Praktische Informatik II SS 2007

Übungsblatt 10

Abgabe: 9.07.2007

Aufgabe 1 Kein blaues Wunder mehr!

Leider besteht ein hoher Prozentsatz an E-Mail aus "überraschenden Angeboten", den "besten Medikamenten", "heissen Börsentippps" und ähnlichen Angeboten. :-(Ihr sollt einen Spam-Filter programmieren, der unter Verwendung des Boyer-Moore-Algorithmus das Vorkommen von typischen Spam-Anzeichen in einem Text überprüft.

Aufgabe 1.1 Suchen

Programmiert eine Methode, die eine E-Mail (char-Array) nach einem Suchbegriff mithilfe der Methode von Boyer und Moore unter Verwendung der *bad character* Strategie durchsucht.

Aufgabe 1.2 Spam oder nicht Spam

Schreibt nun eine Klasse, die ein Array von Suchbegriffen mit einer Bewertung der "Spam-igkeit" enthält. Eine Methode soll einen übergebenen Text nach allen Suchbegriffen durchforsten und eine Gesamt-Spam-Bewertung ermitteln. Suchbegriffe können auch negative Bewertungen haben, also dagegen sprechen, dass es sich bei einer Mail um Spam handelt.

Aufgabe 1.3 Ganz oder gar nicht

Erweitert die Boyer-Moore-Implementierung um die Verwendung der *good suffix* Strategie.

Aufgabe 1.4 Boyer-Moore mit Wildcards

Wie müsste der Boyer-Moore Algorithmus abgeändert werden, damit auch Wörter mit sogenannten Wildcards gesucht und gefunden werden können? Beschränkt Euch dabei auf Wildcards, die genau ein Zeichen repräsentieren.

Beispiel: Text = 'EinSuchbeispielfuerWildcardSuche', Muster = 'Wil?card' → Ergebnis 19.

Aufgabe 1.5 Pro und Contra

Wie kann man nach mehreren Begriffen *gleichzeitig* suchen? Welche Vorteile und welche Nachteile/Probleme entstehen dabei beim Boyer-Moore Verfahren?

Anmerkung: Das interessante Alphabet für diese Untersuchung ist der erweiterte ASCII-Zeichensatz.

Anmerkung 2: Es ist nicht nötig, als "Testdaten" eine Sammlung Eurer schönsten Spam-Mails mitzuliefern.