



# Центр непрерывного образования



---

**Итоговая работа по программе профессиональной переподготовки  
Специалист по Data Science**

**Студент:** Байдаков Илья

**Руководитель:** Косолапов Кирилл

**Тема работы:** Обработка текстов, полученных после системы  
распознавания речи, с целью выделения тем и именованных сущностей

---

## ОБЗОР ЛИТЕРАТУРЫ И АКТУАЛЬНОСТЬ РАБОТЫ

---

История изучения и общие черты тематического моделирования, выделения именованных сущностей и работы с речевыми записями

- **1950 → 2023:** от «мешка слов» к нейронным сетям
- **тематика работ:** методы | языки | датасеты | специфические задачи
- **качество обработки текста** увеличивается
- **потребность в обработке текста** растёт. Сфера применения: медиа | наука | бизнес и пр.
- **тенденция:** применение сетей *RNN-LSTM* и *transformer*
- **суммаризация текста от системы распознавания речи:**
  - Е около 10 опубликованных работ
  - голос → текст → обработка текста
  - экстрактивные и абстрактивные методы

## ЗАДАЧА

---

➤ Цель работы: определение особенностей выделения тем, именованных сущностей и аномалий в текстах, полученных после системы распознавания речи (ASR)

➤ Исходные данные:



*Full MovieLens Dataset*: информация о фильмах ~ 45000 строк

**movielens**

Non-commercial, personalized movie recommendations.



видеофайлы: *mp4* / *avi* / 20 шт.



**opensubtitles.org**

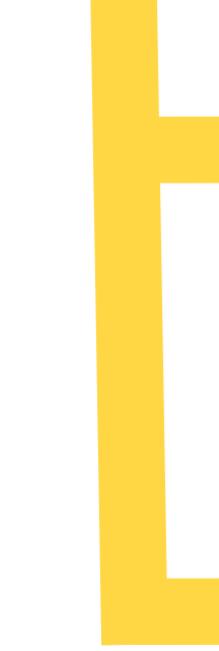


субтитры фильмов: *srt* / 1950 шт.

➤ Задачи:



**Задача выделения именованных сущностей (NER).** Определение метрики F-score для результата ASR и сравнение с WER для текста субтитров



**Задача тематического моделирования** текстов-результатов ASR. Выделение тем методом обучения без учителя



**Задача предсказания признака *keywords*** по тексту субтитров и по результату ASR

# ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ

---

1

**Первичная обработка и объединение исходных датасетов**

pandas, seaborn

2

**Загрузка и обработка субтитров srt → текст**

API OpenSubtitles, pysrt

субтитры

3

**Получение и нарезка аудиодорожки для ASR:**

ffmpeg, pydub

аудио

# ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ

---



речь

**Получение текста из аудиодорожки:**

googlecolab, whisper



ключевые  
слова

**Препроцессинг признака *keywords*:**

фильтрация по количеству, чистка, удаление редких ключевых слов



субтитры

**Препроцессинг текстов субтитров:**

nltk и собственные правила

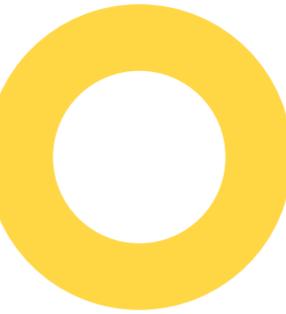
## ЗАДАЧА ВЫДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ (NER)

- ▶ Для задачи NER использована модель *spaCy* «*en\_core\_web\_lg*»
- ▶ Проведена ручная разметка именованных сущностей, общий объём:  
~10 тыс. слов, ~ 1.3 тыс. именованных сущностей  
F-score (субтитры) = 0,84  
F-score (текст ASR) = 0,75
- ▶ Сделана оценка влияния ошибок ASR на качество выделения именованных сущностей

Фильм	Annie Hall	
Источник текста	Субтитры	ASR
Вставки	35	37
Замены	62	50
Пропуск	41	91
Всего NE	589	
F-мера	0,80	0,71

Пример результата ручной разметки

# ЗАДАЧА ВЫДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ (NER)



Ошибки определения NER в текстах, полученных системой ASR.  
Слева — субтитры, справа — ASR

important joke for me is one that's usually attributed to Groucho Marx PERSON . I think it appears originally in Freud PERSON 's Wit and its Relation ORG to the Unconscious It goes like this - I'm paraphrasing. I would never wanna belong to any club... ...that

quickly the the other important joke for me is one that's usually attributed to Groucho Marx PERSON But I think it appears originally in Freud PERSON 's wit and its relation to the unconscious and it goes like this. I'm paraphrasing I would never want to belong to any club that would have someone like me for a

assigned to our school. I always thought my schoolmates were idiots. Melvyn Greenglass PERSON . His fat little face. And Henrietta Farrell PERSON . Just Miss Perfect PERSON all the time. And Ivan Ackerman PERSON . Always the wrong answer. Always.

assigned to our school. I Must say I always thought my schoolmates were idiots Melvin Greenglass PERSON , you know it's fat little face and Henrietta Farrell PERSON just miss perfect all the time and Ivan Ackerman PERSON always the wrong

## Ошибка определения регистра (название книги)

cutted guy... ...looking at me in a funny way and saying, "We have a sale this week DATE on Wagner PERSON ." Wagner PERSON , Max PERSON . Wagner PERSON . I know what

at me in a funny way and smiling And he's saying yes we have a sale this week DATE on Wagner Wagner PERSON . Max Wagner PERSON So I know what he's really trying to tell me

stop listening to him. He's screaming his opinions in my ear. Like all that Juliet PERSON of the Spirits or Satyricon ORG . I found it incredibly... indulgent. He really is. He's one of the most indulgent filmmakers.

listening to him Leading from one to the other He's screaming his opinions in my ear I don't know if he's a Juliet of the Spirits or Satyarakhan PERSON I found it incredibly indulgent you know he really is He's one of the most

## Неверная расстановка пунктуации

## Орфографическая ошибка

## ЗАДАЧА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

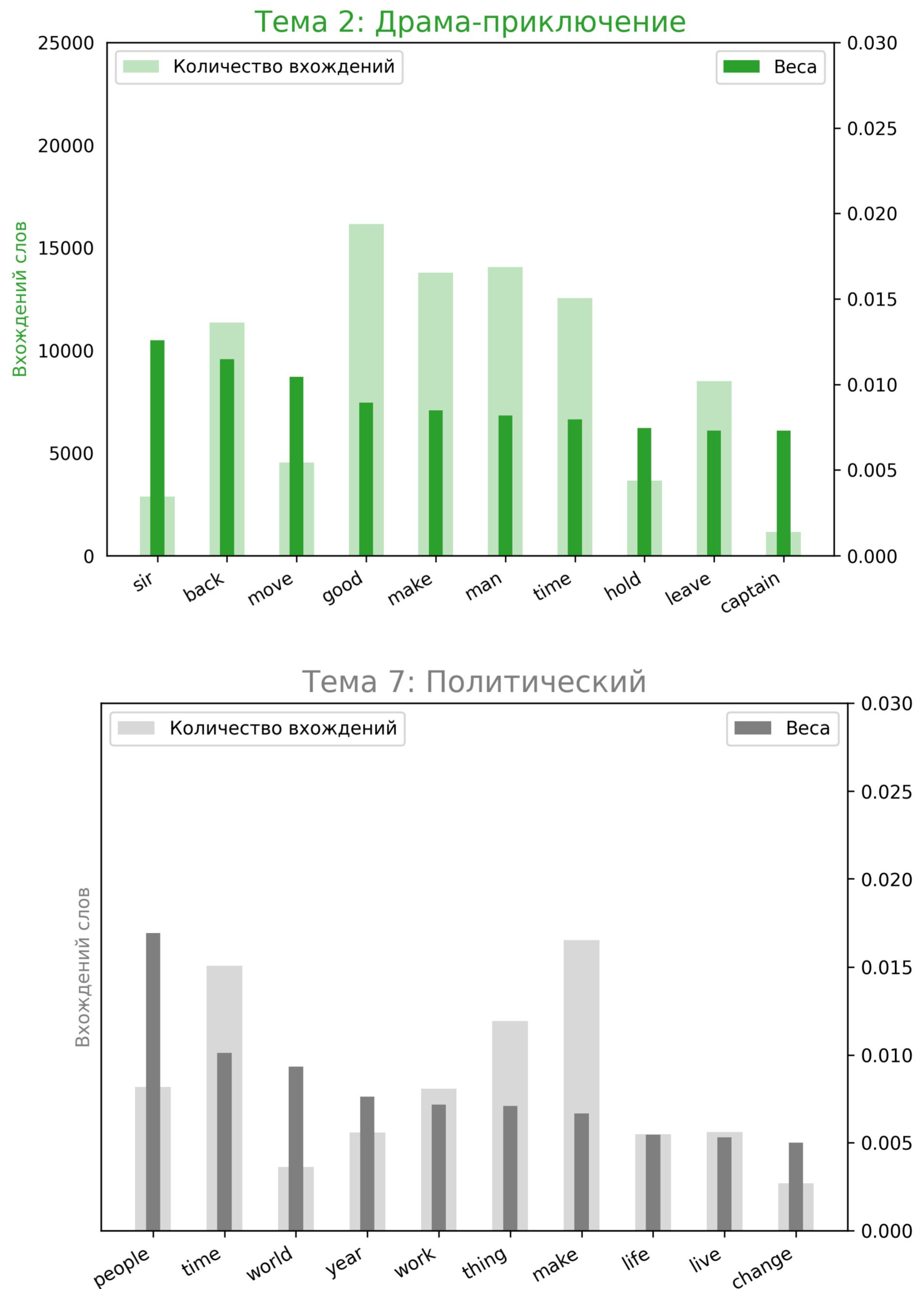
---

- Требуется проверить, насколько успешно можно проводить тематическое моделирования для текстов, полученных ASR
- Использовал алгоритм LDA
- Выбор количества тем проводился на основе кривой изменения метрики *coherence* /и по субъективной оценки смысла токенов
- Выделение тем прошло успешно: темы выражают общее содержание фильма
- Модель *LDA Mallet* обучалась на 1950-ти файлах субтитров
  - Тематическое моделирование текстов-результатов ASR дало тот же результат, как для субтитров
  - Существенное влияние на моделирование оказывают стоп-слова и популярные разговорные слова, общие для всех тем

# ЗАДАЧА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

## Результаты тематического моделирования

moviename	relevant tokens	topic_name
Dancer in the Dark	love, good, make, feel, thing, great, girl, ni...	Мелодрама
9 Songs	love, good, make, feel, thing, great, girl, ni...	Мелодрама
Star Wars: Episode IV - A New Hope	sir, back, move, make, good, captain, time, sh...	Драма-приключение
The Fifth Element	sir, back, move, make, good, captain, time, sh...	Драма-приключение
American Beauty	good, talk, time, back, happen, dad, mom, call...	Молодёжная драма
The Simpsons Movie	good, talk, time, back, happen, dad, mom, call...	Молодёжная драма
Forrest Gump	good, guy, man, back, boy, time, ai, make, big...	Комедийная драма
Back to the Future	good, guy, man, back, boy, time, ai, make, big...	Комедийная драма
Kill Bill: Vol. 1	fuck, man, shit, guy, car, good, give, make, b...	Бытовая драма
A History of Violence	fuck, man, shit, guy, car, good, give, make, b...	Бытовая драма
Metropolis	people, time, world, work, thing, year, make, ...	Политический
Good Bye Lenin!	people, time, world, work, thing, year, make, ...	Политический
Beverly Hills Cop	kill, man, call, find, give, police, car, guy,...	Полицейский детектив
Beverly Hills Cop II	kill, man, call, find, give, police, car, guy,...	Полицейский детектив
Raiders of the Lost Ark	kill, man, die, find, back, dead, fight, make,...	Приключение
Indiana Jones and the Temple of Doom	kill, man, die, find, back, dead, fight, make,...	Приключение
Pirates of the Caribbean: The Curse of the Bla...	man, good, make, give, leave, time, mother, wo...	Мелодрама
Mifunes sidste sang	man, good, make, give, leave, time, mother, wo...	Мелодрама



## ЗАДАЧА ПРЕДСКАЗАНИЯ ПРИЗНАКА *keywords*

---

### Используемые алгоритмы

➤ Классификатор *OneVsRest + SGD*

➤ Функция потерь *log-loss*

➤ *TF-IDF* векторизация текста

➤ Ограничение датасета для обучения

- не более 3500 слов в субтитрах (итого 968/194 текстов для обучения/валидации)
- встречаемость слов в словаре субтитров не менее 3 раз
- встречаемость ключевых слов не чаще 2 раз в словаре ключевых слов
- не менее 15 ключевых слов для каждого фильма
- проверка на ASR: 20 фильмов

## ЗАДАЧА ПРЕДСКАЗАНИЯ ПРИЗНАКА *keywords*. ИСХОДНЫЕ ДАННЫЕ

Признак <i>keywords</i>	Текст субтитров, предобработанный	Название фильма
<b>imdb_id</b>		
76759 ['android', 'spaceopera', 'rebellion', 'planet...']	['hear', 'shut', 'main', 'reactor', 'destroy', ...]	Star Wars: Episode IV - A New Hope
109830 ['basedonnovel', 'love', 'friendship', 'flashb...']	['hello', 'name', 'forrest', 'forrest', 'gump'...]	Forrest Gump
169547 ['nudity', 'femalenudity', 'malenudity', 'comi...']	['need', 'father', 'role', 'model', 'horny', '...']	American Beauty
168629 ['murder', 'friendship', 'smalltown', 'robbery...']	['sweat', 'know', 'im', 'excite', 'though', 's...']	Dancer in the Dark
119116 ['love', 'alien', 'newyorkcity', 'future', 'sh...']	['come', 'come', 'please', 'aziz', 'aziz', 'az...']	The Fifth Element
...	...	...

## ЗАДАЧА ПРЕДСКАЗАНИЯ ПРИЗНАКА *keywords*

---



### Пример работы классификатора



y\_train: murder | basedonnovel | prison | paris | torture | france | church | gypsy | disfigurement | cathedral | bell | hunchback | torment | dignity | bellringing

Предсказание (субтитры): basedonnovel | musical | paris | thcentury | obsession | religion | orphan | dance | sword | judge | fool | witchhunt | mockery | cathedral | bell

Предсказание (текст ASR): basedonnovel | musical | paris | thcentury | obsession | religion | gallery | dance | sword | judge | fool | witchhunt | insanity | cathedral | jewelry

### Результат



Субтитры: F-score = 0.07

Тексты от ASR: F-score = 0.05

### Попытка предсказания с помощью RNN-LSTM: F-score ~ 0.02

## ВЫВОДЫ

---



При извлечении именованных сущностей качество транскрибации играет важную роль. Каждый вид ошибок, допускаемых ASR, снижает качество NER



При тематическом моделировании характер данных имеет первостепенное значение. Влияние ASR на тематическое моделирование выражено слабо.



В случае предсказания смысла текста внимание должно быть уделено релевантности признаков, используемых как смысловые. Использование текстов ASR возможно, но в ущерб точности



Существенный фактор — высокое качество современных ASR моделей

При решении реальной задачи NER или Topic modelling требуется всесторонний подход



БАЙДАКОВ ИЛЬЯ

