

# Анализ текстов на естественных языках

## Контрольная работа

**Задание 1.** (2 балла) Дан язык  $L = \{aa, bb, abab, baba, aaaa, bbbb\}$  над алфавитом  $\Sigma = \{a, b\}$ . Вычислить оценки максимального правдоподобия для параметров биграммной языковой модели, с их помощью вычислить вероятность строки  $abba$ .

**Задание 2.** (5 баллов) Звуки в естественных языках можно разделить на гласные, согласные и непонятные. В зависимости от позиции в слове непонятные звуки могут играть роль как гласных (например, нести ударение), так и согласных (например, закрывать слог и влиять на длительность или оттенок предшествующего гласного). Дан алфавит  $\Sigma = \{a, b, j\}$ , набор помет  $T = \{C, V\}$  и обучающая выборка  $(\mathcal{X}, \mathcal{Y}) = \{(a, V), (ba, CV), (baj, CVC), (ajb, VVC), (jab, CVC), (bja, CVV)\}$ . Требуется определить модель слоговой структуры как биграммную скрытую марковскую модель и с её помощью найти слоговую структуру (наиболее вероятную аннотацию) строки  $jaj$ . Параметры модели задать на основе оценок максимального правдоподобия.

**Задание 3.** (3 балла) Требуется построить систему автоматической расстановки переносов для русского языка на основе скрытой марковской модели. Выпишите спецификацию модели.

**Задание 4.** (2 балла) Дана вероятностная контекстно-свободная грамматика  $\langle N, \Sigma, R, S, q \rangle$ , где  $N = \{A, B, C, S\}$ ,  $\Sigma = \{a, b\}$ ,  $R = \{S \rightarrow CB \mid BC, C \rightarrow BA \mid AB, A \rightarrow a, B \rightarrow b\}$ ,  $q(S \rightarrow CB) = 0.7, q(S \rightarrow BC) = 0.3, q(C \rightarrow BA) = 0.4, q(C \rightarrow AB) = 0.6, q(A \rightarrow a) = 1.0, q(B \rightarrow b) = 1.0$ . Найти вероятность порождения строки  $bab$ .

Для сдачи нужно набрать 7 баллов.