アルゴリズムとデータ構造

グループワーク

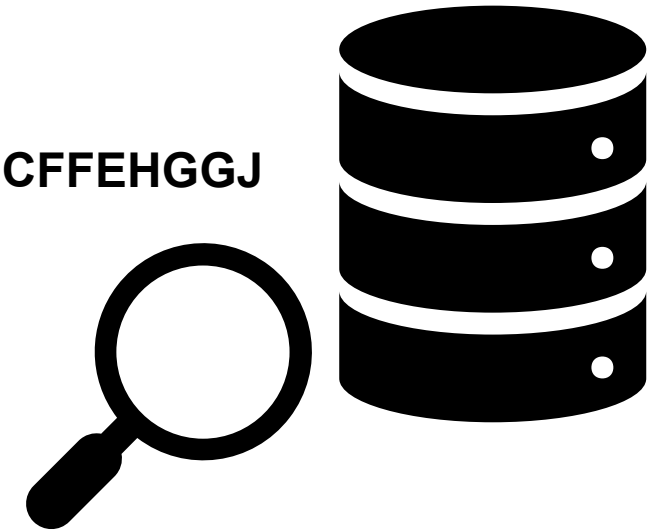
課題

課題：類似キーワードはある？

特定のキーワードと**類似する**
ものが存在するか，できるだ
け早く知りたい

AAAAAABDEDHAIAD
AAAAAACHJACDCCJ
AAAAACCECAGGBCE
AAAAACFDICCCJJJ
AAAAACHBIIFCIDG
AAAAACHHJJECBFD
AAAAAEAGEHCEEDA
AAAAAEEHGDEDEBD
AAAAAEGJAHGGAJI
AAAAAEHEFICFFDD
AAAAAEIFGDJHJDI
…

HGCCBBBCFFEHGGJ



多くのキーワードを格納
したデーベースがある.

- データベースのデータ
  文字列のセット $B = b_1, b_2, \ldots, b_N,\ (b_i \in \{A, \ldots, J\}^{\ell})$
- 索引データ（要作成）
  - 200MByte以内．フォーマット自由.
- クエリデータ
  文字列 $Q = q_1, q_2, \ldots, q_L,\ (q_i \in \{A, \ldots, J\}^{\ell})$
- 検索結果（要作成）
  ビット列 $V = v_1, v_2, \ldots, v_L,\ (v_i \in \{1,0\})$
  - 各クエリについて編集距離3以内のキーワードがBに存在する場合は1,しない場合は0
- N ,L： 10^6, ℓ： 15 とする.

- 評価指標
  - 実行速度： V の計算時間　※ 索引構築時間は含めない
  - 精度：V と正解のハミング距離

データベース，クエリ，結果の例

AAAAAABDEDHAIAD
AAAAAACHJACDCCJ
AAAAACCECAGGBCE
AAAAACFDICCCJJJ
AAAAACHBIIFCIDG
AAAAACHHJJECBFD
AAAAAEAGEHCEEDA
AAAAAEEHGDEDEBD
AAAAAEGJAHGGAJI
AAAAAEHEFICFFDD
AAAAAEIFGDJHJDI
…

IGCFFHHJGDGEIAH
GDHFIEEEDAGGEDH
BADGICJJJCEHDGH
HJCGICFBJEADDCJ
IHAFFBHIBHDACID
BJBGIBAFADJEJJB
FFJGJICACHJEIGJ
HGCCBBBCFFEHGGJ
FIABJFADADEIFFE
FIECDAIHFJDIEJC
AFCJDHEEHHFAIGJ
…

01010111110000101001001O

計測に関して

- 実行速度
  - 計測環境（予定）
  - OS: Ubuntu 24.04.3 LTS， gcc: v13.3.0， CPU: AMD Ryzen Threadripper 3970X， build-essentialのみで開発できるプログラムを対象とする
  - 索引構築：4分で打ち切り
  - 検索：1分で打ち切り
  - メモリ5Gを超えたプログラムの実行は保証しない

成果物提出のルール

- プログラムのソースを二つ提出.
- 提出するプログラムはC言語で記述すること.
- 指定環境（後述）で動作確認すること.
- 使用メモリの上限は5Gbyte.
- 提出：
  - 準備用のプログラムはprep_グループ番号.c，検索用のプログラムはsearch_グループ番号.c，Moodleの「成果物提出」に**グループの代表者が提出**.
    - 例： グループ番号が0の場合は prep_0.c, search_0.c を提出する.

性能評価

- 実行速度（経過時間），精度を評価指標とする．
- 各グループの得点は以下により求める．
  - 各指標の順位の総和を加算．
  - 各指標の１～４位にはそれぞれ，-10，-5，-2，-1を加算．
  - また，「精度」に関してのみ，最下位から数えて3番目までのグループにそれぞれ 50, 30, 10を加算．
  - 複数チームが同順位の場合，順位がx，チーム数がyであった場合，x, x+1,..., x+y-1までの順位に相当する得点の平均を付与します．
    - 例）速度順が，A班，B班 = C班, D班, …であった場合，A班は 1-10，B，C班は{(2-5) + (3-2)}/2 = -1, Dは4-1となります．
  - 中間計測に参加するグループには，-1を加算．
    （不具合確認のためにも，参加をお勧めします．）
- 例えば，実行速度で1位，精度で5位，中間計測に参加した場合の得点は，1 - 10 + 5 -1 = **-5**
- 得点は低いほど良い．

動作確認

- 提出物は，以下の環境で動作することを確かめてから提出すること．
- OS　：Ubuntu 24.04.3 LTS
  開発：build-essential

- Linux環境をお持ちでない方は，**インストールして利用することを推奨**しますが，インストールせずに利用することもできます．（次を参照）

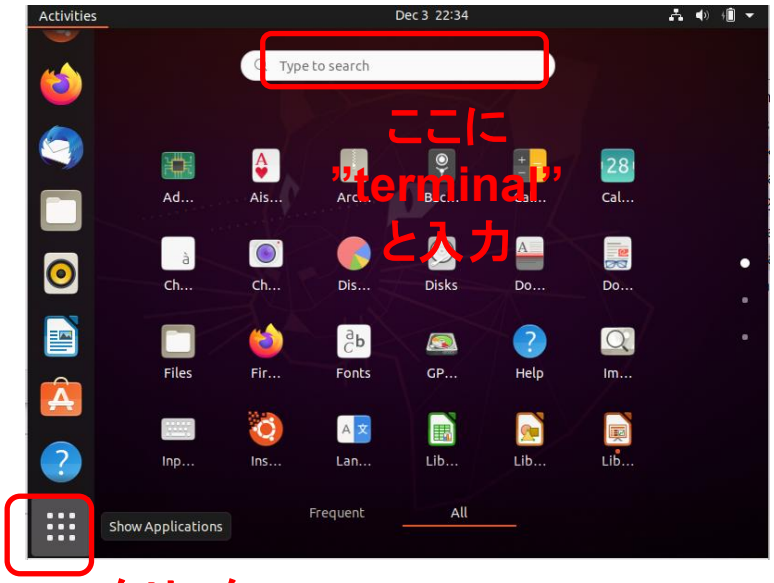
1. https://releases.ubuntu.com　より isoイメージを入手
2. VirtualBoxをインストール
3. 仮想マシンの作成

4. isoイメージを設定

「光学ドライブ」をクリックして
「ディスクファイルを選択」をクリックした後で
イメージを設定

5. 仮想マシンの電源を入れ

6. Live CDを起動



クリック

7. terminalを立ち上げて以下を実行

sudo apt-get install build-essential

生成AIの利用について

- 生成AIの利用は禁止しません.
- ただし，生成AIの出力をそのまま使うことは禁止します．成果物の内容について自分の言葉で詳細に説明できることを条件とします.

どんな方法で解くか？

- 準備時間（索引構築）は実行時間に含まれない.
- 検索を早くする索引を構築できるか？
- 速度と精度のトレードオフはあるか？

- ぜひ活発な議論を
- ソースの共有
  - github (https://github.com/)
  - Dropbox
  - Google Drive

グループワーク進め方

- 進め方は自由ですが，過去に見られたケースをいくつか紹介します．

その１

1. 議論により問題に対する理解を深め，様々なアイデアを出す．
2. 個人で取り組み，一番良い解法をグループの解法とする．

その２

1. 議論により問題に対する理解を深め，様々なアイデアを出す．
2. 個人，あるいはペアで解法を考え，それを元に方針を見出す．
3. 分担して実装し，テストデータでの性能を確認しながら，さらに良い手法がないか議論をする．
4. 発表資料を分担して準備する．

グループディスカッション

- 本日不在のメンバーに後日連絡を取る必要がある場合は，グループフォーラムをご利用ください．