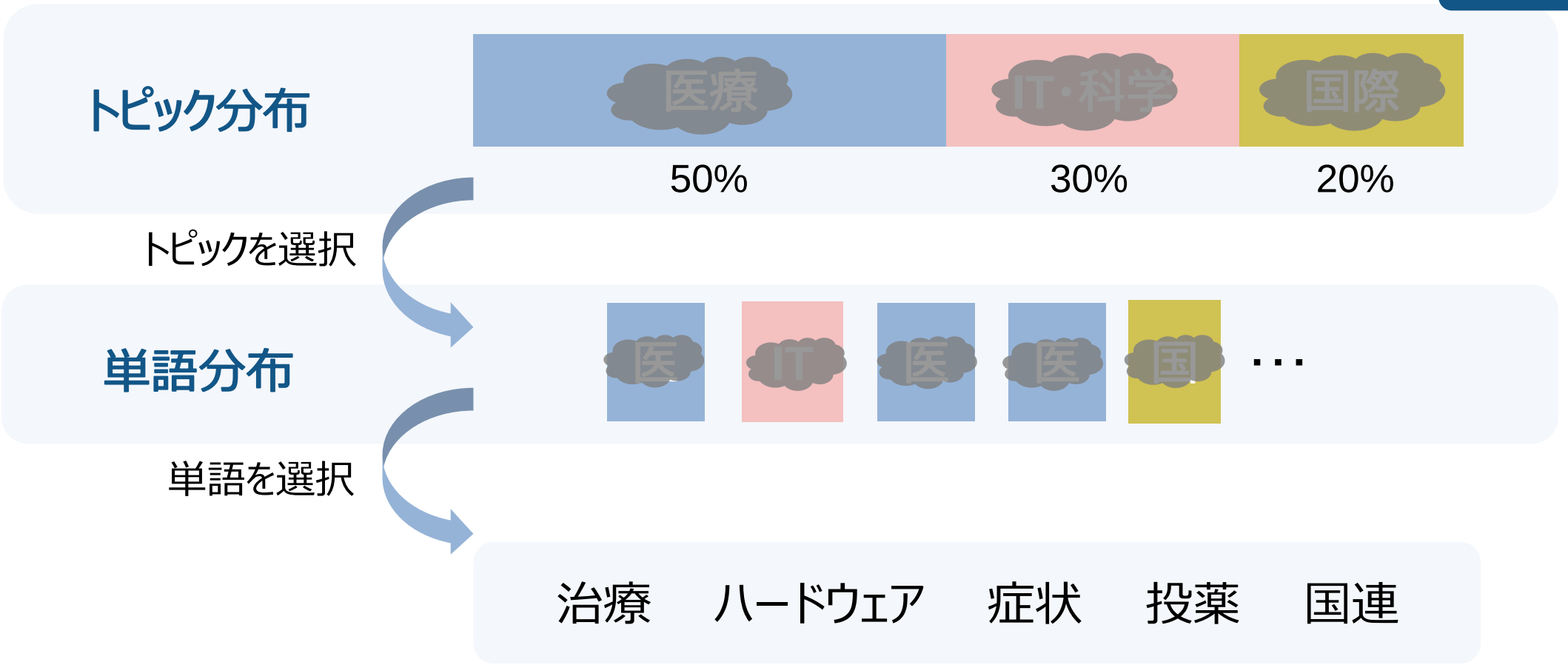
  - 単語分布

医療

治療	診断	症状	血液	…
0.6%	0.2%	0.4%	0.1%	…

潜在的ディリクレ配分法 (Latent Dirichlet allocation; LDA) とは

教師なし学習



トピックモデルの評価指標

- 代表的な評価指標 – **Perplexity**
  - あるモデルがドキュメントをどれくらいうまく表現できるか
  - モデルの予測性能を測る指標
  - 単語の選択肢の数 (perplexity 100 なら 100 単語に候補を絞れたと等価)

- 確率で表現

  perplexity = 1 / p (正解単語 | モデル)

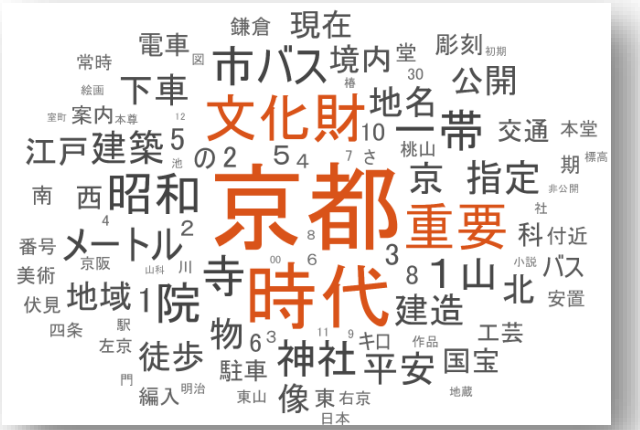
  - 言語モデルを仮定することで正解単語の候補数を減らす
  - 確率 0.01 の正解を選ぶこと
  - 100個の候補単語から正解を選ぶこと

例2: 観光地の類似度判定

日本語テキストを対象とした解析

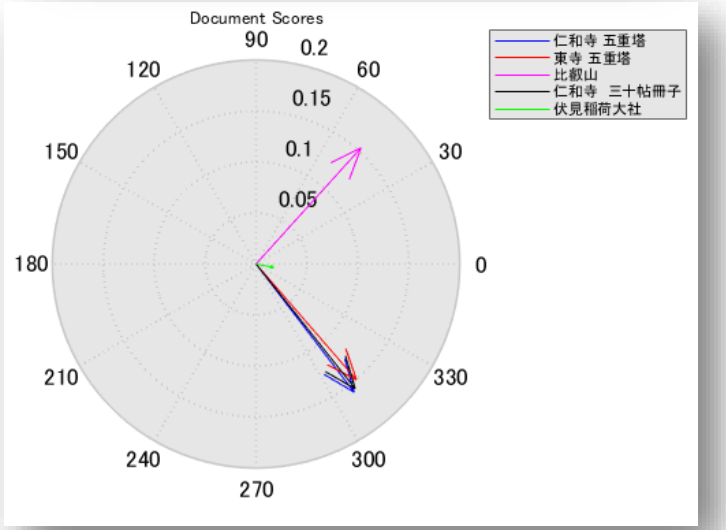
目的

- 観光地の説明文書から類似度を計算し、オススメや特定の状況に関連することを見つけ出す



アプローチ

- 観光地情報の読み込み
- 日本語の形態素解析器を使って特定の**品詞抽出**
- 可視化による前処理の必要性や**ストップワード**の発見
- 文書間の**類似度算出**
- **単語の分散表現**で特定状況に関連する単語の抽出



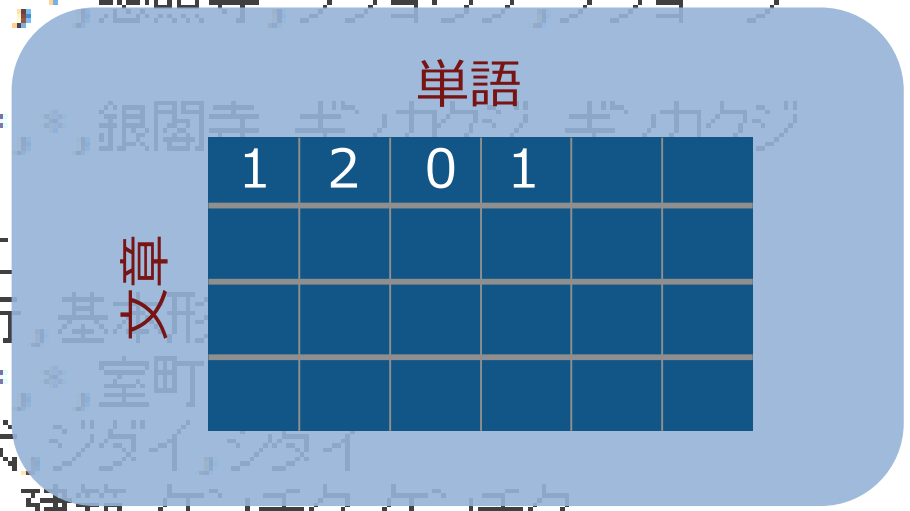

```
>> sight_analysis_live
```

テキストの前処理

- Tips: 一番最初に、最終的な形を決定

単語　　品詞情報

```
allstr =
'慈照寺　　名詞,固有名詞,一般,*,*,*,慈照寺,ジショウジ,ジショージ
(　記号,括弧開,*,*,*,*,(,(,(
銀閣寺　　名詞,固有名詞,組織,*,*,*,銀閣寺,ギンカクジ,ギンカクジ
)　記号,括弧閉,*,*,*,*,),),)
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
ある　　動詞,自立,*,*,五段・ラ行,基本形
室町　　名詞,固有名詞,一般,*,*,*,室町
時代　　名詞,一般,*,*,*,*,時代,ジダイ,ジダイ
建築　　名詞,サ変接続,*,*,*,*,建築,ケンチク,ケンチク
の　助詞,連体化,*,*,*,*,の,ノ,ノ
```



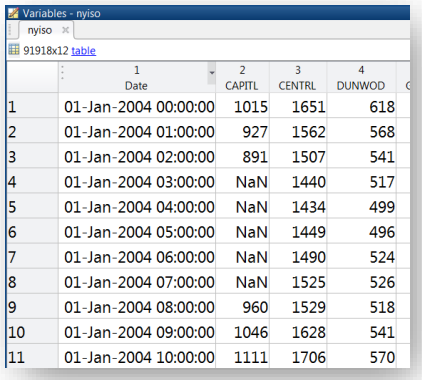
単語の分散表現（単語の埋め込み）

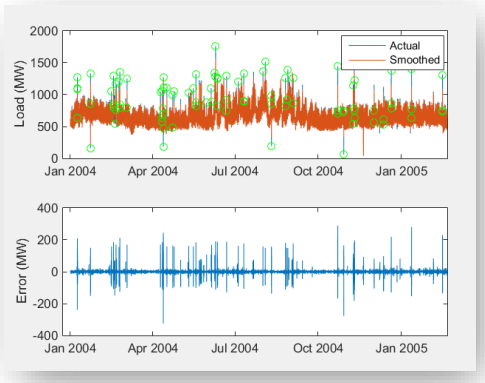
- **One-hot 表現**
  `[イチロー　ジロー　サブロー] = [1　0　0]`

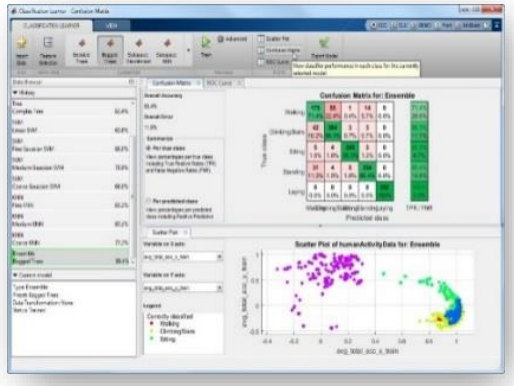
- **分散表現**
  `[スポーツ シアトル 東京 カレー 芸術] = [0.9 0.7 0.1 0.9 0.03]`
  - たくさんの対象物を表現可能
  - 曖昧性をもたせることが可能

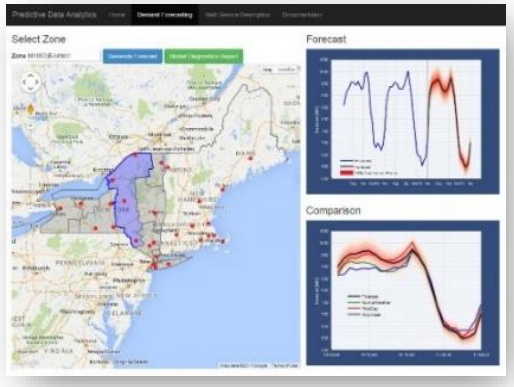
- 分散表現の代表的な方法: **`word2vec`** (Skip-gram)
  - 隠れ層＆出力層の2層ニューラルネットワークで重み学習

Text Analytics Toolbox と他機能の組み合わせ









データアクセス → データの前処理 → 予測モデルの構築 → システムへの統合

データベース

Deep Learning (RNN, LSTM)

exe / Web アプリ化などの展開

並列計算

前提条件: Statistics and Machine Learning Toolbox

Key Takeaways

- データ読み込みから予測モデルの構築までの
  テキストデータ解析のワークフローに対応
- テキストを対象とした幅広い分野の解析が可能
- 機械学習と組み合わせてさらなる分析へ

Check out our blogs!
https://blogs.mathworks.com/loren/2017/09/21/math-with-words-word-embeddings-with-matlab-and-text-analytics-toolbox/

テキスト解析関連用語集

ステミング	言う、言った、言うてはる　など、変形が異なる動詞を同じと扱う方法
ストップワード	が、あれ、この　など、一つでは主要な意味をなさない単語
トークン	意味を成す文字集合。通常は単語を指す
分散表現、単語の埋め込み	単語とその意味を1対1で対応させるのではなく、 単語を複数単語の組み合わせでベクトルとして表現する方法 例: イチロー [スポーツ、シアトル、東京、カレー、芸術] = [0.99 0.7 0.1 0.9 0.03]
word2vec (わーどとぅべっく)	分散表現を実現させる方法の一つ
正規表現	文字列の集合をルールベースで表す表現方法 例: メールアドレス `[a-z_]+@[a-z]+¥.(com|co.jp|ac.jp)`
Bag-of-words	文章内の単語の出現回数を表現した行列（単語の順番は考慮しない）
トピックモデル	文章が複数のトピックから生成されていると仮定したモデル