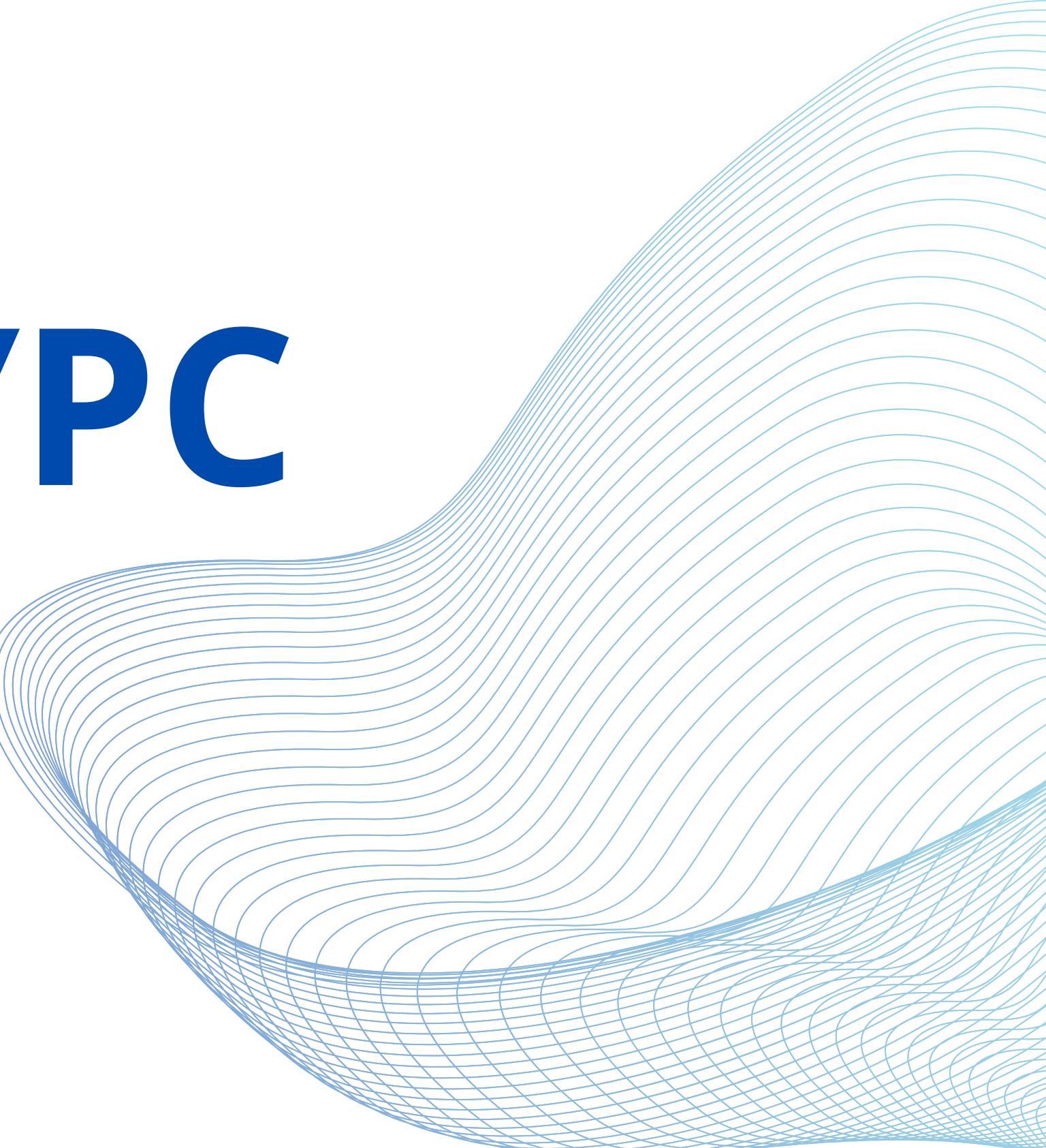




КИНОДИСКУРС АНАЛИЗ N-ГРАММ

Степанишина Елена, 1 курс (Магистратура)



АКТУАЛЬНОСТЬ

- Исследование кинодискурса представляет собой источник информации об обществе.
- Применение современных инструментов для анализа текстовых данных.



КИНОДИСКУРС



1

Объект

Непрофессиональные
русскоязычные интернет-рецензии.

2

Предмет

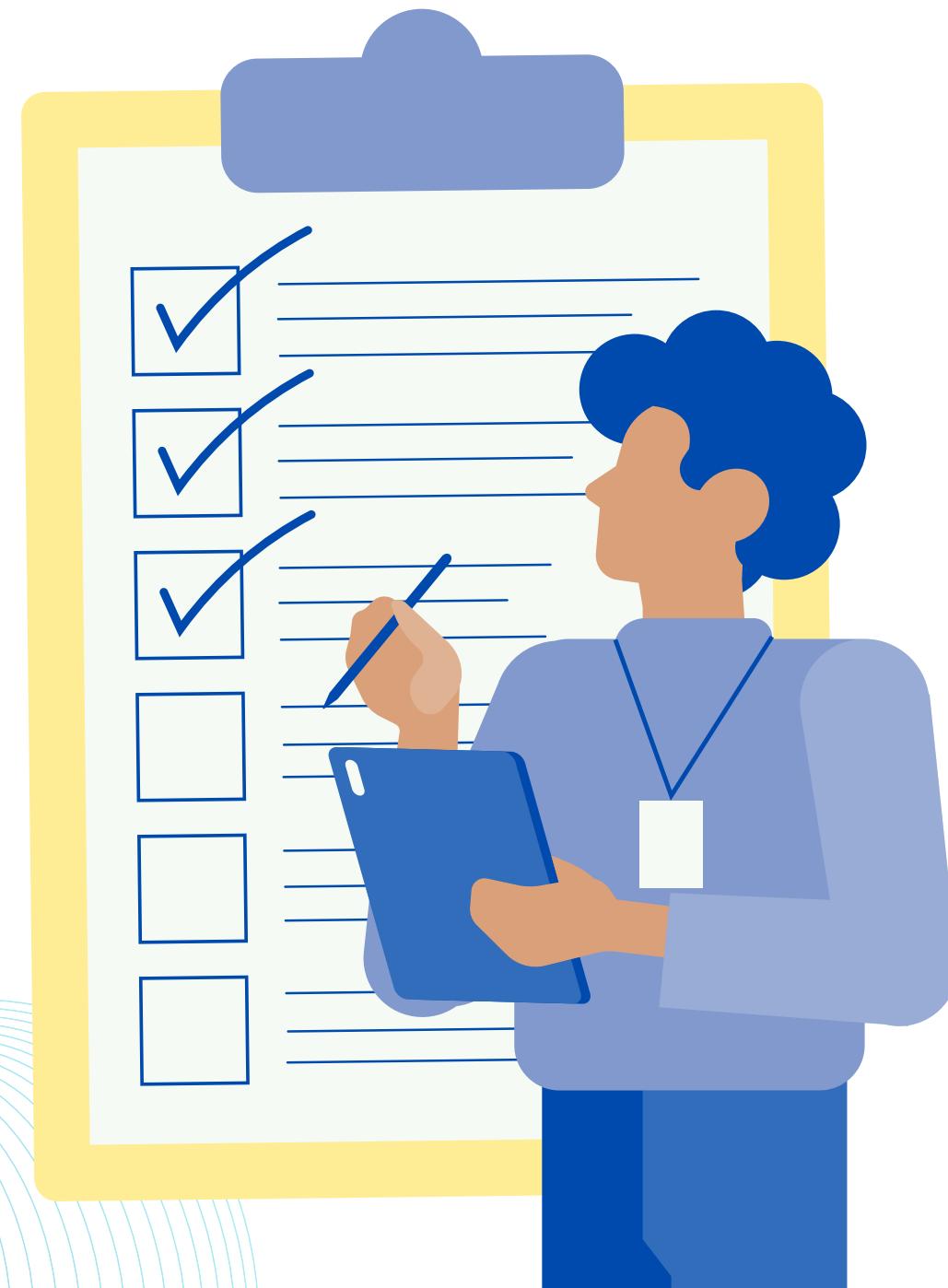
Структурно-смысловая организация
русскоязычных интернет-рецензий.

3

Цель

Исследовать особенности кинодискурса на
основе русскоязычных интернет-рецензий

ЗАДАЧИ



1

Изучить основные
понятия дискурса;

2

Изучить методы
анализа текста и
дискурса;

3

Провести исследование.

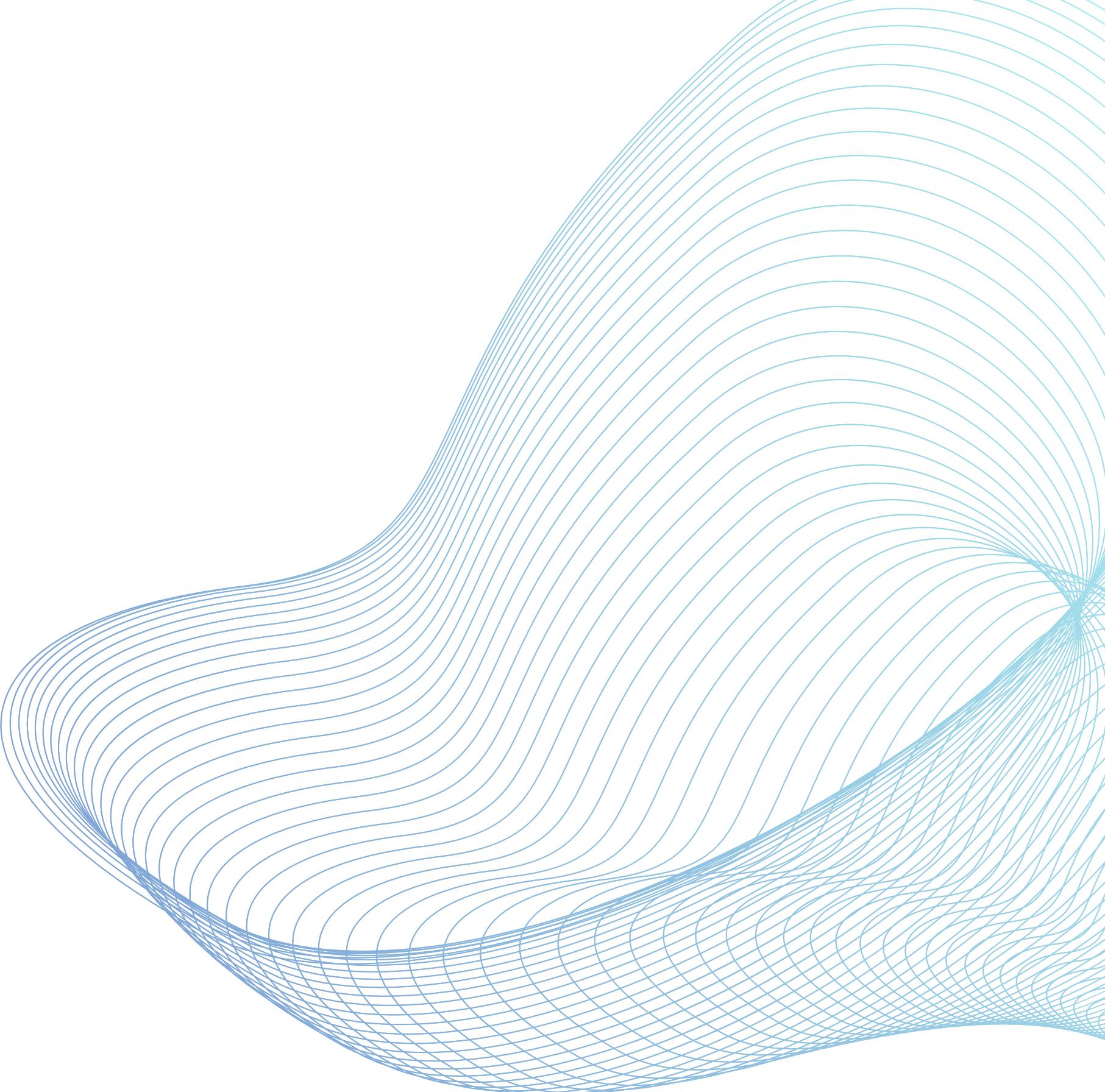
ДИСКУРС - ?

- «Речь, опрокинутая в жизнь»
[Арутюнова];
- Отражение языковой и
социокультурной реальности
[Степанов];
- Сложное единство языковой формы,
значения и действия, которое могло
бы быть наилучшим образом
охарактеризовано с помощью
понятия коммуникативного события
или коммуникативного акта [ван
Дейк].



N-ГРАММЫ – ?

- Последовательность из N-элементов



ИССЛЕДОВАНИЕ

Провести исследование
непрофессиональных
русскоязычных рецензий с
сайта “Кинопоиск”



1 Сбор данных

Собрать корпус интернет-рецензий

2 Предобработка

- Сформулировать теоретическую модель исследования
- Обработать текстовые данные

3 Тематические исследования

- Выявлены наиболее часто встречающиеся словосочетания.
- Построена сеть взаимосвязей

СБОР И ОБРАБОТКА ДАННЫХ

- Данные собраны с сайта “Кинопоиск”;
- Всего в корпусе 4к отзывов разного объема
- Фильмы российского производства;
- Текстовые данные были предобработаны: лемматизация, удаление стоп-слов и разделение на биграммы.



НЕОБРАБОТАННЫЙ ТЕКСТ

review

С большим удовольствием пишу рецензию
на фильм...

Довелось на днях посмотреть новую
экранизацию ...

Настоящее кино - это не просто фильм.
Настояще...

У фильма есть два явных достоинства: это
неопи...

Как увидела трейлер, сразу поняла, что
пойду н...

bigrams_no_lem

[(С, большим), (большим,
удовольствием), (удов...

[(Довелось, на), (на, днях), (днях,
посмотреть...

[(Настоящее, кино), (кино, -), (-, это),
(это,...

[(У, фильма), (фильма, есть), (есть, два),
(дв...

[(Как, увидела), (увидела, трейлер,),
(трейлер...

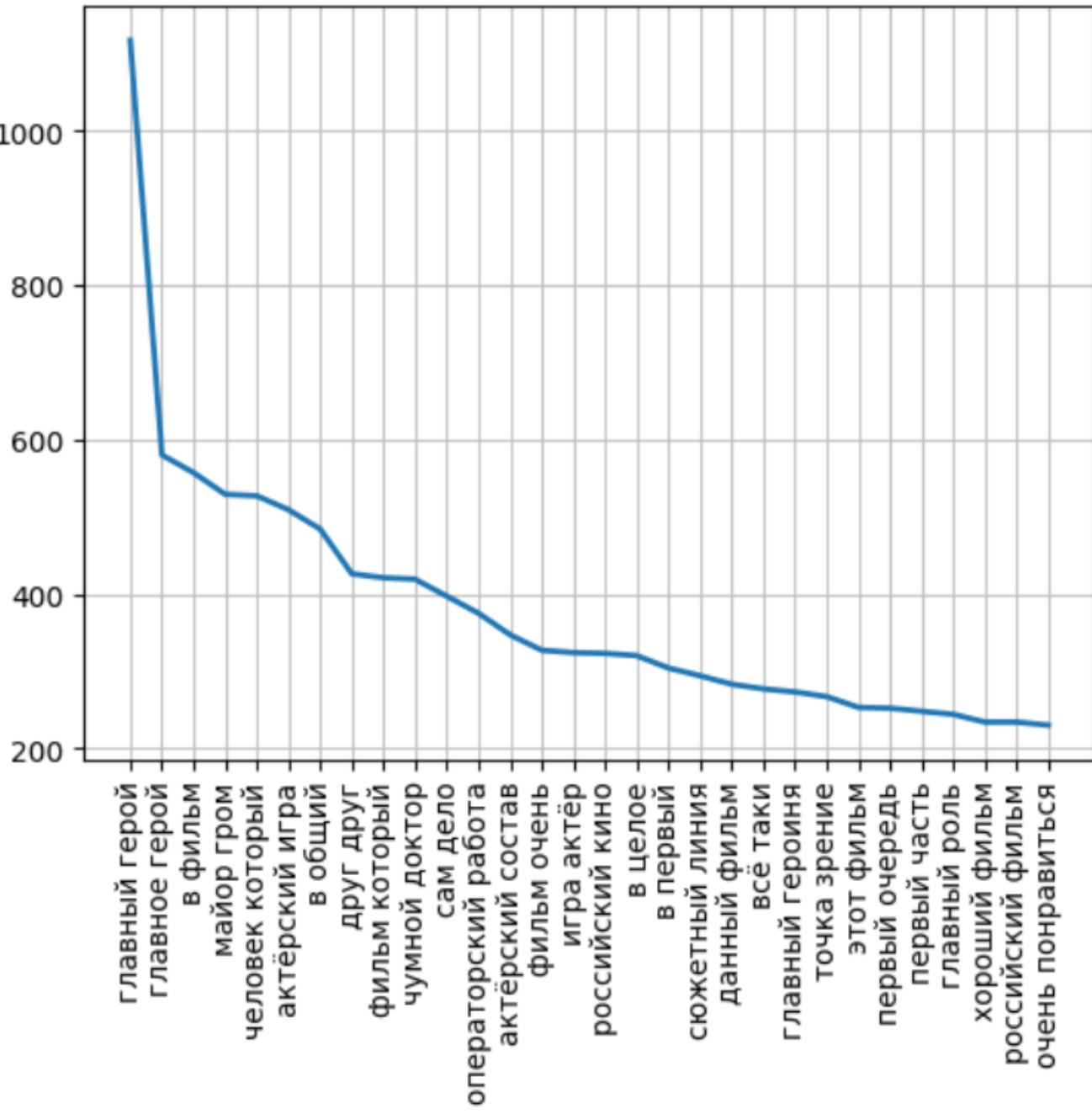
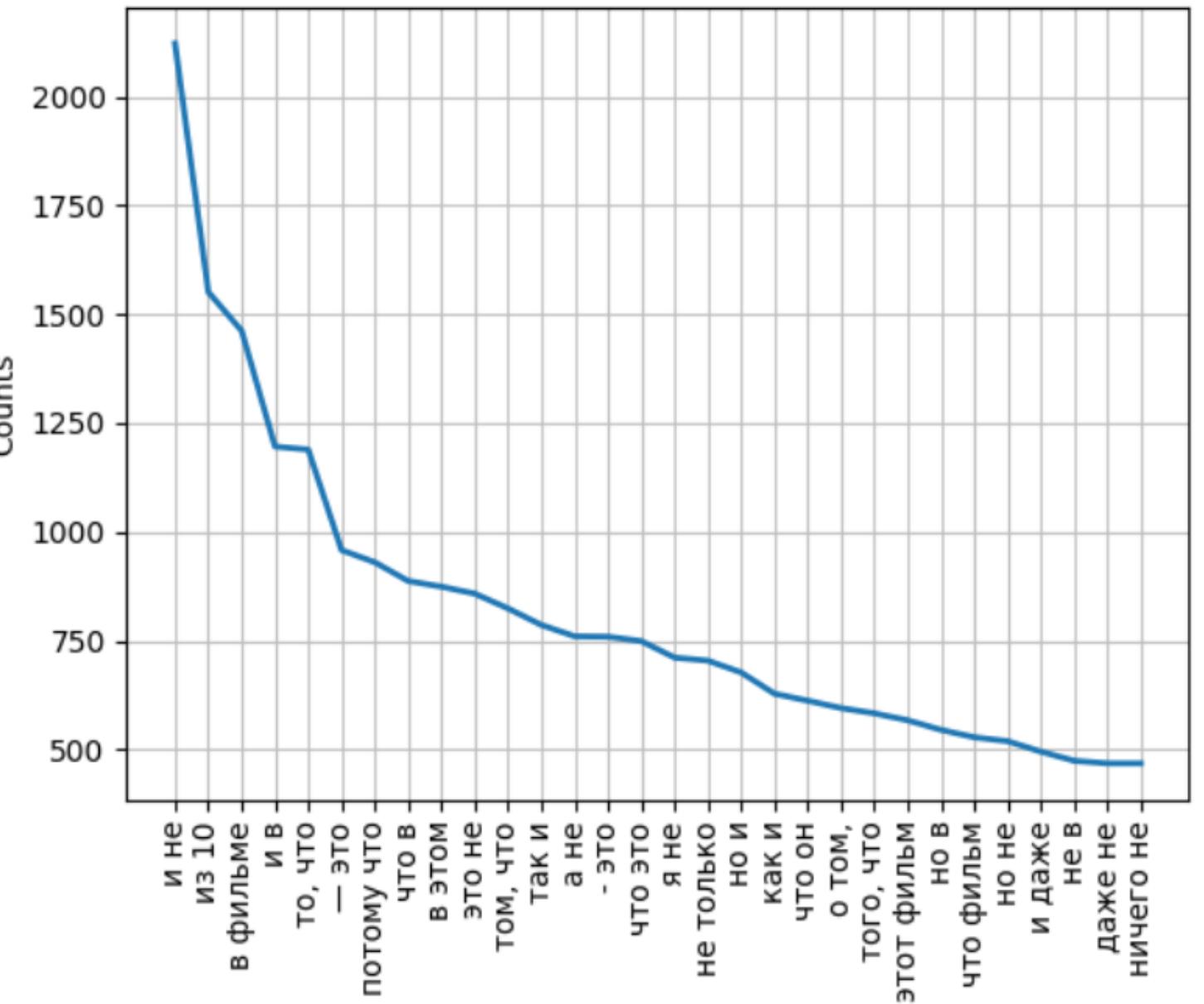
ОБРАБОТАННЫЙ ТЕКСТ

text_lem	bigrams_lem
с больши́й удовольстви́е писа́ть рецензия́	[(с, больши́й), (больши́й, удовольстви́е),
фильм с...	(удово...)
довестись день посмотре́ть новый	[(довестись, день), (день, посмотре́ть),
экранизаци́я ре...	(посмо...)
настоящи́й кино просто фильм настоящий	[(настоящи́й, кино), (кино, просто),
кино реф...	(просто, ф...
у фильма явный достоинство неописуемый	[(у, фильма), (фильм, явный), (явный,
красота ...	достоинст...
как увидеть трейлер сразу понять пойти	[(как, увидеть), (увидеть, трейлер),
фильм б...	(трейлер,...)



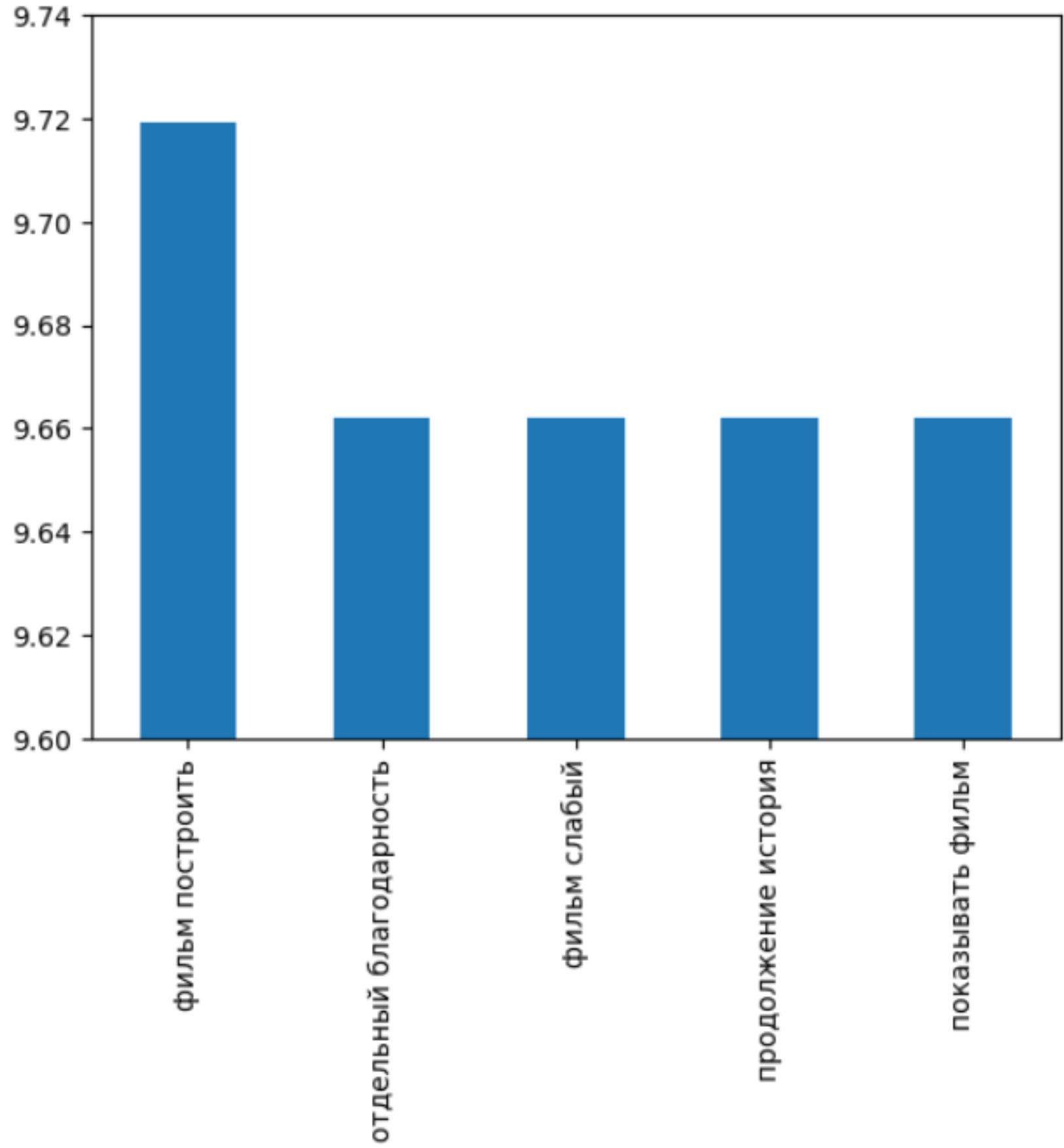
Без обработки

['С большим',
'большим удовольствием',
'удовольствием пишу',
'пишу рецензию',
'рецензию на',
'на фильм-сказку',
'фильм-сказку «По',
'«По щучьему',
'щучьему велению»!',
'велению»! Сюжет',
'Сюжет целиком',
'целиком выдержан',
'выдержан в',
'в стиле',
'стиле русских',
'русских народных',
'народных сказок',
'сказок сборника',
'сборника Александра',
'Александра Николаевича',
'Николаевича Афанасьева.',



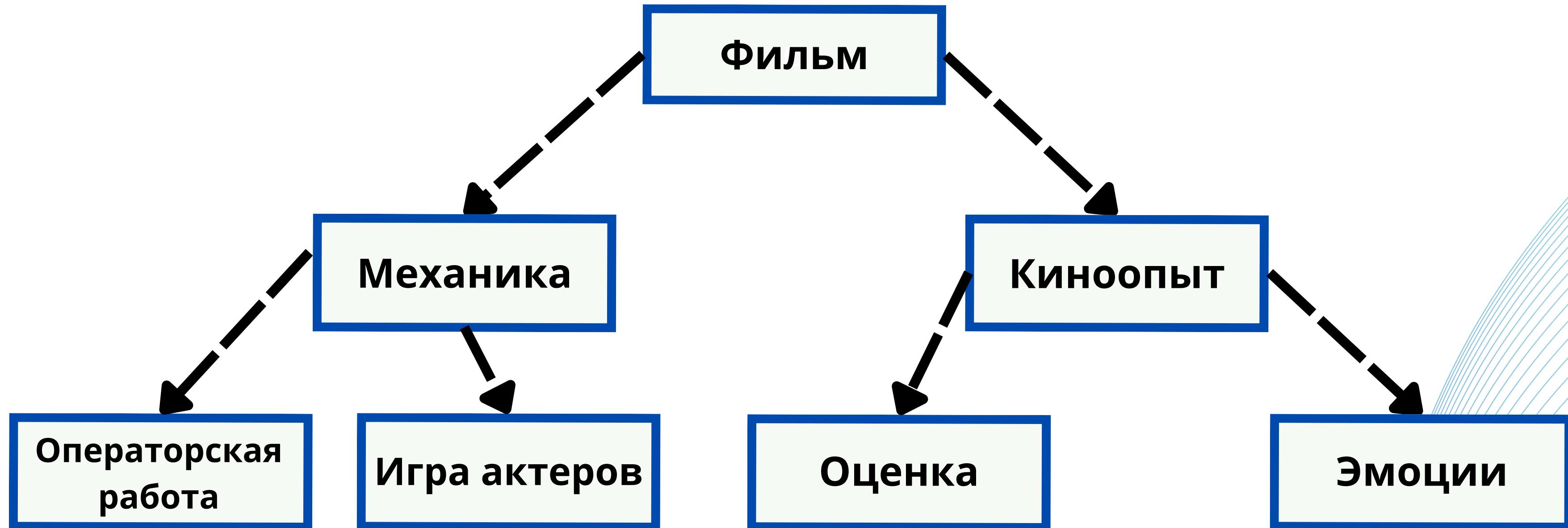
Предобработка

ЧАСТОТА ВСТРЕЧАЕМОСТИ



- **Фильм построить (9.72)**
- **Отдельный благодарность (9.66)**
- **Фильм слабый (9.66)**
- **Продолжение история (9.66)**
- **Показывать фильм (9.66)**

СЕМАНТИЧЕСКОЕ ДЕРЕВО



LIKEHOOD RATIO

“Фильм”

Понравиться
489.294

Снятый
367.82

Получиться
318.41

Катастрофа
268.77

“Актёр”

Играть
589.005

Справиться
128.91

Подобрать
103.05

Отлично
49.97

“Режиссёр”

Сценарист
536.014

Дебютант
69.74

Постановщик
68.76

Продюсер
56.92

“Эмоция”

Просмотр
54.53

Переживание
41.89

Переигрывание
20.56

Бревно
16.01

“Оценка”

Ставить
83.12

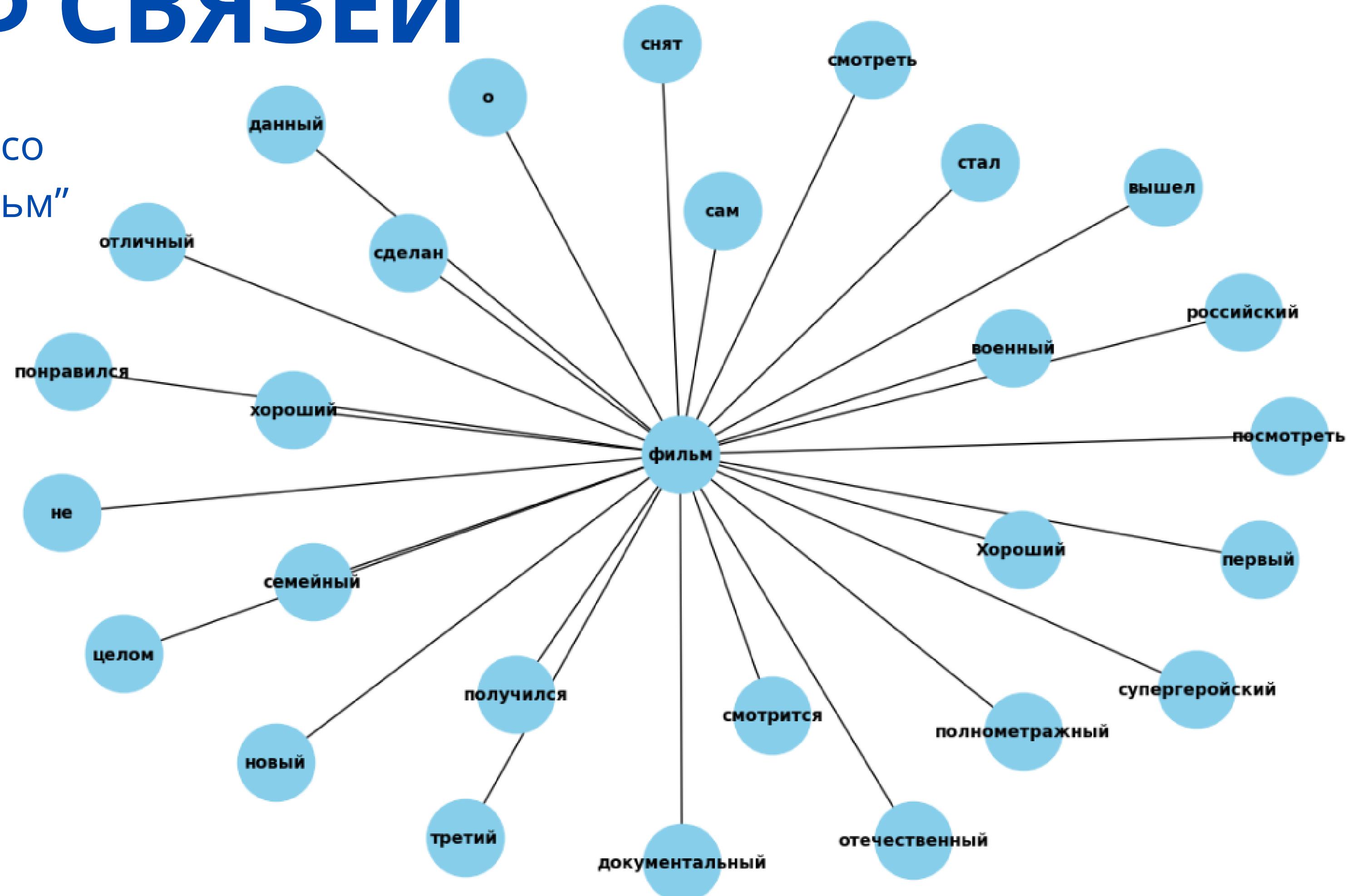
Абстрагироваться
15.56

Невыдержаный
15.114

Субъективный
13.410

ГРАФ СВЯЗЕЙ

На примере
коллокаций со
словом “фильм”



ВЫВОДЫ

01

Наиболее часто встречающиеся словосочетания – словосочетания со словом “фильм”.

02

Категория описания операторской работы имеет больше часто встречающихся позитивно коннотированных словосочетаний;

03

Категория “Киноопыт” работы имеет больше часто встречающихся негативно коннотированных словосочетаний;

ОБЩИЕ ВЫВОДЫ

Перспективы исследования:

- Создание текстового корпуса на основе непрофессиональных интернет-рецензий;
- Тематическое модерование и глубинное тематическое моделирование;
- Увеличение датасета для дальнейшего изучения.

01

N-граммы позволяют исследовать тематическую направленность текста.

02

Алгоритм предобработки текста существенно влияет на результат исследования.

03

Графы также могут быть инструментом для визуализации текстовых исследований.

**СПАСИБО ЗА
ВНИМАНИЕ!**

