Intro to NLP

Сергей Аксенов

Высшая Школа Экономики

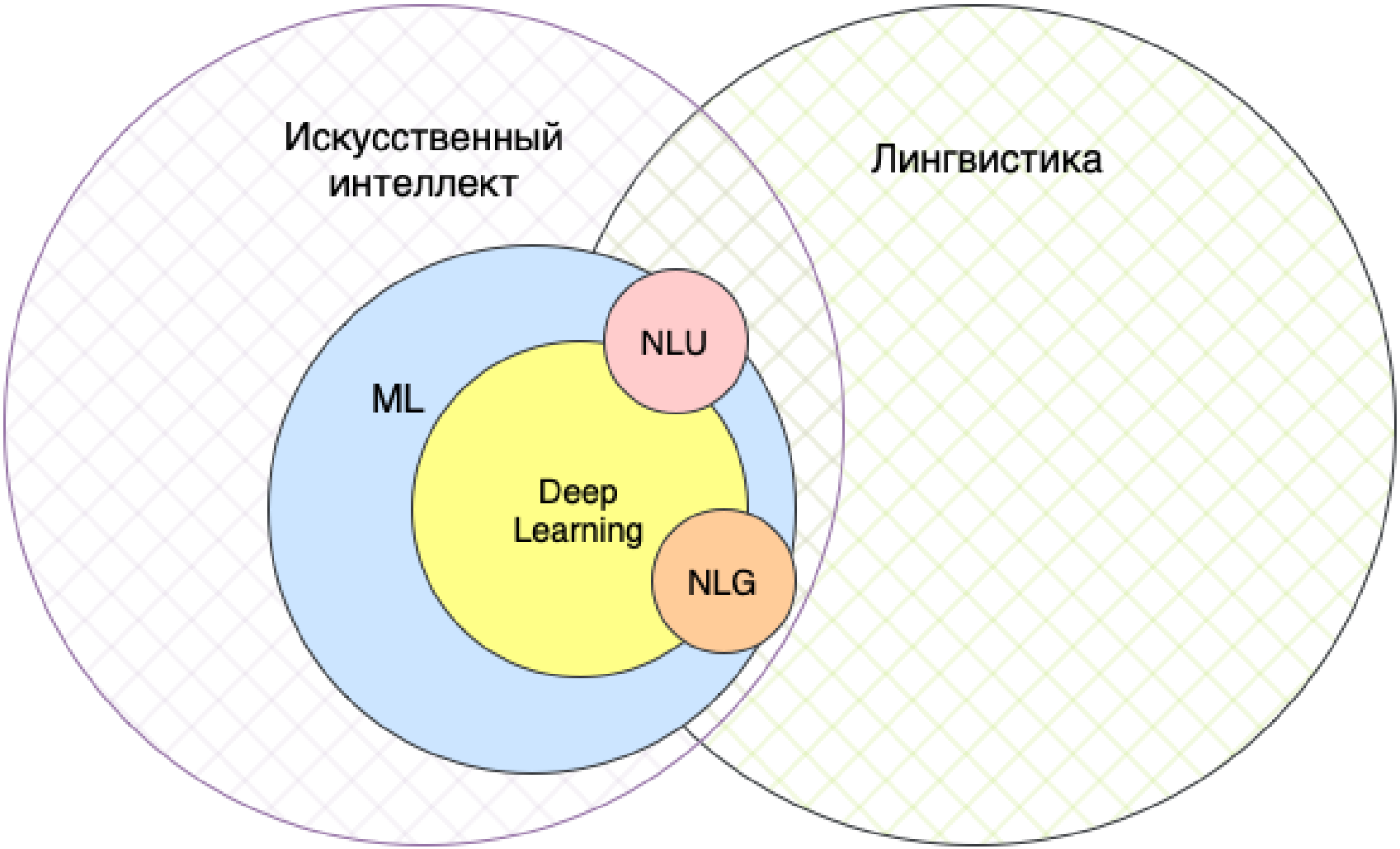
12 сентября 2022 г.

Today

Intro

Об этом курсе

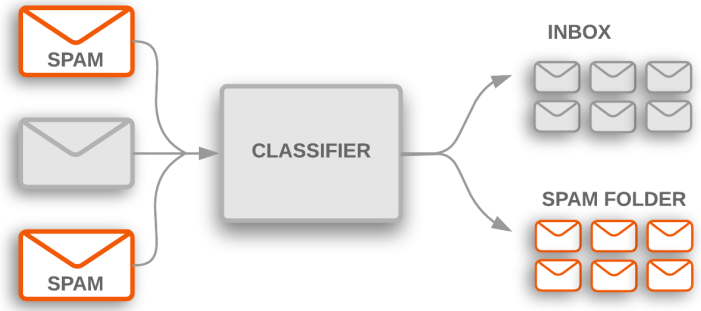
Автоматическая обработка текстов



Примеры задач

Text classification

- Sentiment analysis
- Intent detection
- Spam filtering
- Topic classification



Sequence labelling

- Named entity recognition
- Coreference resolution

Sequence transformation (seq2seq)

- Machine translation
- Question answering

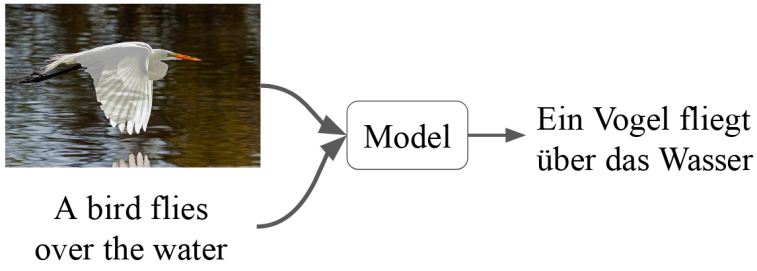
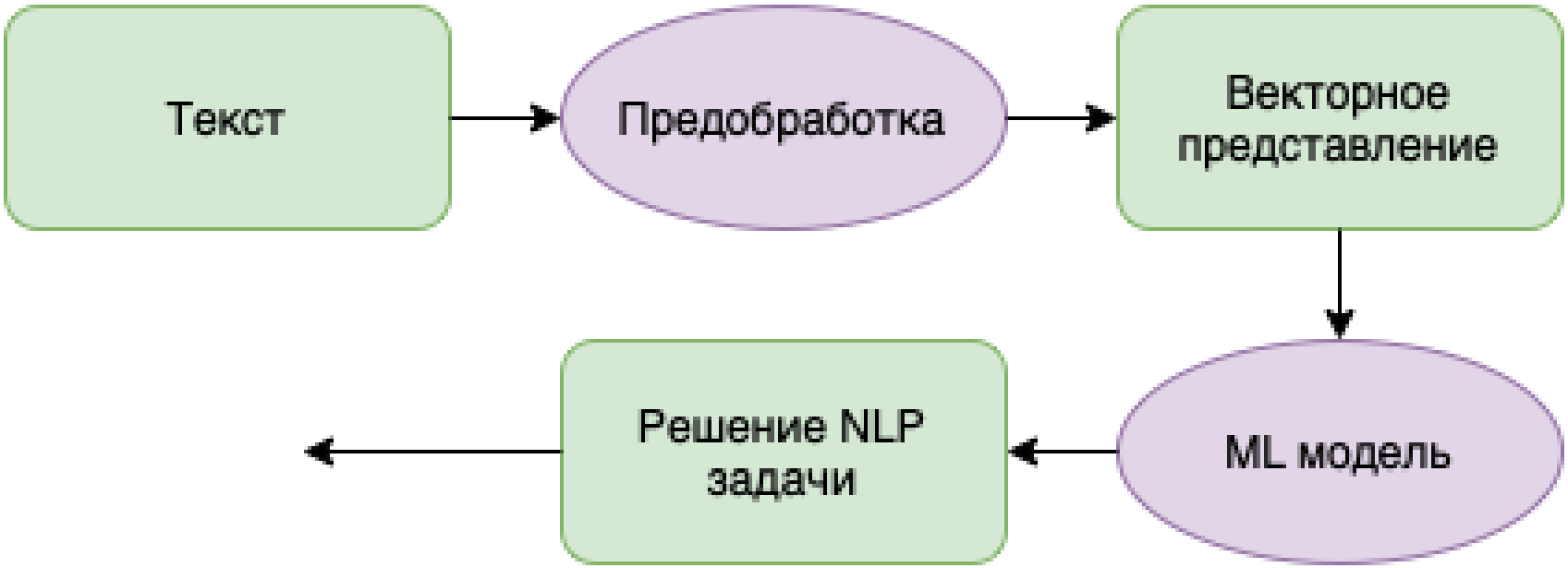


Image sources: developers.google.com, towardsdatascience.com, statmt.org

Машинное обучение в обработке текстов



Почему работать с текстами сложно

Многозначность

- *орган*, *bank*
- *Он был в отличной форме, но на животе она не застегивалась*

Омонимия:

- *the ship*, *to ship*
- *Стекло*

Идиомы:

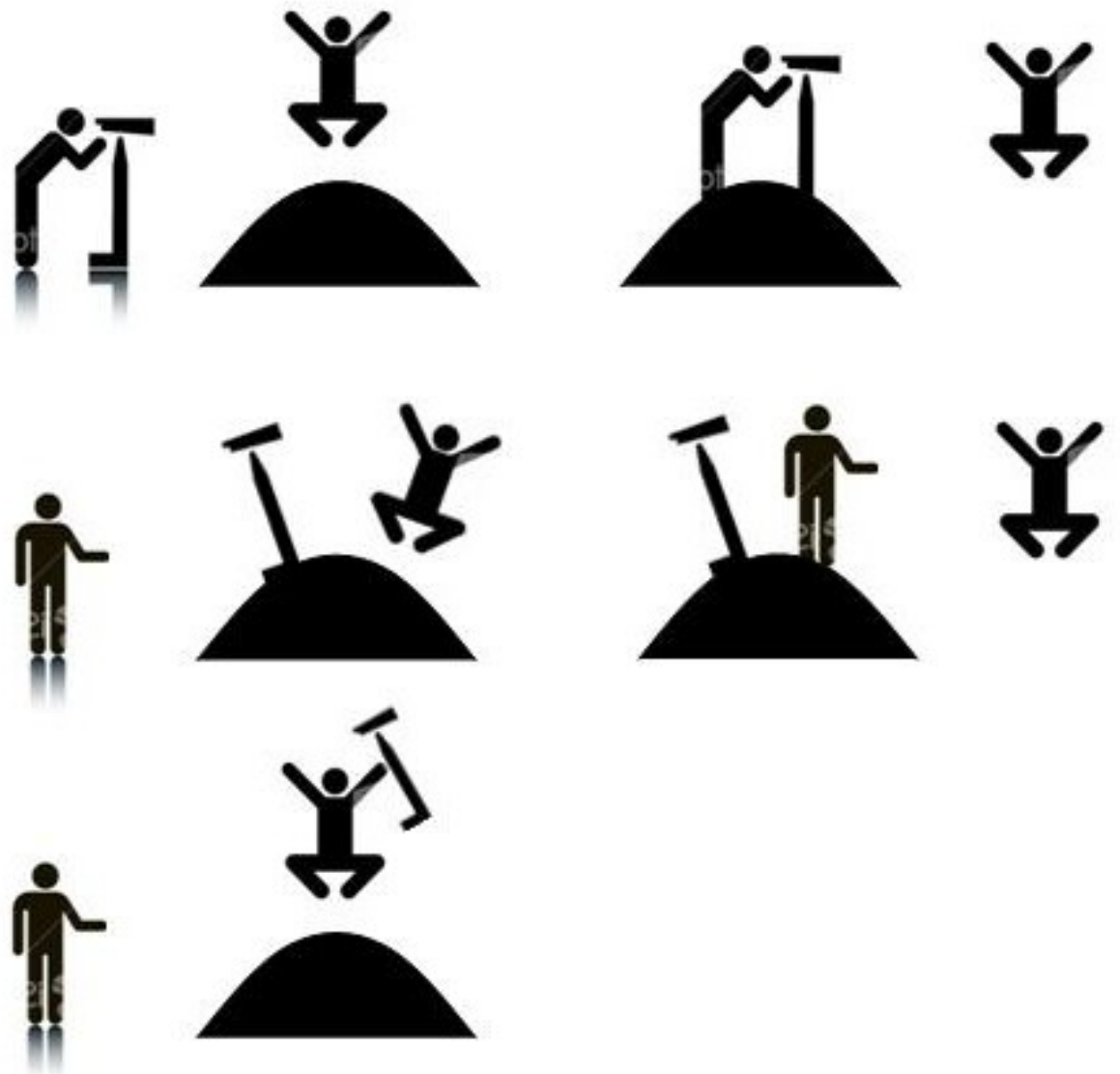
- *Играть с огнем*
- *Вышел из себя*

Синтаксическая неоднозначность:

- *John saw the man on the mountain with a telescope*

Почему работать с текстами сложно

John saw the man on the mountain with a telescope



Почему работать с текстами сложно

Знание об окружающем мире

- *Даша **[обругала/обняла]** Машу, потому что она была расстроена*
- Кто расстроился? **[Даша/Маша]**

Неологизмы:

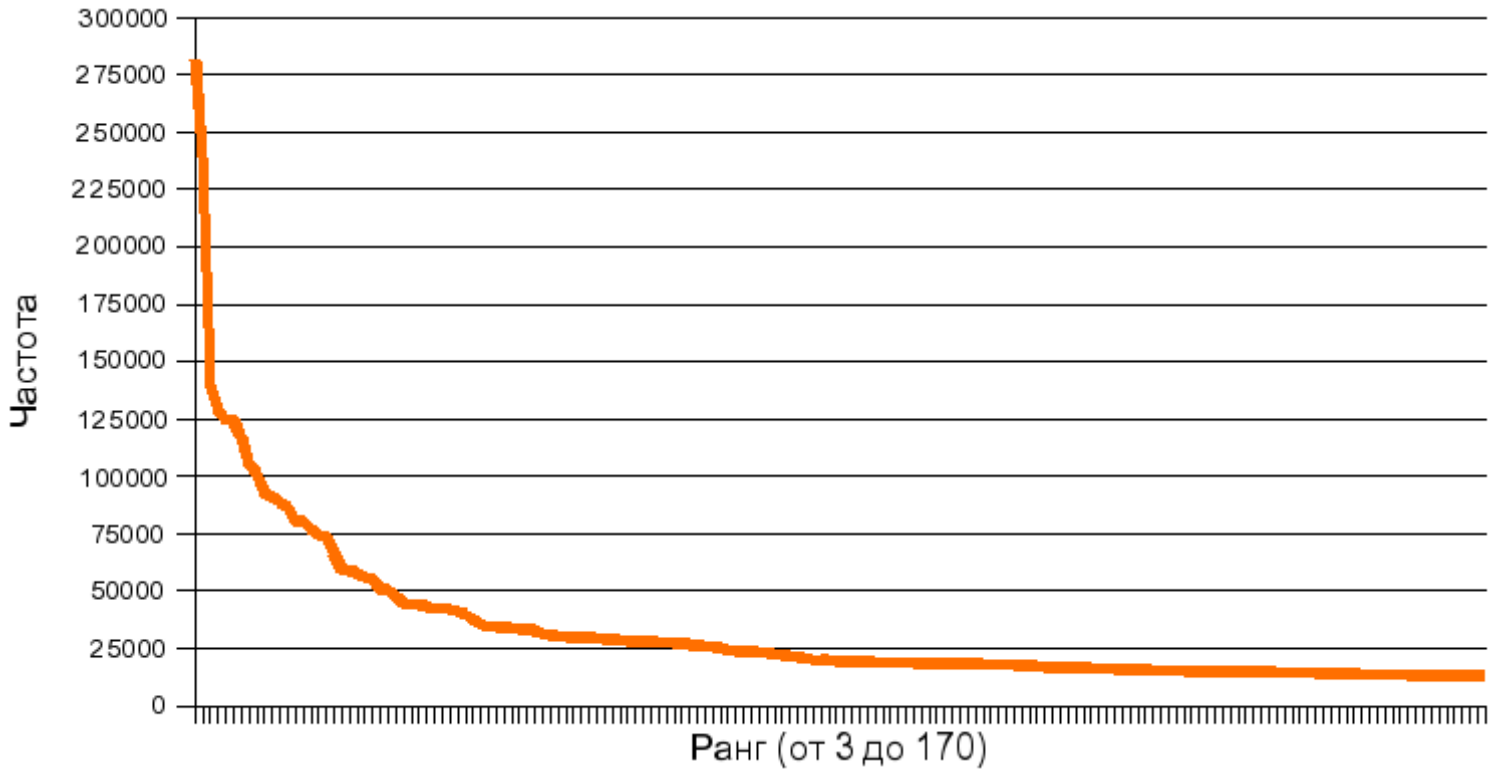
- *Репост*
- *Каршеринг*
- *Митинг*

Сложные именованные сущности:

- *«**Анна Каренина**» стала первым российским **мюзиклом**.*

Почему работать с текстами сложно

- **Слово** - базовая структурная единица языка
- Слово само по себе несет много важной информации
- Возникают большие разреженные пространства признаков
- Закон Ципфа



Today

Intro

Об этом курсе

Об этом курсе

- **Лекции**: Сергей Аксенов
- **Репозиторий**:
  `https://github.com/SergeyAxe/HSE_nlp_2022`
- **Чат**: `https://t.me/+SDCFWAmliloOZDZi`
- **Итоговая оценка**:
  $M_{hw} = \frac{1}{3}(M^1_{hw} + M^2_{hw} + M^3_{hw})$
  $M_{quiz} = \frac{1}{3}(M^1_{quiz} + M^2_{quiz} + M^3_{quiz})$
  $M_{final} = round(0.4M_{exam} + 0.7M_{hw} + 0.2M_{quiz})$
  $M_{exam}, M^i_{hw}, M^i_{quiz} \in [0, ..10]$

План курса

1. Предобработка текстов
2. Векторные языковые модели
3. Классификация текстов
4. Моделирование последовательностей
5. Walk down Sesame Street
6. Синтаксис
7. Машинный перевод
8. Генерация текстов

Предобработка текстов

- Токенизация
- Сегментация предложений
- Удаление пунктуации
- Удаление стоп-слов
- Фильтрация по длине, частоте, регулярному выражению
- Лемматизация – приведение к нормальной форме
- Стемминг – приведение к основе