プログラミング基礎演習
レポート課題2
～ヒント～

2013/2/5

岩成達哉

はじめに

今回の課題は，

- 前回のものを流用できる部分も多い
- 力技で押し切れる
- やることはそんなに難しくない

前回よりも簡単！

やる気になった(?)ところでやることを整理

1. 行毎にドキュメントを保存
2. 1つ目の単語を登録
3. 2つ目の単語を1つ目の単語と関連付けて登録
4. 1,2の単語のセット全てに対してoverlap係数を計算
5. 係数は単語のセットと一緒に配列などで保存
6. 計算した結果をソート
7. 入力された値の数だけ上位から表示

青文字で書いた4つのデータ構造が必要

前回の課題のプログラムを流用しよう

- 行毎にドキュメントにする部分はそのまま
- 1つ目の単語の登録をする部分は使えそうなら使う

今回は

てきとーな構造でも動くプログラムはできる

でも遅い…(初心者はとりあえず動けば良い…？)

データ構造とは

- 2分木とか線形リストとか
- 計算量という概念が重要
- ケース・バイ・ケースで構造を選んで利用する

いろいろ調べて適した構造を使えるようになろう！
（おそらく今回はこちらが狙いかな…）

伏見が別にまとめます！

４つのデータ構造を考える(1)

1. ドキュメントのデータ構造
   前回のでOK(ぶっちゃけ今回は重要ではない)

2. 1つ目の単語のデータ構造
   構造体に最低限必要な変数を考える
   - 単語自体
   - 転置索引
     同じ行なら追加しないの判定
     出現回数にも使える
   - この単語に続く2つ目の単語の一覧(3. へ)
   - データ構造を実現するのに必要なもの
     単方向リストなら次の構造体へのポインタ

計算とか後々のことを考えて変数やデータ構造を考える

(僕は出現回数はintで持ってて，indexでアクセスするため動的配列に)

４つのデータ構造を考える(2)

3. 2つ目の単語のデータ構造

基本的には1つ目の単語の構造体と同じ

- 単語自体
- 転置索引
- データ構造を実現するのに必要なもの
- その他

(僕は同じ単語が入ってる1つ目の単語の配列のindexをもたせてる)

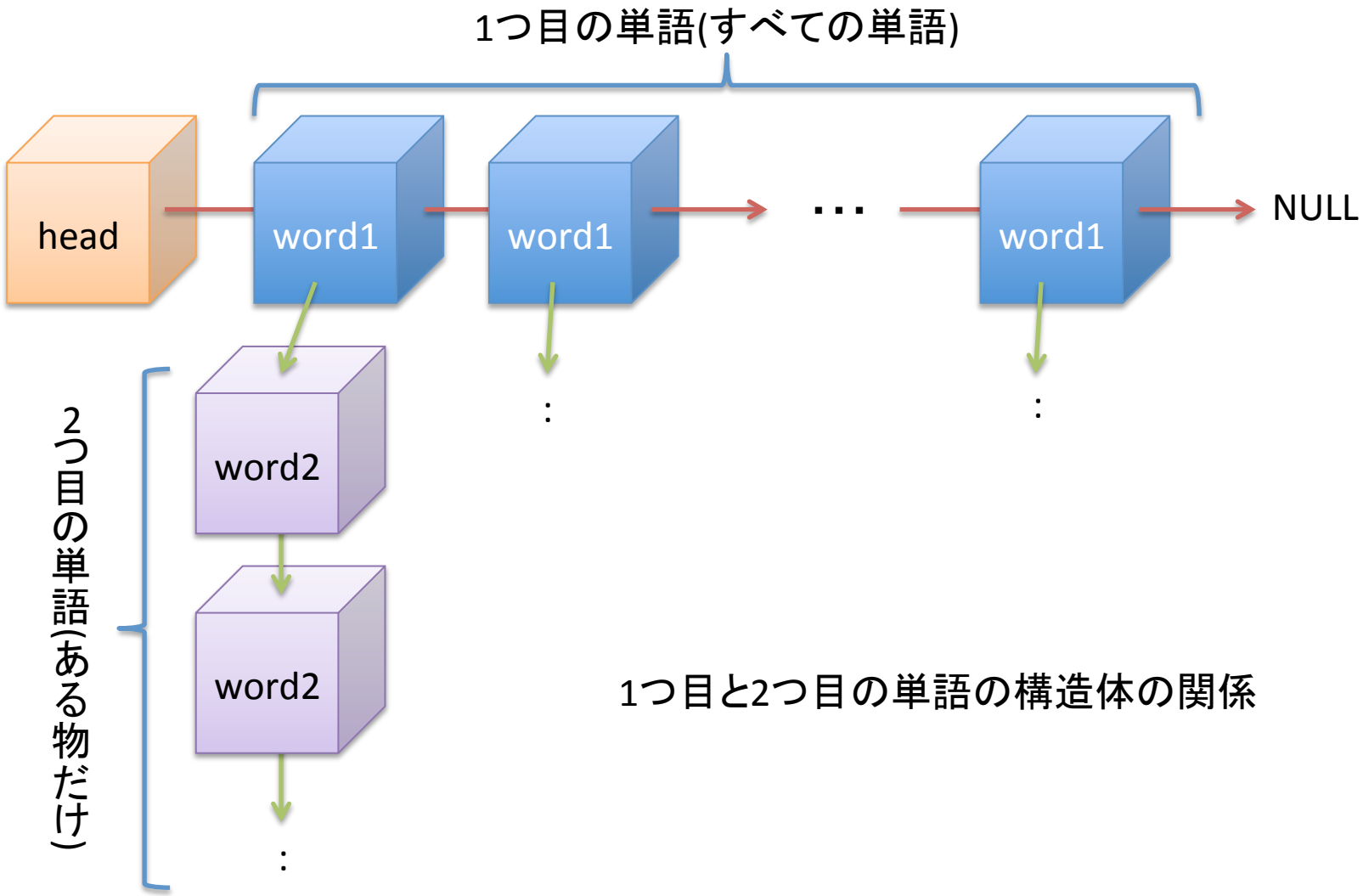
4. 結果のデータ構造

2つの単語とoverlap係数

リストでもいいけどqsort(※)とかを使うんだったら配列が良い

※ http://www.cc.kyoto-su.ac.jp/~yamada/ap/qsort.html

線形リストだとイメージはこんな感じ



やることを整理(再掲)

1. 行毎にドキュメントを保存
2. 1つ目の単語を登録
3. 2つ目の単語を1つ目の単語と関連付けて登録
4. 1,2の単語のセット全てに対してoverlap係数を計算
5. 係数は単語のセットと一緒に配列などで保存
6. 計算した結果をソート
7. 入力された値の数だけ上位から表示

2までは今までのものをいじるだけでも動くが,
データ構造を変更した方がいい場合も多いのでは…？
（それぞれ実装が違うので自分で考えよう…）

2つ目の単語の登録方法の概要

※説明の簡単化のため，線形リストを想定して書く

行の先頭以外(前回の単語がある)なら

- 前回の単語が入っている構造体をリストから探す
- その構造体が持っているリストに2つ目の単語があるか探す

単語がすでに登録されているなら

- 転置索引によって同じ行(ドキュメント)か判定

同じ行なら

転置索引に追加しない(出現回数をカウント)

同じ行でないなら

転置索引に追加

登録されていないなら

要素を作成して転置索引などを追加

次はOverlap係数

- 1つ目の単語のリストを順に見ていく
  1つ目の単語の出現回数を得る |S(a)|
- 2つ目の単語のリストがあるだけ見ていく
  セットで出た時の出現回数を得る |S(a,b)|
- 2つ目の単語を1つ目の単語のリストから探しだす
  2つ目の単語の出現回数を得る |S(b)|

これで計算できる！

結果の保存

Overlap係数は求まったので
「2つの単語と係数のセット」を保存していこう

ソートの計算量のことも考えてソートを選ぶと
結果を保存するデータ構造がおのずと決まる

qsortを使うときは動的配列を使うと良い

動的配列？

- Cは配列の要素数が固定になってる

  ex) int a[3]; // 要素数3で変更できない

- 配列を動的に確保する

  soft2の勾配法のところでやったことある

  double *g = (double *)malloc(dim * sizeof(double));

- 足りなくなったらreallocする(要素数を持っておく)

これを構造体に対して行う！

ソート

- ソートもいろいろある
  - 選択，挿入，クイック，マージ，ヒープ，ボゴ…
- こいつも計算量の概念が重要
- ケース・バイ・ケース

伏見が別にまとめます！

まとめ

データ構造とソートを計算量を考えて選ぶのが大事

→ 今回の課題の目的だと思う

2つの単語を空白区切りでくっつけて2分木をつくるってのも面白いです(by 安東)

news.txtくらいなら線形リストでもOK

わからない人は最強の織田さんに聞こう！

(twbtarai.blackpepper@gmail.com)

プログラミング講習会のMLに投げても答えてくれます！