

「データ工学」期末レポート課題（締め切り：7/29 23:59、提出先：e-class）

以下の課題1～7を解きなさい。ただし、途中経過が分かるように丁寧に解答すること。

課題1.

情報抽出が役立つ場面（どのようなテキストからどのような情報を抽出するのか、それがどのように役立つか）を、一つ具体例をあげ説明しなさい。ただし、授業で取り上げた例（意見抽出と異動情報抽出）を使ってはいけない。

課題2.

ある検索システムである検索質問による検索を行ったところ、以下の表の通りの結果となった。次の(1)から(5)の各問い合わせなさい。四捨五入する際は小数点第三位を四捨五入すること。

- (1) 精度を求めなさい。
- (2) 再現率を求めなさい。
- (3) 正解率を求めなさい。
- (4) F値 ($\alpha = 1/2$) を求めなさい。
- (5) 検索システムを評価する際、評価指標として正解率を使うのが不適切な場合が多い。その理由を答えなさい。

	検索された文書数	検索されなかった文書数
適合文書数 (検索結果として妥当な文書数)	300	200
非適合文書数 (検索結果として不適切な文書数)	100	2500

課題3.

テキスト検索において索引語を作成する際、接辞処理を行うことがある。接辞処理とはどのような処理かを答えなさい。また、接辞処理を行う目的を答えなさい。

課題 4.

以下の 3 つのテキスト (D1～D3) と検索質問 (Q) を考える。次の(1)と(2)に答えなさい。

テキスト 1 (D1) transfer of data blocked by a gate

テキスト 2 (D2) debug of soft conducted by a soft engineer

テキスト 3 (D3) data transfer conducted by an engineer

検索質問 (Q) gate soft engineer

(1) 3 つのテキスト (D1～D3) と検索質問 (Q) の特徴ベクトルを tf・idf 法で求めなさい。解答する際にはベクトルの各次元が何の索引語を表すかを明確にすること。索引語には単語を用いること。単語はスペースで区切られているものとし、ステミングは行わない。ただし、不要語リストとして「a」と「an」が登録されているものとする。また、tf の計算の際には、文書中の単語数での正規化は行わなくてよい。 \log の底は 10 とし、 $\log_{10} 2 = 0.3$ 、 $\log_{10} 3 = 0.5$ として計算すること。

(2) (1)で作成した特徴ベクトルを用いて検索質問 (Q) と各文書 (D1～D3) との類似度を内積で求めなさい。

課題 5.

検索対象のテキストとして以下の文字列を考える。次の(1)と(2)に答えなさい。

検索対象テキスト： わにはにわには X

ただし、X には、あなたの苗字の最初のひらがな 1 文字を入れること（例えば、田村の場合、「わにはにわにはた」となる）。

(1) 上記検索対象テキストの文字列に対して、文字ユニグラム索引転置ファイルと文字バイグラム索引転置ファイルを示しなさい。ただし、位置情報はバイト単位とし、ひらがな 1 文字は 2 バイトとする。また、テキストの最初の「わ」の位置は 0 とすること。

(2) (1)で作成した文字ユニグラム索引転置ファイルを用いて、「はにわ」という文字列を検索する過程を示しなさい。図を用いる場合は、図だけではなく文による説明も行い、検索過程が分かるように丁寧に解答を行うこと。

課題 6.

(1) 検索文字列集合が{aa, ba, baa, bba, bac}の場合のマシン AC (goto 関数、failure 関数、output 関数) を構築しなさい。

(2) (1)で構築したマシン AC を用いて、テキスト「abbaac」から検索文字列を検索する様子と結果を示しなさい。

課題 7.

- (1) 検索文字列「hot」を誤り 1 以下で受理する非決定性有限オートマトンを構築しなさい。
- (2) (1)で構築した非決定性有限オートマトンを決定性有限オートマトンに変換しなさい。
- (3) 「hat」というテキストに対して、検索文字列「hot」の近似文字列照合を考える。(2)で構築した決定性有限オートマトンを用いて、誤り 1 以下で照合に成功するか否かを判定しなさい。