

Intro to NLP

Сергей Аксенов

Высшая Школа Экономики

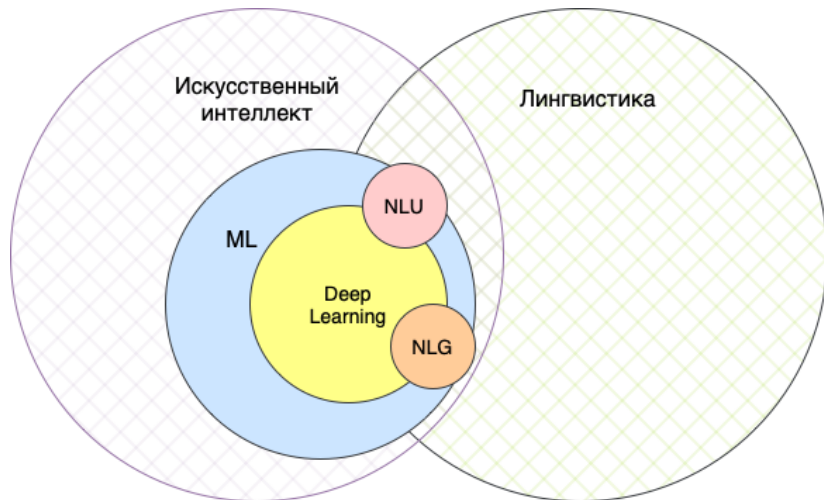
12 сентября 2022 г.

Today

Intro

Об этом курсе

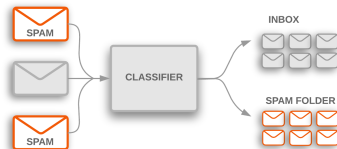
Автоматическая обработка текстов



Примеры задач

Text classification

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



Sequence labelling

- Named entity recognition
- Coreference resolution

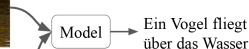
contentShip to site index@helios.futurinvest.org In Today's [Pope's Advertisements Supported](#) [ORG](#) by [B.I. Agent](#) [Polar Stock](#) [PERSON](#).
[Who Collected Trump](#) [PERSON](#) a [Text](#) in [Fandango](#) [PERSON](#) [Trump](#) [PERSON](#) were uncovered, was fired [Credit](#) [J. Kumpack](#) [PERSON](#) for [The New York](#)
[Financially Adam Goldstein](#) [ORG](#) and [Michael S. Schmieding](#) [PERSON](#) [13](#) [CARDINAL](#) [2018](#) [WASHINGTON](#) [CARDINAL](#) [Polar Stock](#)
[PERSON](#) the [F.B.I.](#) [ORG](#) senior counterintelligence agent who disparaged President [Trump](#) [PERSON](#) in inflammatory text messages and helped
oversee the [Hobby Club](#) [PERSON](#) email and [Hobby](#) [ORG](#) investigations, has been fired for violating bureau policies, Mr. [Stock](#) [PERSON](#)'s lawyer
said [Monday](#) [DATE](#). Mr. Trump and his allies seized on the texts — exchanged during the [2016](#) [DATE](#) campaign with a former [F.B.I.](#) [ORG](#) lawyer,
[Live Page](#) — [F.](#) [PERSON](#) ensnaring the [Greece](#) [ORG](#) investigation as an illegitimate "catch hunt" Mr. [Stock](#) [PERSON](#) who rose over [20](#) [years](#)
[DATE](#) at the [F.B.I.](#) [ORG](#) is become one of its most experienced counterintelligence agents, was a key figure in [the early months](#) [DATE](#) of the
inquiry. Along with writing the texts, Mr. [Stock](#) [PERSON](#) was accused of sending a highly sensitive search warrant to his personal email account. The
[F.B.I.](#) [ORG](#) had been under immense political pressure by Mr. [Trump](#) [PERSON](#) to dismiss Mr. [Stock](#) [PERSON](#), who was removed [last summer](#)
[DATE](#) from the staff of the special counsel, [Robert S. Mueller II](#) [PERSON](#). The president has repeatedly denounced Mr. [Stock](#) [PERSON](#) in posts on

Sequence transformation (seq2seq)

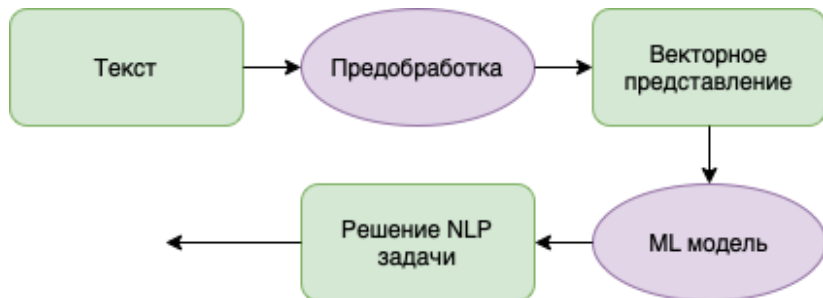
- Machine translation
- Question answering



A bird flies
over the water



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:

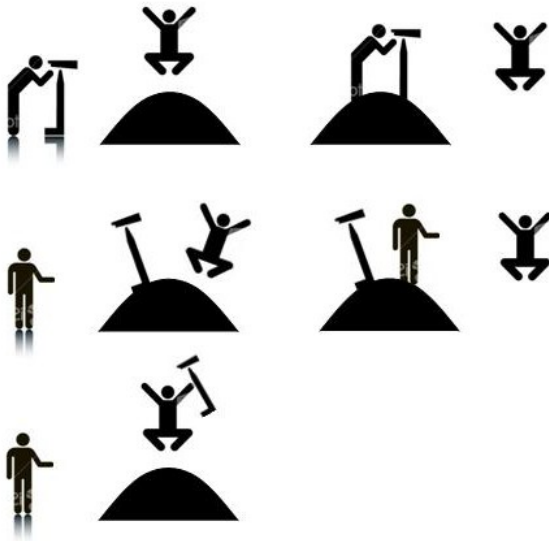
- ▶ *Играть с огнем*
- ▶ *Вышел из себя*

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:

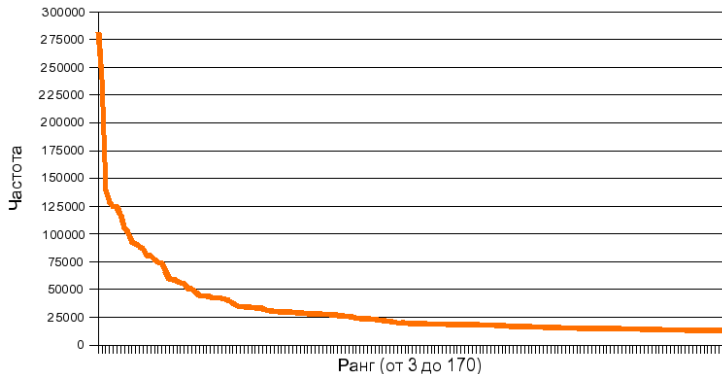
- ▶ Репост
- ▶ Каршеринг
- ▶ Митинг

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/+SDCFWAmllilo0ZDZi>

► Итоговая оценка:

$$M_{hw} = \frac{1}{3}(M_{hw}^1 + M_{hw}^2 + M_{hw}^3)$$

$$M_{quiz} = \frac{1}{3}(M_{quiz}^1 + M_{quiz}^2 + M_{quiz}^3)$$

$$M_{final} = \text{round}(0.4M_{exam} + 0.7M_{hw} + 0.2M_{quiz})$$

$$M_{exam}, M_{hw}^i, M_{quiz}^i \in [0, ..10]$$

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