# Контрастивный анализ текстов RusDraCor

Евтодиева Анна

# Ссылка на репозиторий с проектом

https://github.com/Gratisfo/Parentents-and-children/blob/main/Contrastive_analysis_RusDraCor.ipynb

# Постановка задачей

Провести контрастивный анализ текста

Подобрать наилучший классификатор реплик родителей и детей

# Постановка задачи

Подзадачи

- контрастивный анализ включает распределения
  - количества токенов, предложений
  - длин предложений
  - глубины синтаксического дерева
  - частей речи
- сравнение нескольких видов классификаторов

# Описание данных

Тексты из корпуса RusDraCor закодированные в TEI-5

В теге `<listRelation>` указаны все связи между персонажами

Были выбраны пьесы с параметром “`parent_of`”

Атрибут passiv указывает на имя ребенка, а active на имя родителя

```
<listRelation type="personal">
<relation passive="#viktor" active="#tumansky" name="parent_of"/>
<relation passive="#okaemov" active="#masha" name="related_with"/>
</listRelation>
```

# Методы решения. Данные и анализ

**Сбор данных:** API RusDraCor, request, класс и функции

**Контрастивный анализ:**

для визуализации: `matplotlib, seaborn`

препроцессинг: `razdel, nltk, pymorphy2, re`

глубина синтаксических деревьев: `ipymarkup, navec, slovnet`

# Методы решения. Классификатор

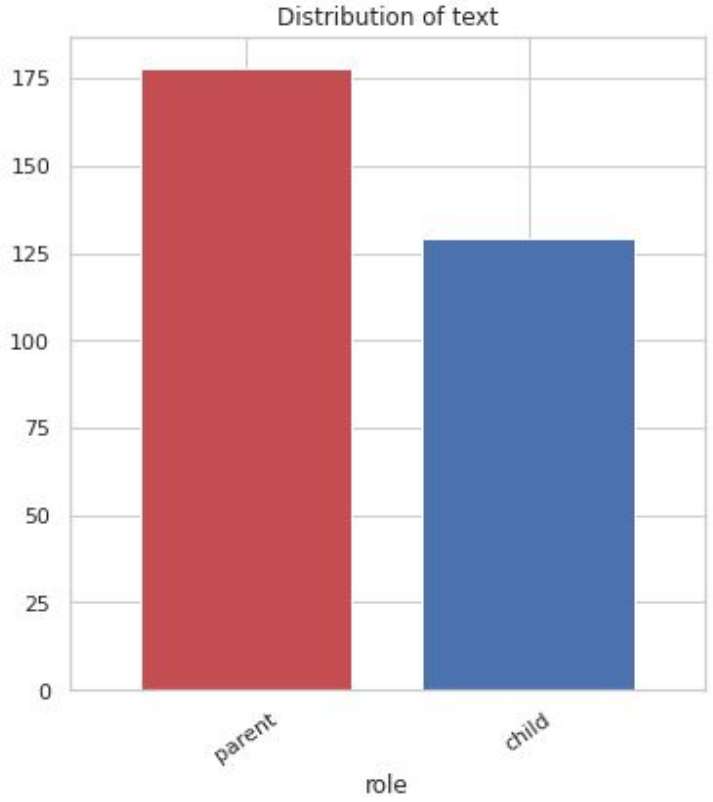
Были использованы разные классификаторы, а именно

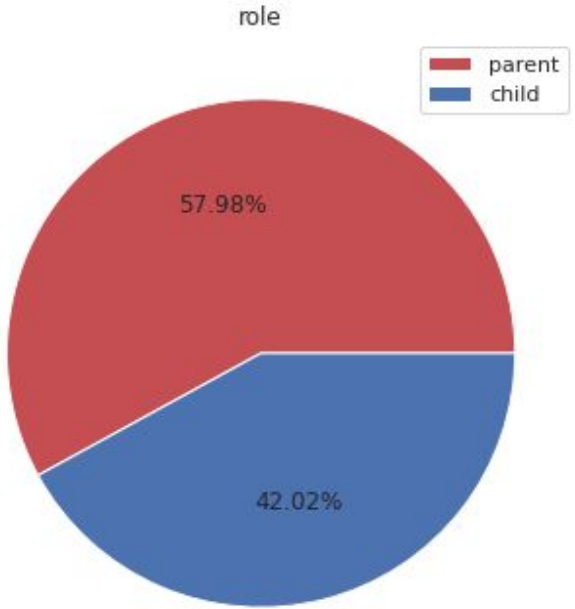
- Multinomial Naive Bayes
- Stochastic gradient descent (SGD)
- Logistic Regression
- KNeighbors

С помощью **Grid Search** подбирались наилучшие параметры

Обучение производилось как на обработанных данных, так и на сыром тексте

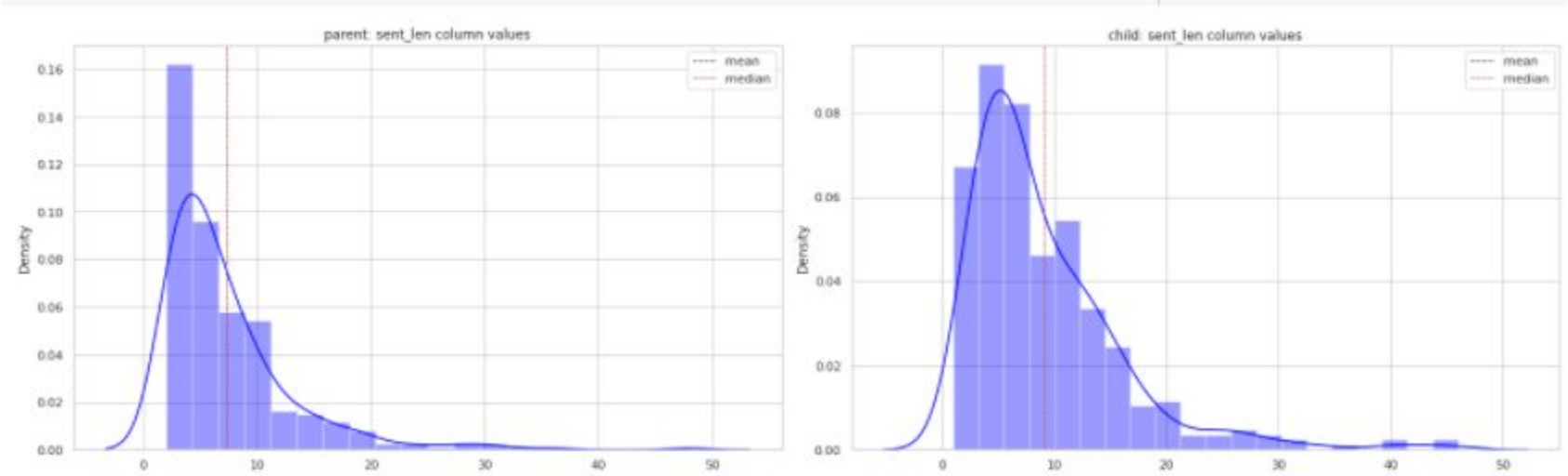
# Результаты. Контрастивный анализ



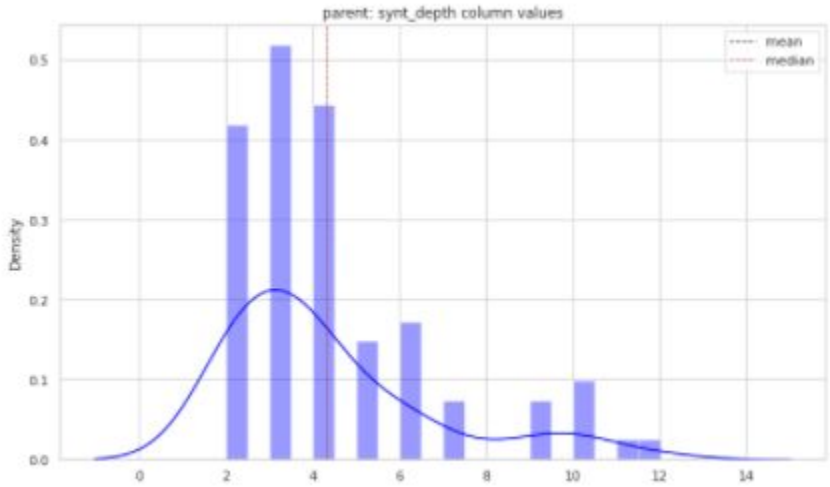
# Результаты. Контрастивный анализ

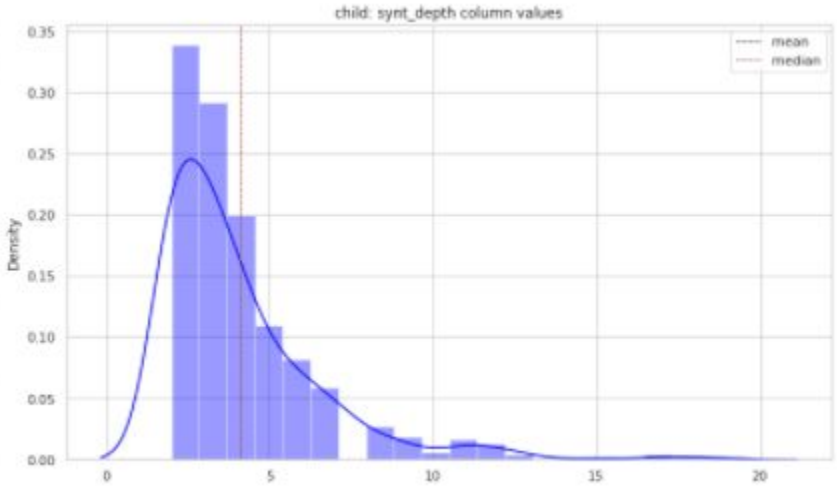
## Распределение по длинам предложений



# Результаты. Контрастивный анализ

## Распределение по глубине синтаксических деревьев



# Результаты. Качество классификаторов по accuracy

| Classificators | row text | row text + Grid Search | preprocessed text |
|---|---|---|---|
| **MultinomialNB** | 0.7402597402 | 0.7272727272 | 0.72727272727 |
| **SGDClassifier** | 0.7272727272 | 0.766233766 | 0.75324675324 |
| **Logistic Regression** | 0.7532467532 | 0.7532467532 | 0.68831168831 |
| **KNeighborsClassifier** | 0.6623376623 | 0.6233766233 | 0.636363636363 |

# Выводы

Предложения родителей немного длиннее предложений детей, =>

глубины предложений для них тоже больше

**Лучший классификатор:**

Логистическая регрессия на необработанных текстах

# Возможные улучшения

Использование word2vec embeddings

Попробовать классифицировать BERTом

Использовать больший объем данных