

СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КУРСОВ ПРОЕКТ

ПО СИСТЕМИ, ОСНОВАНИ НА ЗНАНИЯ

Тема:

Система за проверяване на правописни грешки

Студенти:

Мартин Василев Живков, 3-та група, №72071

Веселин Василев Живков, 3-та група, №72103

София, януари 2023 г.

1. Формулировка на задачата

Задачата, която решава програмата е поправката на сгрешена по правопис дума. При подадена сгрешена по правопис дума програмата трябва да може да изведе коректната дума, която е близка по правопис на сгрешената.

1. Използвани алгоритми

Извикването на метода correction(w) се опитва да намери най-добрата правописна корекция за дадена дума w. Това става, като се разглеждат всички възможни правописни корекции (кандидати) и се избира тази, за която има най-голяма вероятност да бъде предвидената корекция. За да се намери тази вероятност, се използва теоремата на Бейс, която включва два вероятностни модела: един за това колко вероятно е да се направи правописна поправка и един за това колко вероятно е да се направи грешка. Използва се и механизъм за подбор, за да се избере кандидат-корекцията с най-висока комбинирана вероятност.   
Източник: <https://norvig.com/spell-correct.html>

1. Описание на програмната реализация

Програмната ни реализация се състои от 3 файла – index.ts, spellChecker.ts и текстов файл, чрез който да се зарежда речник от правилно написани думи, от които да се учи програмата.

Във файла index програмата се стартира, като са представени и няколко теста за демонстрация.

Във файла spellChecker на първо място използваме външната библиотека ‘fs’, която ни позволява да прочитаме файл от операционната система. Имаме дефиниран класът SpellChecker, в който е имплементирана цялата логика на поставена задача. SpellCheck съдържа няколко важни метода, които ни помагат да реализираме задачата.

Метод ‘correct’ - при подаден низ (сгрешена дума), връща като резултат коригираната дума.

Метод ‘trainFromFile’ – при подаден път до файл, зарежда данните от файла, с които да се учи програмата

Метод ‘getCandidates’ – помощен метод, който връща множество от предложени кандидати за коригирана дума

Метод ‘max’ – връща най-подходящия кандидат от множеството предложени кандидати (най-често срещания близък кандидат)

Метод ‘edit’ – намира всички възможни кандидати, като извършва различни комбинации по сгрешената дума като например изтриване на символ, транспониране на символ, заместване на символ и вграждане на символ.

Метод ‘known’ – проверява дали предложена дума всъщност е съществуваща дума в реалния свят

1. Примери, илюстриращи работата на програмната система

Първо си задаваме няколко теста за демонстрация, като се масив от двойка от сгрешена дума и очаквана дума:

Text

Description automatically generated

Изход на зададените случаи:

A picture containing text, plaque

Description automatically generated

1. Литература

How to Write a Spelling Corrector - <https://norvig.com/spell-correct.html> от Питър Норвиг, 2007г.