プログラミング基礎演習　最終レポート課題

目的

プログラミング基礎演習では毎週の課題と２回のレポートによって成績を決定する．本レポートはそのための第 2 回目のレポート課題である．プログラムに関してはコンパイル出来ないもの，計算結果が明らかに間違っているもの，課題と全く違うプログラムには点数を与えない．尚，ソースが読み難いもの（コメントがない，インデントしていない，ループや条件分岐がトリッキー，変数名からその役割が分からないもの）は減点対象となる．一方，追加機能を取り入れたもの，など工夫が見られる場合は，加点を行う．

もちろん，誰かが用意した模範解答や Web から発見したプログラムをコピーしたようなものはカンニングと見なされ**ゼロ点**となる．但し、調査をしながら解答するのは悪いことではない．その場合，出典を明らかにし，どの部分が自分のオリジナルなのか明確にする必要がある．この先，卒論研究で「誰も書いたことのないプログラムを自分で書く」ことになるので，そのための訓練だと思って自分のオリジナルなプログラムにして下さい．

課題：共起表現を見つけよ！

目的

複数の単語が一つの連結された表現として使われることを共起表現(collocation)という．例えば，fast food, train ticket, school bus など英語では数多く共起表現が存在する．共起表現の一つの特徴として共起表現に含まれる単語それぞれが本来持っている意味と異なる意味を表す場合や，本来の意味を更に特徴づける場合がある．与えられた文書の中から共起表現を抽出するのが本課題の目的である．

課題概要

文が行単位に分割されているテキストファイル（例：news.txt）を考える．このファイルに含まれている全単語の集合を V とする．V に含まれる単語２個からなる共起表現を見つけよ．見つけた共起表現をその共起の強さでソートし，上位の n 個を出力するプログラムを作成せよ．(n の値はコマンドライン引数で指定できるようにする)

ヒント

V に含まれる２つの異なる単語 a と b を考える．更に，入力ファイルの各行に整数の ID が振られていると仮定する．a が出現する文の ID（番号）の集合を S(a)とし，b が出現する文の ID(番号)の集合を S(b)とする．更に，a の直後に b が出現する文の ID 番号の集合を S(a,b)とする．a と b がどれくらい強く共起しているかを表す一つの尺度として，次式で定義される overlap 係数がある．

a と b の共起の強さ = |S(a,b)| / min(|S(a)|, |S(b)|)
ここでは，|S|は集合 S の要素数を表し，min(x,y)は x と y の内，小さいもの(minimum)を返す関数である．
例えば，fast food を含む文の数を 10, fast を含む文の数を 14, food を含む文の数を 12 とすると，上記の計算式は 10/min(14,12) = 10/12 = 0.83 となる．
この尺度以外にも 2 つの単語からなる共起表現の共起の強さを評価するための尺度を自分で定義して使って良い．

条件

1. news.txt を入力ファイルとして使うこと．
2. news.txt に含まれる 2 単語からなる表現（2 単語連結）全てを対象とし，共起の強さを計算し，ソートし，上位(共起の強さが最も高い)n 個（n の値はユーザーが与える)のみを出力すること．

提出方法及び締切:

作成したプログラムのソースファイルを学籍番号.c とする．ソースファイルの先頭に氏名，学年と学籍番号をコメントとして書いておくこと．

尚，プログラムのコンパイル仕方，実行仕方（パラメータなどあれば），どのような方法で共起を計算したか，結果はどうなったかを学籍番号.txt というテキストファイルに書くこと．

プログラムソースコード(学籍番号.c)と説明ファイル(学籍番号.txt)を zip ファイルとしてまとめる．尚，zip ファイルの名前は学籍番号.zip とすること．

上記作成した zip ファイルを CFIVE から提出する．

締切： 2013 年 02 月 17 日（日曜日) 23:50