Классификатор классифицирует тексты по двум классам:

1. «Разработка» - тексты, связанные с разработкой программного обеспечения;
2. «Другое» - тексты, не связанные с разработкой программного обеспечения.

Для классификации текстов используется метод опорных векторов или SVM.

Для векторизации текстов используется bag-of-words, нормализованная TF-IDF.

Процесс получения вектора признаков:

1. Все буквы текста приводятся к нижнему регистру;
2. Выделение отдельных слов в тексте;
3. Исключение «стоп слов» из текста;
4. Проведение лемматизации;
5. Подсчет частоты появления слова в документе;
6. Нормализация векторов с помощью TF-IDF.

Всего в коллекции документов находится порядка 1389 статей двух категорий. Из них 70% статей используется для обучения модели и 30% статей для тестирования работы обученной модели. Перед разбиением статей коллекция перетасовывается, то есть точное количество текстов обеих категорий неизвестно.