**Результаты тестирования ресурса на предмет нахождения в текстах местоимений «я-мы» группы**

Материал: тексты ресурса «ЯПлакалъ»:

1. Автор KalinAKalina (https://www.yaplakal.com/members/member258716.html) – около 5000 слов
2. Автор SESHOK (https://www.yaplakal.com/members/member328581.html) – около 5000 слов

Ошибки, обнаруженные в работе программы:

|  |  |
| --- | --- |
| Ошибка | Комментарий |
|  | Программой выделяется область, совпадающая по длине с местоимением, но выделение происходит «с опережением» |
|  | Программа не опознала местоимение группы «я-мы» в первом случае из-за наличия тире в составе орфографического слова (в остальных случаях ПО среагировало на появление этой формы) |
|  | Программа не выделяет форму «моём», написанную через «е». Вероятно, это следствие омографии с формой 1 лица множественного числа настоящего времени от глагола «мыть» |
|  | Вместе с местоимениями «я-мы» группы программой выделились местоимения «ты-вы» группы |
|  | В первом случае местоимение не выделилось из- за пунктуационного знака в составе орфографического слова. Два или более знаков препинания, вероятно, автоматически означают конец слова. |
|  | ПО не выделило местоимения с постфиксом «-то» и форму, ошибочно не отделённую пробелом от предыдущего слова |

Возможные варианты решения обозначенных проблем:

1. Заменять знаки препинания пробелами, после чего проводить токенизацию по промежуткам из одного или нескольких пробелов. Слова с дефисным написанием находить отдельно и вносить в специальный список форм, не разбиваемых при токенизации.
2. Проверить список местоимений и их форм (возможно попадание туда местоимений «ты-вы группы»)
3. Предусмотреть частеречную разметку загружаемых текстов, проверять формы на наличие омографов, при наличии выбирать частеречный тег по принципу наибольшей вероятности (вычислять при помощи скрытых марковских моделей (например, алгоритм Витерби))