**Планирование эксперимента**

Цель эксперимента – убедиться в том, что предложенный алгоритм обладает предсказательной способностью.

Эксперимент заключается в проверке алгоритма Random Forest запущенного на финансовых отчетах и новостях фондовой биржи. Тестирование алгоритма проходило на основе 9-ти акций и изменения их цен в период с 2002 по 2009 года. Котировки данных которые использовались приведены в таблице:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Название фирмы | Google | Apple | Amazon | Adobe | HP | IBM | Intel | MicroSoft | NVidia |
| Листинг на бирже | GOOG | AAPL | AMZN | ADBE | HPQ | IBM | INTC | MSFT | NVDA |

Ниже представлены методы которые использовались для извлечения признаков из текста:

* Unigram
* NMF(50)
* NMF(100)
* NMF(200)
* Ансамбль из этих признаков

Общее число признаков в представленных моделях составило порядка 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**

При помощи валидационных данных с 2009 по 2010 года, была вычислена точность предсказания.

Самым точным методом из всех 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 |  |