Анализ текста нескольких авторов

Выполнил: Клейман Вадим
Руководитель: Шпильман Алексей

Цель

В мире существует довольно много произведений написанные нескольким авторами (Братья Стругацкие, Ильф и Петров), под групповыми псевдонимами (Козьма Прутков, Николя Бурбаки).

Цель:

Получить спектральную картину авторства таких произведений.

Что уже есть?

Фреймворк для определения авторства, где в качестве модели для обработки текста используется Bag of Words

Задачи

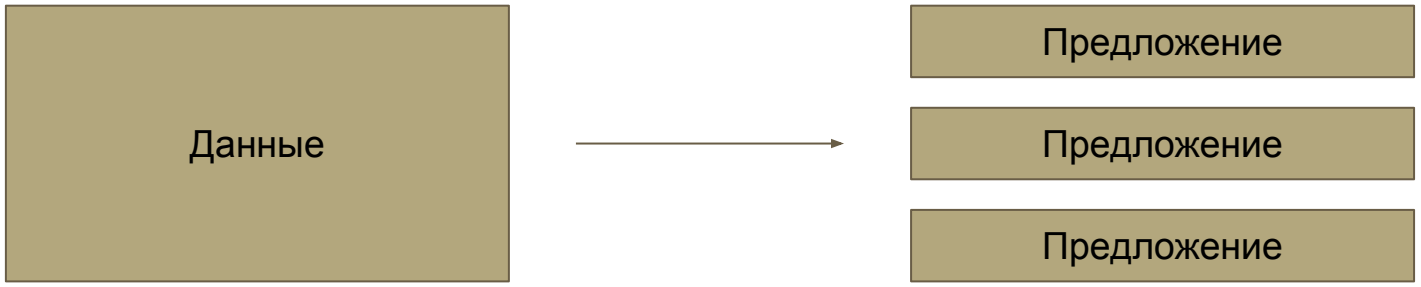
1. Обучить модель Word2vec

2. Используя Word2vec обучить несколько классификаторов:
   a. LinearSVC
   b. Random Forest
   c. XGBoost
   d. Passive-Aggressive
   e. Perceptron
   f. BernoulliNB
   g. Ridge Classifier

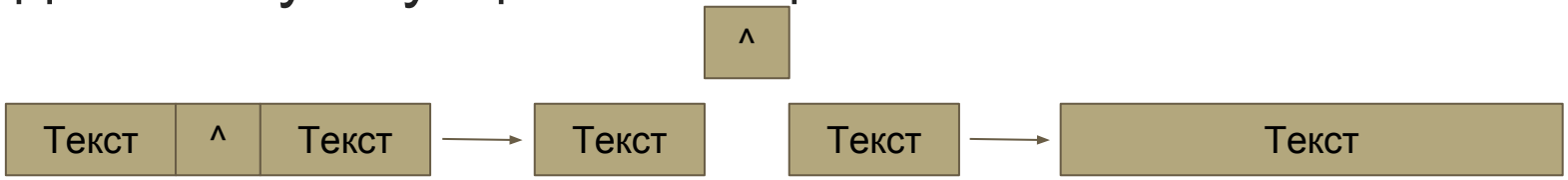
3. Объединить Word2vec и Bag of Words и оценить результаты.

Подготовка данных

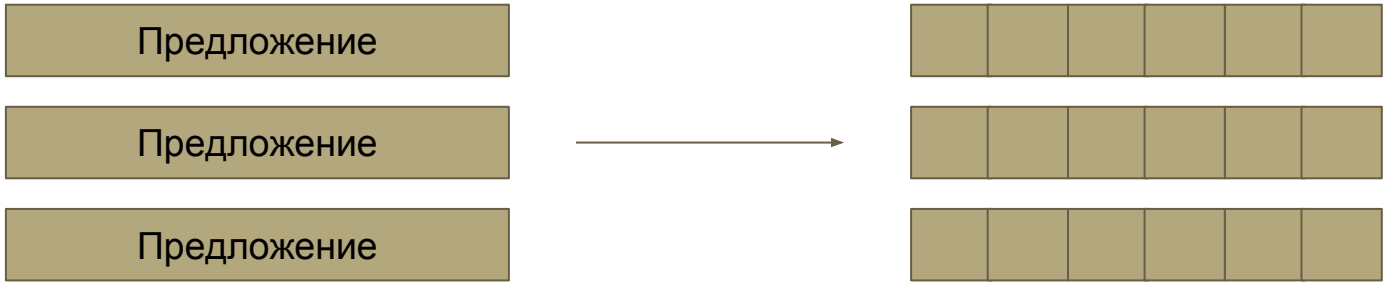
1. Разбить исходный текст на предложения.



2. Удалить пунктуацию и спец символы.

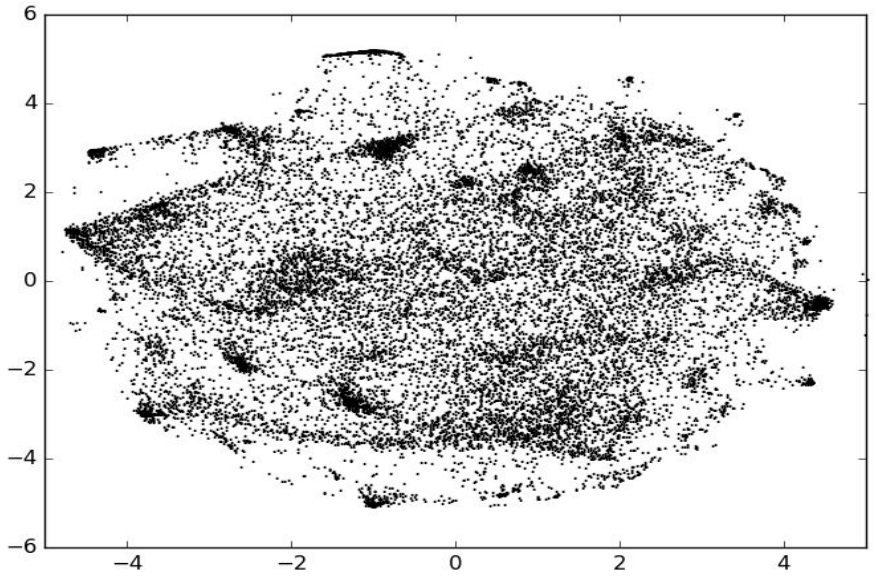


3. Предложения разбить на слова.



Результаты обучения модели Word2vec

Word2vec map



Десять слов, наиболее близких по значению к слову “sword”.

['crossbow', 0.6490396857261658], [‘axe’, 0.5989659428596497], [‘blade’, 0.5988671183586121],

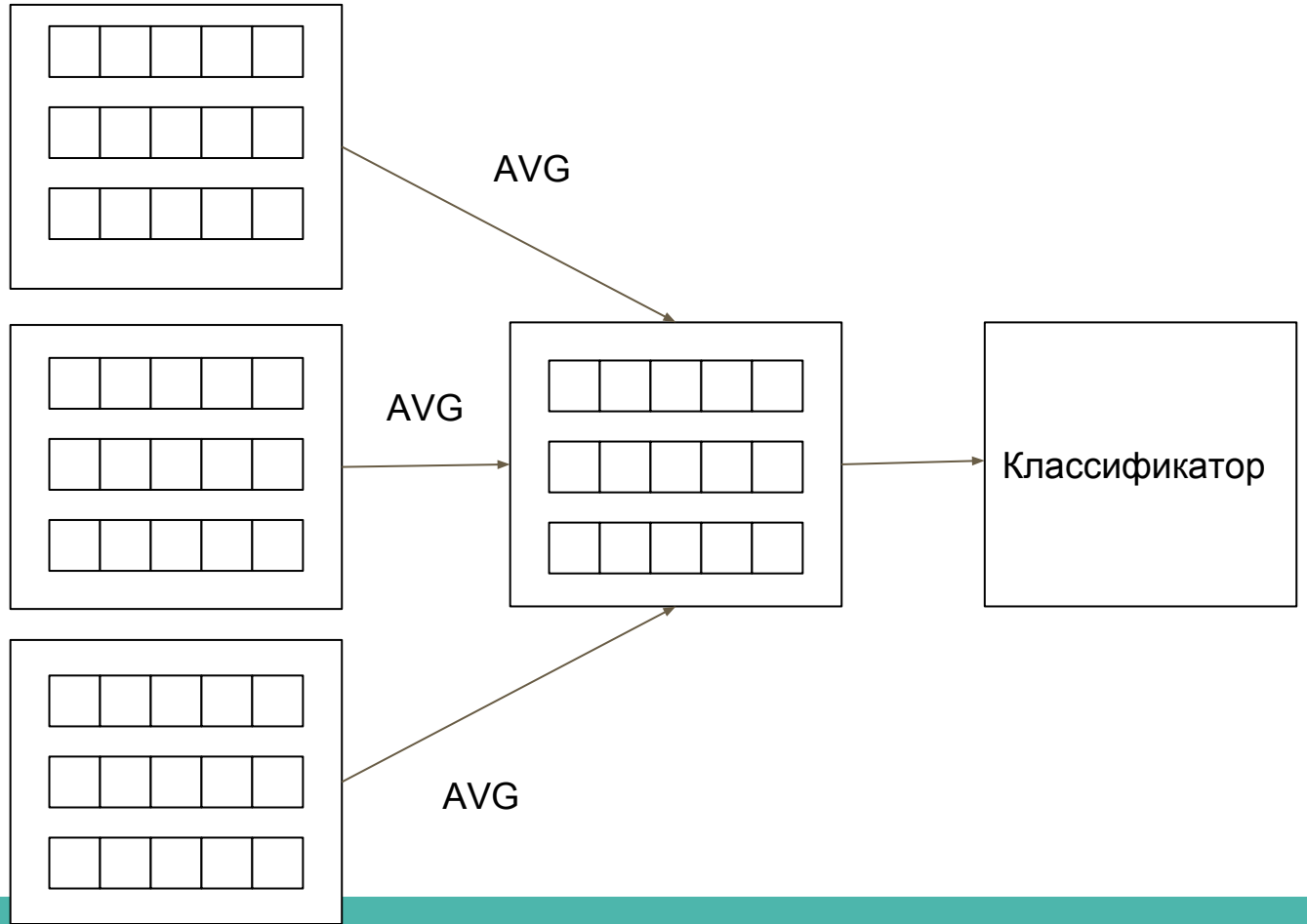
[‘knife’, 0.583248496055603], [‘nsword’, 0.568102240562439], [‘truncheon’, 0.562075853347778],

[‘dagger’, 0.559165298938751 2], [‘bow’, 0.5572289824485779], [‘scythe’, 0.5490979552268982],

[‘pike’, 0.53576153516769 41]

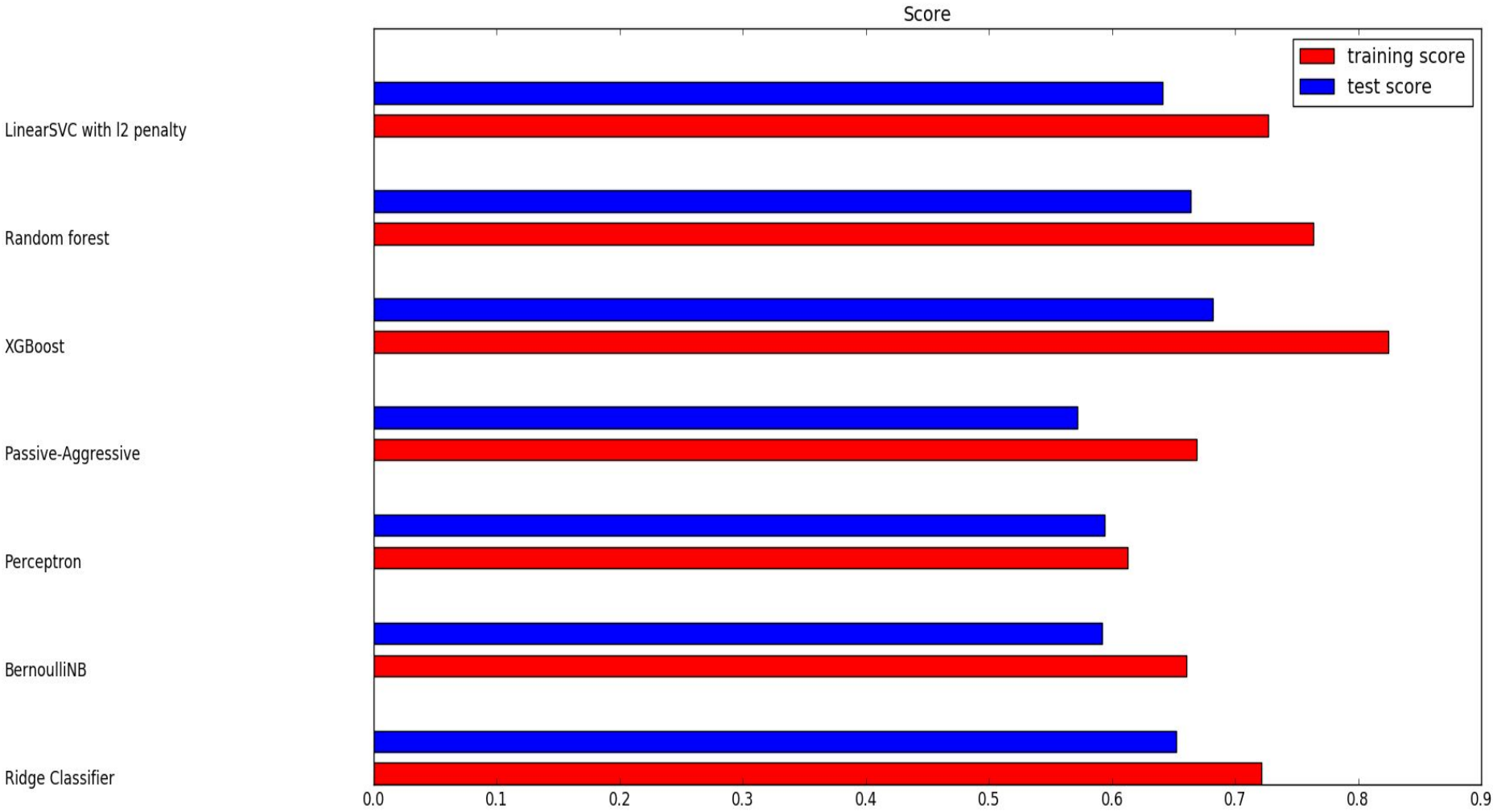
Подготовка данных

1. Разбиваем текст на блоки. В каждом блоке 100 слов.
2. В каждом блоке усредняем все вектора слов. В итоге каждый блок будет представляет собой вектор.
3. Полученные данные передаем в классификатор.



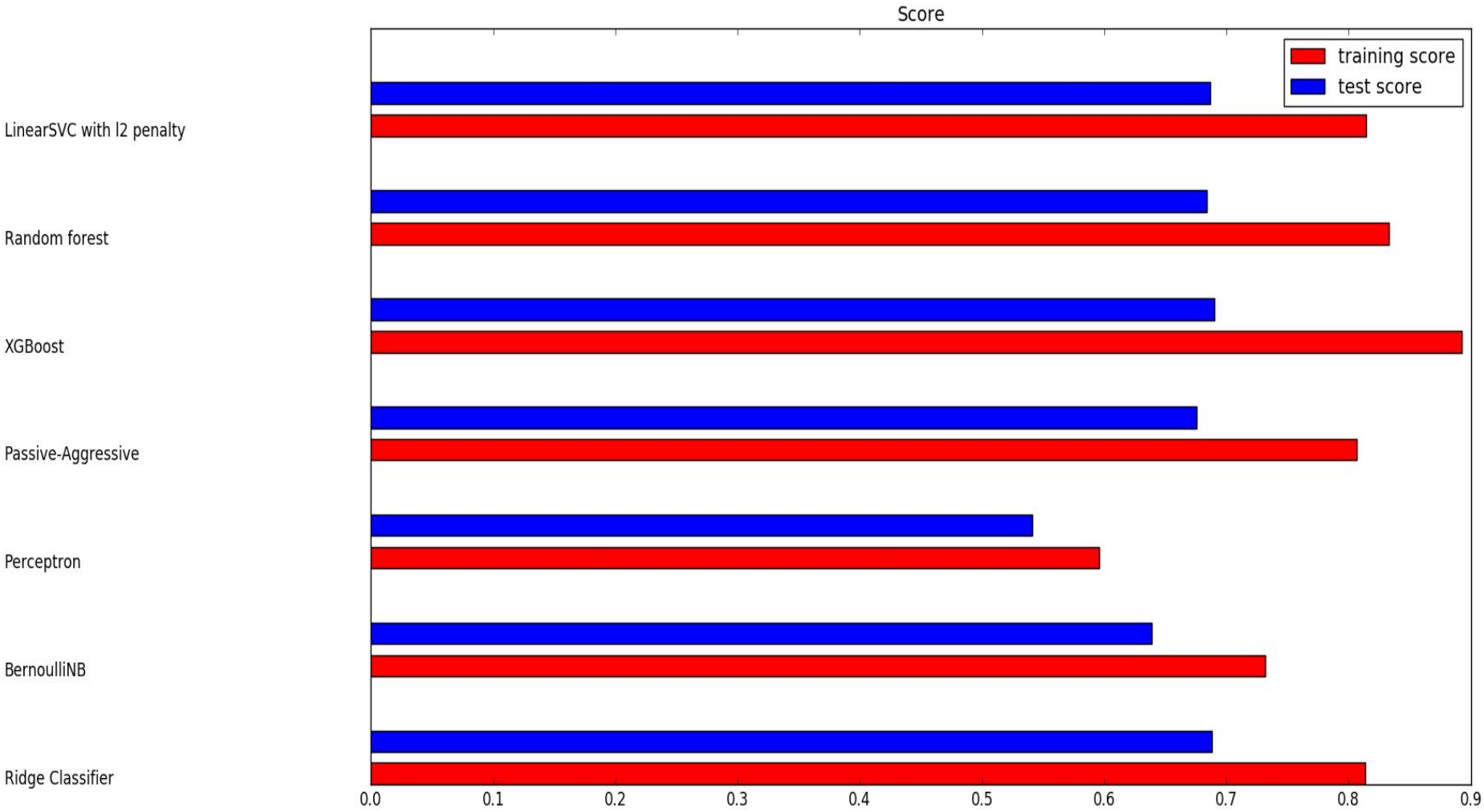
Результаты

Обучение классификаторов(количество слов в блоке - 100, блоки не пересекаются)



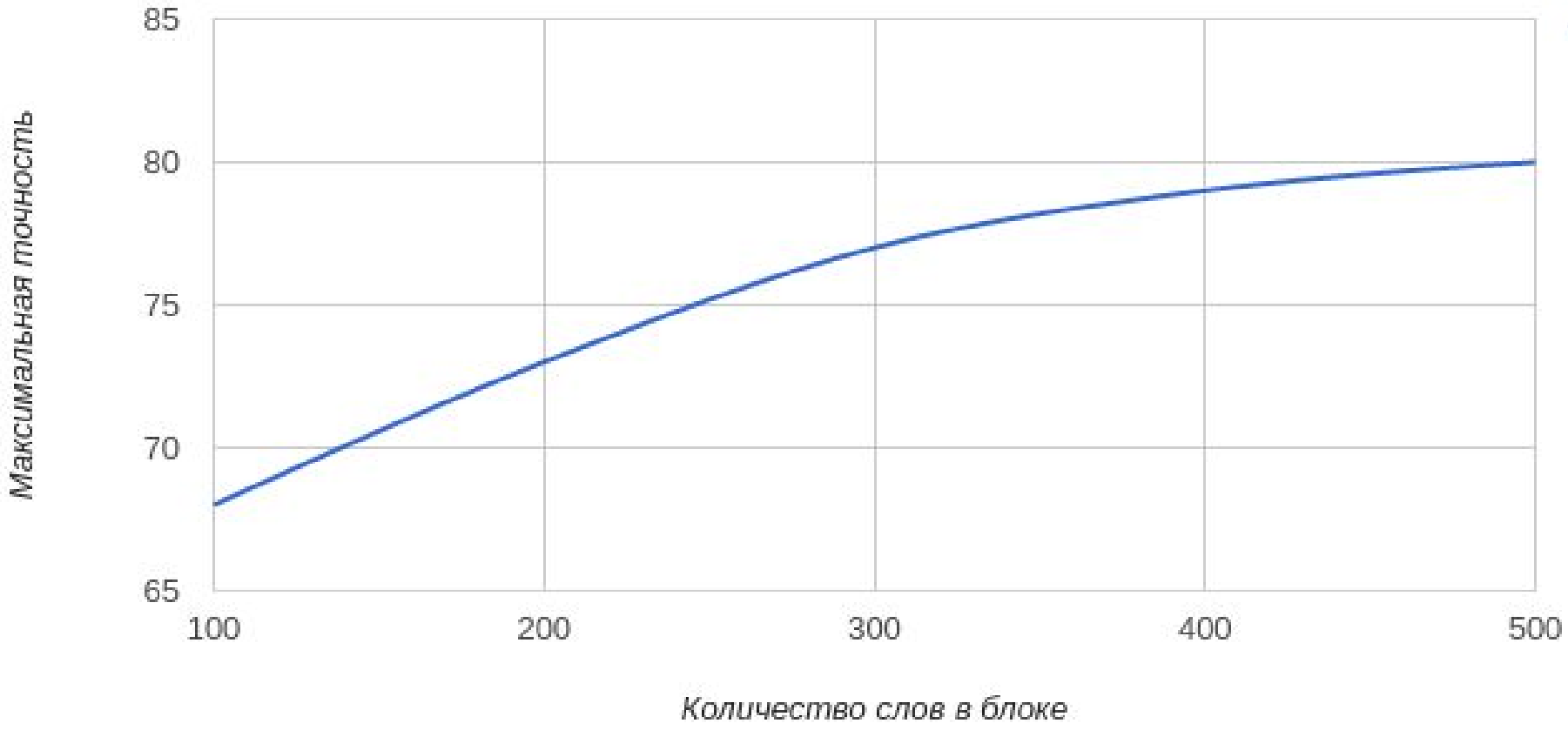
Результаты

Обучение классификаторов(количество слов в блоке - 100, блоки пересекаются)



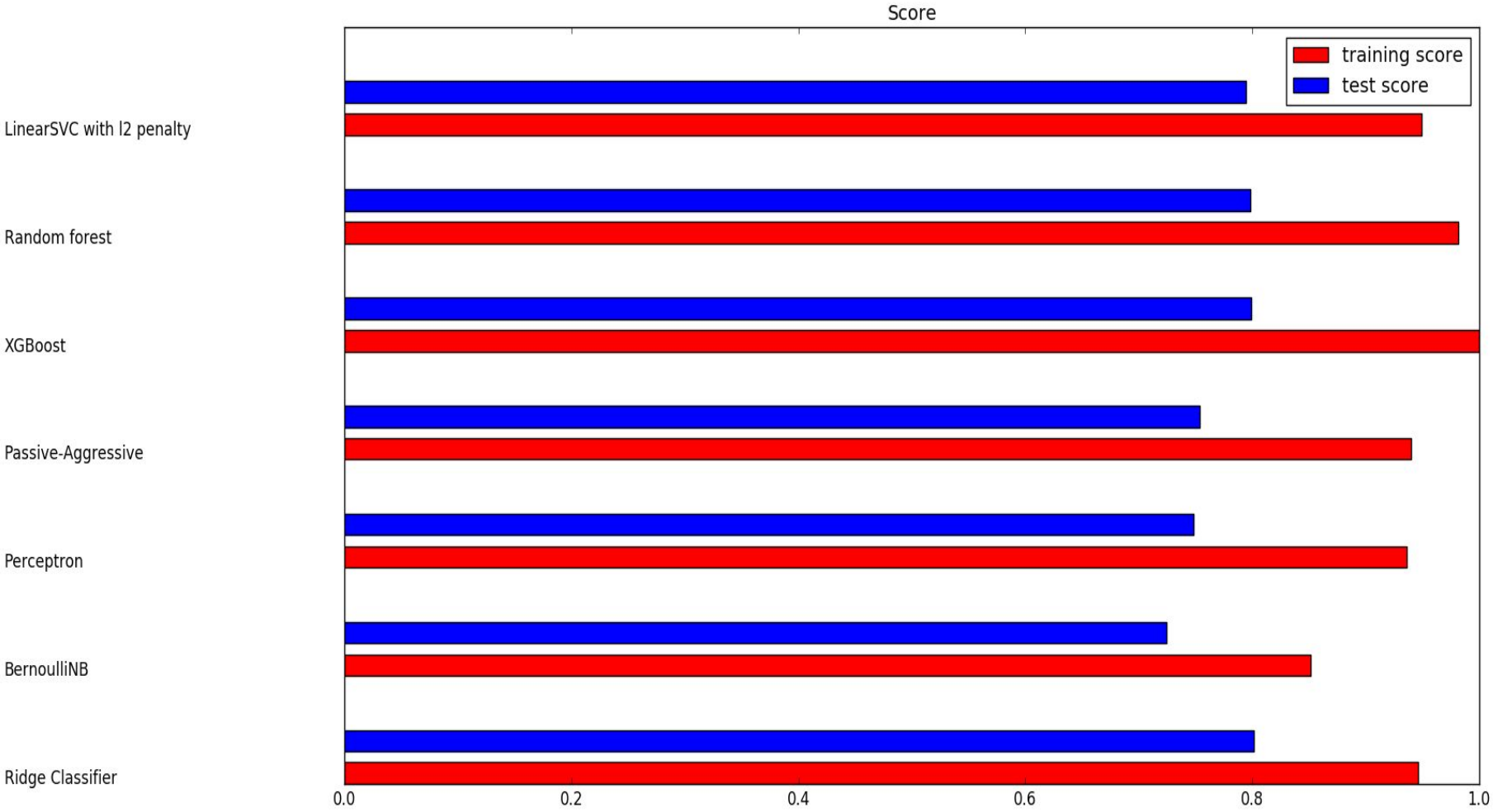
Низкая точность! Что делать?

Попробуем постепенно увеличивать количество слов в блоке и посмотреть как меняется точность классификаторов

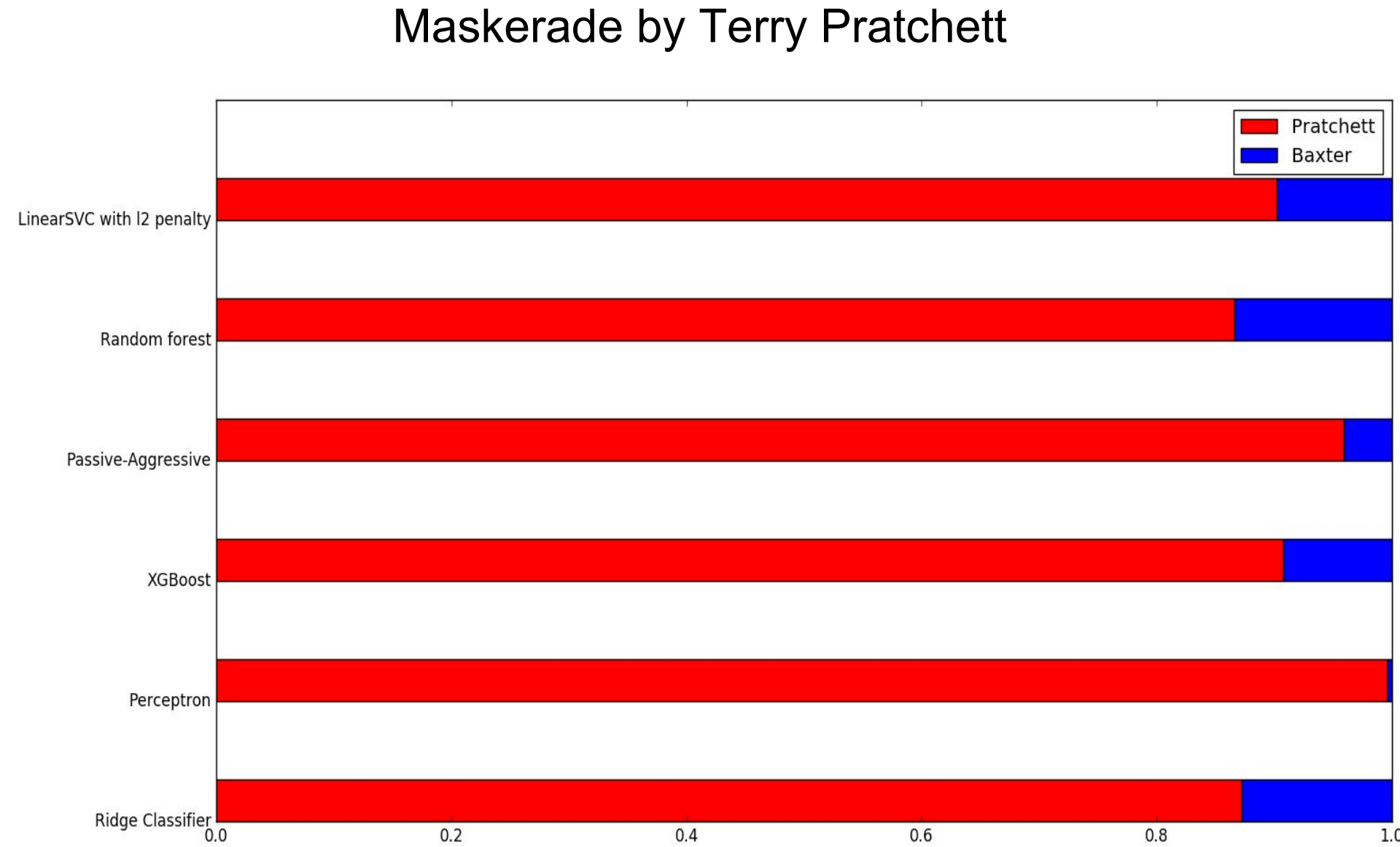


Результаты

Обучение классификаторов(количество слов в блоке - 500, блоки пересекаются)

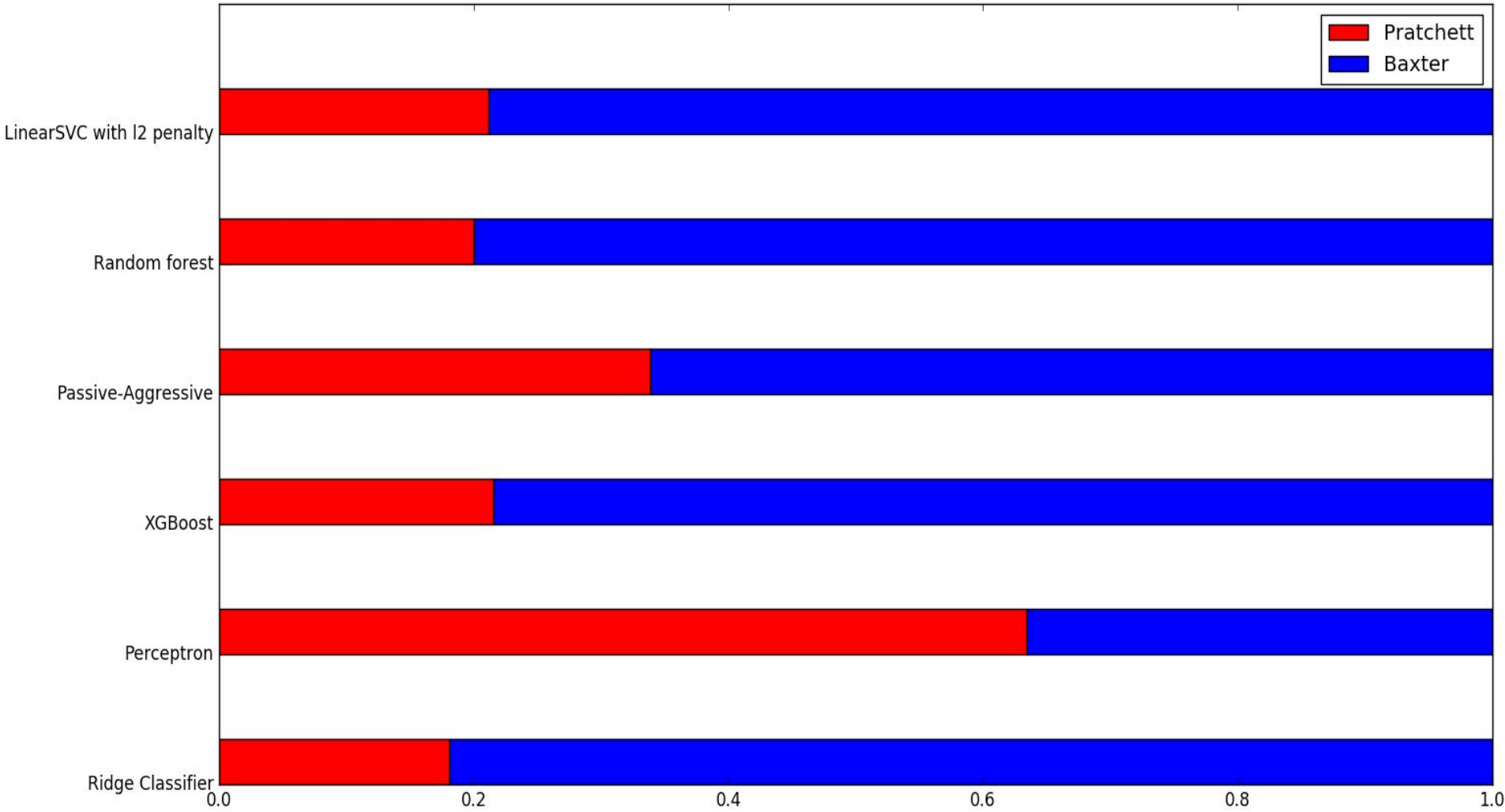


Результаты


Maskerade by Terry Pratchett
Pratchett
Baxter
LinearSVC with l2 penalty
Random forest
Passive-Aggressive
XGBoost
Perceptron
Ridge Classifier
0.0
0.2
0.4
0.6
0.8
1.0

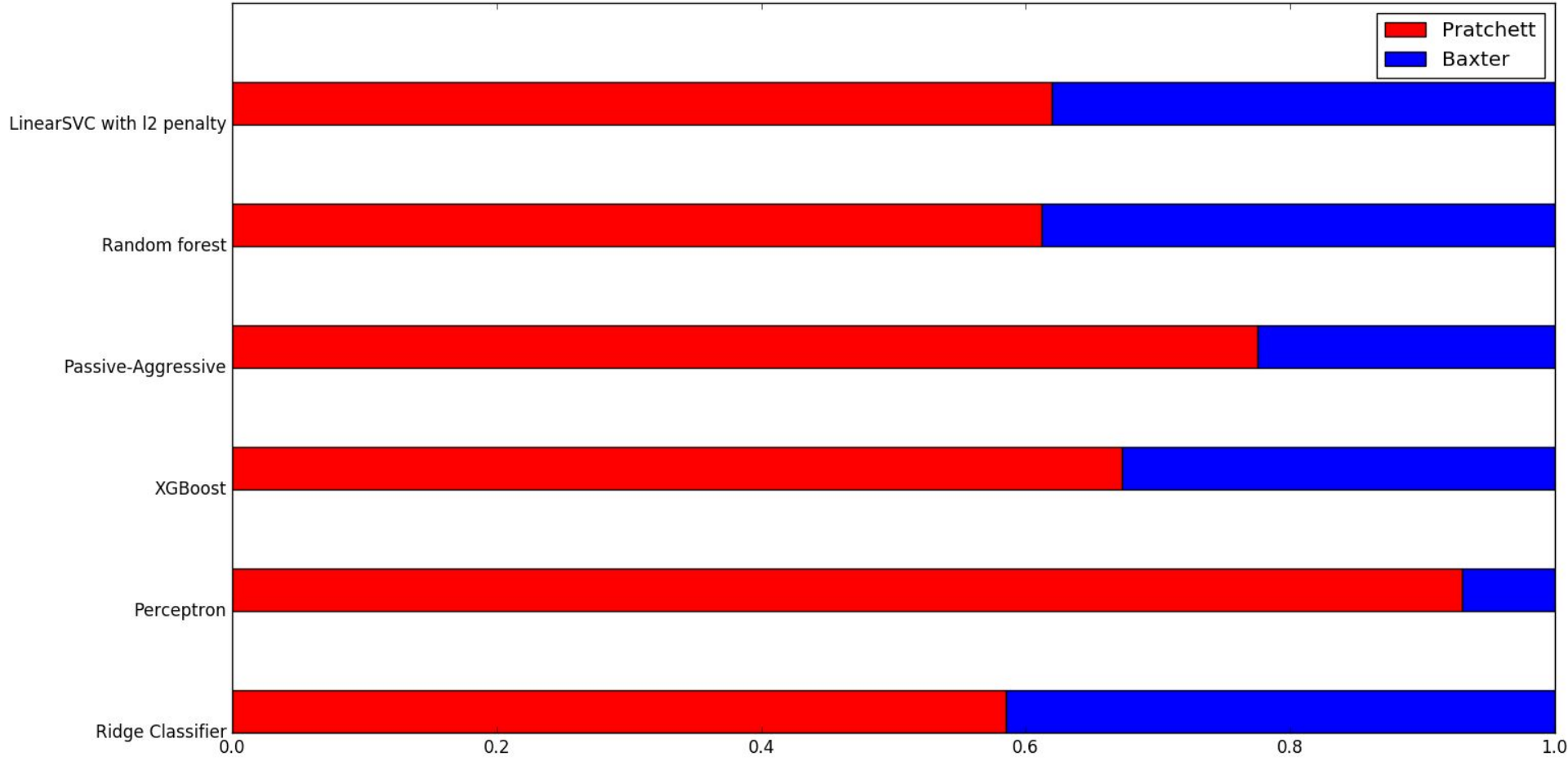
Результаты

Raft by Stephen Baxter



Результаты

The Long Earth by
Terry Pratchett and Stephen Baxter



Результаты

Ансамбль классификаторов: Linear SVC, Random Forest, XGBoost, RidgeClassifier

Maskerade by
Terry Pratchett

Raft by
Stephen Baxter

The Long Earth by
Terry Pratchett and Stephen Baxter

Word2vec + BagOfWords

Bag of Words - 88%

Word2vec - 71%

Word2vec + BagOfWords - 91%

Технологии

Язык программирования: Python

Библиотеки

Что хочется сделать?

1. Добавить алгоритмы кластеризации, например K-means и посмотреть на точность классификаторов

2. Посмотреть на спектральную картину для книг написанных более чем 2-я авторами.

3. Интегрировать более сложные алгоритмы классификации

Спасибо за внимание!