2 возможные стратегии (можно использовать обе)

**1.**

1. С помощью обученной модели word2vec и списка seed немного расширить список оценочной лексики

2. Из расширенного списка составить небольшой датасет с учётом правого и левого контекста (найти его по корпусу текстов) по следующему принципу:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X (data) | | | | | | y (target) |
| слово из списка | слово справа | часть речи слова справа | слово слева | часть речи слова слева | оценка всего текста | оценочная ли это лексика (бинарный, 1 или 0) |

2.2 Часть речи можно получить с помощью pymorphy

3. В пункте 2 - тренировочный датасет; по такому же принципу составить датасет из всего корпуса тестов (распарсив его перед этим с помощью beautifulsoup или чего-то подобного и токенизировав)

4. Предобработка: векторизовать (с помощью bag of words, tf-idf, hashing и т.д.) слова в датасете, части речи привести к числовому виду с помощью label encoder'а - это тестовый датасет

5. Обучить какую-нибудь модель (можно подобрать, опираясь на прошлый опыт) на тренировочном датасете

6. Предсказать для каждого слова из тестового датасета, является ли оно оценочным

7. Просмотреть результаты, выбрать некоторое количество удачных, добавить их к тренировочному сету и снова обучить модель

8. Снова сделать предсказание для тестового сета и с помощью результатов расширить тренировочный

9. Повторять это, пока не соберется достаточного размера тренировочная выборка и последний раз расширить её с помощью получившегося предсказания

10. В итоге мы имеем список оценочный лексики + коллокации (т.к. собран правый и левый контекст)

11. То же самое можно проделать для определения тональности слов и коллокаций

**2.**

Мы планируем создать и применить semi-supervised алгоритм, основанный на синтаксических наблюдениях над оценочной лексикой, используемой в отзывах.

1. Предикативность/аттрибутивность прилагательных

Гипотеза:

(1) В текстах отзывов все или почти все прилагательные в предикативной позиции будут оценочными.

Эта гипотеза эмпирически подтвердилась на нескольких случайно взятых отзывах.

Как мы планируем ее использовать:

1. составляем список прилагательных adjlist, встречающихся в некотором подкорпусе.

2. Для каждого словоупотребления каждого прилагательного определяем, предикативной или аттрибутивной является эта позиция. Мы предполагаем, что это можно сделать с достаточно высокой точностью с помощью несложных правил.

3. Для каждого прилагательного из adjlist мы считаем процент предикативных и аттрибутивных употреблений. Согласно гипотезе (1), для оценочных прилагательных процент предикативных употреблений (далее pred\_ind) будет значительно выше, чем для неоценочных.

4. Расширяем seed случайными неоценочными прилагательными из списка adjlist.

5. Считаем средний pred\_ind для оценочных (eval\_av\_pred) и для неоценочных (noneval\_av\_pred) прилагательных из seed.

6. Для каждого прилагательного из adjlist, не вошедшего в seed, определяем, к чему ближе его pred\_ind: к eval\_av\_pred или к noneval\_av\_pred и соответствующим образом классифицируем. Для прилагательных, располагающихся от eval\_av\_pred и noneval\_av\_pred на примерно одинаковом расстоянии, можно применить другие методы, либо пометить их как «иногда оценочные».

Таким образом можно разделить все прилагательные на оценочные и неоценочные. Эти результаты мы планируем использовать для дальнейшего выделения оценочной лексики. Как мы планируем это сделать:

1. Корпус делится на некоторые синтагмы (предложения/клаузы).

2. Для каждой синтагмы, содержащей больше некоторого количества прилагательных (это будет гиперпараметр adj\_num), определяется процент оценочных прилагательных в ней. Этот процент мы будем называть synt\_adj\_eval.

3. Берем синтагмы с очень высоким и очень низким synt\_adj\_eval. Сравниваем частотную лексику, употребляемую в каждой из этих групп. Лексика, частотная для группы с высоким synt\_adj\_eval, но не частотная для группы с низким, будет, предположительно, оценочной.