単語共起頻度データベース（Version 1.1）

2009/12/24　version 1 初版
2010/03/31　verions 1 第２版（ファイル容量の追記）
2011/01/24　version 1.1 初版（非単語を除去．詳細は下記参照）

※ 概要

本データベースは，大量のウェブ文書を用いて，様々な条件で２つの単語が共に出現する頻度（共起頻度）を計算し，各単語について，３種の共起スコアの高い順に，単語とそのスコアを記録したものです．

３種類の共起スコアとは，Dice 係数，ディスカウンティングファクター有りの相互情報量[Pantel 04] (以降，Dpmi とする)，共起頻度です．

[Pantel 04]　P. Pantel and D. Ravichandran: Automatically Labeling Semantic Classes, In Proc. of HLT / NAACL, pp. 321-328 (2004).

Dice 係数は，２つの単語, t1, t2 があった場合, C(t1,t2)を t1 と t2 の共起頻度, C(t1)を t1 の出現頻度，C(t2)を t2 の出現頻度とした場合，

$$Dice(t1,t2) = \frac{C(t1,t2)}{C(t1)+C(t2)}$$

となります．

Dpmi は, 相互情報量(Point-wise Mutual Information)が低頻度の単語に正のバイアスがかかることが知られているため，その影響を減らしたスコアです．N は，頻度の計算対象のデータ総数です．

$$dpmi(t1,t2) = pmi(t1,t2) \times \frac{C(t1,t2)}{C(t1,t2)+1} \times \frac{\min(C(t1),C(t2))}{\min(C(t1),C(t2))+1}$$

$$pmi(t1,t2) = \log\left(\frac{\frac{C(t1,t2)}{N}}{\frac{C(t1)}{N} \times \frac{C(t2)}{N}} \right)$$

共起頻度は，C(t1,t2)そのものです.
例えば，「野球」の Dice 係数の上位の単語は，

サッカー:0.362974　格闘技:0.227781　プロ野球:0.220464 ゴルフ:0.210349　テニス:0.208742　試合:0.173582　選手:0.158105 高校野球:0.157891 バスケットボール:0.144332　競馬:0.136342 スポーツ:0.135528　バレーボール:0.133510 阪神:0.12301 巨人:0.115695

となっています．このように，関連の深い語が上位に来ています.

ただし，上記はあくまでも 1 例で，全単語について，上記のように関連の深い語が上位に来ているとは限りません．詳しくは，本ドキュメント「利用に関する注意」もご参照ください.

本データベースには，ダウンロード配布の上記データベースの他に，その元となるデータも含まれます. ただし，それらデータは非常にデータ量が多いため，USB の外付け HDD で配布いたします．基本的には，HDD を送付し，データをコピーして返却いただく形になります.

HDD 配布による全データをご希望の方は，以下の言語資源サイトの「言語資源の入手の手順」の「言語資源取得申請書」に，必要事項をご記入の上，高度言語情報融合フォーラム事務局まで送付してください.

http://alaginrc.nict.go.jp/resources/nictmastar/resource-info/2010-07-27-08-01-47.html

また，本データベースを便利に利用するためのツールを公開いたします．ツールは下記からダウンロードできます．使用法も記載しておりますので，ご参照ください.

http://alaginrc.nict.go.jp/SortedFileSearch/

※ ファイル(version 1)

本データベースで，ダウンロード配布するデータベースは以下の通りです． 1 GB を超えるファイルは，ファイル名の右に容量を示しています．ダウンロード配布の全ファイルの記憶には，圧縮前で約 5G，展開後で約 12G のディスク容量が必要です．

- 1m-0.1k.100m-docs.dice　(gzip 圧縮：約 1.1G　展開後：約 2.4G)
  - 約 100 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，dice 係数の上位 100 語とそのスコアを記録.
- 1m-0.1k.100m-docs.dpmi　(gzip 圧縮：約 1.1G　展開後：約 2.3G)
  - 約 100 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，dpmi の上位 100 語とそのスコアを記録.
- 1m-0.1k.100m-docs.freq　(gzip 圧縮：約 0.4G　展開後：約 1.2G)
  - 約 100 万語，約 1 億文書を用いて，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，共起頻度の上位 100 語とそのスコアを記録.
- 1m.100m-docs.tf
  - 約 100 万語，約 1 億文書を用いて，各単語の出現頻度を記録.
- 1m.100m-docs.df
  - 約 100 万語，約 1 億文書を用いて，各単語の文書頻度を記録.
- 500k-0.1k.100m-docs.w4.dice　(gzip 圧縮：約 0.5G　展開後：約 1.2G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，dice 係数の上位 100 語とそのスコアを記録
- 500k-0.1k.100m-docs.w4.dpmi　(gzip 圧縮：約 0.5G　展開後：約 1.1G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，dpmi の上位 100 語とそのスコアを記録
- 500k-0.1k.100m-docs.w4.freq　(gzip 圧縮：約 0.2G　展開後：約 0.6G)
  - 約 50 万語，約 1 億文書を用いて，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，共起頻度の上位 100 語とそのスコアを記録

- 500k-0.1k.100m-docs.w0.dice　　（gzip 圧縮：約 0.4G　展開後：約 1.1G）
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，dice 係数の上位 100 語とそのスコアを記録
- 500k-0.1k.100m-docs.w0.dpmi　　（gzip 圧縮：約 0.4G　展開後：約 1G）
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，dpmi の上位 100 語とそのスコアを記録
- 500k-0.1k.100m-docs.w0.freq　　（gzip 圧縮：約 0.2G　展開後：約 0.5G）
  - 約 50 万語，約 1 億文書を用いて，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，共起頻度の上位 100 語とそのスコアを記録
- 500k.100m-docs.tf
  - 約 50 万語，約 1 億文書を用いて，各単語の出現頻度を記録.
- 500k.100m-docs.df
  - 約 50 万語，約 1 億文書を用いて，各単語の文書頻度を記録.

HDD 配布では，さらに以下が加わります. HDD 配布の全ファイルの記憶には，圧縮前で約 1.2T，展開後で約 3.7T のディスク容量が必要です.

- 1m-1m.100m-docs.data　（gzip 圧縮：約 750G　展開後：約 2.4T）
  - 約 100 万語，約 1 億文書を用いて，全ての単語の組み合わせについて，2 つの単語が共に出現する文書の頻度を記録.
- 500k-500k.100m-docs.w4.data　（gzip 圧縮：約 66G　展開後：約 196G）
  - 約 50 万語，約 1 億文書を用いて，全ての単語の組み合わせについて，2 つの単語が近接 4 文内に共に出現する文書の頻度を記録
- 500k-500k.100m-docs.w0.data　　（gzip 圧縮：約 19G　展開後：約 61G）
  - 約 50 万語，約 1 億文書を用いて，全ての単語の組み合わせについて，2 つの単語が 1 文内に共に出現する文書の頻度を記録
- 1m-10k.100m-docs.dice　　（gzip 圧縮：約 91G　展開後：約 238G）
  - 約 100 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，dice 係数の上位 1 万語とそのスコアを記録.
- 1m-10k.100m-docs.dpmi　　（gzip 圧縮：約 89G　展開後：約 211G）
  - 約 100 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，dpmi の上位 1 万語とそのスコアを記録.

- 1m-10k.100m-docs.freq　(gzip 圧縮：約 70G　展開後：約 200G)
  - 約 100 万語，約 1 億文書を用いて，C(t1,t2)を t1，t2 が共に出現する文書の頻度として，共起頻度の上位 1 万語とそのスコアを記録.
- 500k-5k.100m-docs.w4.dice　(gzip 圧縮：約 20G　展開後：約 49G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，dice 係数の上位 5 千語とそのスコアを記録
- 500k-5k.100m-docs.w4.dpmi　(gzip 圧縮：約 20G　展開後：約 43G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，dpmi の上位 5 千語とそのスコアを記録
- 500k-5k.100m-docs.w4.freq　(gzip 圧縮：約 12G　展開後：約 30G)
  - 約 50 万語，約 1 億文書を用いて，C(t1,t2)を t1,t2 が近接 4 文内に共に出現する文書の頻度として，共起頻度の上位 5 千語とそのスコアを記録
- 500k-5k.100m-docs.w0.dice　(gzip 圧縮：約 13G　展開後：約 34G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，dice 係数の上位 5 千語とそのスコアを記録
- 500k-5k.100m-docs.w0.dpmi　(gzip 圧縮：約 13G　展開後：約 30G)
  - 約 50 万語，約 1 億文書を用いて，C(t1)，C(t2)をそれぞれ，t1，t2 の出現文書頻度，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，dpmi の上位 5 千語とそのスコアを記録
- 500k5k.100m-docs.w0.freq　(gzip 圧縮：約 7.4G　展開後：約 21G)
  - 約 50 万語，約 1 億文書を用いて，C(t1,t2)を t1,t2 が 1 文内に共に出現する文書の頻度として，共起頻度数の上位 5 千語とそのスコアを記録

※ ファイル(version 1.1)

Version 1 のデータは，自動解析された結果から単語を抽出しているため，実際には意味をなさない文字列（非単語）が単語として認識されていることがありました．そこで，約 100 万単語候補から人手で確認した約 33,000 の非単語を削除したデータを作成しました．ただし，全ての非単語を除去した保証はなく，残った単語に非単語が存在することがあります．また，正しい単語を非単語として削除している可能性もございますのでご了承下さい．Version 1.1 のダウンロード配布の全ファイルの記憶には，圧縮前で約 5G，展開後で約 12G のディスク容量が必要です．データ名は，元の version 1 のファイルに".no.bad.word"

を付加したものになっています. 各データの説明は前述の version 1 のファイルをご参照ください.

- 1m-0.1k.100m-docs.dice.no.bad.word （gzip 圧縮：約 0.98G　展開後：約 2.3G）
- 1m-0.1k.100m-docs.dpmi.no.bad.word （gzip 圧縮：約 0.97G　展開後：約 2.2G）
- 1m-0.1k.100m-docs.freq.no.bad.word （gzip 圧縮：約 0.37G　展開後：約 1.1G）
- 500k-0.1k.100m-docs.w4.dice.no.bad.word （gzip 圧縮：約 0.47G　展開後：約 1.1G）
- 500k-0.1k.100m-docs.w4.dpmi.no.bad.word （gzip 圧縮：約 0.46G　展開後：約 1G）
- 500k-0.1k.100m-docs.w4.freq.no.bad.word （gzip 圧縮：約 0.21G　展開後：約 0.5G）
- 500k-0.1k.100m-docs.w0.dice.no.bad.word （gzip 圧縮：約 0.46G　展開後：約 1.1G）
- 500k-0.1k.100m-docs.w0.dpmi.no.bad.word （gzip 圧縮：約 0.45G　展開後：約 0.97G）
- 500k-0.1k.100m-docs.w0.freq.no.bad.word （gzip 圧縮：約 0.21G　展開後：約 0.5G）

HDD 配布では, さらに以下が加わります. HDD 配布の全ファイルの記憶には, 圧縮前で約 1.1T，展開後で約 3.5T のディスク容量が必要です.

- 1m-1m.100m-docs.data.no.bad.word （gzip 圧縮：約 710G　展開後：約 2.3T）
- 500k-500k.100m-docs.w4.data.no.bad.word （gzip 圧縮：約 63G　展開後：約 189G）
- 500k-500k.100m-docs.w0.data.no.bad.word （gzip 圧縮：約 20G　展開後：約 59G）
- 1m-10k.100m-docs.dice.no.bad.word （gzip 圧縮：約 86G　展開後：約 224G）
- 1m-10k.100m-docs.dpmi.no.bad.word （gzip 圧縮：約 84G　展開後：約 199G）
- 1m-10k.100m-docs.freq.no.bad.word （gzip 圧縮：約 66G　展開後：約 189G）
- 500k-5k.100m-docs.w4.dice.no.bad.word （gzip 圧縮：約 19G　展開後：約 47G）
- 500k-5k.100m-docs.w4.dpmi.no.bad.word （gzip 圧縮：約 19G　展開後：約 42G）
- 500k-5k.100m-docs.w4.freq.no.bad.word （gzip 圧縮：約 12G　展開後：約 29G）
- 500k-5k.100m-docs.w0.dice.no.bad.word （gzip 圧縮：約 14G　展開後：約 32G）
- 500k-5k.100m-docs.w0.dpmi.no.bad.word （gzip 圧縮：約 13G　展開後：約 29G）
- 500k5k.100m-docs.w0.freq.no.bad.word （gzip 圧縮：約 7.7G　展開後：約 20G）

※ ファイルフォーマット

.dice,.dpmi,*.freq のファイルは，以下のフォーマットとなっています.
ただし，ダウンロード配布のファイルは，gzip 圧縮されています.

文字コードが UTF8 で，正規表現で表すと，一行が以下のフォーマットで書かれたファイルが生成されます.

<単語>([::space::]<単語>:<スコア>)+¥n

ここで，<単語>は，任意の単語の文字列，[::space::]は，空白とします．

(<単語>:<スコア>)は，行頭の単語との共起スコアの降順に並んでいます．つまり，行頭の単語にとって，n 番目の（<単語>:<スコア>）は各種共起スコアの n 位の単語とその共起スコアとなります．

*.data のファイルは，文字コードが UTF8 で，正規表現で表すと，一行が以下のフォーマットで書かれています．

<単語>([::space::]<単語>:<共起頻度>)+¥n

行頭の単語は unix 系 OS の”env LC_ALL=C sort”のソート順と同じ並びとなっています．また，その単語に対応する各(<単語>:<スコア>)も，単語について，同じソート順で並んでいます．2 分探索など，ソート順が関係する探索プログラムを作成する場合に留意してください．

,tf,.df のファイルは，文字コードが UTF8 で，一行が

<単語>[::space::]<頻度>¥n

となっています．

行頭の単語は unix 系 OS で”env LC_ALL=C sort”のソート順と同じ並びとなっています．

※ 生成方法

本データベースの作成の対象の文書データは，Tsubaki[Shinzato 2008]で収集された日本語の約 1 億ウェブページです．

[Shizato 2008]K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi, Tsubaki: An open search engine infrastructure for developing new Information access, In proc of the 3rd IJCNLP, pp. 189-196, 2008.

本データベースの「文」とは，Tsubaki の Standard Format の<S>です．
カウント対象の「単語」は，Alagin フォーラムで公開されている「文脈類似語データベース version 1」に準拠しています．つまり，本データベースの 100 万語，50 万語は，基本的には，それぞれ以下と同様です．

100 万語:　1m-2k.s1.data ,1m-2k.s1.data, 1m-2k.s2.data,1m-2k.s1+s2.data
1m-rv100k.data
50 万語:　old.500k-2k.data

ただし，文書の解析に用いられた形態素解析(Juman)，構文解析(Knp)のバージョン違いや，単語のフィルタ規則の違いによって，100 万語の方は約 2000 語，50 万語の方は約 1 万語，それぞれ抽出されなかったため，必ずしも全単語は含まれません．

※ 利用条件

本データベースの利用には，（独）情報通信研究機構と利用許諾契約を結ぶ必要があります．詳しくは，

http://www.alagin.jp

を御覧下さい．

※ 利用に関する注意

本データベースは，インターネットホームページ等，（独）情報通信研究機構以外の第三者が作成した文書等のデータから，語彙の抽出及び統計処理等によって作成されたものです．そのため本データベースの内容は，（独）情報通信研究機構の主体的な意思決定・判断を示すものではありません．本データベースの生成は，電子的な方法又は一様の選別基準による機械的判定によって行われています．そのため本データベースの内容の正確性，真実性及び相当性は一切保証されません．以上の理由により，（独）情報通信研究機構は，本データベースの内容について，責任を負いかねます．本データベースの使用に関連して生ずる損失，損害等についても，一切責任を負いかねます．本データベースには，意図せず，第三者への誹謗中傷，差別用語，個人情報などが含まれている場合があります．本データベースを利用の際はこれらによる権利侵害に十分な注意

をお願いいたします．利用者においては，本データベースの以上の特質をよくご理解の上で，本データベースをご利用下さい．

※ 本データベースに関する問い合わせ先

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
MASTAR プロジェクト　言語基盤グループ
Email: alagin-lr@khn.nict.go.jp