Natural Language Processing

Яковенко Ольга

Распознавание именованных сущностей

Нахождение и классификация упоминаний объектов в неструктурированном тексте по заранее определенным категориям, таким как:

- имена людей,
- организации,
- местоположения,
- медицинские коды,
- выражения времени, количества, денежные значения,
- проценты и т. д.

Распознавание именованных сущностей

Jim bought 300 shares of Acme Corp. in 2006.

[Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

Распознавание именованных сущностей

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Для чего это нужно?

Когда мы хотим избавиться от ИС в тексте?

Для определения общего смысла текста нам нужны только общеупотребительные слова

Задачи:

- Тематики
- Сентимент

Когда мы хотим ИС найти в пределах определённого текста?

Найти ИС конкретного типа и с каждым провести отдельный анализ, зависящий от типа

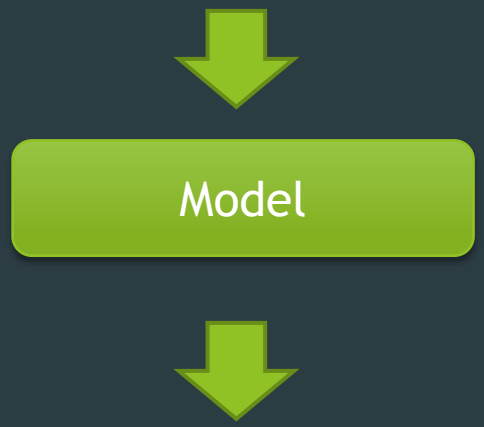
Задачи:

- Автоматическое извлечение информации (information retrieval)
- Вопросно-ответные системы/чат-боты

ML алгоритмы

Фиксированный input

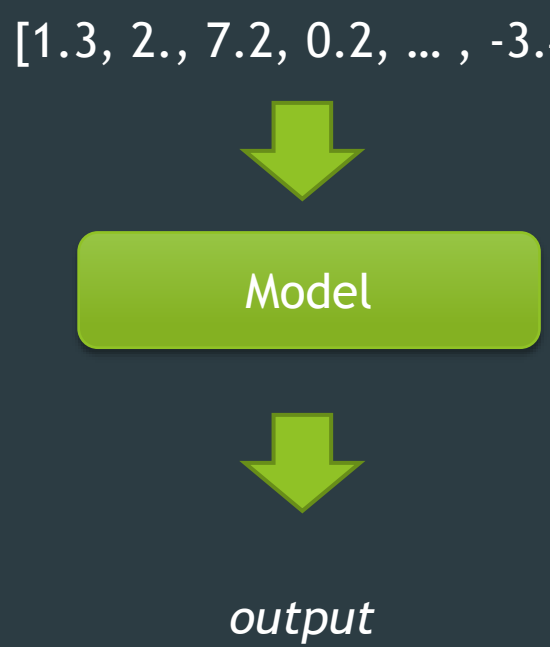
[1.3, 2., 7.2, 0.2, … , -3.4]



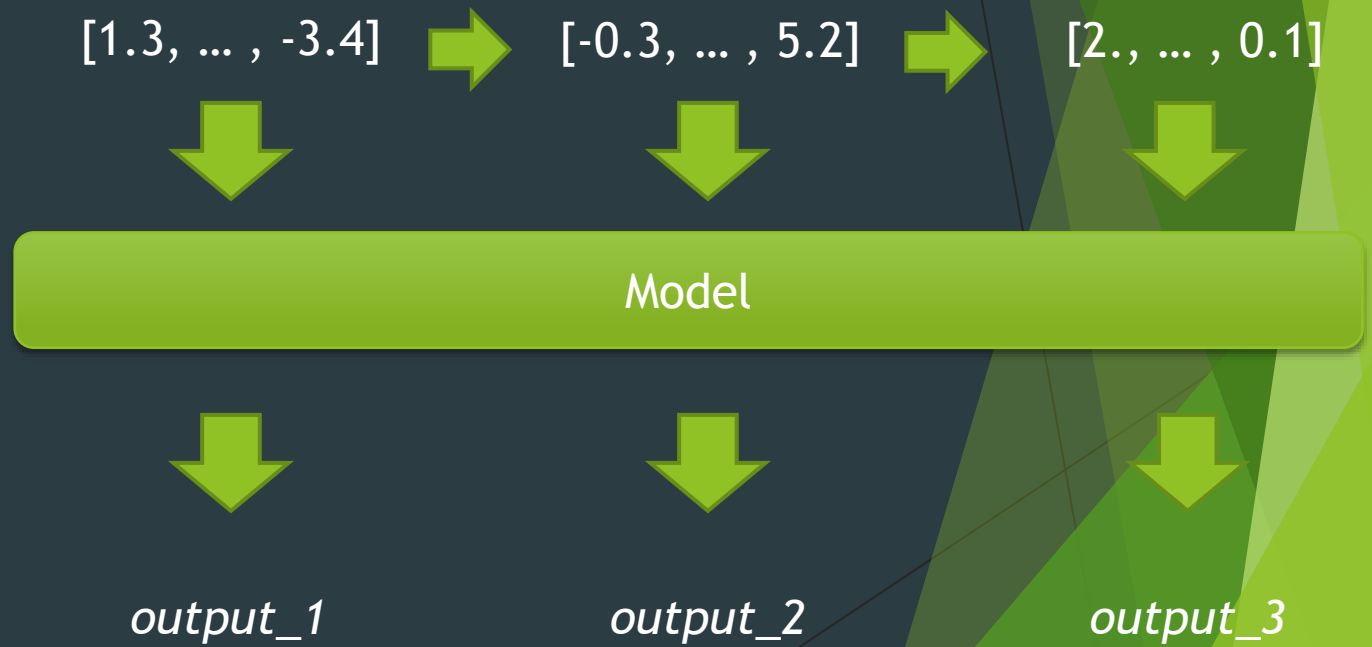
output

ML алгоритмы

Фиксированный input



Input переменной длины



ML алгоритмы

Фиксированный input

- Decision trees
- Logistic regression
- Multilayer perceptron (MLP)
- Convolutional neural network (CNN)
- …

Input переменной длины

- Recurrent neural network (RNN)
- Conditional random fields (CRF)
- Markov models + Hidden Markov models (MM + HMM)
- Finite-state automaton (FSA)

ML алгоритмы

Фиксированный input

- Decision trees
- Logistic regression
- Multilayer perceptron (MLP)
- Convolutional neural network (CNN)
- …

Input переменной длины

- Recurrent neural network (RNN)
- Conditional random fields (CRF)
- Markov models + Hidden Markov models (MM + HMM)
- Finite-state automaton (FSA)

Задание

Датасет: https://github.com/dialogue-evaluation/factRuEval-2016/tree/master/devset

crfsuite: https://pypi.org/project/python-crfsuite/

Пример NER с crf-suite: https://github.com/scrapinghub/python-crfsuite/blob/master/examples/CoNLL%202002.ipynb