テキストマイニングの解説

マイニング

マイニングとは，英単語の "mine"(採掘坑 , 鉱山 , 鉱床 , 採鉱する , 坑道を掘る）の mining から由来。　つまり，鉱物資源，例えば，ダイヤモンドや金を採鉱する事を意味する。　これになぞらえ，膨大なデータの中から重要な情報を掘り出すことをデータマイニングと呼ぶ。　ここでは，次の段階を踏み，貴重で膨大な歴史資料から重要な情報を掘り出そうとするのが，テキストマイニングである。

作業手順

- （1）各アーカイブ（図書館等）から歴史資料をコピー又は，画像データを入手
- （2）紙媒体の場合，スキャナを利用し，画像データ化（画像データ入手の場合には不要)。
- （3）画像データからテキストデータ化
    - OCR（Optical Character Reader：光学式文字読取装置）ソフトを活用し，テキストデータ化
- （4）その他のテキスト変換処理
- （5）テキストの整形処理
    - テキストデータ化したデータから文字認識ミスの不要文字を取り除く
- （6）ステミング（語幹処理）
    - ステミング
- （7）統計ソフト R を利用し，テキストデータを解析
    - 重要な情報を掘り出す。

テキストマイニングの作業方法

各アーカイブ（図書館等）からの資料入手

スキャナを利用し，紙媒体資料を画像データ化

- スキャナを利用し，紙媒体資料を JPEG 等の画像データ化を行います。
    - ADF（Auto Document Feeder：原稿自動送り装置）が装備されているスキャナ必須。

・FUJITSU 製の SCANSNAP を利用するのも一法
　・http://scansnap.fujitsu.com/jp/

画像データからテキストデータ化

■ OCR ソフトを活用し，テキストデータ化

・OCR ソフト：メディアドライブ株式会社　e.typist v12.0 を利用。
・http://mediadrive.jp/products/et/

その他のテキスト変換処理

・HTML → TXT
　・HTML ファイルは，Web 等で利用されているマークアップ言語。
　・HTML ファイルを TXT 変換を行うには，松尾登志也氏が開発された HtoX32C というソフトを利用する。ただし，日本語での出力を前提としている為，英語以外の欧文には対応していない。
　　・http://win32lab.com/fsw/htox.html

・PDF → TXT
　・pdftotext.exe は，Unix の Xwindow System 上で PDF ファイルを閲覧するビューワー Xpdf に付属するソフト。欧文特殊文字を含むドイツ語やフランス語のテキストも問題なく変換できる。
　　・http://www.foolabs.com/xpdf/download.html
　　・その他参考資料 http://www.geocities.co.jp/SiliconValley-Bay/1992/tips/pdf2text.html

・PS → TXT
　・PS ファイル，PDF ファイル閲覧ソフト GSview に付属の pstotxt3.exe を利用する。
　・GSview を活用するには，まず，GhostScript をインストールする必要がある。

　　・GhostScript http://pages.cs.wisc.edu/~ghost/
　　・上記サイトより，次の２つのプログラムをダウンロードする。
　　・GPL Ghostscript 8.71
　　・GSview 4.9
　　・（2010.7.16. 現在）

　・ちなみに，.ps.gz ファイルを解凍するには，gzip.exe を利用する。
　　・http://www.gzip.org/

・DOC → TXT
　・Word の DOC ファイルをテキストファイルに変換するソフトは，antiword を利用する。
　　・http://www.informatik.uni-frankfurt.de/~markus/antiword/index.html

・DOC,PDF,Excel, 一太郎→ TXT
　・hishida 氏が開発した xdoc2txt を利用する。
　　・http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html

・RTF → TXT
　・針谷壮一氏が開発した RTF（リッチテキストファイル）コンバータを利用する。
　・文字コード変換にも利用できるソフトである。
　　・http://www5b.biglobe.ne.jp/~harigaya/rtfcnv.html

・改行コード変換
　・OS により改行コードは，割り当てられる制御コードが相違する為，異機種間（UNIX,Windws，Mac）でテキストデータをやり取りする場合には，改行コードの変換が必要となる。

・文字コード変換
　・日本語環境　JIS，EUC，Shift-JIS，Unicode
　・西ヨーロッパ言語　ISO-8859-1，Windows ANSI，MacRoman，UTF-8
　・RTF コンバータ（rtfconv）は，複数の文字コードに対応している為，これを活用する。
　　・その他，詳細については，参考書籍を参照の事。

・参考書籍　コーパス言語学の技法　　言語データの収集とコーパスの構築

テキストの整形処理

■テキストデータ化したデータから文字認識ミスの不要文字を取り除く

・正規表現
　・正規表現とは，「特殊な記号を使って文字列のパターンを作り，条件に当てはまる複数の文字列を一度に検索したり置換するための表現方法」のことである。

・ワイルドカードと正規表現
　・似通ったファイル名を検索する場合には，次のようなマーク”＊”（アスタリスク・マーク）を利用する。

これにより，サンプルのような似通ったファイルが検索される。

サンプル　似通ったファイルを探す
aaa1.txt
aaa2.txt
bbb1.txt
bbb2.txt

***1.txt を検索すると・・・**
aaa1.txt
bbb1.txt
が，探し出される。

・「ワイルドカードでは，＊や？に特別な意味を持っていた。正規表現では，こうした特別な意味を持たせた記号の事をメタキャラクタ，あるいはメタ文字という。メタキャラクタを使うことによって，それまでは一つひとつ別々に探さなければならなかった文字列を一度に探せるようになる。・・・正規表現では，あるパターンがある特定の文字列に一致することをマッチ (match) するという。」

・以下，代表的なメタキャラクタを例示する。
　・任意の一文字を表す　．（ピリオド）
　・直前のパターンの０回以上の繰り返しを表す　＊（アスタリスク）
　・直前のパターンの１回以上の繰り返すを表す　＋（プラス）
　・直前のパターンの０回または１回の繰り返しを表す　？（クエスチョンマーク）
　・右端のメタキャラクタをエスケープする　￥（円マーク）
　・文字列の先頭を表す　＾（キャレット）
　・文字列の最後を表す　＄（ドルマーク）
　・選択を表す　｜（縦線）
　・パターンをグループ化する　()
　・キャラクタクラスを表す　[]

・制御文字を表すメタキャラクタ

制御文字	制御内容
\f	改ページ
\n	改行
\r	リターン
\t	タブ
\s	空白文字

OS によって改行は，相違する為，注意を要す。OS による \n，\r の取扱いは，次の通り

	Windows	UNIX/MacOSX	MacOS9
\n	CR+LF	LF	CR
\r	CR	CR	CR

CR は，キャリッジリターン，LF は，ラインフィード。

・参考書籍　コーパス言語学の技法　　テキスト処理入門

ステミング（語幹処理）

■ステミング

・英語のステミング処理で最も利用されているアルゴリズムは，Porter のステミングアルゴリズム。
・このアルゴリズムは，ケンブリッジ大学の Martin Porter 氏が発明したもので，『An algorithm for suffix stripping』という論文で紹介された。
・参考書籍　実用 Perl プログラミング第 2 版　オライリー・ジャパン

・参考資料
　・http://tartarus.org/~martin/PorterStemmer/

統計ソフト R を利用し，テキストデータを解析

R のページをご覧下さい。

参考資料

・weka http://www.cs.waikato.ac.nz/ml/weka/
　・オープンソースデータマイニングプログラム