重要単語抽出プログラム（映画用）説明書

1. プログラム概要

　映画の概要内の単語の重要度を算出するプログラムです．重要度を算出することでその映画にとって重要な単語を知ることができます．

1. プログラムを使用するために

　プログラムを使用するには形態素解析ソフト茶筅[1]をインストールする必要があります． dwango\_gijutu\_appealフォルダ内の「chasen233\_031208.exe」を実行してインストールします．インストール先はC:\Program Files (x86)\ChaSenとなります．

1. 実行結果

　Visual Studio等を使用して、重要単語抽出プログラム（映画用）フォルダの「project.sln」を開き、「デバッグ開始」コマンド等で実行します。動作確認はMicrosoft Visual Studio 2010で行いました。

プログラムを実行するとtfidfフォルダ内に「\*\*\*.txt」（\*\*\*には000～134までの3桁の数字が入る）という135個のファイルが作成されます．3桁の数字は「概要.xlsx」の2～136行目の映画と対応しています．

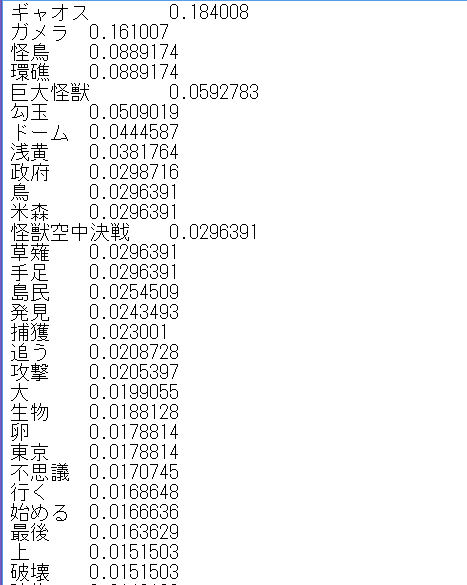


図1　 tfidfフォルダ内の000.txt

図1のtfidfフォルダ内の000.txtは「ガメラ 大怪獣空中決戦」という映画の概要内の単語を重要度の高いものから順に並べています．上位に「ギャオス」や「ガメラ」，「巨大怪獣」等，この映画に特徴的な単語が上がっています．

形態素解析ソフト茶筅の設定の違いにより、プログラムを実行しなおすと上記の結果と異なる場合があります。

1. その他ファイル説明

　「概要\_改.txt」は「概要.xlsx」の映画のタイトルと概要を繋げてテキスト化したファイルです．「概要\_改2.txt」は「概要\_改.txt」から改行文字を取り除いたものです．重要度の算出は「概要\_改2.txt」から算出しています．

　tfフォルダ内にもtfidfフォルダと同様に「\*\*\*.txt」（\*\*\*には000～134までの3桁の数字が入る）という135個のファイルが作成されています．これは各映画の概要内の単語の出現頻度を保存しています．tfidfフォルダのファイルの単語の重要度はこの単語の出現頻度等をもとに算出しています．

1. 重要度の説明

重要度は情報検索における重み付け手法の1つとして幅広く使われている*tf・idf*重み付けによって算出しています．*tf・idf*は単語の出現頻度である*tf*と逆文書頻度である*idf*の2つの指標に基づいて算出します．

5.1. tf（Term Frequency）

*tf*とは単語の出現頻度を意味し，単語の網羅性を示す値です．文書中にある単語がどれだけ多く出現するかを示しています．何度も繰り返し使われる語は，重要であると考えられ，*tf*は式(1)で定義されます．

 (1)

は文書*d*中に出現する単語*t*の頻度を表す．は文書*d*中の全単語数を表します．よって，は文書*d*内の単語*t*の相対頻度を表します．

5.2. tf（Term Frequency）

*idf*とはある単語がどれだけ限定された文書内で出現するか，という特定性を表す尺度です．特定の文書内にしか出現しない単語は重要度が高く，どんな文書にも出現する一般的な単語は重要度が低いと考えられ，*idf*は式(2)で定義されます．

 (2)

*N*は全文書数を表します．は単語*t*が出現する文書数です．対数をとるのはの変化を小さくするためです．は，多数の文書に出現する場合小さくなり，特定の文書にしか出現しない場合大きい値をとります．よって，は単語*t*がある文書を特徴付けている程度を表します．

5.3. tf・idf

単語出現頻度である*tf*と単語の特定性を表す*idf*を組み合わせた*tf・idf*を文書*d*中の単語*t*の重みとし，式(3)のように*tf*と*idf*の積で定義します．

 (3)

参考文献

1. 形態素解析ソフト茶筅，http://chasen-legacy.sourceforge.jp/