Ответы на вопросы

1 В чем недостатки этого способа построения спеллчекера?

1. Время потраченное на получение файла словаря примерно 3-4 часа.
2. Символы пунктуации просто удаляются из текста при проверке, можно сделать более умную проверку.
3. Т.к. я просто получаю текст при помощи библиотеки gensim, в словарь попадают иногда не совсем корректные слова, например основы некоторых слов без окончания, но иногда такие слова могут реально использоваться в сообщении от пользователя, поэтому это одновременно и плюс и минус.

2 Быстро ли работает ваш спеллчекер? Можно ли его ускорить и если да, то как?

1. Сам спеллчекер работает быстро, т.к. я просто один раз подгружаю готовый файл словаря, а дальше просто проверяю наличие слова в списке.
2. Возможно можно захэшировать словарь и тогда будет еще быстрее, но разница не будет сильно заметной, т.к. итак работает достаточно быстро.
3. Самый долгий этап - получение словаря из dump файла, возможно можно делать это более оптимизированно, например написав самостоятельно умный парсер для xml.

3 Приведите интересные примеры правильных слов, которые этот спеллчекер считает ошибочными и наоборот, слов с опечаткой, которые считаются им правильными.

1. Слова с опечаткой, но которые считает правильными: приет, прие, нон, сонце, ветв.
2. Правильные слова, но которые считает ошибочными: все слова длины один - (а, с, я, в), больше слов не нашел.