AutoNLP

With CatBoost

Препроцессинг

- Токенизация
- Удаление стоп-слов и пунктуации
- Лемматизация
- Стемминг

Стемминг и лемматизация

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

Функция предобработки текста

```python
def normalize_text_with_morph(self, x):
    x = x.lower().replace("ё", "е")

    words = ''.join([[" ", i][i in self.alphabet] for i in x]).lower().split()  # токенизация
                                                                                # и удаление пунктуации
    words = [w for w in words if w not in self.stop_words] # удаление стоп слов
    words = [self.morph.normal_forms(w)[0] for w in words]  # лемматизация
    words = [self.stemming(w) for w in words]  # стемминг
    return ' '.join(words)
```

Векторизация

Основные подходы

- TF-IDF
- Bag of words
- Word embeddings

Датасет

Проблема - данных нет, зато мы знаем соотношение классов

Решение - поиск аналогичного датасета и костомизация под изучаемый

Тестирование и подбор гиперпараметров

Проблема - ограничение по времени

Решение - подбор на схожих данных

```
grid = {'learning_rate': [0.1, 0.3, 0.5],
        'depth': [6, 8, 10],
        'l2_leaf_reg': [1, 3, 5, 7, 9]}

model = CatBoostRegressor()
result = model.randomized_search(grid,
                         X=_train,
                         y=train['target'],
                         )
```

Спасибо за внимание

github: github.com/sir-timio