|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Планирование эксперимента**  Цель эксперимента - удостовериться, что предложенный далее алгоритм имеет предсказательную способность  Задачей эксперимента является проверка алгоритма Random Forest (параметром 2000 деревьев) запущенного на финансовых отчетах и новостей фондовой биржи. Алгоритм тестировался по 9-ти акциям и их изменениям цены с 2002 по 2009 года. Использовались данные по следующим котировкам:   |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Название фирмы | Google | Apple | Amazon | Adobe | HP | IBM | Intel | MicroSoft | NVidia | | Листинг на бирже | GOOG | AAPL | AMZN | ADBE | HPQ | IBM | INTC | MSFT | NVDA | |
|  | Использовались следующие методы для извлечения признаков из текста   * Unigram * NMF(50) * NMF(100) * NMF(200) * Ансамбль из этих признаков   Общее число признаков в разных моделях составило около 16000 |
|  |  |
|  |  |
|  | В работе использовались данные из статьи :*“Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the Importance of Text Analysis for Stock Price Prediction.”* <http://nlp.stanford.edu/pubs/stock-event.html>  **Полученные результаты**  На разных признаках была вычислена точность предсказания при помощи валидационных данных с 2009 по 2010 года.  Самым точным среди остальных оказался ансамблевый метод |
|  |  |
|  |  |
|  | Описание алгоритма Random Forest доступно по ссылке:  [https://scikit-learn.org/stable/modules/tree.html#tree](https://scikit-learn.org/stable/modules/tree.html%23tree) |
|  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ось X (модель): RF\_unigram | RF\_NMF\_50 | RF\_NMF\_100 | RF\_NMF\_200 | RF\_Ensemble |
| Ось Y (ошибка): accuracy\_score | F1 | AUCROC |  |  |
| Ось Z (выборки): 2002-2009 | 2002-2009 | 2010-2011 | 2012-2013 |  |