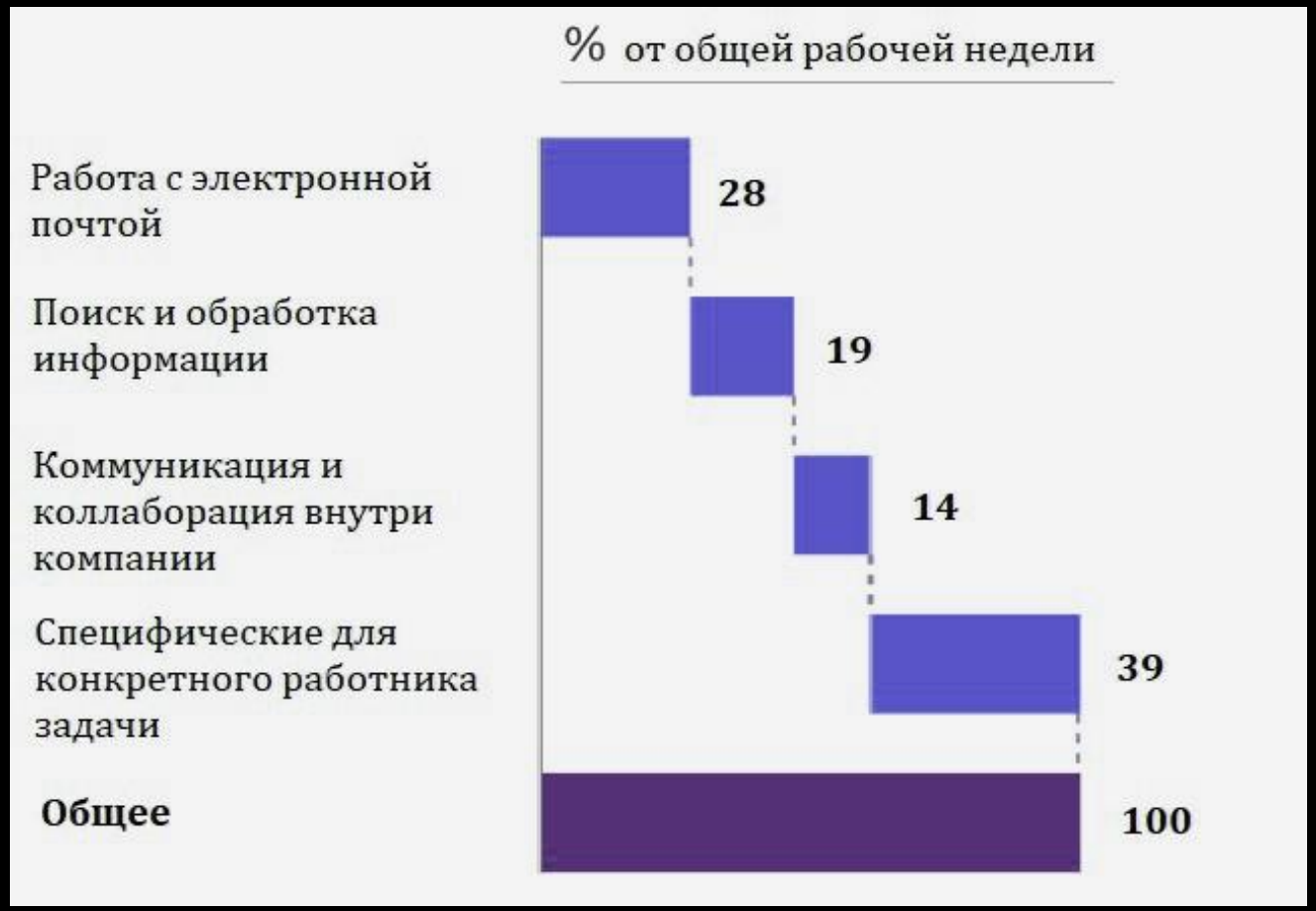
Московский Авиационный Институт
(Национальный исследовательский институт)

«Информатика: проблемы, методы, технологии» (IPMT-2021)

Исследование методов нечеткого сравнения строк и их применение в алгоритме поиска опечаток в тексте

магистрант, каф.319 Потапова А.А.

Автоматизация поиска опечаток



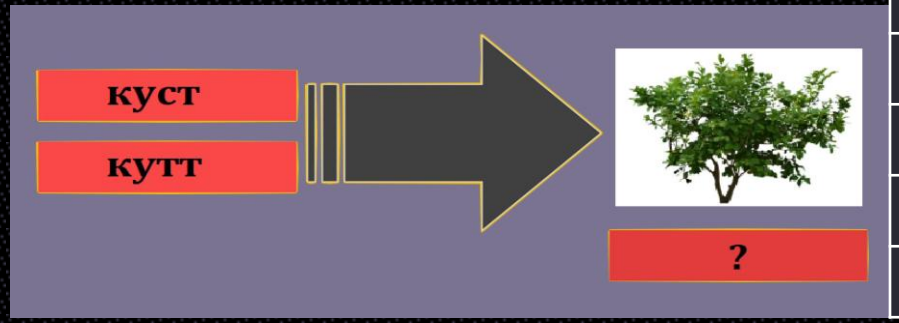
Данные об ошибках

Номер	Вид ошибки	Частота (в %)
1	Замена “и” - “е”	27
2	Замена “а” - “о”	25
3	Лишний пробел, слово должно быть написано слитно	9.1
4	Отсутствие пробела, вместо одного слова два	8.5
5	Потеря одной из удвоенных букв	6.6
6	Замена глухой буквы звонкой и наоборот	3.6
7	Гласные после ц	2.7
8	Удвоение одиночной буквы	2.6
9	Потеря “ь”	1.3
10	Лишний “ь”	0.6
11	Замена “ё” - “е”	0.1

Методы сравнения строк

- Soundex
- Расширение выборки

- Алгоритм Вагнера-Фишера:
  - мера Левенштейна
  - мера Левенштейна-Дамерау

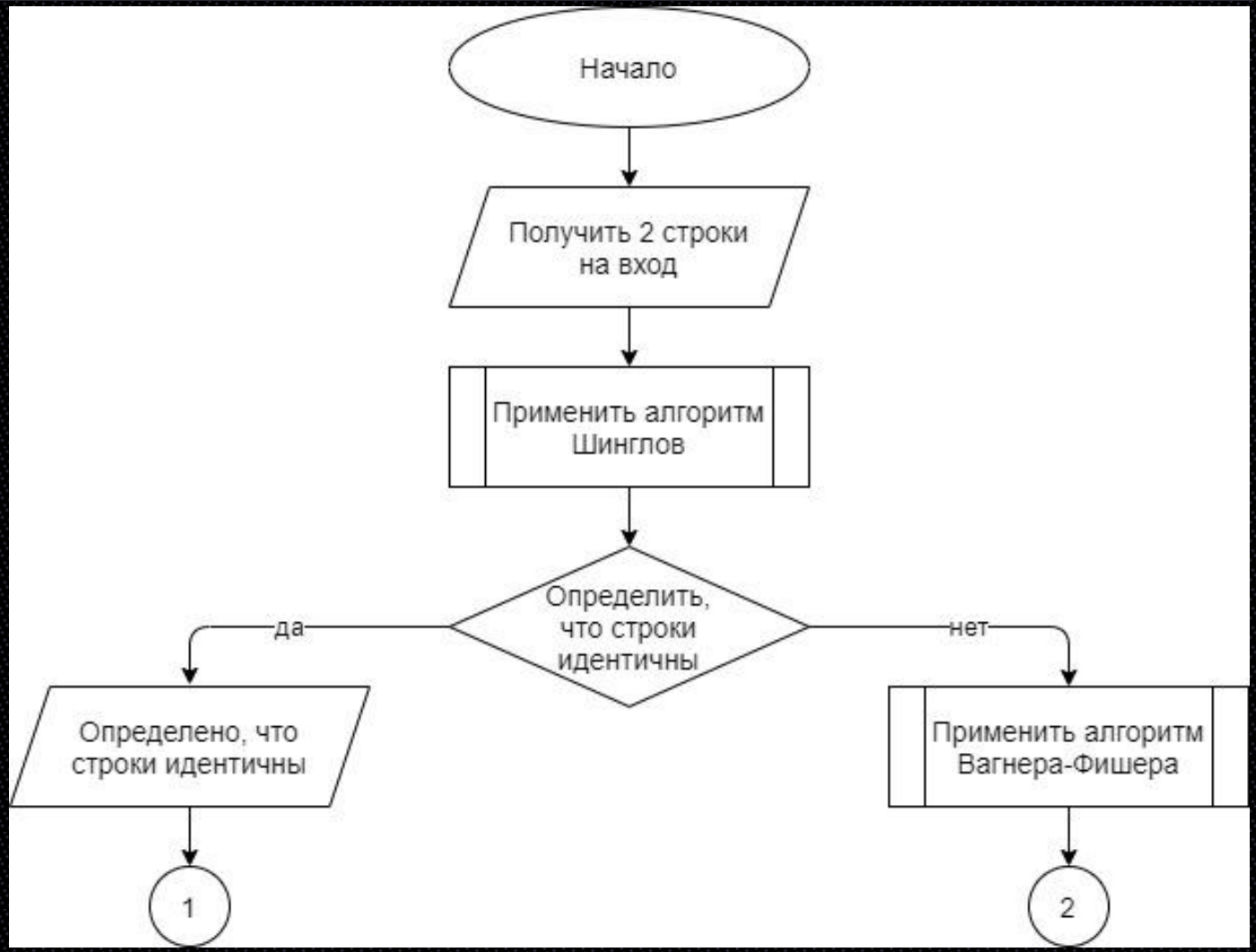


		п	о	л	и	в	а	ю		к	у	с	т
	0	1	2	3	4	5	6	7	8	9	10	11	12
п	1	0	1	2	3	4	5	6	7	7	8	9	10
о	2	1	0	1	2	3	4	5	6	7	8	9	10
л	3	2	1	0	1	2	3	4	5	6	7	8	9
в	4	3	2	1	1	1	2	3	4	5	6	7	8
а	5	4	3	2	2	2	1	2	3	4	5	6	7
ю	6	5	4	3	3	3	2	1	2	3	4	5	6
	7	6	5	4	4	4	3	2	1	2	3	4	5
к	8	6	6	5	5	5	4	3	2	1	2	3	4
у	9	7	7	6	6	6	5	4	3	2	1	2	3
т	10	8	8	7	7	7	6	5	4	3	2	2	3
т	11	8	9	8	8	8	7	6	5	3	3	3	2

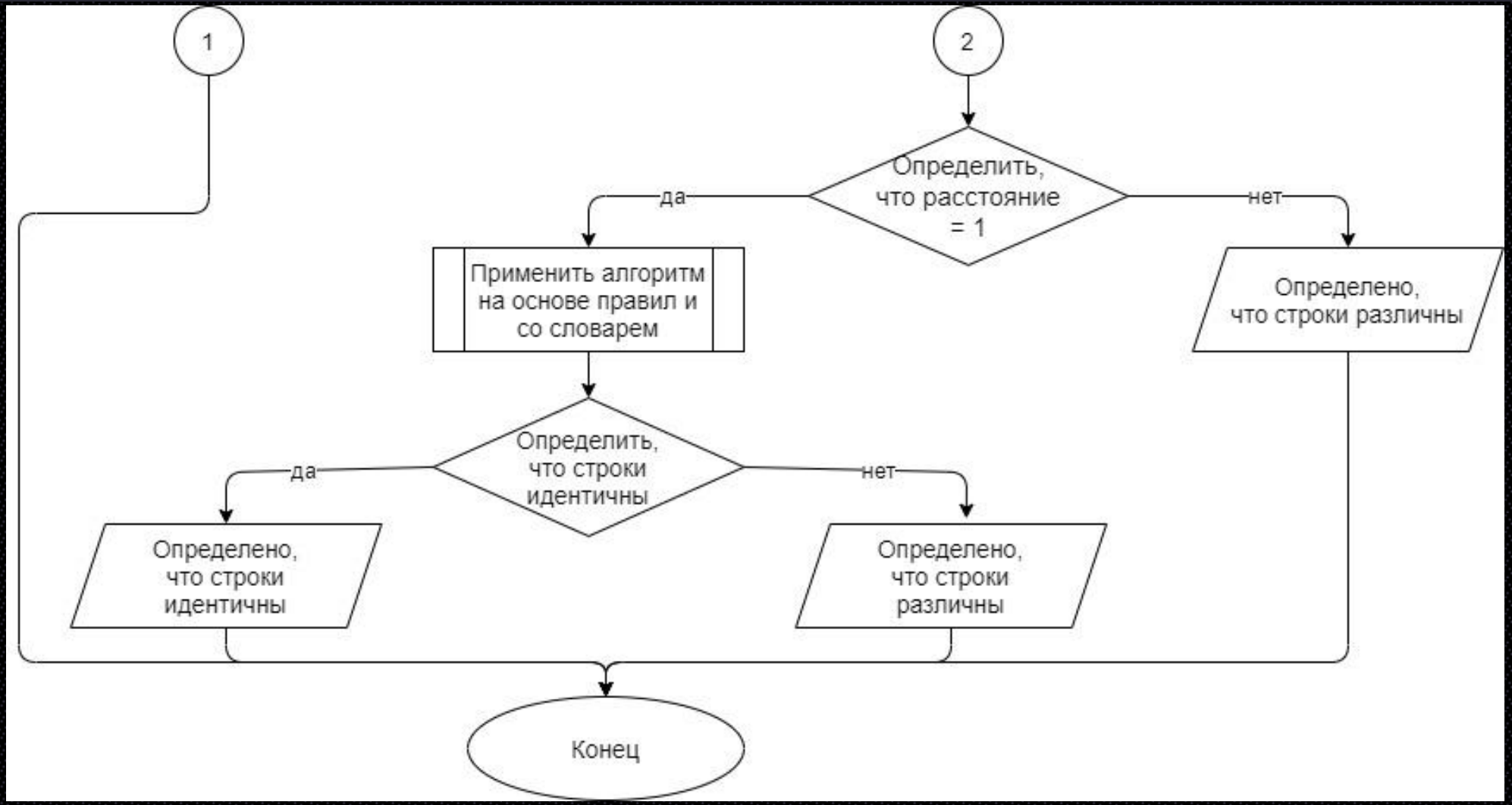
Алгоритм поиска опечаток

- Производится поиск редакционного расстояния
- Если расстояние равно 1, то в строке может быть опечатка

- Фредерик Дамерау заметил, что 80% ошибок при наборе текста человеком это транспозиция



Алгоритм поиска опечаток



Выводы

- Создан список на основе статистических данных об ошибках для создания правил проверки
- Проведен анализ алгоритмов нечеткого сравнения строк
- Предложен алгоритм, который позволит улучшить качество работы существующих систем автоматизированного анализа текста за счет поиска и исправления опечаток во входных текстах

Спасибо за внимание!

Исследование методов нечеткого сравнения строк и их применение в алгоритме поиска опечаток в тексте

магистрант, каф.319 Потапова А.А.