プログラミング基礎演習　中間レポート課題

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１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（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：検索エンジンを作ろう！

目的

今日，Web 検索エンジンは Web から効率良く，情報を収集するための重要なツールとなっている．Web スケールで検索を実装するには大変複雑なシステムが必要であるが，本課題ではその最も基本的なデータ構造である「転置索引」を C 言語で作成する．

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．このファイルのそれぞれの行が別々の文書（ドキュメント）とする.このファイルを読んで，それぞれの行への転置索引を作成せよ．次に，作成した転置索引を使って複数の単語からなる検索クエリに関する文書（行）を出力できるようにせよ．

条件

1. 検索対象とするファイル名をコマンドライン引数で指摘できるようにする．尚，ファイルに含まれる行数，総単語数は未知とする．
   例えば，プログラム名は engine とし，コマンドプロンプトを$で表すと，
   $ ./engine -f news.txt
   として，コマンドライン引数(f)で入力対象のファイルを指定できるようにすること．
2. 検索クエリに含まれる単語数は未知とし，入力された単語全てを含む AND 検索に関する文が出力できるようにすること．例えば，プログラムが実行されたら，

Enter Query:
のような入力メッセージが表示され，
john said and
のような複数単語からなるクエリに関して john, said, and という 3 つ単語全て含む(AND 検索)文を news.txt から出力できるようにすること．仮に，検索結果がゼロの場合（クエリに含まれる単語が全て出現する文が news.txt 中に存在しない場合）は何らかのメッセージでユーザーにそのことを伝えるようにする．

上記の条件さえ満たしていれば特に次に述べるステップ（ヒント）を使わなくても正しいプログラムとみなす．但し，上記の条件を満たさない場合（news.txt に関する行数，総単語数など既知でないと動かないプログラムや検索クエリに含まれる単語数が既知（固定数）でないと動かないプログラム）は減点対象となる．

ステップ（ヒント）

1. -f でコマンドラインから指定されたファイルを開き，その一行ずつ読み，それらの行に固有の整数 id を振り，連結リスト(linked list)としてメモリ上に保存するプログラムを書け．例えば，文(body)とその ID(id)をまとめる構造体として次が考えられる．

```
struct DOCUMENT{
   int id;
   char * body;
   struct DOCUMENT * next;
};
```

2. (1)のプログラムで作成した DOCUMENT の連結リストを処理し，ある単語 w とその単語が出現するドキュメント達の id のリストへの対応を表す転置索引(inverted index)を作成する．例えば john という単語が 0 番目，2 番目と 3 番目の行に出現するならば，
john → [0,2,3]という対応を記録する転置索引を作成する．
もちろん，単語の総和が未知であり，更に，ある単語が出現する行の数も未知であるためこの転置索引を動的に作成する必要がある．

3. 転置索引を作成したらユーザーから検索クエリを求める入力プロンプトを表示し，検索クエリを受け取る．受け取った検索クエリを単語に分割し，それに含まれるそれぞれの単語 u について u が含まれる文の ID リストを転置索引から読みとる．最後に，全ての検索単語の AND 検索を行うためにはこれらの ID リストに共通して現れる ID を探せばよい．検索結果の出力として，該当する文の ID と文そのものを標準出力に出力する．

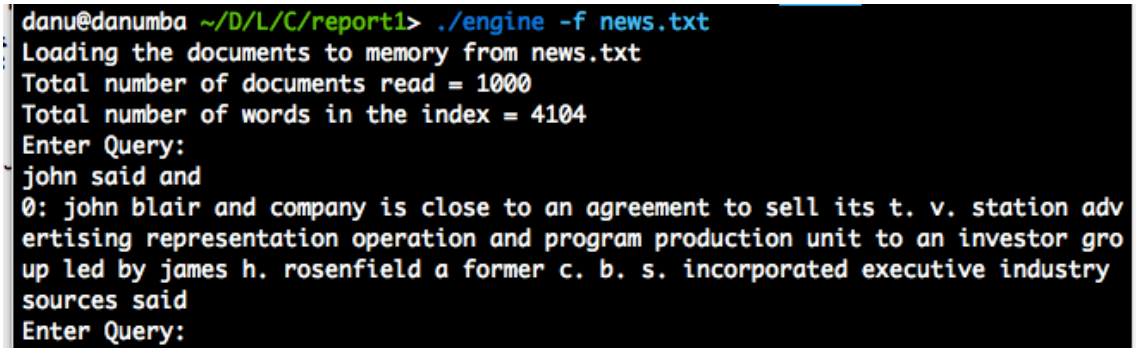

図 1. サンプル出力

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．ソースファイルを CFIVE の report1 から提出する．

締切： 2013 年 01 月 06 日（日曜日）23:50