Методы машинного обучения для классификации электронной почты в системе защиты от массовых несанкционированных рассылок

выполнил студент 320 группы
Конов Михаил Алексеевич

Научный руководитель:
Царев Дмитрий Владимирович

Введение

- **С**пам **–** массовая рассылка корреспонденции лицам, не выражавшим желания её получить
- Определение спама – субъективно, т.е важно построение персонифицированных систем фильтрации спама
- Техники фильтрации спама:
  - На основе отправителя(blacklist, whitelist)
  - На основе содержания(методы на основе наборов правил, методы машинного обучения)
- За 2019 год доля спама увеличилась на 4% и составила 56.51%
- Потери компаний из-за снижения производительности сотрудников, которые вынуждены отвлекаться на спам, за 2018г. оцениваются в 257 млрд. долларов
- Негативное влияние спама также включает фишинг, распространение вирусов, рекламу незаконных услуг и т.д.

Постановка задачи

- Задачей данной работы является исследование и разработка методов машинного обучения для построения персонифицированных моделей классификации в системе защиты от массовых несанкционированных рассылок электронной почты

Обзорная часть. Извлечение признаков письма

- **MIME** – стандарт, описывающий передачу различных типов данных по электронной почте
- Базовые типы MIME: application, audio, image, message, multipart, text, video и др.
- Анализируемые типы:
    1) **message**(заголовок сообщения)
    2) **text**(текстовая информация)
- Для извлечения векторов признаков из текста существует набор методов:
    1) Мешок слов**(BOW**) – каждому слову ставится в соответствие количество его упортеблений в тексте
    2) **TF/IDF** - каждому слову ставится в соответствие его частота в тексте, умноженная на обратную частоту документов набора с этим словом
    3) **Word2vec –** для получение векторных представлений используется нейронная сеть, учитывающая семантические значения слов

Обзорная часть. Алгоритмы выбора признаков

- Алгоритм **TFDCR** сопоставляет каждому признаку вес на основе формул (1) или (2) и выбирает признаки с наибольшим весом

$$Weight_{term} = |CountTerm_{spam} - CountTerm_{leg}| * \frac{CountDocs_{spam}}{N_{spam}} * \frac{N_{leg}}{CountDocs_{leg}} \quad (1)$$

$$Weight_{term} = |CountTerm_{spam} - CountTerm_{leg}| * \frac{CountDocs_{leg}}{N_{leg}} * \frac{N_{spam}}{CountDocs_{spam}} \quad (2)$$

- Алгоритм **WOA(whale optimization algorithm)** выполняет поиск оптимального решения при помощи заданного количества поисковых агентов (вектора из 0 и 1 с размерностью, равной числу признаков: 0 – признак не выбран, 1 – признак выбран). На каждой итерации находится агент, являющийся оптимальным решением. Затем, все поисковые агенты делают шаг к оптимальному решению либо по спирали, либо напрямую(рисунки 1 и 2)

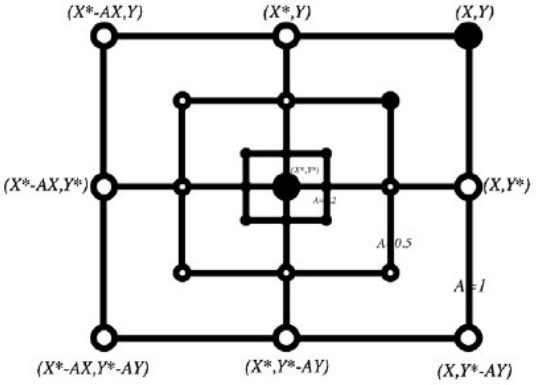


Рисунок 1 – движение поисковых агентов напрямую к оптимальному решению

- Алгоритм хи-квадрат**(CHI2)** считает зависимость классовой метки от каждого признака согласно значению критерия хи-квадрат и выбирает признаки с наибольшей зависимостью

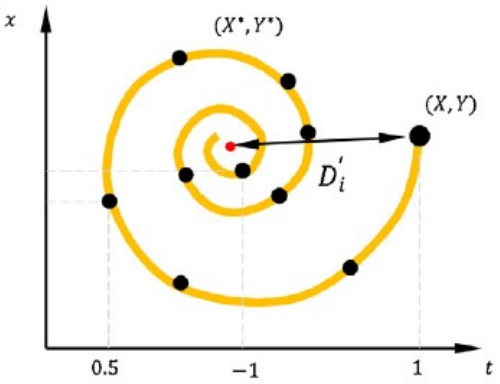


Рисунок 2 – движение поисковых агентов по спирали к оптимальному решению

Обзорная часть. Алгоритмы классификации

- **J48(C4.5)** – алгоритм классификации, строящий дерево решений на основе критерия прироста информации (разность энтропий множества примеров до и после разбиения).
- **CART** - алгоритм классификации, строящий бинарное дерево решений на основе индекса Джини.
- **Rotation Forest** – ансамблевый алгоритм, строящий деревья J48 на основе трансформированного методом главных компонент случайного разбиения изначального пространства признаков.
- **SVC** – алгоритм строит поверхность с наилучшим разделением классов с использованием метода внутренней точки.
- **SMO** - алгоритм строит поверхность с наилучшим разделением классов при помощи выбора и последовательной оптимизации двух параметров разделяющей поверхности.

Обзорная часть. Критерии оценки

- Критерии оценки наборов данных: год создания набора, число примеров в наборе, доля примеров спама
- На основе 7 публикаций за 2016-2020 годы были выбраны алгоритмы: WOA, TFDCR, SVM(SMO), J48, Rotation Forest, Naive Bayes
- Критерии оценки алгоритмов: Accuracy, Precision, Recall(ф-лы (1)-(3)), ROC AUC

Характеристики наборов данных

	Год	Число примеров	Процент примеров спама
Spambase	1999	4601	39%
Spam Assassin	2002	4198	33%
PU	2000, 2003	7101	57%
ENRON	2006	33090	52%
TREC 2007	2007	75419	67%
CEAS 2008	2008	140772	80%

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Обзорная часть. Алгоритмы

Характеристики алгоритмов классификации и выбора признаков(*)

	Accuracy	Recall	Precision
Naive Bayes	0.91(SD)[3] 0.85(SB)[3] 0.885(SB)[4]	0.83(SD)[3] 0.86(SB)[3] 0.885(SB)[4]	0.83(SD)[3] 0.88(SB)[3] 0.885(SB)[4]
Rotation Forest	0.942(SB)[2, 4] 0.969(EN)[2]	0.942(SB)[2, 4] 0.969(EN)[2]	0.942(SB)[2, 4] 0.969(EN)[2]
WOA + Rotation Forest	0.9989(SB)[2] 0.9943(EN)[2]	0.999(SB)[2] 0.9944(EN)[2]	0.999(SB)[2] 0.9944(EN)[2]
J48	0.923(SB)[4] 0.986(SA)[5]	0.923(SB)[4] 0.989(SA)[5]	0.923(SB)[4] 0.996(SA)[5]
TFDCR + SVM	0.939(EN)[6] 0.954(PU)[6]	-	-
TFDCR + Incr. SVM	0.975(EN)[6] 0.97(PU)[6]	-	-

(*) Сокращения: EN – ENRON; SD – SpamData; SB-SpamBase; SA – Spam Assassin;

Обзорная часть. Выводы

- Для получения векторных представлений будут использованы методы TF/IDF и BOW
- Для выбора признаков в практической части будут использованы алгоритмы: TFDCR и CHI2
- Для классификации были выбраны алгоритмы: SVC, CART Rotation Forest
- Для обучения и тестирования были выбраны наборы данных: ENRON, CEAS, TREC и SpamAssassin

Построение решения. Декомпозиция.

- Решение задачи будет состоять из нескольких стадий:

  1)Извлечение векторов признаков, представляющих письма (feature extraction)

  2)Выбор наиболее информативных признаков из данных векторов (feature selection)

  3)Обучение и тестирование алгоритмов классификации. Оценка результатов классификации. Выбор алгоритмов классификации и выбора признаков для использования в системе фильтрации спама

Параметры алгоритмов классификации

SVC	max_iter=1000, n_features=10000(4000 для CEAS и TREC)
CART	n_features=10000(4000 для CEAS и TREC)
Rotation Forest	max_depth=50, n_estimators=5, n_features=10000 (для TREC max_depth=10, n_estimators=2, n_features=500, на CEAS не тестировался)

Построение решения. Spam Assassin; ENRON

Метрика ROC AUC на наборах ENRON и SA(BOW)

	SA	EN1	EN2	EN3	EN4	EN5	EN6
TFDCR+SVC	0.7797	0.9651	0.9519	0.7196	0.9259	0.9338	0.8964
CHI2+SVC	0.7354	**0.9693**	**0.9574**	0.7080	0.9272	0.9346	0.8919
TFDCR+CART	0.9166	0.93	0.9293	**0.9523**	0.9502	0.9338	0.9302
CHI2+CART	0.9068	0.9331	0.9267	0.9410	0.9489	**0.9559**	**0.9387**
TFDCR+RotF	0.9132	0.9291	0.9489	0.9426	**0.9553**	0.9363	0.9334
CHI2+RotF	**0.9217**	0.9366	0.9462	0.94	0.9457	0.9550	0.9297

Метрика ROC AUC на наборах ENRON и SA(TF/IDF)

	SA	EN1	EN2	EN3	EN4	EN5	EN6
TFDCR+SVC	0.9627	0.9819	0.9776	0.9926	0.9592	0.9829	0.9610
CHI2+SVC	**0.9664**	**0.9883**	**0.9911**	**0.9931**	**0.9659**	**0.9872**	**0.9765**
TFDCR+CART	0.9187	0.9382	0.9393	0.9314	0.96	0.9525	0.9314
CHI2+CART	0.9115	0.9475	0.9439	0.9337	0.9577	0.9571	0.9386

Метрика Accuracy на наборах ENRON и SA(BOW)

	SA	EN1	EN2	EN3	EN4	EN5	EN6
TFDCR+SVC	0.8615	**0.9666**	0.9726	0.8543	0.9611	0.9631	0.9475
CHI2+SVC	0.8211	0.9649	**0.9741**	0.8340	0.9636	0.9614	0.9449
TFDCR+CART	0.9149	0.9461	0.9483	**0.9593**	0.9611	0.9479	0.9454
CHI2+CART	0.9098	0.9443	0.9503	0.9494	0.9621	**0.9654**	0.9561
TFDCR+RotF	0.9242	0.9402	0.9622	0.9533	**0.9667**	0.9491	0.95
CHI2+RotF	**0.9307**	0.9414	0.9581	0.9544	0.9626	0.9649	**0.9571**

Метрика Accuracy на наборах ENRON и SA(TF/IDF)

	SA	EN1	EN2	EN3	EN4	EN5	EN6
TFDCR+SVC	0.9719	0.9818	0.9871	**0.9951**	0.9793	0.9895	0.9803
CHI2+SVC	**0.9755**	**0.9877**	**0.9948**	0.9934	**0.9828**	**0.9930**	**0.9864**
TFDCR+CART	0.9228	0.9467	0.9534	0.9494	0.9692	0.9602	0.9485
CHI2+CART	0.9120	0.9549	0.9586	0.9505	0.9727	0.9649	0.9581

Построение решения. TREC

Метрика ROC AUC на наборе TREC(BOW)

	TR1	TR2	TR3	TR4	TR5	TR6	TR7	TR8
TFDCR+SVC	0.9482	0.9410	0.8813	0.9422	0.9372	0.9429	0.9401	0.9358
CHI2+SVC	0.9389	0.9419	0.8830	0.9469	0.9299	0.9521	0.9318	0.9418
TFDCR+CART	0.9688	**0.9662**	**0.9704**	**0.9663**	**0.9701**	0.9574	0.9665	0.9622
CHI2+CART	**0.9690**	0.9621	0.9688	0.9631	0.9668	**0.9728**	**0.9682**	0.9621
TFDCR+RotF	0.9603	0.96	0.9602	0.9588	0.9666	0.9691	0.9644	**0.9695**
CHI2+RotF	0.9634	0.9489	0.9592	0.9560	0.9681	0.9590	0.9648	0.9620

Метрика ROC AUC на наборе TREC(TF/IDF)

	TR1	TR2	TR3	TR4	TR5	TR6	TR7	TR8
TFDCR+SVC	0.9781	0.9764	0.9786	0.9785	**0.9806**	0.9813	0.9786	0.9809
CHI2+SVC	**0.9817**	**0.9822**	**0.9817**	**0.9854**	0.9805	**0.9855**	**0.9871**	**0.9843**
TFDCR+CART	0.9652	0.9651	0.9668	0.9603	0.9680	0.9662	0.9724	0.9640
CHI2+CART	0.9738	0.9689	0.9654	0.9687	0.9660	0.9751	0.9676	0.9637

Метрика Accuracy на наборе TREC(BOW)

	TR1	TR2	TR3	TR4	TR5	TR6	TR7	TR8
TFDCR+SVC	0.9614	0.9569	0.9196	0.9579	0.9563	0.9566	0.9566	0.9550
CHI2+SVC	0.9550	0.9585	0.9193	0.9598	0.9495	0.9662	0.9534	0.9585
TFDCR+CART	0.9733	**0.9720**	**0.9736**	**0.9698**	0.9724	0.9614	0.9714	0.9656
CHI2+CART	**0.9736**	0.9685	0.9727	0.9666	0.9711	**0.9781**	**0.9724**	0.9675
TFDCR+RotF	0.9685	0.9692	0.9695	0.9682	0.9717	0.9742	0.9708	**0.9749**
CHI2+RotF	0.9704	0.9611	0.9691	0.9646	**0.9736**	0.9672	0.9714	0.9692

Метрика Accuracy на наборе TREC(TF/IDF)

	TR1	TR2	TR3	TR4	TR5	TR6	TR7	TR8
TFDCR+SVC	0.9842	0.9833	0.9852	0.9849	**0.9855**	0.9868	0.9846	0.9868
CHI2+SVC	**0.9871**	**0.9868**	**0.9868**	**0.9891**	**0.9855**	**0.9891**	**0.9904**	**0.9884**
TFDCR+CART	0.9672	0.9717	0.9707	0.9653	0.9724	0.9701	0.9762	0.9688
CHI2+CART	0.9772	0.9727	0.9727	0.9724	0.9711	0.9785	0.9749	0.9694

Построение решения. CEAS

Метрика ROC AUC на наборе CEAS(BOW)

	CE1	CE2	CE3	CE4	CE5	CE6	CE7	CE8
TFDCR+SVC	0.9213	0.9358	0.9187	0.9442	0.8905	0.9437	0.9477	0.9248
CHI2+SVC	0.9356	0.9483	0.8979	0.89	0.8979	0.9446	0.9497	0.9499
TFDCR+CART	0.9838	**0.9821**	0.9788	0.9823	0.9781	**0.9807**	**0.9819**	**0.9842**
CHI2+CART	**0.9861**	0.9795	**0.9836**	**0.9835**	**0.9836**	0.9762	0.9807	0.9787

Метрика ROC AUC на наборе CEAS(TF/IDF)

	CE1	CE2	CE3	CE4	CE5	CE6	CE7	CE8
TFDCR+SVC	0.9811	0.9776	**0.9856**	**0.9810**	0.9766	0.9822	0.9768	0.9815
CHI2+SVC	**0.9815**	**0.9840**	0.9846	0.9790	**0.9880**	**0.9827**	**0.9798**	**0.9838**
TFDCR+CART	0.9709	0.9630	0.9733	0.9697	0.9681	0.9719	0.9606	0.9682
CHI2+CART	0.9722	0.9578	0.9721	0.9645	0.9625	0.9670	0.9708	0.9667

Метрика Accuracy на наборе CEAS(BOW)

	CE1	CE2	CE3	CE4	CE5	CE6	CE7	CE8
TFDCR+SVC	0.9213	0.9358	0.9187	0.9442	0.8905	0.9437	0.9477	0.9248
CHI2+SVC	0.9356	0.9483	0.8979	0.89	0.8979	0.9446	0.9497	0.9499
TFDCR+CART	0.9838	**0.9821**	0.9788	0.9823	0.9781	**0.9807**	**0.9819**	**0.9842**
CHI2+CART	**0.9861**	0.9795	**0.9836**	**0.9835**	**0.9836**	0.9762	0.9807	0.9787

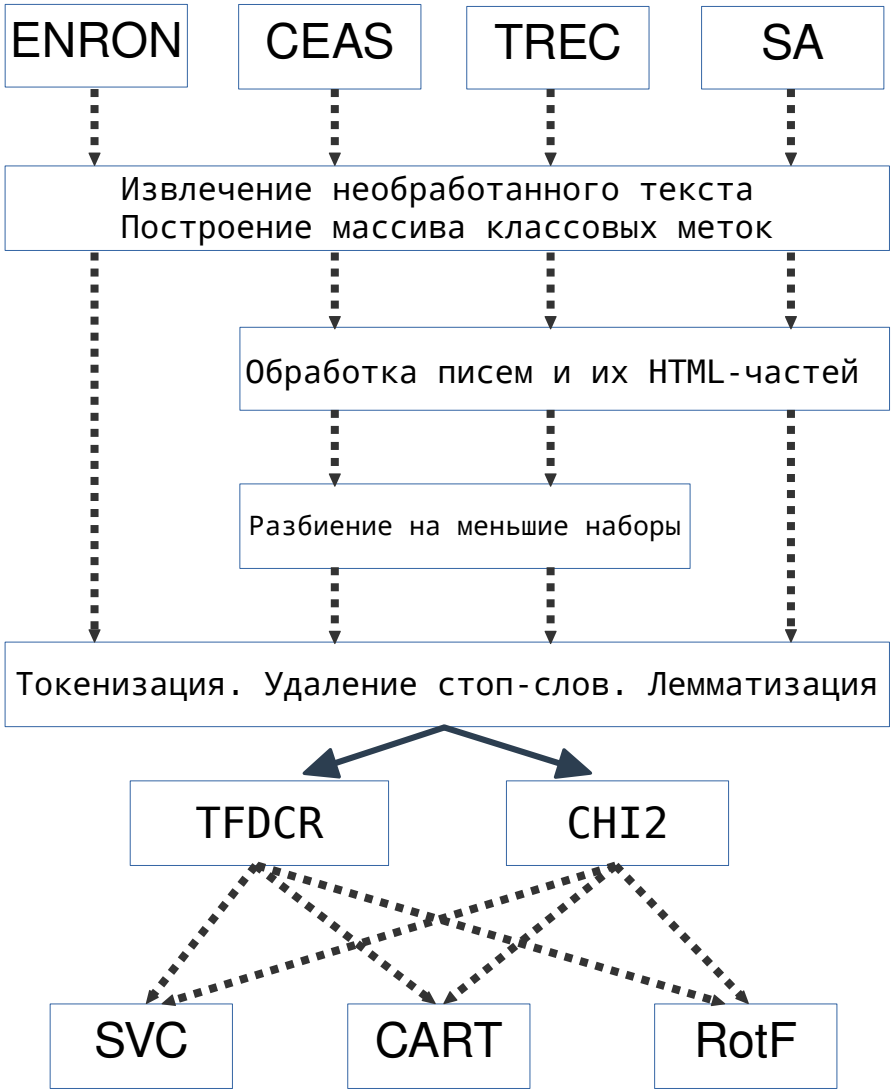
Метрика Accuracy на наборе CEAS(TF/IDF)

	CE1	CE2	CE3	CE4	CE5	CE6	CE7	CE8
TFDCR+SVC	0.9912	0.9905	0.9938	**0.9923**	0.9895	**0.9921**	0.9905	0.9907
CHI2+SVC	**0.9914**	**0.9926**	**0.9941**	0.9909	**0.9943**	0.9916	**0.9907**	**0.9919**
TFDCR+CART	0.9843	0.9816	0.9861	0.9845	0.9836	0.9842	0.9793	0.9838
CHI2+CART	0.9857	0.9797	0.9850	0.9838	0.9804	0.9823	0.9843	0.9833

Построение решения. Вывод

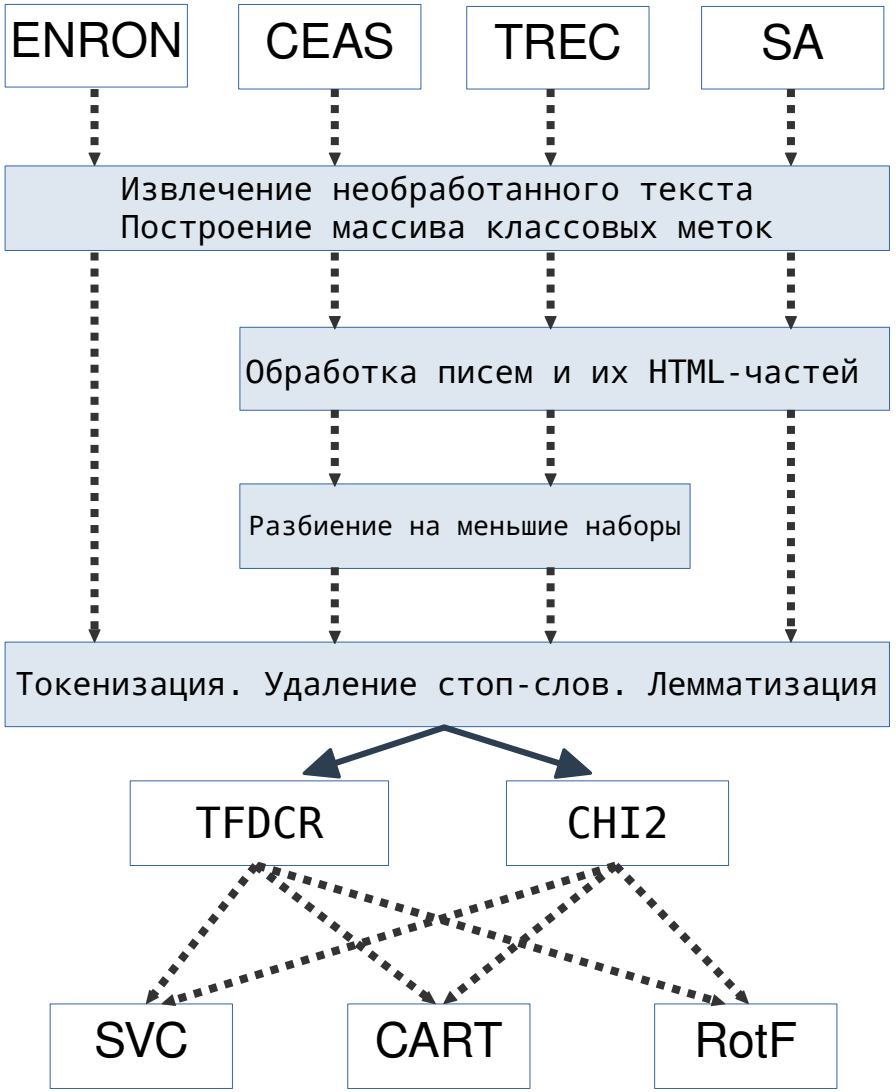
- На основе проведенных экспериментов было выявлено что лучше всего для использования в системе фильтрации спама подходит алгоритм CHI2+SVC с методом выбора признаков TF/IDF

Описание практической части. Итоговый алгоритм



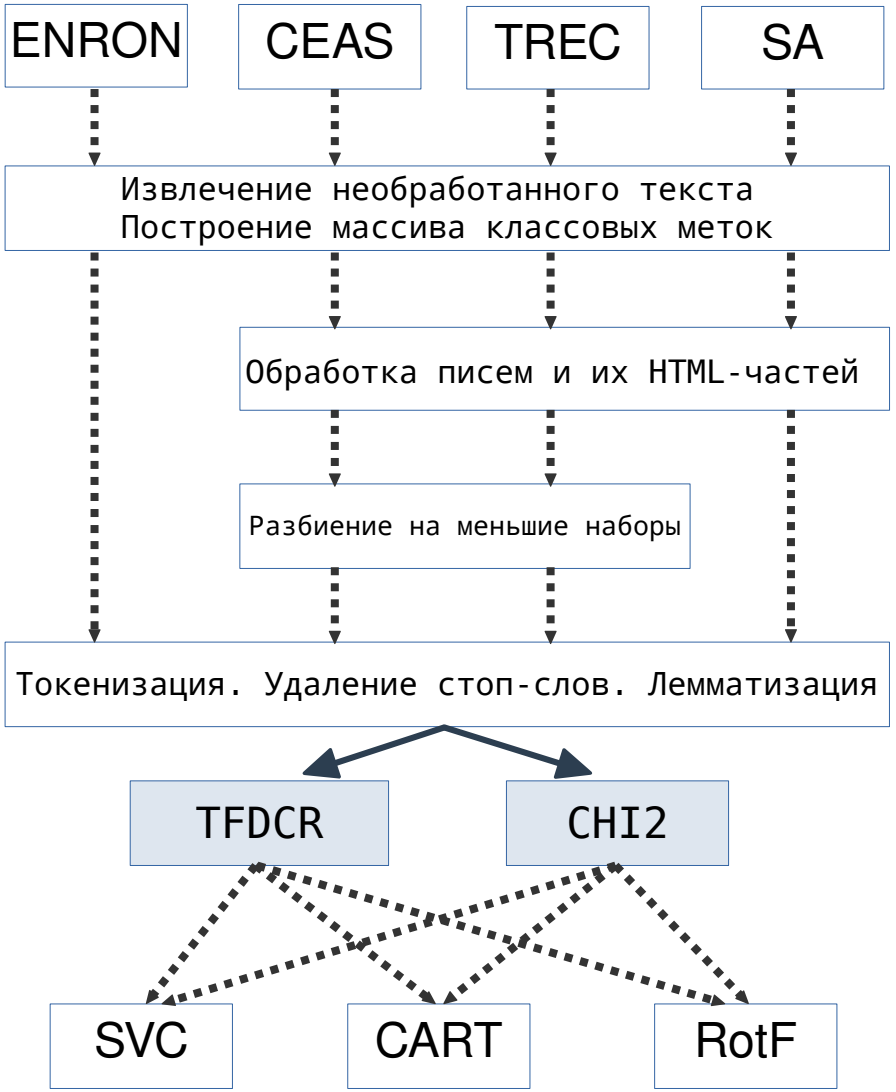
- Для написания практической части использовался язык **Python 3** и среда **Google Colaboratory**. Для запуска кода были использованы 2 процессора **Intel Xeon CPU @ 2.30GHz** и графическая карта **Nvidia K80 / T4 GPU**. Максимальный объем оперативной памяти: 12 ГБ

Описание практической части. Предобработка



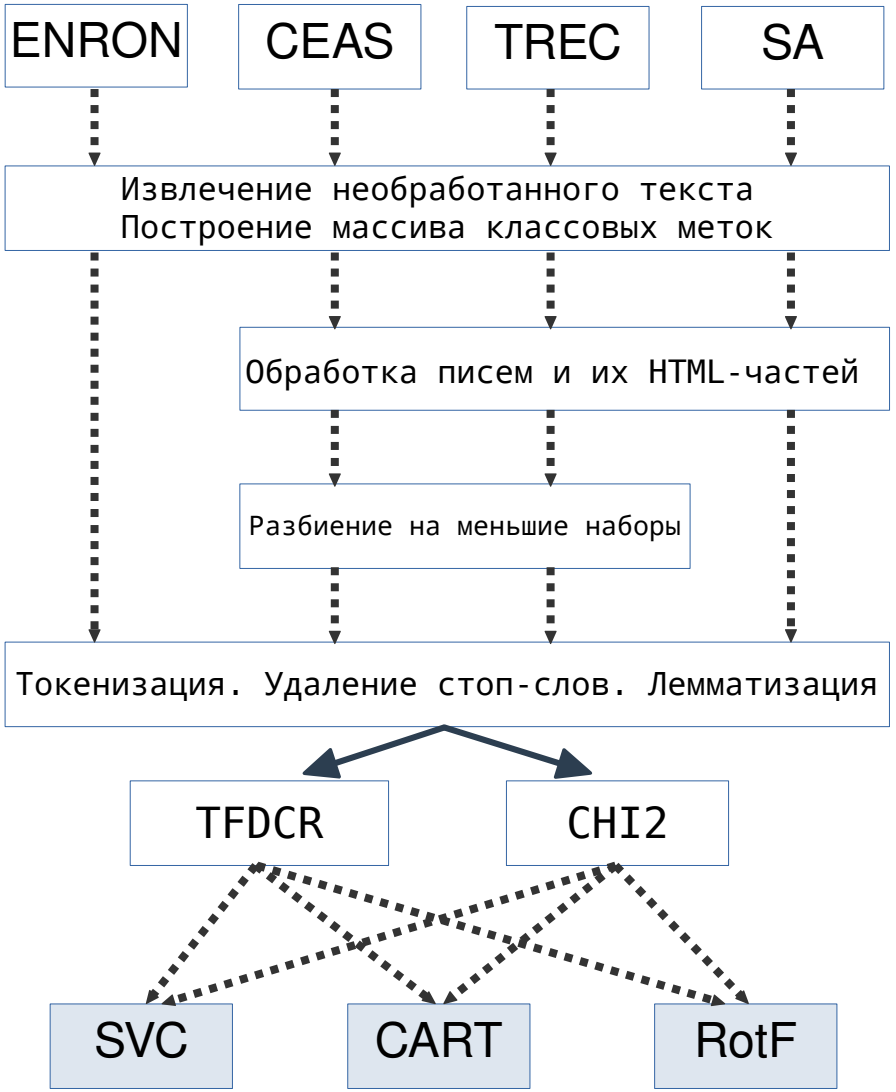
- Извлечение: **tarfile, zipfile**
- Обработка писем: **email**
- Обработка HTML: **beautifulsoup**
- Токенизация, удаление стоп-слов: **nltk**
- Лемматизация: **nltk.wordnet**
- Размер модуля: 206 строк

Описание практической части. Выбор признаков



- Для получения векторных представлений использовались алгоритмы TF/IDF и “мешка слов” из **scikit-learn**
- Хи-квадрат: **scikit-learn**
- TFDCR: собственная реализация
- Размер модуля: 120 строк

Описание практической части. Классификация



- Использовалось разбиение на обучающую и тестовую выборки(размеры 0.67, 0.33)
- SVC, CART: **scikit-learn**
- Rotation forest: **rotation-forest**
- Размер модуля: 123 строки

Результаты

- На основе проведенного обзора современных алгоритмов для классификации спама и выбора признаков для экспериментального исследования были выбраны алгоритмы: TFDCR, CHI2, SVM(SVC), CART, Rotation Forest
- Построен программный стенд, позволяющий оценить работу выбранных алгоритмов на наборах данных TREC, CEAS, ENRON и Spam Assassin
- Планы на будущее:

  1)Применение нейросетевых классификаторов(ANN, RWN)

  2)Рассмотрение других алгоритмов получения векторных представлений(Word2vec)

  3)Рассмотрение оптимизационных алгоритмов выбора признаков(WOA, Antlion optimization)

  4)Интеграция алгоритмических разработок в систему фильтрации спама

Спасибо за внимание!

Список литературы

[1] - Bhuiyan H., Ashiquzzaman A., Juthi T., Biswas S., Ara J. A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques // Global Journal of Computer Science and Technology: Software & Data Engineering. 2018. [2.] N 18. P. 21-29 [PDF] (https://computerresearch.org)

[2] - Shuaib M., Abdulhamid S.M., Adebayo O.S. et al. Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification. // SN Appl. Sci. 2019. [1.] N 390 [PDF] (https://doi.org/10.1007/s42452-019-0394-7)

[3] - Rusland N.F., Norfaradilla W., Shahreen K., Hanayanti H. Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets // IOP Conf. Ser.: Mater. Sci. Eng. 2017. N 226. [PDF] (https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012091)

[4] - Shuaib M., Osho O., Alhassan J., Abdulhamid S., Ismaila I. Comparative Analysis of Classification Algorithms for Email Spam Detection. // International Journal of Computer Network and Information Security. 2018. N 1. P. 60-67. [PDF] (http://www.mecs-press.org/ijcnis)

[5] - Al-Shboul B., Hakh H., Faris H., Aljarah I., Alsawalqah H. Voting-based Classification for E-mail Spam Detection. // Journal Of ICT Research And Applications. 2016. [10.] N 1. P. 29-42. [PDF] (http://journals.itb.ac.id)

[6] - Sanghani G., Kotecha K. Incremental personalized E-mail spam filter using novel TFDCR feature selection with dynamic feature update // Expert Systems With Applications. 2019. N115. P. 287–299.

[7] - Gbenga Dada E., Bassi S.J., Chiroma H., Abdulhamid S.M., Adetunmbi A.O., Ajibuwa O.E Machine learning for email spam filtering: review, approaches and open research problems // Heliyon 2019. [6.] N 5. [PDF] (http://www.heliyon.com)