Морфологический анализатор

Тестовое задание

ЗАДАЧИ

Разработать библиотеку классов и консольный интерфейс к ней. Библиотека должна выполнять функции графематического и морфологического анализа. Графематический анализатор выделяет в исходной строке отдельные слова — последовательности кириллических символов. Морфологический анализатор выполняет две функции: определяет морфологические характеристики слова и ставит слово в указанную форму.

Консольный интерфейс должен предлагать два сценария работы:

1. Морфологический анализ всех слов во введённой строке. Пользователь вводит строку (например, одно предложение на русском языке). Графематический анализатор выделяет в строке все слова. Морфологический анализатор выдаёт все возможные варианты разбора каждого слова:

   Исходная строка: *Папа мыл пол.*

   Результат работы: *«Папа» — начальная форма «папа», существительное, одушевлённое, мужской род, единственное число, именительный падеж; «мыл» — начальная форма «мыть», глагол, несовершенный вид, переходный, прошедшее время, мужской род; «пол» — начальная форма «пол», существительное, неодушевлённое, мужской род, единственное число, именительный или винительный падеж.*

2. Склонение одного введённого слова. Пользователь вводит строку: слово в начальной форме и желаемые морфологические характеристики. Морфологический анализатор склоняет слово в указанную форму:

   Исходная строка: *тряпка: творительный падеж, множественное число.*

   Результат работы: *тряпками.*

ТЕОРИЯ

Графематический анализ — это первый этап при анализе любого текстового материала, неважно, написан он на естественном языке или формальном. На этом этапе строка (список символов) разбивается на сегменты: абзацы, предложения, устойчивые выражения, слова и специальные конструкции. В зависимости от задачи, специальными конструкциями могут быть имена, даты, формулы — любые последовательности слов и символов, которые должны обрабатываться как единое целое.

В нашем случае процедура графематического анализа очень проста. Нам достаточно выделять последовательности кириллических символов (как в верхнем, так и в нижнем ре-

гистре). Для успешного выполнения этой части задания будет полезно ознакомиться с регулярными выражениями.

Морфологический анализ — это определение по выделенному слову его атрибутов: начальной формы, части речи и различных морфологических характеристик (род, число, падеж, вид глагола и т. д.). Кроме того, в задачи морфологического анализатора часто входит склонение слов.

Качественные морфологические анализаторы строятся на основе словарей (что и предлагается в тестовом задании). Про строение морфологических словарей есть неплохая статья на Хабре. В нашей задаче словарь очень прост — в нём всего 10 слов. Кандидату предлагается самостоятельно выбрать формат представления словаря и реализовать поиск по нему.

АЛГОРИТМ

К заданию приложены два файла: words.txt и flexia.txt. В них приведён морфологический словарь на десять слов. Первым этапом работы библиотеки должна быть загрузка морфологического словаря из этих файлов. Файл words.txt содержит основы[1] слов, их постоянные морфологические характеристики и идентификатор системы склонения. Файл flexia.txt содержит системы склонения слов: здесь каждому идентификатору сопоставляется список окончаний и морфологические характеристики, соответствующие данному склонению слова. Расшифровка морфологических характеристик приведена в Приложении.

После загрузки в оперативную память морфологического словаря, система готова к обработке пользовательских запросов. Если пользователь желает получить результаты морфологического анализа для текста на русском языке, система выделяет из текста все слова, а затем каждое слово анализируется с помощью словаря. Если слова в словаре не оказалось, вместо его морфологических характеристик следует написать «неизвестное слово». Если слову соответствует несколько наборов морфологических характеристик, следует определить, какие характеристики присущи всем вариантам разбора, а какие характеристики являются омонимичными. Омонимичные характеристики группируются:

> *«пол» — начальная форма «пол», существительное, неодушевлённое, мужской род, единственное число,* ***именительный или винительный падеж****.*
>
> *«мамы» — начальная форма «мама», существительное, одушевлённое, женский род,* ***единственное число и родительный падеж или множественное число и именительный падеж****.*

Если пользователь хочет склонять одно слово, графематический анализ не производится. Слово сразу ищется в морфологическом словаре, а затем выбирается соответствующая

[1] Следует отметить, что термины «основа слова» и «окончание» здесь употребляются не в грамматическом смысле. Под основой тут понимается та часть слова, которая не изменяется в процессе склонения, а под окончанием — вся изменяемая часть. Например, для слова *мыть* в основу войдёт только буква *м*, а всё остальное будет изменяться при склонении.

форма этого слова. Если введённое пользователем слово не содержится в словаре, выводится сообщение о том, что слово не найдено.

ТРЕБОВАНИЯ

1. При реализации тестового задания разрешается пользоваться только стандартной библиотекой .NET. Нельзя использовать сторонние решения для графематического и морфологического анализа.
2. Требуется разработать структуру классов для морфологического словаря, графематического и морфологического анализа.
3. Вся логика приложения должна содержаться в библиотеке классов. Консольное приложение — это только интерфейс к функциям библиотеки.
4. Нужно написать модульные тесты для всех открытых методов библиотеки.
5. Все классы, поля и методы библиотеки должны быть прокомментированы.
6. При разработке следует использовать систему контроля версий Git. Разработанную программу нужно загрузить в приватный репозиторий на сайте Bitbucket. Когда разработка будет окончена, добавить пользователя DFirstov к репозиторию.
7. При обнаружении ошибок в исходных данных, следует их исправить и добавить в репозиторий файл readme.txt с описанием того, что было исправлено.

ПРИЛОЖЕНИЕ

Расшифровка морфологических характеристик в словаре:

сущ	существительное
гл	глагол
мест	местоимение
союз	союз
пред	предлог
од	одушевлённое
не	неодушевлённое
жр	женский род
мр	мужской род
ср	средний род
нв	несовершенный вид
пе	переходный глагол
личн	личное местоимение
1л	первое лицо
2л	второе лицо
3л	третье лицо
ед	единственное число
мн	множественное число
ип	именительный падеж
рп	родительный падеж
дп	дательный падеж
вп	винительный падеж
тп	творительный падеж
пп	предложный падеж
мп	местный падеж
инф	инфинитив
нв	настоящее время
пв	прошедшее время