

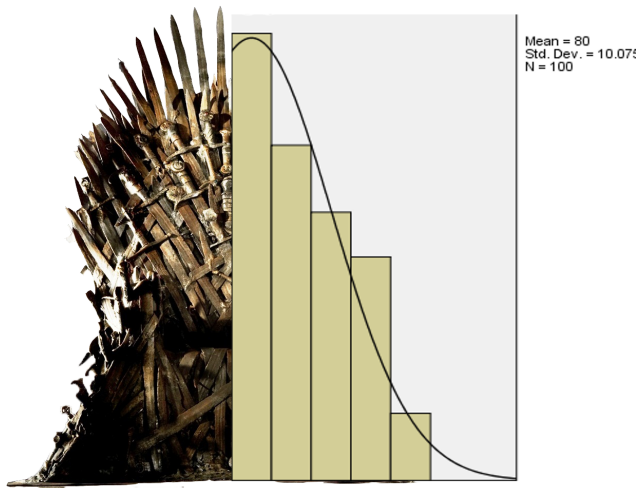
# Дисперсионный анализ и Железный трон

Борис Орехов

НИУ Высшая школа экономики

*nevmenandr@gmail.com*

5 июля 2016



## 1 Вводные замечания

- Правда ли все дело в дисперсионном анализе?
- Почему именно Дж. Мартин?
- «Игра престолов» и статистика
- Как встретиться литературоведению и анализу данных?

## 2 Задачи, материал, методы

## 1 Вводные замечания

- Правда ли все дело в дисперсионном анализе?
- Почему именно Дж. Мартин?
- «Игра престолов» и статистика
- Как встретиться литературоведению и анализу данных?

## 2 Задачи, материал, методы

# Дисперсионный анализ?

Довольно много людей, записавшихся на этот тьюториал (и те, кто в итоге сейчас здесь, и те, кого мы не смогли пригласить), говорили, что им интересно именно применение дисперсионного анализа к художественному тексту.

Должен признаться: упоминание именно дисперсионного анализа в тексте — все-таки рекламный трюк. На самом деле речь будет идти об анализе нарратива любым разумным статистическим способом. Если нам для этого понадобятся графы или векторные модели, которыми занимаются наши коллеги на других тьюториалах, то нас это не остановит.

# Чем хороши книги Дж. Р. Р. Мартина?

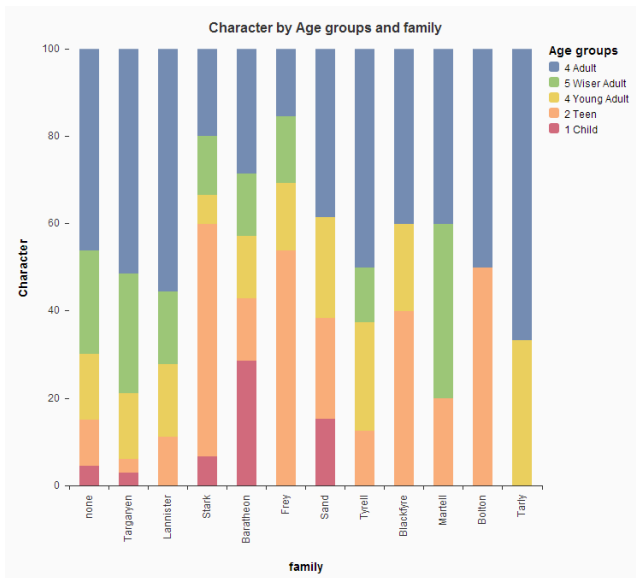
- Это **большие** книги → большая выборка. Общий объем: 1.782.723 слова.
- Цикл не дописан и (как в естественных науках или как с идеями Соссюра о реконструкции) можно будет получить независимое **опровержение или подтверждение** гипотезы, когда выйдут новые книги.
- Благодаря сериалу книги **популярны**, и многие неплохо представляют, что там происходит.

Люди обожают статистику. Вероятно, неполный список статистических этюдов об «Игре престолов»:

- [Guy Yachdav et al, 2016]
- [Азат Хузияхметов, 2016]
- [Allen Downey, 2015]
- [Unknown, 2015]
- [Ember Twygg, 2015]
- [Éric Ledu, 2014]

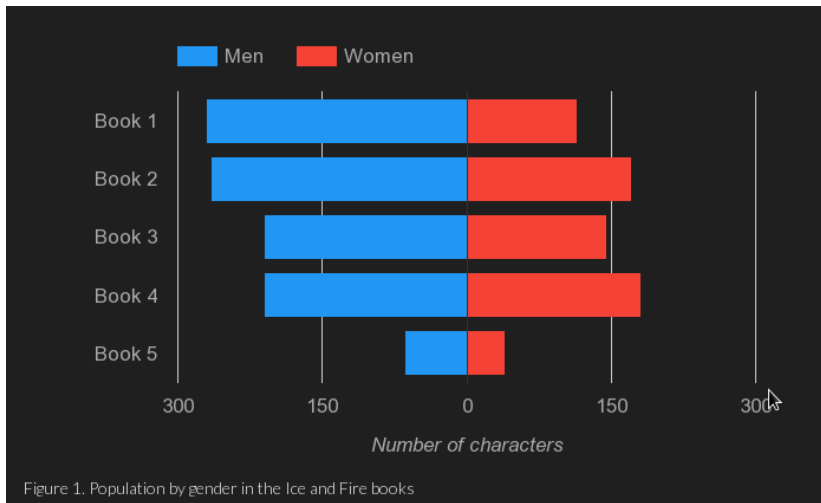
Отдельно укажем аналитику с использованием машинного обучения:

[Guy Yachdav et al, 2016a], [Maxime, 2016], [Erik Germani, 2016]



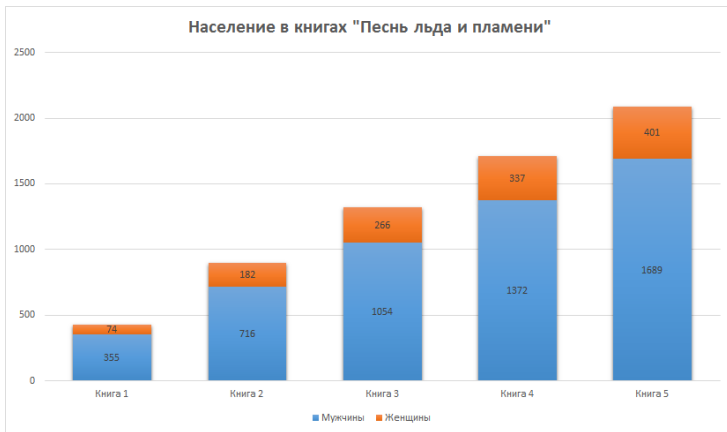
[Éric Ledu, 2014]





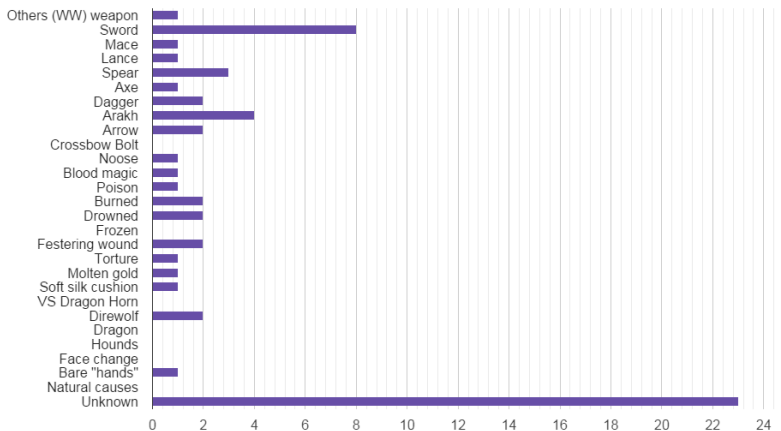
[Guy Yachdav et al, 2016]

Ср. данные из другого источника:



[Азат Хузияхметов, 2016]

Weapon of Death AGOT- By Twygg



Number

<https://statisticsofireandfire.wordpress.com/>

[Ember Twygg, 2015]

Чем это плохо?

Ничем. Это хорошо 😊.

Но, кажется, таким образом мы упускаем литературоведческую фактуру. Получается слишком **distant** reading.

При таком подходе

- на месте «Песни льда и огня» могла быть какая угодно книга;
- данные часто (хотя и не всегда) берутся не из собственно текста, а из энциклопедии вымышленного мира  
<http://awoiaf.westeros.org/>.

## Example (Feature list в [Guy Yachdav et al, 2016a])

House to which a character belongs

Social group to which a character belongs

Male or female

Character's appearance in the book (все книги по отдельности)

Number of dead characters to whom a character is related

Whether the character is married

...

Что не так с этим списком?

# Что же не так?

Эти признаки не отражают **поэтику** произведения.

Перечисленные признаки, разумеется, не случайны. Они взяты из практики применения анализа данных в жизни. Для **человека** с точки зрения статистики признаки принадлежности *семье, социальной группе, пол, состояние в браке* — осмысленны.

Для **персонажа** это не обязательно так.

Здесь мы видим отражение отношений **в вымышленном мире**, а не отражение поэтики. Из-за этого исследователь, который подошел к тексту с таких позиций, выглядит как вульгарный социолог, потому что не видит текста и не обращает внимание на то, как он построен.

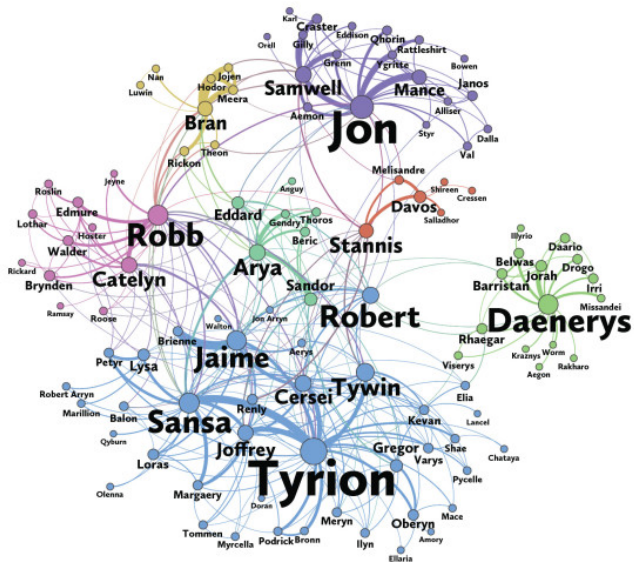
Конечно, устройство вымышленного мира тоже отражает поэтику, но косвенно

# Что же не так?

## Положительный пример

[Andrew Beveridge and Jie Shan, 2016] За основу своего исследования авторы работы взяли электронную версию книги «Буря мечей», так как, по их словам, именно в этой книге повествование достигло достаточного развития. В качестве «связи» (фактически ребром графа) между персонажами выступало упоминание имен друг друга в диапазоне 15 слов. Таким образом, связь героев друг с другом вовсе не означает, что они друзья, — скорее, это просто говорит о том, что они взаимодействуют, говорят друг о друге или упоминаются вместе. В результате, они получили граф, отражающий значимость персонажей.

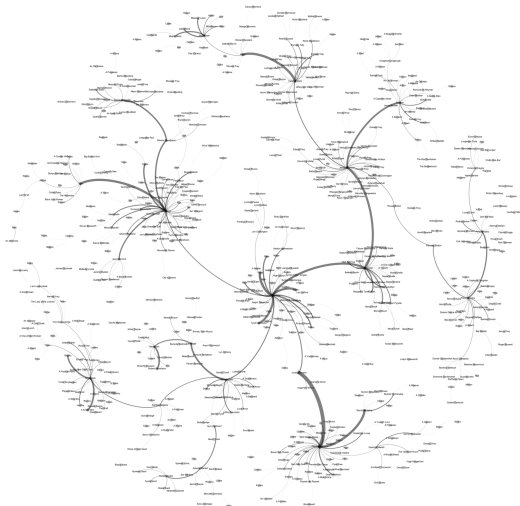
Взаимодействие в тексте — это уже поэтика.



[Andrew Beveridge and Jie Shan, 2016]



Ср. другой социальный граф, отражающий диалоги персонажей:



[Азат Хузияхметов, 2016a]

# Чем будем заниматься мы?

## Литературоведение и анализ данных

Основная задача нашей работы в том, чтобы литературоведение и анализ данных встретились, и их объединение было **не механическим**. Мы должны найти в тексте такие параметры для анализа, которые бы

- 1 соотносились с поэтикой произведения (а не просто были бы привычны для аналитиков, использующих статистику),
- 2 извлекались из текста (а не из вторичных источников),
- 3 могли получить литературоведческую интерпретацию (почти то же, что первое, но не совсем)
- 4 были статистически верно обработаны (само собой)

## 1 Вводные замечания

- Правда ли все дело в дисперсионном анализе?
- Почему именно Дж. Мартин?
- «Игра престолов» и статистика
- Как встретиться литературоведению и анализу данных?

## 2 Задачи, материал, методы

# Что будем считать мы?

## Исследовательские сюжеты

- 1 Цвета
- 2 Субъекты повествования
- 3 Персонажи
- 4 Топонимы

Сначала определимся с тем, что эти направления анализа не случайны и информативны именно для поэтики «Песни льда и пламени».

## Дж. Р. Р. Мартин в интервью:

Фэнтези — серебро и багрянец, индиго и лазурь, обсидиан с прожилками золота и лазурита. А реальность — это фанера и пластик, окрашенные в грязно-коричневые и желтовато-зеленые тона. Фэнтези имеет вкус хабанеры и меда, корицы и гвоздики, превосходного красного мяса и вина, сладкого, словно лето. Реальность — это бобы и тофу, а в конечном итоге — прах.

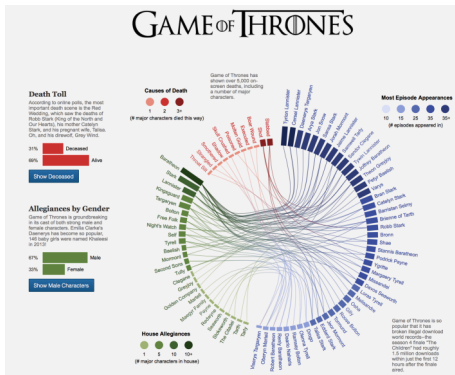
- Цвета навязчиво встречаются в каждом описании:
  - Jon's eyes were a **grey** so dark they seemed almost **black** (GoT, Bran I)
  - She wore a garland of **pale blue** roses, and her eyes wept blood (GoT, Eddard XIII)
- Многие значимые имена и названия включают цвета: Серый Ветер, Чёрный замок.

# Субъекты повествования (POV)

- Важный конструктивный элемент текста.



- Важная составляющая поэтики нарративного текста.



<http://thronesviz.github.io/>

- Способ организации художественного пространства.
- Правда ли это отличается от принадлежности дому, которую можно взять из энциклопедии?





## Компьютерная лингвистика и анализ данных!

Репозиторий с кодом и данными:

[https://github.com/nevmenandr/Martin\\_tutorial](https://github.com/nevmenandr/Martin_tutorial)

# Чего хотят литературоведы?

Литературоведы интересуются сложными и плохоформализуемыми категориями: смыслами, сложноорганизованными структурами текста и т.д.

Но мы можем подобраться к ним только через атомарные частицы низкого уровня: слова.

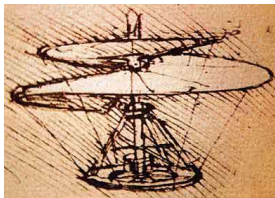
# Через слова и «напрямую»

- 
- Вот как, - сказал Чапаев. - Какие вы интересные знаете слова.
- О, до вас в этой области мне далеко. Кстати, не объясните ли вы, что такое зарука?
- Как? - наморщился Чапаев.
- Зарука, - повторил я.
- Где это вы услышали?
- Если я не ошибаюсь, вы сами только что говорили с трибуны о своей командирской заруке.
- А, - улыбнулся Чапаев, - вот вы о чем. Знаете, Петр, когда приходится говорить с массой, совершенно не важно, понимаешь ли сам произносимые слова. Важно, чтобы их понимали другие. Нужно просто отразить ожидания толпы. Некоторые достигают этого, изучая язык, на котором говорит масса, а я предпочитаю действовать напрямую.

В. Пелевин «Чапаев и Пустота»

# Почему компьютерная лингвистика?

Компьютерная лингвистика — это вертолёт.



Нужно ли пытаться сконструировать свой, если можно уже летать на существующем?







Guy Yachdav et al (2016)

Song of Ice and Data

<https://got.show/>



Allen Downey (2015)

Bayesian survival analysis for «Game of Thrones»

<http://alldowney.blogspot.ru/2015/03/bayesian-survival-analysis-for-game-of.html>



Éric Ledu (2014)

Data Geek III - Analyzing Games Of Thrones Data for the GoT Challenge

<http://scn.sap.com/community/lumira/blog/2014/12/11/data-geek-iii-analyzing-games-of-thrones-data-for-the-got-challenge>



Guy Yachdav et al (2016a)

How do we predict likelihood of death?


<https://got.show/machine-learning-algorithm-predicts-death-game-of-thrones>





Maxime (2016)


Summer is Coming - Game of Thrones Analytics


<http://www.dataiku.com/blog/2016/04/23/got-analytics.html>

 Beveridge, Shan (2016)  
Network of Thrones  
*Math Horizons* Vol. 23, No. 4 (April 2016), pp. 18-22

 Unknown (2015)  
(Spoilers All) I made a word count of the whole ASOIAF books  
[https://www.reddit.com/r/asoiaf/comments/3froem/spoilers\\_all\\_i\\_made\\_a\\_word\\_count\\_of\\_the\\_whole](https://www.reddit.com/r/asoiaf/comments/3froem/spoilers_all_i_made_a_word_count_of_the_whole)

 Ember Twygg (2015)  
Statistics of Ice and Fire  
<https://statisticsoficeandfire.wordpress.com/>

 Азат Хузияхметов (2016)  
Игра Престолов в числах  
<https://geektimes.ru/post/275466/>

 Азат Хузияхметов (2016а)  
Теория графов в Игре Престолов  
<https://habrahabr.ru/post/302936/>



Erik Germani (2016)

A Study of Ice and Fire

<http://atseajournal.com/asoiaf/>



That's all Folks!