RocketScience.ai

TENSOR.BY

ML-course

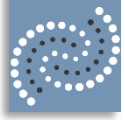
6. Natural Language Processing

Александр Фридман (Data Scientist),
alexandef@epica.ai

План

- Задачи и приложения **NLP**
- Задача классификации текстовой информации
- **FastText** – золотая пуля при решении задачи классификации текстов
- «Классические» **способы векторизации** текстовой информации
- Решение задачи классификации текстов средствами **Scikit-Learn**
- Методы предобработки текстовой информации. **NLTK**
- **spaCy**

Задачи и приложения NLP



- Классификация текстовой информации
- Тематическое моделирование (Topic Modelling)
- «Сокращение» текста (Text Summarization)
- Выявление имен собственных (Named Entity Recognition, NER)
- Определение частей речи (Part Of Speech Tagging, POS)
- *Построение ответов на вопросы* (Question Answering)
- *Машинный перевод* (Machine Translation)
- Paraphrase Detection (определение, несут ли два предложения один и тот же смысл)
- Speech Recognition
- Character Recognition
- Проверка правописания (Spell Checking)

FastText

Бесплатная, **легковесная** библиотека с **открытым исходным кодом**, позволяющая пользователям решать задачи классификации текстовой информации.

Не требует высокопроизводительного **железа** (работает на CPU).

Построенные модели могут быть значительно ужаты без значимой потери в качестве.

https://fasttext.cc/

FastText

Имеются **предобученные модели** для множества языков.

Открыт общественности в 2016 году компанией Facebook.

[1] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, *Bag of Tricks for Efficient Text Classification*

[2] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, *FastText.zip: Compressing text classification models*

«Классические» **способы векторизации** текстовой информации

Определения:

- Корпус (Corpora)
- Токен
- N-грамма
- Словарь

«Классические» способы векторизации текстовой информации

Мешок слов (Bag of Words)

- Порядок следования слов в документе не учитывается
- Алгоритм:
  1. Построение словаря
  2. Выбор слов, которые будут использоваться в качестве признаков (например, исключить очень редкие и очень частые), и их индексация
  3. Векторизация выборки

	also	love	programming
love programming	0	1	1
programming also love	1	1	1

«Классические» **способы векторизации** текстовой информации

TF-IDF (Term Frequency Inverse Document Frequency)

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Методы предобработки текстовой информации

С целью повышения качества работы модели можем предварительно «почистить» данные:

- Исправление опечаток
- Замена редких слов на их синонимы
- Замена фраз с отрицанием (не добрый -> злой)
- Удаление «стоп-слов» (предлоги, союзы, местоимения)
- Лемматизация, стемминг

Q & A

Thank you!