● 概要

本データベースは，独立行政法人情報通信研究機構 旧知識処理グループ（情報信頼性プロジェクト）によって開発され，高度言語情報融合フォーラム（ALAGIN）からオープンソースソフトウェアとして配布されている「意見（評価表現）抽出ツールVersion 1」※のための意見解析用モデルファイルと評価表現辞書から構成されます．

意見（評価表現）抽出ツールは，1行につき1文が書かれたテキストファイルを入力として，機械学習を使って意見や評判および評価（以下，これらをまとめて「評価」と呼びます）がテキストファイル中のそれぞれの文に存在するかどうかの判定を行い，その文に評価が存在すると認められた場合，以下の情報を出力するツールです．

(1) 評価を表す表現の抽出（評価表現抽出）
(2) 評価の意味的な分類（評価タイプ分類）
(3) 評価が肯定的なニュアンスを表すのか，否定的なニュアンスを表すのかの判定（評価極性判定）
(4) 評価を発信する主体の抽出（評価保持者同定）

本データベースには(1)-(4)の判定，分類に応じて4種類のモデルファイルが含まれています．このモデルファイルは，[文献1,2,3,4] の手法を用いて，Web上の2万文を対象にタグ付与されたデータを元に学習が行われています．また，評価表現辞書は，約35,000語の評価極性を記載した辞書（dictionary.dic）と，文全体の評価表現を反転させるような語を250語記載した反転辞書（reverse.dic）の2つの辞書から構成されます．これらの辞書は[文献1,2,3,4]とは異なり，すべて独立行政法人情報通信研究機構 情報分析研究室が構築しています．

本データベースを意見（評価表現）抽出ツールと共に用いることにより，生文を入力とする評価表現抽出，分類が可能になります．

※ http://alaginrc.nict.go.jp/opinion/index.html

● 注意事項

本データベースに含まれるモデルファイルは統計的機械学習によって構築されています．また，評価表現辞書には自動処理で構築された語が含まれています．したがって，その性質上，特定の個人，法人，団体の名称や，用法が適切でない場合に差別的表現や誹謗中傷ととられる表現が含まれている可能性が否定できません．その結果，本データベースを「意見（評価表現）抽出ツール」で利用した場合，ツールの出力に，例えば，特定の商品に対して誤って否定的な評判情報が抽出される可能性があります．これらは極端な場合には誹謗中傷と取られる可能性もないわけではありませんので，自動処理，機械学習の結果であることを明記する等，本データベース（「意見（評価表現）抽出ツール用モデル」）の契約の手引き，契約書をよくご覧の上，十分なご注意の上お使いください．また，用法が適切でない場合に差別的表現とみなされる表現が，否定的な評判情報として抽出されることにより，差別的または，人格侵害であると捉えられてしまう可能性があり得ますので十分ご注意の上ご活用下さい．

本データベースの内容，および，意見（評価表現）抽出ツールを本データベースを基に利用した結果について，その正確性，真実性，相当性についての保証はなく，また，独立行政法人 情報通信研究機構の主体的な意思決定・判断を示すものでもありません．本データベースの使用に関連して生ずる損失，損害等について，いかなる場合においても一切責任を負いません．上記点に関しましても本データベース（「意見（評価表現）抽出ツール用モデル」）の契約の手引き，契約書をよくご覧の上，十分なご注意の上お使いください．

● ファイル構成

```
README.utf.txt     このファイル
model/         モデルファイル群
dic/           評価表現辞書群
```

● 使用方法
1．「意見（評価表現）抽出ツール」をインストールします．
   - インストールについては以下をご参照ください．
http://alaginrc.nict.go.jp/opinion/index.html
2．本データベースを以下の要領で展開します．
```
 % tar zxvf extraopinion_model-1.0.tar.gz
```
3．展開後，extraopinion_model-1.0 内のmodel/，dic/の2つのディレクトリを適当な場所にコピーします．
4．意見（評価表現）抽出ツールの conf.sh 内で定義されている環境変数model，dic※

を編集し，model/，dic/をコピーした場所を指定します．
※ デフォルトでは，モデルファイルの場所は
extractopinion-1.0/modeldata/sample/が指定されています．辞書ファイルは
exptractopinion-1.0/dic/ が指定されています．conf.shを書き換えない場合は，
上記ディレクトリにモデルファイルと辞書を上書きしてコピーして下さい．

● 動作環境

OS: Linux，CentOS 5.5で動作確認
メモリ: 4GBで動作確認

● 解析精度

今回配布するモデルの精度は以下のようになっております．以下の解析精度の評価はいずれも10分割交差検定で行っています．

・評価極性判定

Precision（肯定）:	0.8732
Recall（肯定）:	0.8932
F値（肯定）:	0.8831
Precision（否定）:	0.8658
Recall（否定）:	0.8415
F値（否定）:	0.8535
精度:	0.8701

注）評価表現が正しく抽出され，評価極性のある評価タイプが抽出されたと仮定して評価．Precision,Precisionは肯定の場合と否定の場合に分けて評価を行っています．Precisionは本ツールが出力した肯定（否定）の極性のうち，正しい極性を出力した割合を表します．Recall はテストデータ中の肯定（否定）の極性のうち，本ツールが正しい極性を出力した割合を表します．精度はテストデータ中で正しい出力が得られた事例の割合を表します．

・評価表現抽出（抽出範囲の主辞の一致）

Precision	0.6350
Recall	0.3893
F値	0.4826

注）Precisionは本ツールが抽出した評価表現のうち正しく抽出された評価表現の割合を表します．Recallはテストデータ中の正解評価表現のうち本ツールが正しく抽出した評価表現の割合を表します．人手で構築された学習用評価情報コーパスは2名の作業者によって行われ，適宜基準について合議を行い進められました．2名の評価者の一方のアノテーションを正解とみなし，もう一方をシステムの出力とみなした場合における，精度は以下の表のようになります．従って一見上記のシステムの精度は極端に低いように見える可能性がありますが，Precisionに関しては人間の判断に近いものであると考えられます．

コーパスアノテーションの一致率

Precision	0.71
Recall	0.67

・評価タイプ分類

精度	0.6519

注）評価表現が正しく抽出されたと仮定して評価．精度はテストデータの中で正しい出力が得られた事例の割合を示す．

・評価保持者抽出

精度	0.6748

注）評価表現が正しく抽出されたと仮定して評価．精度はテストデータの中で正しい出力が得られた事例の割合を示す．

・文全体の極性判定

入力文全体の極性を出力するタスク，すなわち，1文中において抽出されるポジティブな評価表現1つにつき，+1を与え，ネガティブな評価表現1つにつき，-1を与え，入力文中の全ての評価表現の値の和が正，負，ゼロのどれであるかを出力するタスクを考えます．

この場合は，評価表現抽出が必ずしも正解と完全に一致しなくても文全体の極性の判定は正しく行える場合があるため，その精度は，0.72となります．

● 参考文献

[文献1] Nakagawa, T., Inui, K. and Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 786–794 (2010).
[文献2] 中川哲治，乾健太郎，黒橋禎夫：事例の重み付けに基づく自動獲得されたコーパスの効果的な利用法と評価極性分類への応用，言語理解とコミュニケーション研究会，(2009).
[文献3] Tetsuji Nakagawa, Takuya Kawada, Kentaro Inui, Sadao Kurohashi: Extracting Subjective and Objective Evaluative Expressions from the Web, In Proceedings of the Second International Symposium on Universal Communication (ISUC 2008), pp.251-258 (2008).
[文献4] 川田拓也，中川哲治，森井律子，宮森恒，赤峯享，乾健太郎，黒橋禎夫，木俵豊：Web テキストにおける評価情報の整理・分類およびタグ付きコーパスの構築，言語処理学会第 14 回年次大会論文集，pp.524-527 (2008).