НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Компьютерная лингвистика и информационные технологии

Неделя 1: Предобработка данных

Описание курса

- Блоки “Информационный поиск” и “Компьютерная лингвистика”;
- Преп.: Олег Сериков (@oserikov), Мария Пономарева (@MashPo), Влад Михайлов (@vmkhlv);
- Ассист.: Дарья Самсонова (@SamsonovaDaria), Кирилл Коновалов (@kirillknv);
- https://github.com/vmkhlv/hse_compling_and_it;
- 1 модуль: 0.8 * ДЗ + 0.2 * Тесты;
- 3 модуль: 0.5 * ДЗ + 0.1 * Тесты + 0.2 * Контроль (Модуль 2) + 0.2 * Экзамен (Модуль 3);

Блок “Компьютерная лингвистика”

- Предобработка данных;
- Базовые понятия машинного обучения, алгоритмы машинного обучения для задач классификации, кластеризации и регрессии;
- Архитектуры нейронных сетей: FFN, CNN, RNN, LSTM;
- Sequence-to-sequence, Sequence Tagging, Language Modeling;
- Статические эмбеддинги: Word2Vec, FastText;
- Контекстуальные эмбеддинги: BERT, ELMo.

Фреймворки


Keras
A deep learning library


PYTORCH
Deep Learning with PyTorch

About 634,000,000 results (0.62 seconds)

Russian – detected ⇄ English

Можно поспать?

Mozhno pospat'?

Can I sleep?

Feedback

Говорят ты сегодня нагрубил

21:56

hows life baby gorilla

babygirl*

Олег

Роза моего искусственного интеллекта цвела не ради этого вопроса.

21:56

Как учится машина? Основные понятия

- Объект (картинка, текст, аудио и д.р.);
- Целевая функция / таргет (класс объекта / чиселка);
- Признаки объекта (возраст, пол, уровень дохода и т.д.);
- Матрица объектов-признаков;
- Обучающая / валидационная / тестовая выборки (train / val / test);
- Модель / алгоритм – это решающая функция;
- Функция потерь / ошибок;
- Метрика.

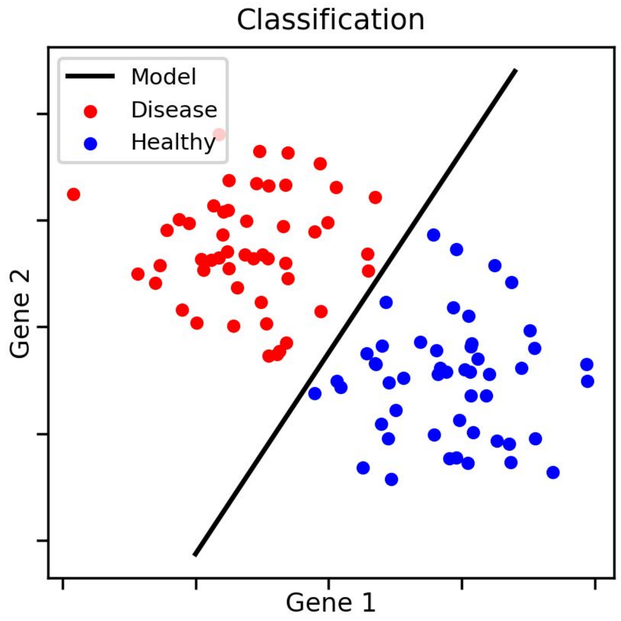
Целевая функция определяет задачу

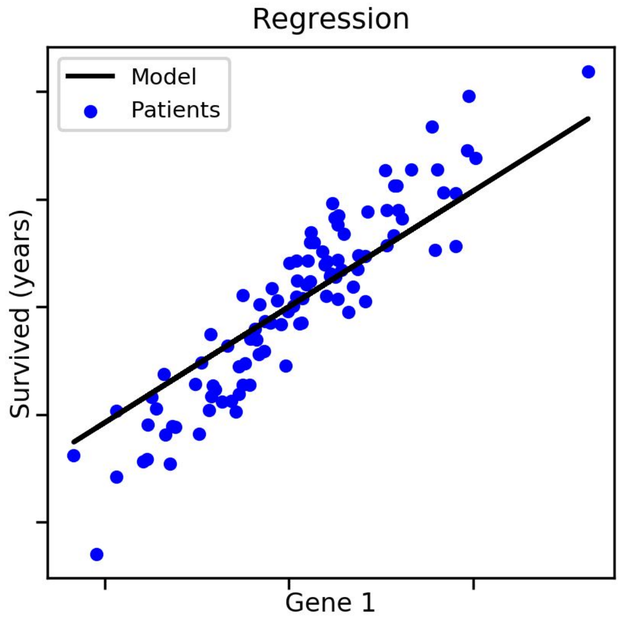
Классификация

- Фильтрация спама;
- Определение темы сообщения;
- Определение языка.

Регрессия

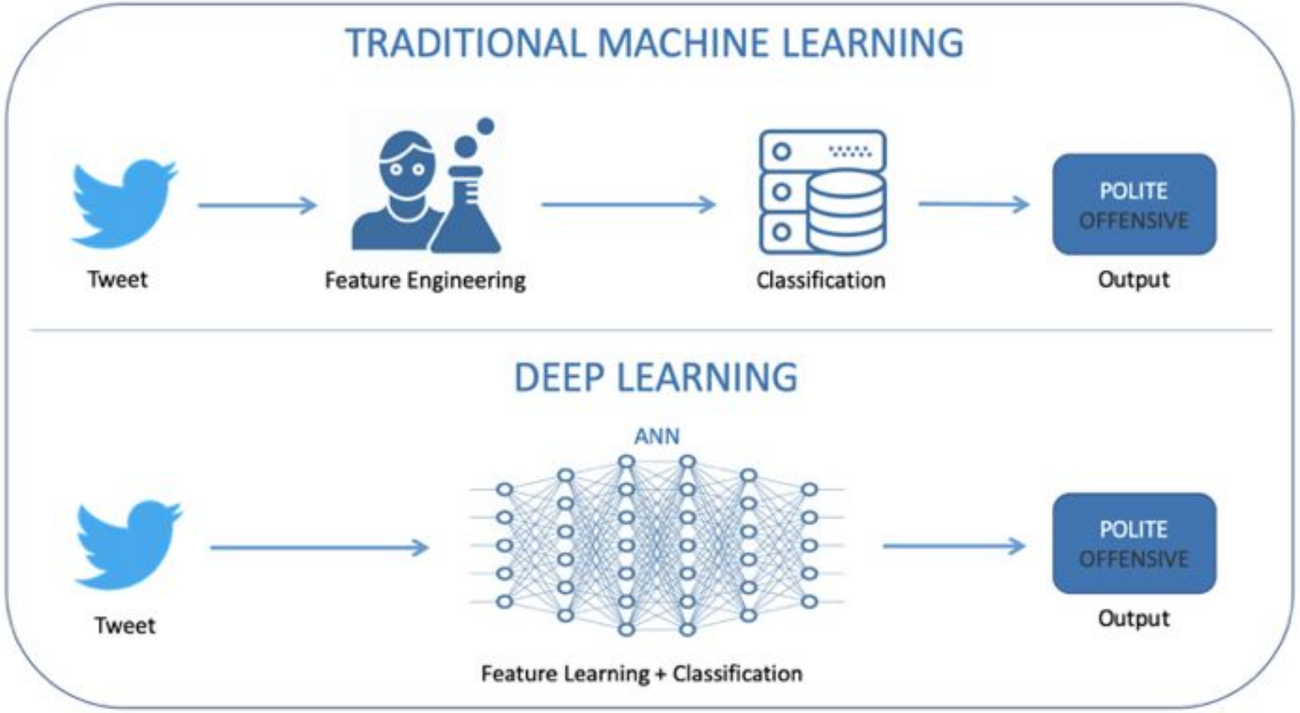
- Количество лайков;
- Количество зараженных COVID-19.





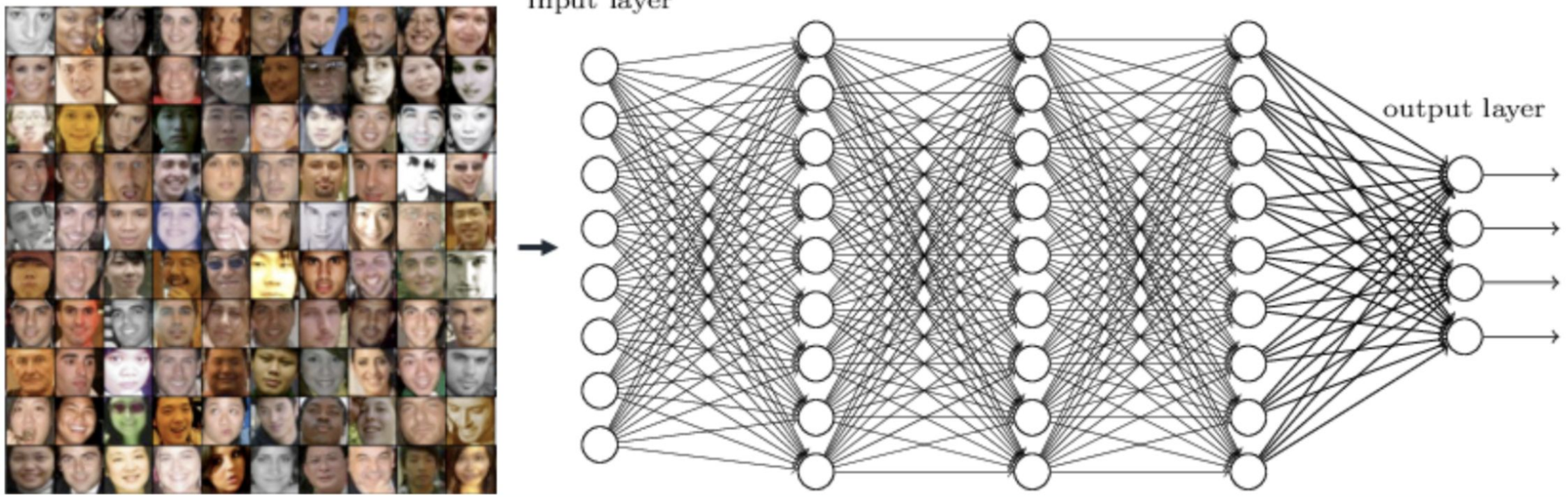
Важность признаков

- Конструирование;
- Извлечение;
- Отбор;
- Определение важности;
- Интерпретация поведения.



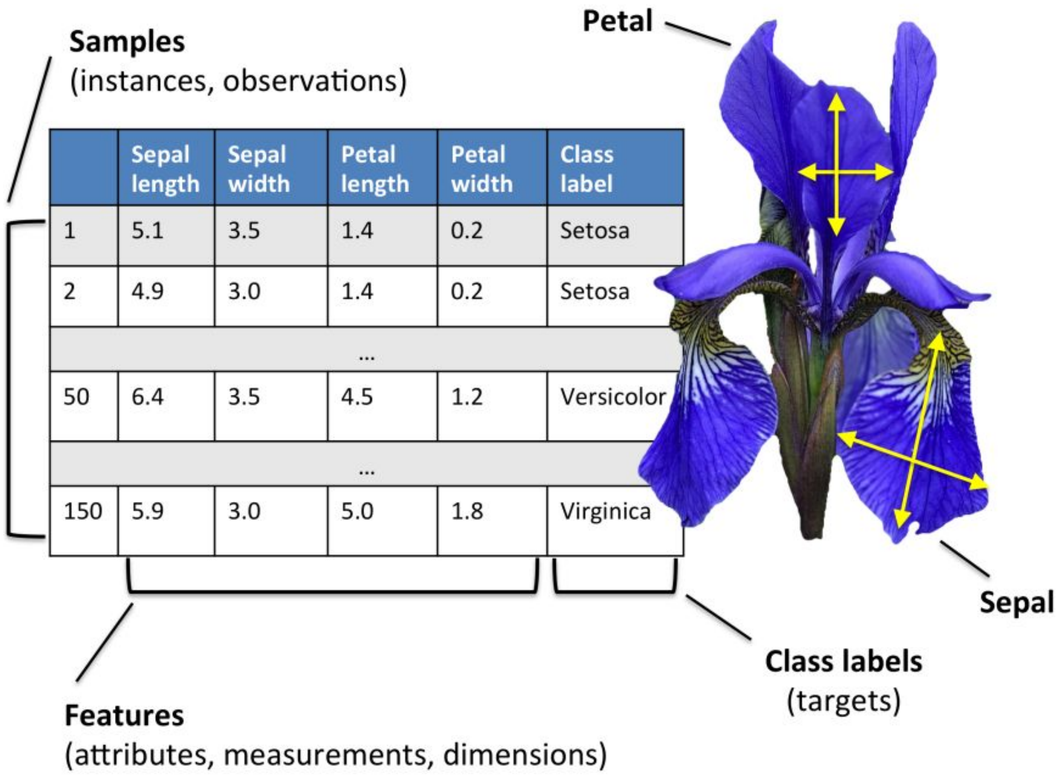
Deep neural networks learn hierarchical feature representations

hidden layer 1 hidden layer 2 hidden layer 3



Типы признаков

- Чиселки (numerical) – доход, объем, масса;
- Бинарные (binary) – пришли / не пришли на пару;
- Категориальные (categorical) – пол, уровень английского;
- Порядковые – курс.



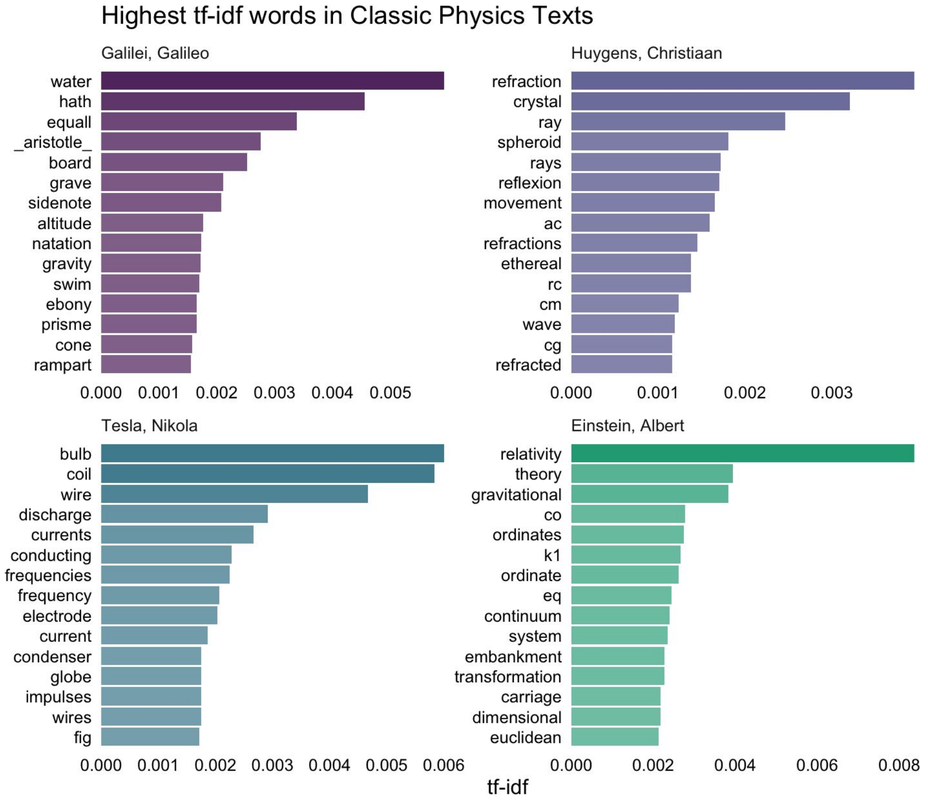
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Текстовые признаки

Текст – это упорядоченная *последовательность*.

Например:

- **Средняя длина;**
- **Средняя частотность по коллекции;**
- **IPM (instance per million);**
- **Среднее количество сущностей / документ;**
- **Частотность пос-тегов по коллекции;**
- **Частотность N-грамм пос-тегов / лемм;**
- **Automated Readability Index (ARI).**



Решаем ЕГЭ

- CatBoostClassifier;
- POS N-grams;
- Bag of words;
- Ensemble + Voting.

.

1. Задание #T29871

Расставьте знаки препинания.

> *Он стоял перед дворцом во время обеда государя (1) глядя в окна дворца (2) ожидая чего-то ещё и (3) завидуя одинаково и сановникам (4) подъезжавшим к крыльцу (5) и (6) мелькавшим в окнах (7) камер-лакеям.*

Укажите цифру(-ы), на месте которой(-ых) в предложении должна(-ы) стоять запятая(-ые).

Ответ

Проверить ответ

Показать разбор и ответ

Определение автора

- **Topic or Style? Exploring the Most Useful Features for Authorship Attribution (Sari et al., 2018)**

.

Type	Group	Category	#	Description
Style	Lexical	Word-level	2	Average word length, number of short words
		Char-level	2	Percentage of digits, percentage of upper-case letters
		Letters	26	Letter frequency
		Digits	10	Digit frequency
		Vocabulary richness	2	Richness (hapax-legomena and dis-legomena)
	Syntactic	Function words	174	Frequency of function words
		Punctuation	12	Occurrence of punctuation
Content	Word *n*-gram	Words unigrams	100	Frequency of 100 most common word unigrams
		Words bigrams	100	Frequency of 100 most common word bigrams
		Word trigrams	100	Frequency of 100 most common word trigrams
Hybrid	Char *n*-gram	Char bigrams	100	Frequency of 100 most common character bigrams
		Char trigrams	100	Frequency of 100 most common character trigrams

Table 3: Authorship attribution feature sets.

Предобработка

- Сегментация: *rusenttokenize, razdel*
- Токенизация: *spacy_russian_tokenizer, razdel*
- Стоп-слова: *nltk stopwords*
- Морфология: *pymorphy2, mystem, slovnet, RNNMorph*
- Синтаксис: *UDPipe, slovnet, stanza, GramEval2020*

	raw_word	stemmed_word
0	..trouble..	..trouble..
1	trouble<	trouble<
2	trouble!	trouble!
3	<a>trouble	<a>trouble
4	1.trouble	1.troubl

Зачем все это?

- Предварительный анализ данных;
- Фильтрация данных;
- Составление датасетов;
- Аугментация данных;
- Конструирование признаков;
- Составление выборок;
- Анализ поведения модели и ошибок.