

III Московско-тартуская  
школа  
4-7 октября 2018

# ПЛАН СОДЕР- ЖАНИЯ

Тьютор: Борис Орехов  
Ассистент: Анна Зуева  
Консультант: Артем Шеля  
Студенты: Юлия Семенова  
Анна Крюкова  
Евгения Черноусова  
Ольга Донина  
Екатерина Колеватова  
Виктория Жданова

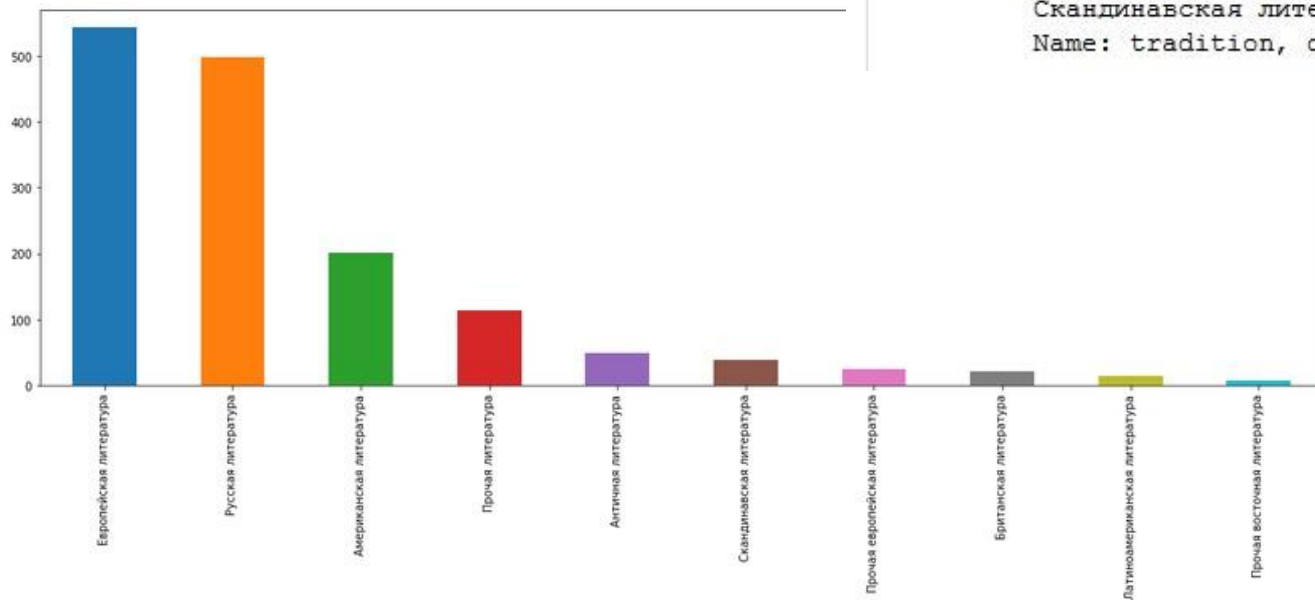
<https://github.com/nevmenandr/brief-content>

# Что у нас было?

1510 пересказов

русской и зарубежной прозы

Out[6]: <matplotlib.axes.\_subplots.AxesSubplot at 0x7fdc44ed3240>



```
In [15]: df.groupby('tradition')['tradition'].count()
```

```
Out[15]: tradition
Американская литература      201
Античная литература         49
Британская литература        21
Европейская литература      543
Латиноамериканская литература  14
Прочая восточная литература   7
Прочая европейская литература  24
Прочая литература           113
Русская литература           499
Скандинавская литература      39
Name: tradition, dtype: int64
```

# Что мы знали о текстах?

Имя автора

Название произведения

Год создания

Годы жизни автора

Имя/nick пересказчика

Традиция, страна

Время чтения текста

Время чтения пересказа

шведская

Мельник — колдун, обманщик и сват	Аблесимов	1742–1783	Читается за 6 минут	40 мин				Русская лите
Клоун Як	Бергман	1883–1931	Читается за 10 минут					Скандинавска
Мариамна	Лагерквист	1891–1974	Читается за 9 минут	1,5 ч	Б. А. Ерхов	erkhov		Скандинавска

# Стилеметрия

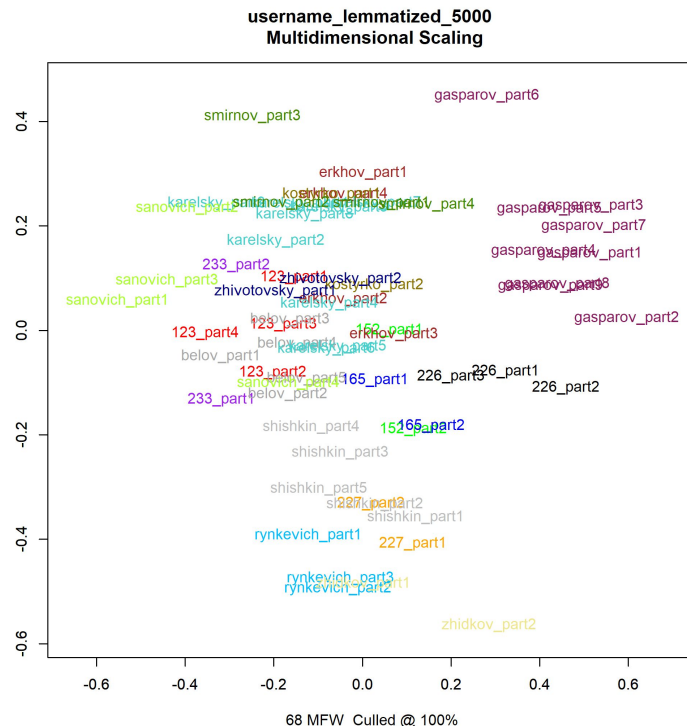
Stylo

Склеили тексты одного  
пересказчика по 5к слов и по 10к

4 корпуса:

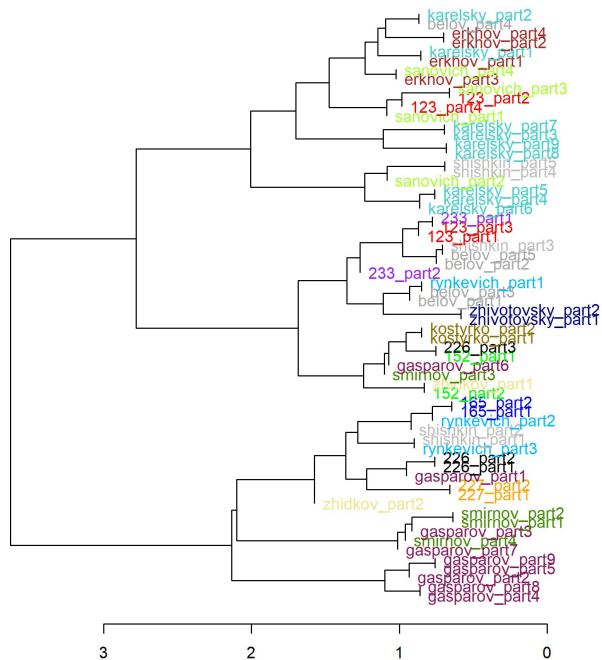
- username\_lemmatized\_5000
- username\_lemmatized\_10000
- username\_tokenized\_5000
- username\_tokenized\_10000

Склеить авторов оригинальных  
текстов не получилось:  
не набралось объема



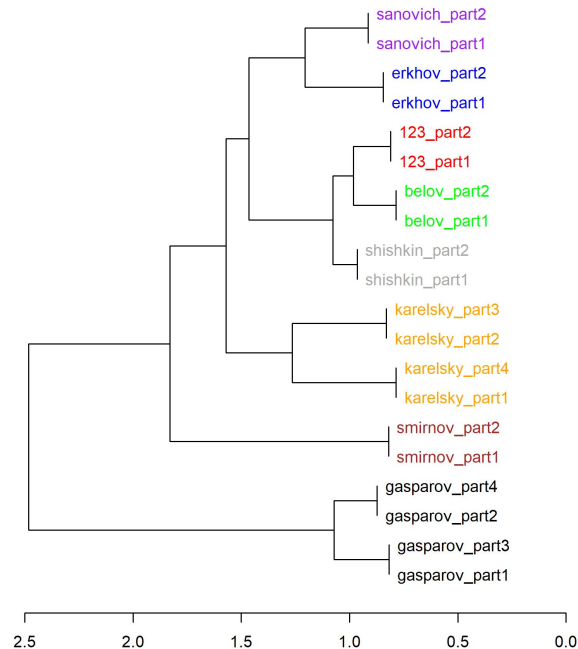
# Стилеметрия: словоформы

username\_tokenized\_5000  
Cluster Analysis



53 MFW Culled @ 100%  
Classic Delta distance

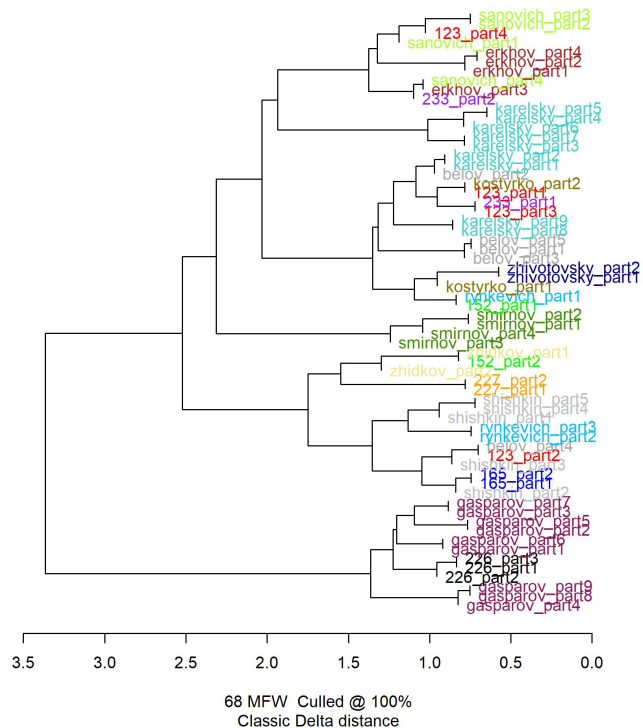
username\_tokenized\_10000  
Cluster Analysis



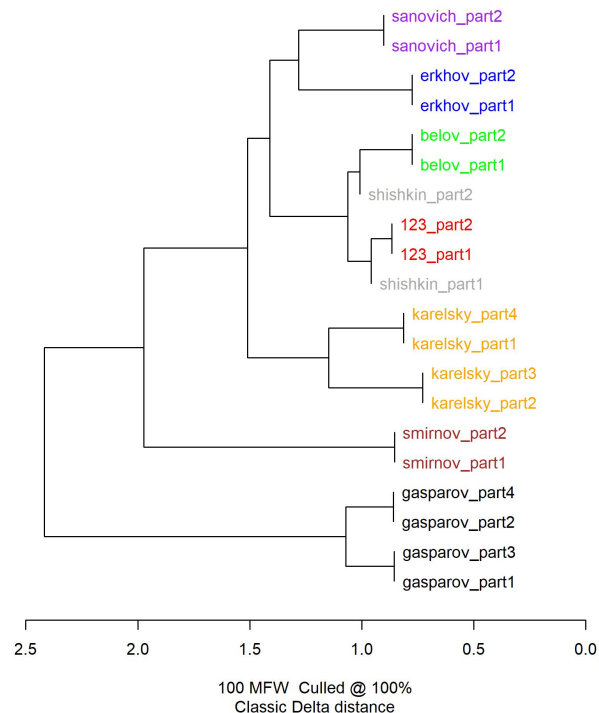
100 MFW Culled @ 100%  
Classic Delta distance

# Стилеметрия: леммы

username\_lemmatized\_5000  
Cluster Analysis



username\_lemmatized\_10000  
Cluster Analysis



# Стилеметрия: результаты

- Авторский сигнал пересказчика хорошо себя проявляет в пересказе.
- Мы не знаем, есть ли авторский сигнал оригинального автора, потому что недостаточен объем.
- Вариативность словесной репрезентации (на нелемматизированных текстах 53 общих слова).

# Текстовый классификатор

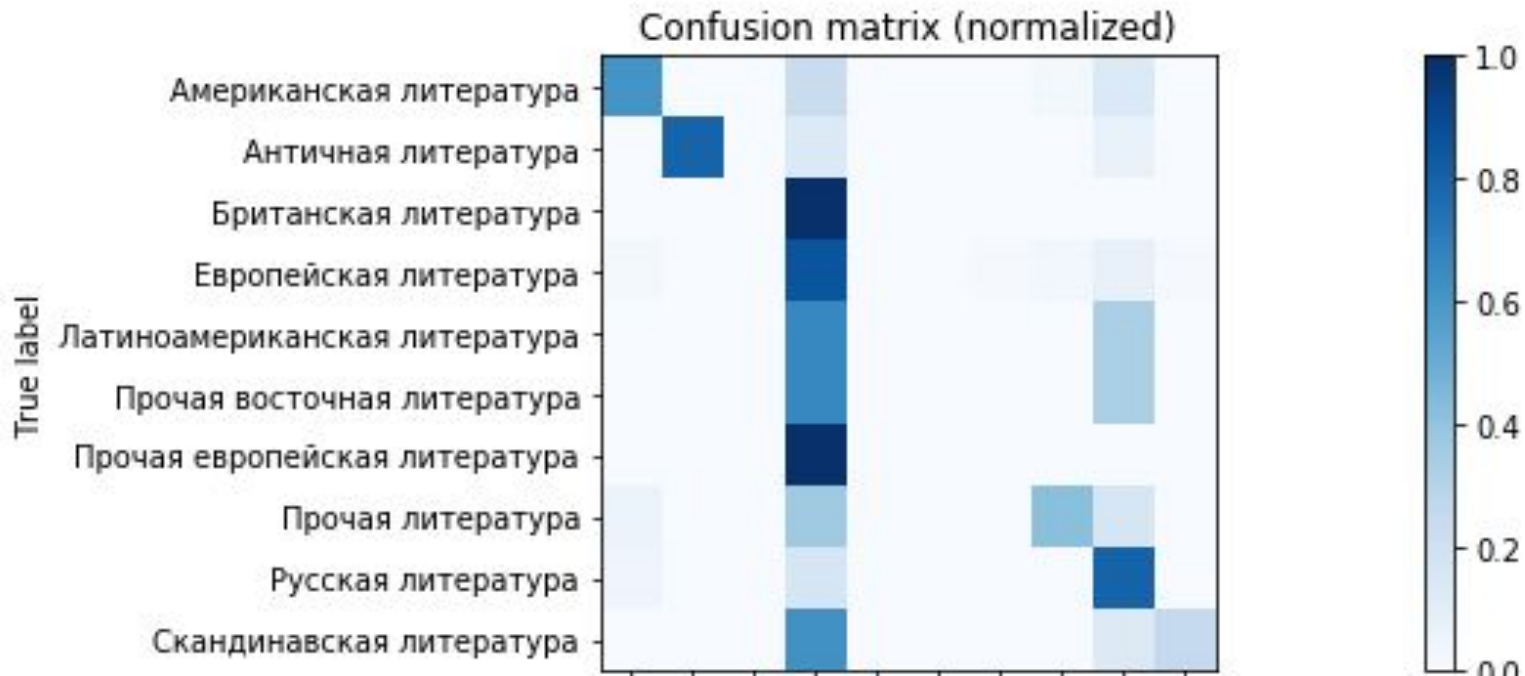
## Задача:

построить классификатор по литературным традициям и проверить, насколько хорошо он будет работать.

Bag of words лемматизированный:	0.692
Bag of words нелемматизированный:	0.718
Character N-grams:	0.692
TF-IDF:	0.764



# Текстовый классификатор BOW-lemmatized



# Классифицирующие слова

Американ. л	Античная л	Британская	Европейска	Латиноамер	Пр. восточ	Пр. европейска	Пр. литера	Русская лит	Скандинавска
часть	раб	владимир	лорд	студент	сиавуш	слепой	столица	русский	пастор
доллар	твой	стивен	сэр	глава	словно	ход	чжан	иван	томас
банк	это	роман	толпа	книга	поход	адвокат	бен	москва	пеппи
ферма	хор	мэри	животное	часть	пустыня	тереза	гора	новелла	замок
индеец	зевс	доктор	покидать	пустыня	воин	дядя	правитель	страшный	несколько
штат	сцена	рентс	полк	библиотека	племя	неля	император	россия	вилфред
мост	пусть	орра	зал	мочь	могила	затем	подруга	автор	кристин
тома	комедия	питер	называть	юноша	престол	сумасшедший	государь	мужик	роза
дерево	жизнь	квентин	постепенно	мария	отец	клара	детектив	егор	храм
находиться	море	оказываться	служить	конец	печаль	сообщать	алан	зверь	превращаться

# Скандинавская литература

['нильс', 'граф', 'эрик', 'фру', 'крестьянин', 'датчанин', 'усадьба', 'бог', 'священник', 'молодой']

['**пастор**', 'фру', 'столяр', 'камергер', 'гостиная', 'париж', 'сын', 'муж', 'привидение', 'мать']

['перо', 'лес', 'тролль', 'дед', 'яхта', 'парень', 'море', 'избушка', 'гость', 'отпускать']

['царь', 'иерусалим', 'дворец', 'родственник', 'стараться', 'мальчишка', 'враг', 'жестокость', 'царица', 'отвращение']

['замок', 'башня', 'привратник', 'эрик', 'роза', 'незримо', 'корона', 'сокровище', 'старый', 'песня']

['шарлотта', 'карл', 'адриан', 'карла', '**пастор**', 'анна', '**пасторский**', 'молодой', 'шапочка', 'поэтому']

['томас', 'симон', 'габриэль', 'симона', 'уходить', 'дверь', 'чердак', 'немец', 'трактир', 'пистолет']

['нора', 'фру', 'письмо', 'деньги', 'увольнять', 'италия', 'чудо', 'муж', 'директор', 'жаворонок']

['фритьоф', 'хельга', 'храм', 'бог', 'бел', 'меч', 'бонд', 'сын', 'страна', 'север']

['жизнь', 'сын', 'человек', 'дом', 'мать', 'отец', 'девушка', 'время', 'год', '**пастор**']


# Результаты

- На сайте-источнике не слишком логичное деление на «традиции-страны», возможно, из-за учебных планов.
- Классификатор переобучился на именах собственных: «Владимир» в британской литературе есть только у Беккета («В ожидании Годо») (в русской в 34 текстах).
- Гипотеза про реалии в восточных литературах (пересказчик упускает реалии, поэтому на них не обучается) не подтвердилась — реалии остаются.

# Тематическое моделирование

## Что делает ТМ?

Находит слова, которые часто встречаются в текстах вместе, и делает вывод, что они характеризуют одну тему.

Гипотеза: ТМ + наш датасет = 

## Почему:

В больших текстах много «лишних» слов, которые запутывают алгоритмы, а с краткими пересказами всё должно быть хорошо.

# Два популярных алгоритма: LDA и NMF

Для всей выборки, 5 тем:

LDA	NMF
<p>[человек, время, становится, мочь, жизнь, друг, отец, дом, год, день]</p> <p>[сальери, равик, моцарт, итен, вокульский, рентс, джонатан, калинович, тилем, неля]</p> <p>[ширин, хосров, сиавуш, уинстон, хакон, таэко, кмицица, фритьоф, шах, колдуэлл]</p> <p>[продукт, клиент, компания, уэсли, ваш, юити, рынок, дэвис, потребитель, бизнес]</p> <p>[иван, дон, становится, мальчик, день, решать, идти, время, человек, дом]</p>	<p>[человек, жизнь, герой, время, становится, дом, друг, отец, год, мочь]</p> <p>[ваш, клиент, компания, продукт, пример, бизнес, сотрудник, человек, решение, использовать]</p> <p>[дон, хуан, жуан, донья, карлос, луис, мануэль, кихот, изабелла, родриго]</p> <p>[эдип, фивы, креонт, антигона, этеокла, иокастый, хор, царь, полиник, лайй]</p> <p>[иван, царь, олеся, васильевич, конь, африканович, яга, иванович, жар, волк]</p>

# Что мы увидели: темы = книги

## Американская литература:

['скарлетт', 'эшли', 'ретт', 'мелани', 'батлер', 'тара', 'уикерш', 'атланта', 'лансинг', 'миссис'] -

**Маргарет Митчелл, “Унесённые Ветром”**

['сноупс', 'флем', 'гэвин', 'минк', 'линда', 'джефферсон', 'рэтлиф', 'юла', 'де', 'варнер'] - **Уильям**

**Фолкнер, “Особняк”**

['дороти', 'оз', 'король', 'дровосек', 'волшебница', 'изумрудный', 'железный', 'озм', 'гном', 'биллин'] -

**Лаймен Фрэнк Баум, “Удивительный волшебник из страны Оз”**

['рудольф', 'томас', 'гретхен', 'листок', 'зеленый', 'дверь', 'негр', 'приключение', 'джин', 'уитби'] -

**Ирвин Шоу, «Нищий, Вор»**

['квентин', 'марго', 'джейсон', 'кэдди', 'бенджи', 'компсон', 'джейс', 'ластер', 'миссис', 'бен'] - **Уильям**

**Фолкнер, “Шум и Ярость”**

['гарри', 'зверобой', 'джудит', 'охотник', 'хетти', 'хаттер', 'ковчег', 'чингачгук', 'воин', 'лодка'] - **Джеймс**

**Фенимор Купер, “Зверобой”**

# После удаления имён

Всё работает лучше: 7 топиков, США.

Акцент на бизнес-литературе.

['отец', 'человек', 'дом', 'друг', 'год', 'сын', 'ребенок', 'жизнь', 'становиться', 'жена']

['ваш', 'человек', 'решение', 'пример', 'компания', 'работа', 'сотрудник', 'бизнес', 'идея', 'использовать']

['продукт', 'клиент', 'компания', 'бренд', 'ваш', 'потребитель', 'рынок', 'товар', 'покупатель', 'пример']

['племя', 'лидер', 'идея', 'член', 'общение', 'человек', 'apple', 'аутсайдер', 'изменение', 'фитнес']

['проект', 'ваш', 'задача', 'список', 'команда', 'сторонник', 'пусть', 'идея', 'этап', 'корзина']

['переговоры', 'сторона', 'позиция', 'решение', 'критерий', 'цена', 'оба', 'оппонент', 'угроза', 'объективный']

['команда', 'член', 'доверие', 'decisiontech', 'конфликт', 'друг', 'цель', 'конструктивный', 'командный', 'решение']



# Пять тем по всем текстам

С именами собственными	Без имён
[человек, жизнь, герой, время, становиться, дом, друг, отец, год, мочь]	[человек, отец, жизнь, становиться, друг, время, год, девушка, сын, жена]
[ваш, клиент, компания, продукт, пример, бизнес, сотрудник, человек, решение, использовать]	[ваш, клиент, компания, продукт, пример, бизнес, сотрудник, человек, решение, использовать]
[дон, хуан, жуан, донья, карлос, луис, мануэль, кихот, изабелла, родриго]	[царь, бог, царица, хор, богиня, царский, царство, царевна, сын, дворец]
[эдип, фивы, креонт, антигона, этеокла, иокастый, хор, царь, полиник, лаий] = <b>Фиванский цикл!</b>	[король, принц, де, королева, герцог, граф, принцесса, рыцарь, замок, королевский]
[иван, царь, олеся, васильевич, конь, африканович, яга, иванович, жар, волк]	[дон, донья, слуга, шпага, король, донна, капитан, поединок, командор, косме]

# Результаты

- Мы не знаем, действительно ли наш датасет подходит для ТМ больше, чем полные тексты.
- Для наших целей NMF работает лучше.
- После удаления имён результаты становятся интереснее.
- Некоторые темы образуются одним произведением.
- Мы не получили именно то, что хотели.

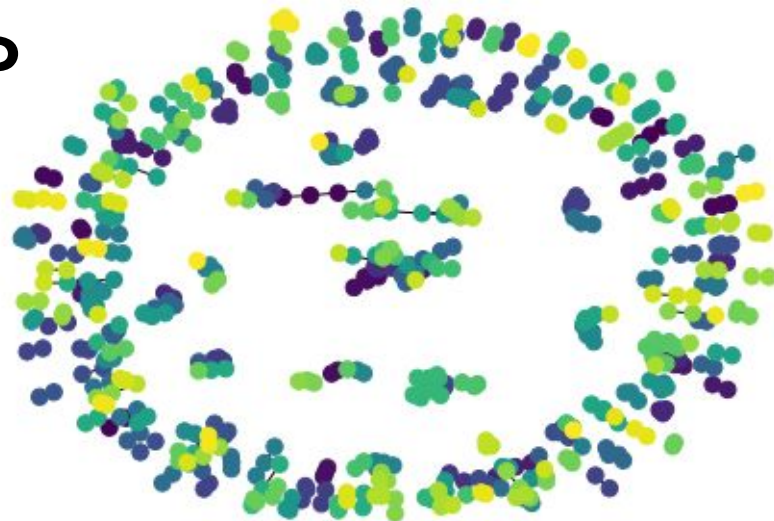
# Текстовая близость

Jaccard similarity mean = 0.0825

TF-IDF cosine similarity mean = 0.0324

Graph = 0.4 - 0.99

Наиболее центральные узлы графа:



**Кавасаки** Стартап: 11 мастер-классов от экс-евангелиста Apple и самого дерзкого венчурного капиталиста Кремниевой долины

**Аулет** Наука предпри-нимательства: 24 шага к успешному старту

**Холмс** Совершенная машина продаж: 12 проверенных стратегий эффективности бизнеса

**Гильбо** Стартап за 100\$: создай новое будущее, делая то, что ты любишь

**Молина** Севильский озорник, или Каменный гость

**Келлер** Зелёный Генрих

**Манн** Молодые годы короля Генриха IV

# Сетевой анализ: соседи по графу

**Кавасаки** Стартап: 11 мастер-классов от экс-евангелиста Apple и самого дерзкого венчурного капиталиста Кремниевой долины

**Аулет** Наука предпри-нимательства: 24 шага к успешному старту

**Гильбо** Стартап за 100\$: создай новое будущее, делая то, что ты любишь

**Годин** Фиолетовая корова: сделайте свой бизнес выдающимся!

**Кавасаки** Как очаровывать людей: искусство влиять на умы и поступки

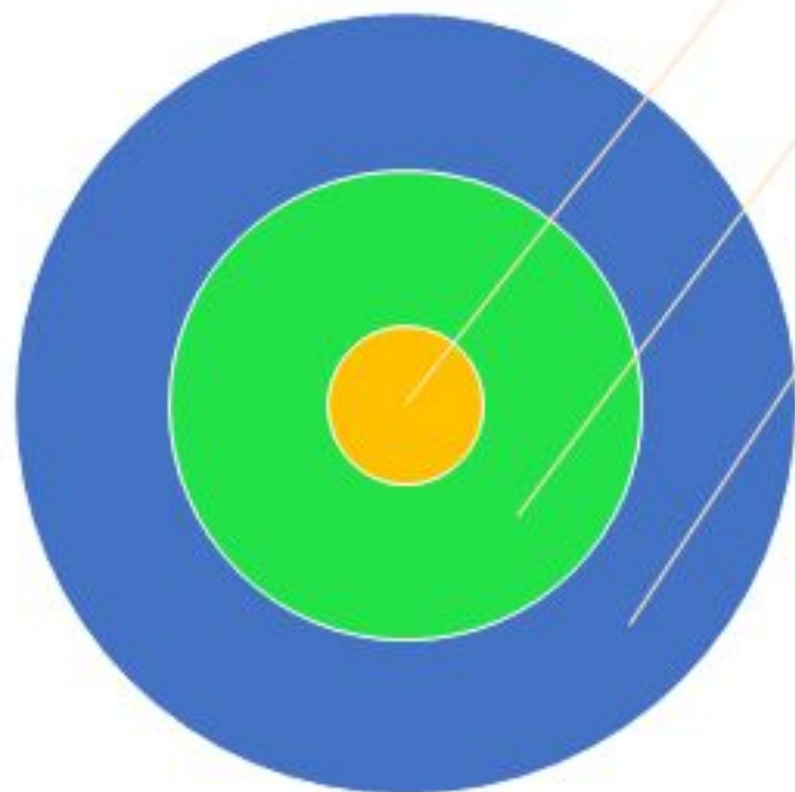
**Кауфман** Сам себе MBA: самообразование на 100%

**Уоррилоу** Созданный на продажу: постройте бизнес, который может процветать без вас

**Феррис** Как работать по 4 часа в неделю и при этом не торчать в офисе «от звонка до звонка», жить где угодно и богатеть

**Холмс** Совершенная машина продаж: 12 проверенных стратегий эффективности бизнеса

**Шей** Доставляя счастье: от нуля до миллиарда



клиент  
пример  
продукт  
люди

бизнес  
компании  
внимание

время  
делайте  
преимущества  
работа  
сотрудники  
услуги  
цели  
ценности

# Проклятие имен собственных

**Бласко Ибаньес** Кровь и песок

**Кальдерон** Дама-невидимка

**Кальдерон** Врач своей чести

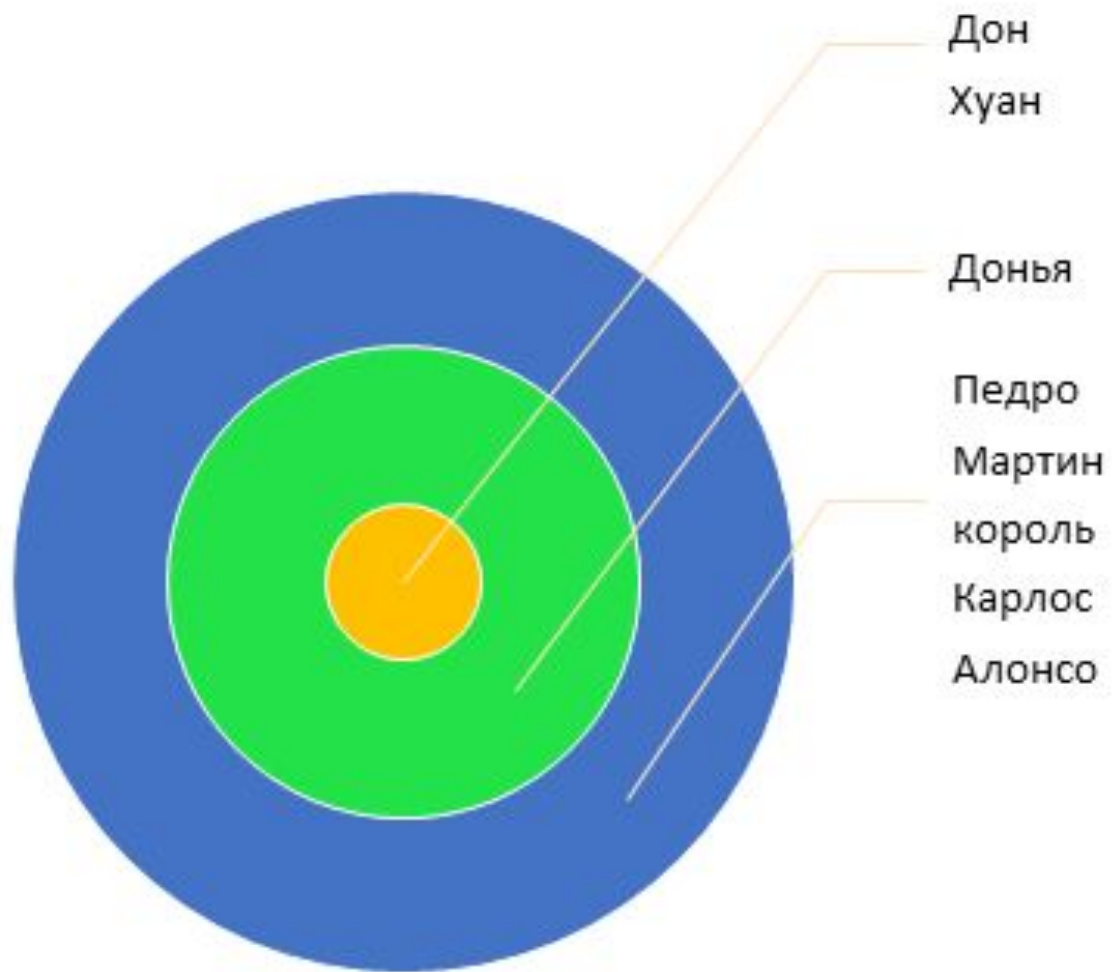
**Кальдерон** Спрятанный кабальеро

**Молина** Дон Хиль Зеленые штаны

**Мольер** Дон Жуан, или Каменный гость

**Скаррон** Жодле, или Хозяин-слуга

**Молина** Севильский озорник, или Каменный гость



# Сетевой анализ: соседи по графу

**Манн** Молодые годы короля Генриха IV

**Гауптман** Потонувший колокол

**Дюма** Королева Марго

**Келлер** Зелёный Генрих

**Новалис** Генрих фон Офтердинген

**Пиранделло** Генрих IV

**Сартр** Дьявол и Господь Бог

**Шварц** Голый король





# Текстовая близость: результаты

- Было предположение, что тексты будут похожи больше.
- Но: низкий коэффициент Жаккарда и низкое значение косинусной близости
- Пересказы бизнес-литературы оказались очень похожи. Возможно, из-за отсутствия имен собственных

# Общие выводы

- Имена собственные все испортили.
- Без них получаются интересные результаты: мы проверили!
- Количество имен собственных можно объяснить спецификой целевой аудитории кратких пересказов: студенты/школьники, которые не читали произведение.
- Бизнес-литература сильно отличается: в том числе из-за отсутствия имен
- А еще, когда работает команда, можно увидеть много разных интересных мелочей!