



Руслан Талипов

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В 2017

Rideró

Для тех, кто смотрит в записи

Тематическое моделирование – удобный инструмент для интерпретации больших коллекций текстов и их «мягкой кластеризации».

Базовые алгоритмы тематического моделирования давно коммодитизированы в большом количестве библиотек с низким порогом входа.

В докладе речь пойдёт об опыте тематического моделирования в проекте Ridero.

Ridero

Вместо введения



Вместо введения



План

Коротко о проекте

—

Что такое тематическое моделирование?

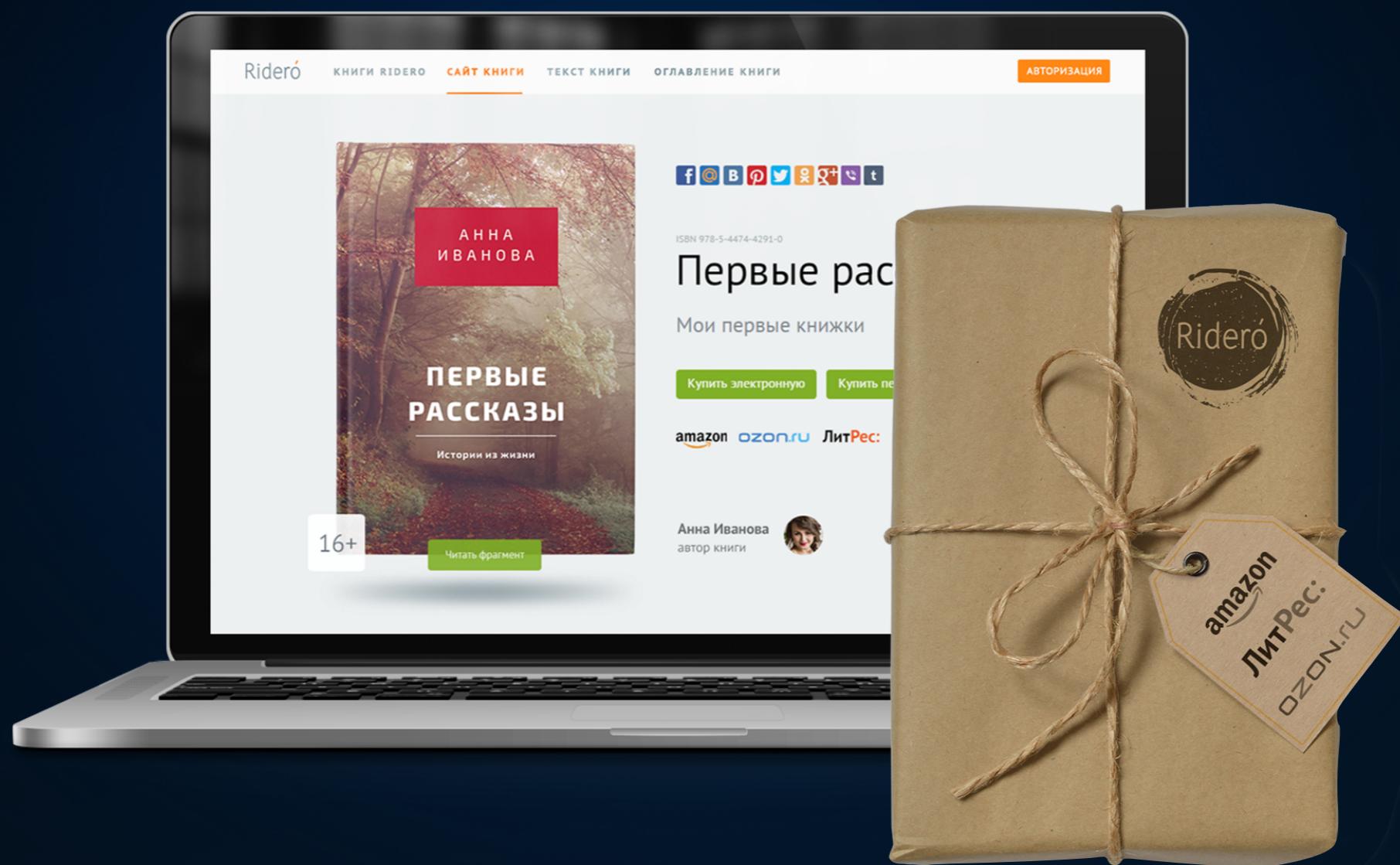
—

Как и какие инструменты мы используем

Riderо

Ridero

Издательский сервис для независимых авторов



Rideró

Текст. Он повсюду

Текст книги

Текст обращения в техподдержку

Текст отзывов в соцсетях

...

Riderо

Как понять, о чём пишут?

Кластеризовать тексты (объединить их в группы).

Понять, о чём каждый из кластеров.

Иметь возможность относить новые тексты
к одному из кластеров.

Как кластеризовать текст?

Нужно уметь вычислять близость двух различных текстов.

- Расстояние Левенштейна
(Редакционное)
- Расстояние Минковского
(Манхэттенское, евклидово)

Пример

```
texts = [  
    'Мне нужна помощь с моей книгой!',  
    'А как добавить иллюстрацию в книгу?',  
    'Как отправить книгу на модерацию?',  
    'Где мне искать картинки похожие на нужную иллюстрацию?',  
    'Как сменить иллюстрацию на обложке книги?',  
    'Что значит "из-за обложки не прошла модерацию"?',  
    'Спасибо!',  
    'Спасибо за помощь!',  
]
```

Матрица «термин-документ»

	doc1	doc2	doc3	doc4	doc5	doc6	doc7	doc8
книга	1	1	1		1			
иллюстрация		1		1	1			
moderация			1			1		
спасибо							1	1
помощь	1							1
на			1	1				
как		1	1		1			
...								
обложка					1	1		

Матрица «термин-документ»

Столбцы и строки это вектора.

Можно применять известные со школы методы измерения расстояния между векторами.

Можно кластеризовывать
(k-means++, иерархическая и т. д.).

Проблемы

Слишком громоздкое представление.

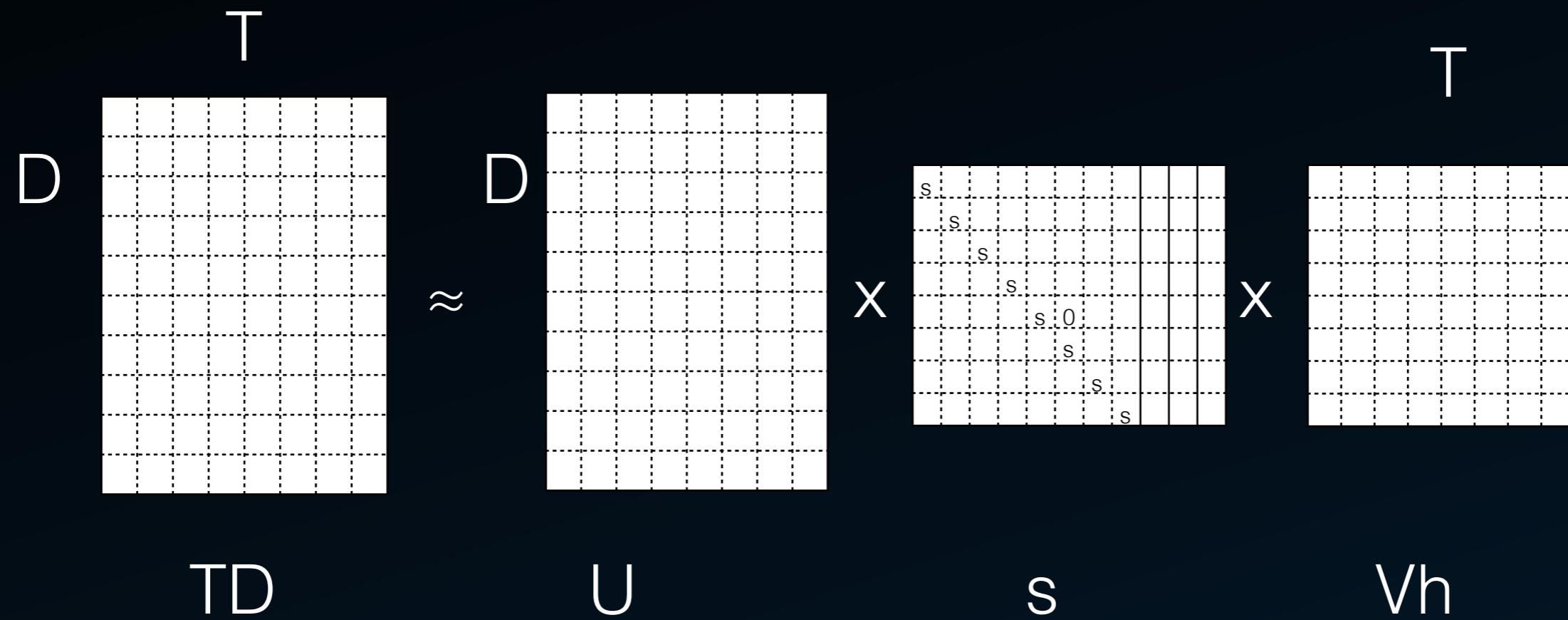
Количество документов ~100К, ~1М, ~10М.

Количество слов ~1М.

Большая часть матрицы заполняется нулями

Требуется низкоразмерное представление

Разложение матрицы



Используем SVD разложение для понижения размерности

Разложение матрицы

$$D^T \approx D^T \times U \times S \times V^T h$$

The diagram illustrates the Singular Value Decomposition (SVD) of a matrix D . On the left, a large square grid labeled D is shown with a dashed grid overlay. Above it, the transpose symbol T is placed above the letter D . To the right of the grid, the symbol \approx indicates approximation. Following this, the decomposition components are shown: D^T (with T above), a tall vertical rectangle labeled U , a horizontal rectangle labeled S containing a diagonal matrix with entries s_{ii} , and a wide horizontal rectangle labeled V^T (with T above) followed by the label h .

Используем SVD разложение для понижения размерности

Построение матрицы

```
texts = [  
    'Мне нужна помощь с моей книгой!',  
    'А как добавить иллюстрацию в книгу?',  
    'Как отправить книгу на модерацию?',  
    'Где мне искать картинки похожие на нужную иллюстрацию?',  
    'Как сменить иллюстрацию на обложке книги?',  
    'Что значит "из-за обложки не прошла модерацию"?',  
    'Спасибо!',  
    'Спасибо за помощь!',  
]
```

```
corpus = build_corpus(text)
print_corpus(corpus)
```

LSA

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

la = np.linalg
U, s, Vh = la.svd(corpus) # corpus.shape (8, 27)
plt.figure(figsize=(16,8))
plt.axis('off')
for i in range(len(texts)):
    plt.text(U[i,0], U[i,1], texts[i], size=18)
```

Результат

Где мне искать картинки похожие на нужную иллюстрацию?

Мне нужна помощь с моей книгой!

Спасибо за помощь!
Спасибо!

А как добавить иллюстрацию в книгу?

Как сменить иллюстрацию на обложке книги?
Как отправить книгу на модерацию?

Что значит "из-за обложки не прошла модерацию"?

LSA

Плюсы

Используем базовые методы линейной алгебры.

Минусы

Для интерпретации кластеров необходимо
«вчитываться» в текст.

Нет статистического обоснования.

Indexing by **L**atent **S**emantic **A**nalysis

<http://lsa3.colorado.edu/papers/JASIS.lsi.90.pdf>

Другой подход

Нужно повысить интерпретируемость

Другой подход



Другой подход

Topics

gene 0.04
dna 0.02
genetic 0.01
...
...

life 0.02
evolve 0.01
organism 0.01
...
...

brain 0.04
neuron 0.02
nerve 0.01
...
...

data 0.02
number 0.02
computer 0.01
...
...

Documents

Seeking Life's Bare (Genetic) Necessities

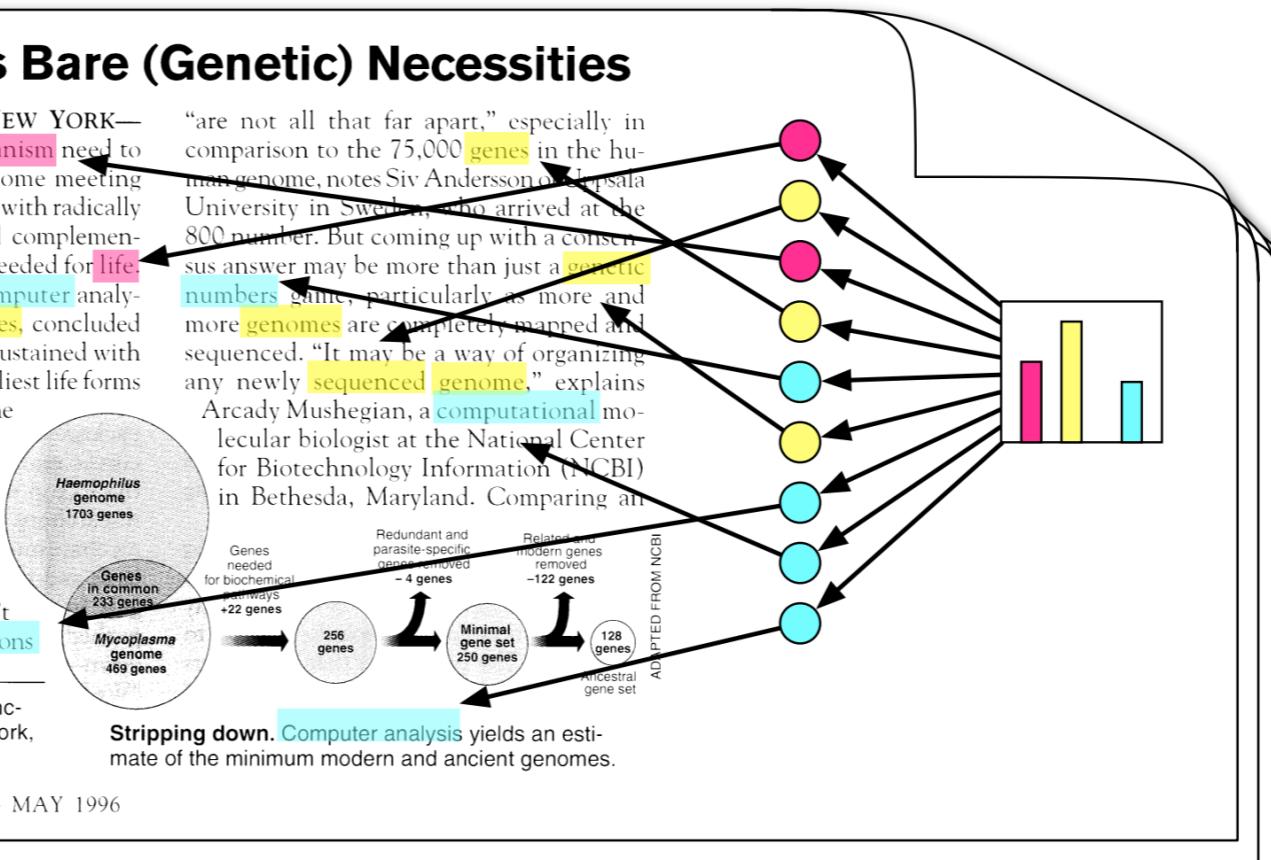
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



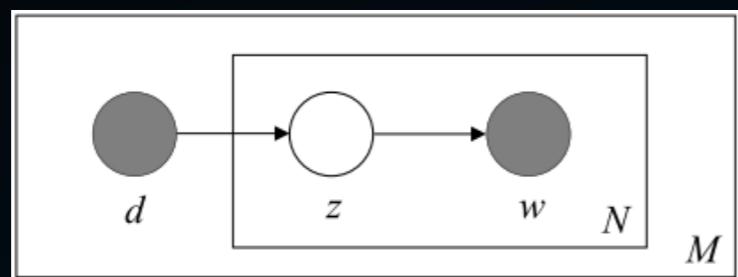
Другой подход

Каждое слово в документе генерируется с какой-то вероятностью из какой-то темы.

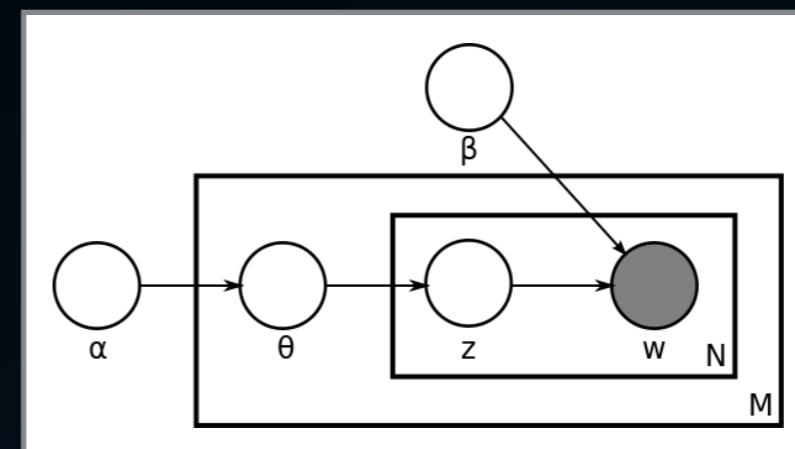
Тем в документе не должно быть много.

Для каждой из тем характерны особые ключевые слова.

PLSA и LDA



PLSA

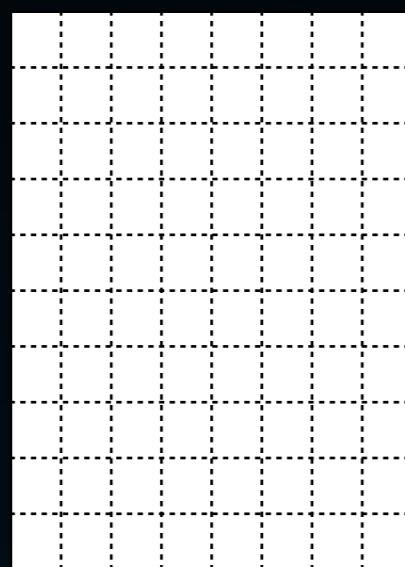


LDA

<http://www.iro.umontreal.ca/~nie/IFT6255/Hofmann-UAI99.pdf>

<https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf>

PLSA



\approx



\times



$d \times w$

$z \times w$

$d \times z$

Rideró

Новации PLSA

Темы представлены ключевыми словами
с весами – проще интерпретировать.

Руками можем контролировать количество тем.

Rideró

Недостатки PLSA

Возможно множество решений (задача не корректно поставлена).

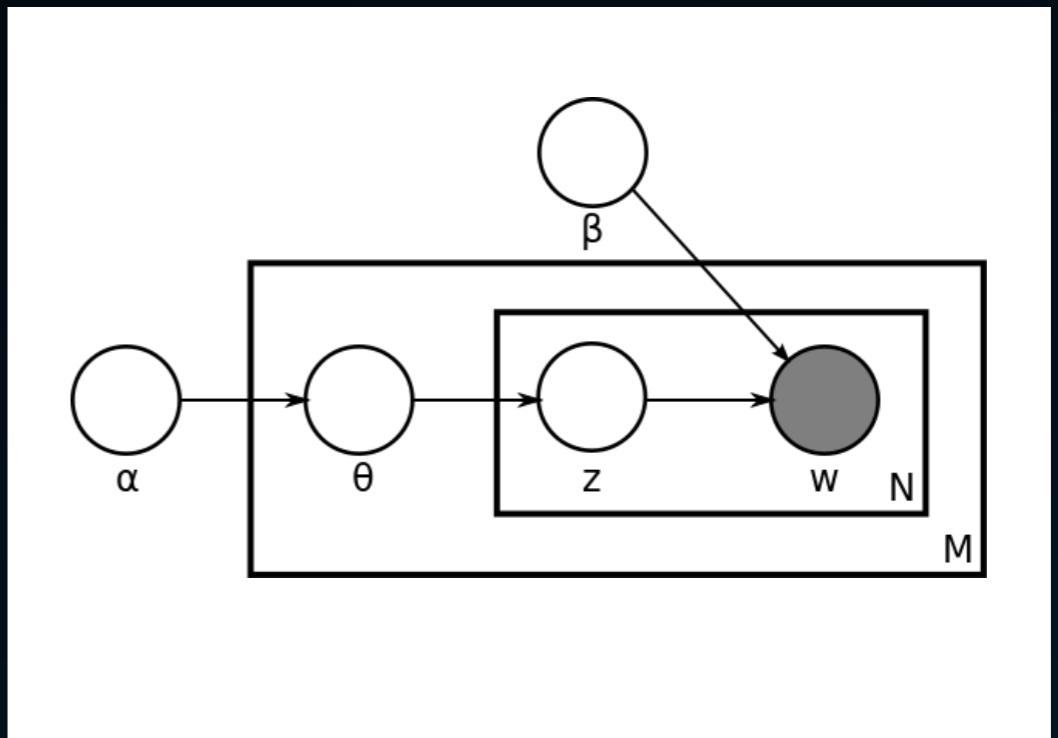
Нет возможности управлять распределением тем в коллекции. Хочется ограничивать количество тем в документе.

Нужно накладывать больше ограничений на модель – добавлять регуляризаторы.

Новации LDA

Параметры α и β
позволяют настраивать
решение.

Меньшие значения
параметров – меньше тем
в документе и слове.



Что я делаю на самом деле

```
>> from gensim.models import LdaModel
```

Rideró

А можно что-нибудь ещё?

sklearn.decomposition.LatentDirichletAllocation

vowpal_wabbit lda

pyspark.mllib.clustering.LDA

...

TensorFlow

Riderо

Бонусы Gensim

gensim.summarization.summarize

Суммаризация текста – выделение ключевых предложений текста.

gensim.summarization.keywords

Извлечение ключевых слов.

Как это было

Задача: определить темы книжек.

Решение: а что, если мы СРАЗУ, ТУПО впихнём тексты на вход.

Ожидание: получить разбивку книжек по ключевым понятиям, которые в них затрагиваются.

Очень просто

```
dictionary = corpora.Dictionary(texts)

corpus = [dictionary.doc2bow(text) for text in texts]

lda = LdaModel(corpus, num_topics=70)
```

Результат...

RiderÓ

Итерация 0

1. это, который, которая, смог
2. и, этот, когда, недавно
3. тот, он, тогда, сам
4. ...
5. а, но, видимо, несколько

Исправляемся

Отфильтровываем наречия и служебные части речи (руmorphy2)

Все слова в словаре приводим в начальную форму. (руmorphy2)

Склеиваем «не» с последующим словом
(не_прийти, не_сделать)

Разделяем цифры и буквы (100г, 25ый), убираем пунктуацию.
Иногда это убивает термины: «1С:Предприятие»

Итерация 1

1. **Алексей**, князь, **Мария**, **Дмитрий**, хан, русь...
2. **Иван**, лес, отец, **Анна**, собака, **Витя**...
3. **Даниил**, рука, клятва, меч...
4. ...
5. **Себастьян**, **Стив**, крыса, два, обезьянка...

Темы по именам

Собрали список имён.

Каждое имя в тексте заменили на МУЖСКОЕ_ИМЯ
или ЖЕНСКОЕ_ИМЯ.

Итерация 2

1. 100, резать, 200, вода, 1, 300, 500
2. Война, 19, посольство, правительство 20, император,
1917, __МУСКОЕ_ИМЯ__, 1905
3. месяц, кормление, 9, малыш
4.
5. 90, 80, __МУСКОЕ_ИМЯ__, Россия, СССР, 1991

Объединить числа и числительные в группы

1, 2, 3, 4_10, 10_50, 50_100, 100_250 ... 1500_1800, 1800-1900,

1900_1918, 1918_1938, 1938_1940, 1941, 1942, 1943, 1944, 1945,

1946_1953, 1953_1990, 1990_2000, 2000_2008, 2008_2012,

2012_2016,

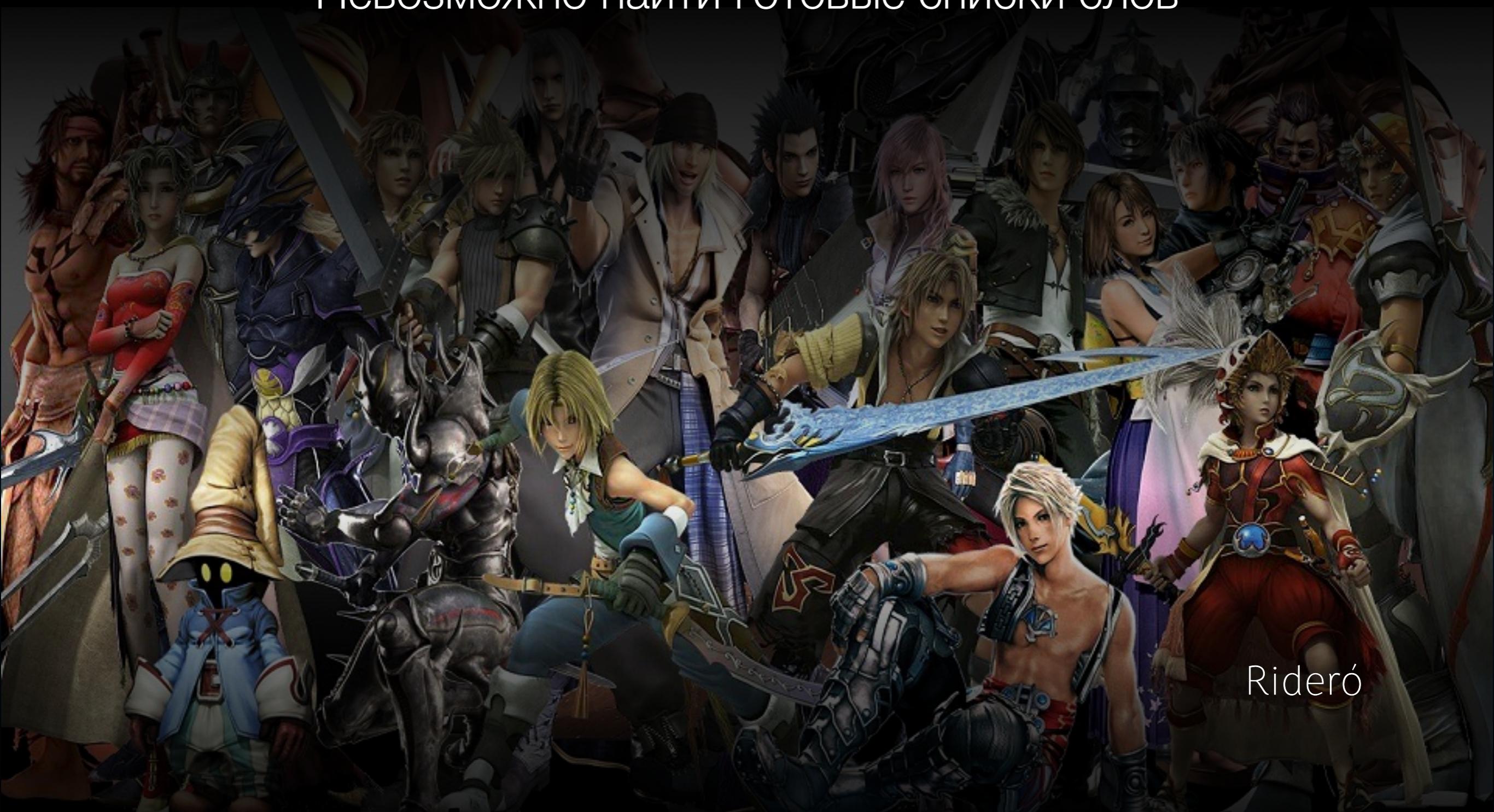
2016_3000, 3000-5000, 5000-10000, over10000.

Итерация 3

1. **Арам**, дракон, **Иса**, **Сильвион**, меч, мальчик
2. **Калинастр**, лазер, искра, **Деймонс**, сержант
3.
4. Глаз, рука, слеза, **Эмриша**, **Тайгер**

Фэнтезийные имена

Невозможно найти готовые списки слов



RiderÓ

Word2Vec

```
>> from gensim.models.word2vec import Word2Vec
```

Word2Vec



- солнце - день + ночь = луна
- день - утро + ночь = вечер
- ночь - тьма + свет = утро/вечер
- час - минута + миллиметр = сантиметр
- час - минута + сантиметр = фут (30 см)
- метр - сантиметр + час = полчаса
- копейка - цент + доллар = рубль
- секс - красота + уродство = пьянство

Word2Vec

Эспрессо Латте

Кофе

Капучино

Чай

Каркаде

Сок

Нектар

Word2Vec

Накатили word2vec на корпус книг.

Нашли слова, употребляемые в том же контексте,
что и простые имена.

Дополнили ими список имен.

Хорошие темы

1. __МУЖСКОЕ_ИМЯ__, сцена, писатель, герой, автор, книга художник, поэт, писать, искусство
2. язык, слово, время, образ, форма,
__МУЖСКОЕ_ИМЯ__, культура, значение, наука
3. __МУЖСКОЕ_ИМЯ__, бог, человек, господь, церковь, христос, храм, __ЖЕНСКОЕ_ИМЯ__.
4. город, поле, дорога, автобус, путешествие, страна, центр, водитель, эфир, автостоп, виза, место, поезд, ехать, скорость.
5. капитан, корабль, __МУЖСКОЕ_ИМЯ__, планета, экран, каюта, планета, команда, борт, система, экипаж

Что ещё нужно учесть.

Исправление опечаток.

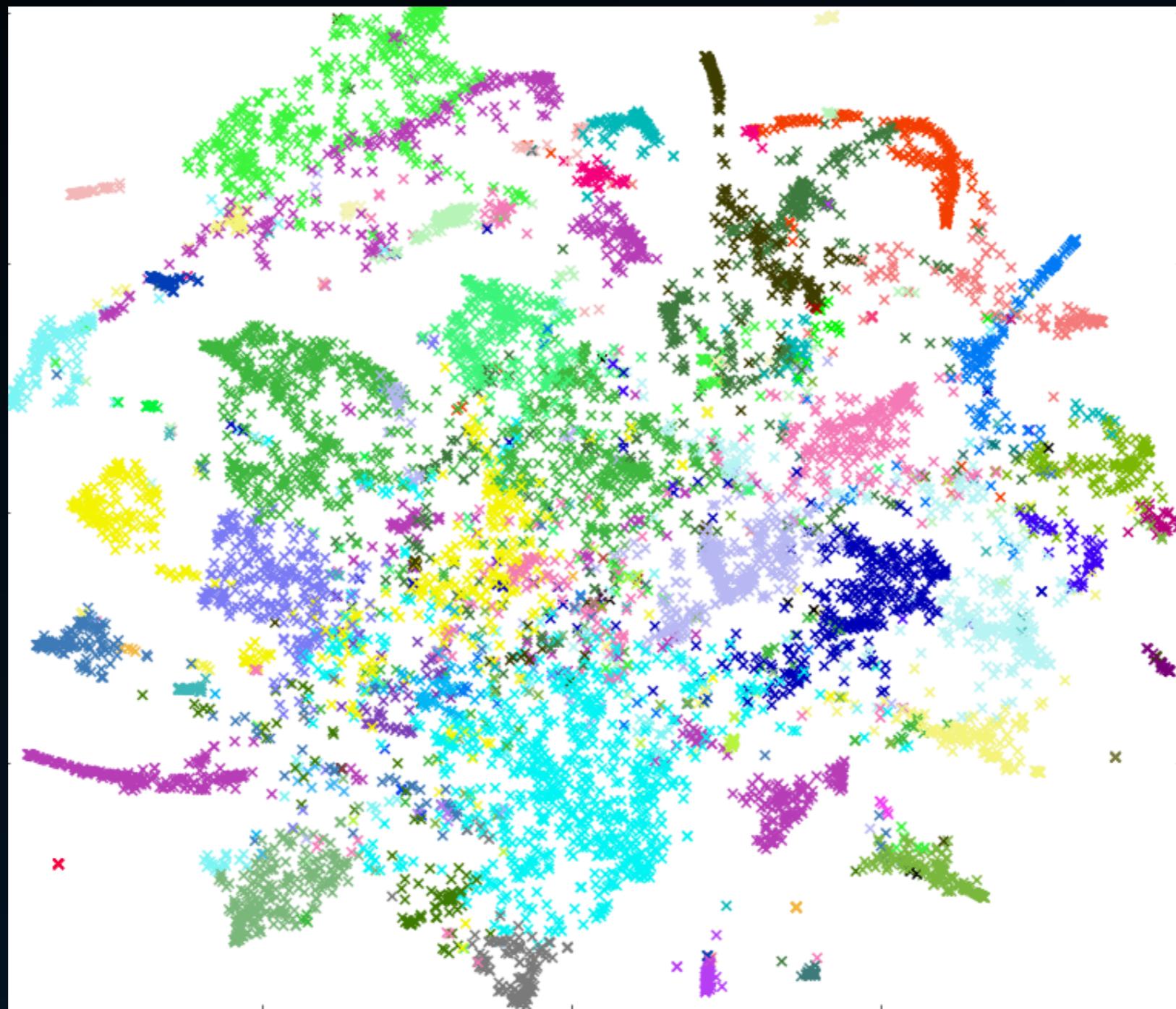
Выделение устойчивых конструкций и идиом
в качестве цельного термина:

Нижний_Новгород, Fullstack_Разработчик.

Препроцессинг

ЭТО ОЧЕНЬ ВАЖНО

Результат

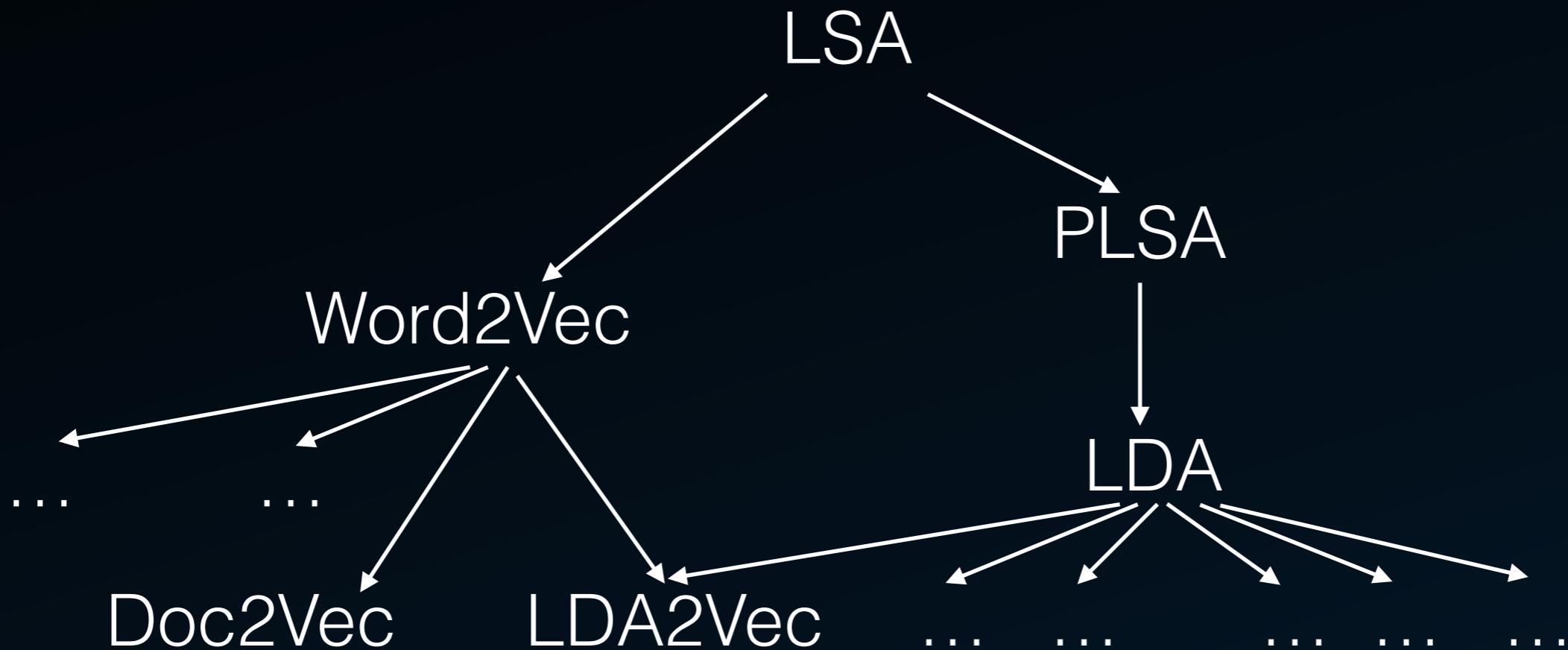


Погоди минуту...

...Gensim и LDA – это же что-то достаточно старое

Rideró

СВЯЗЬ

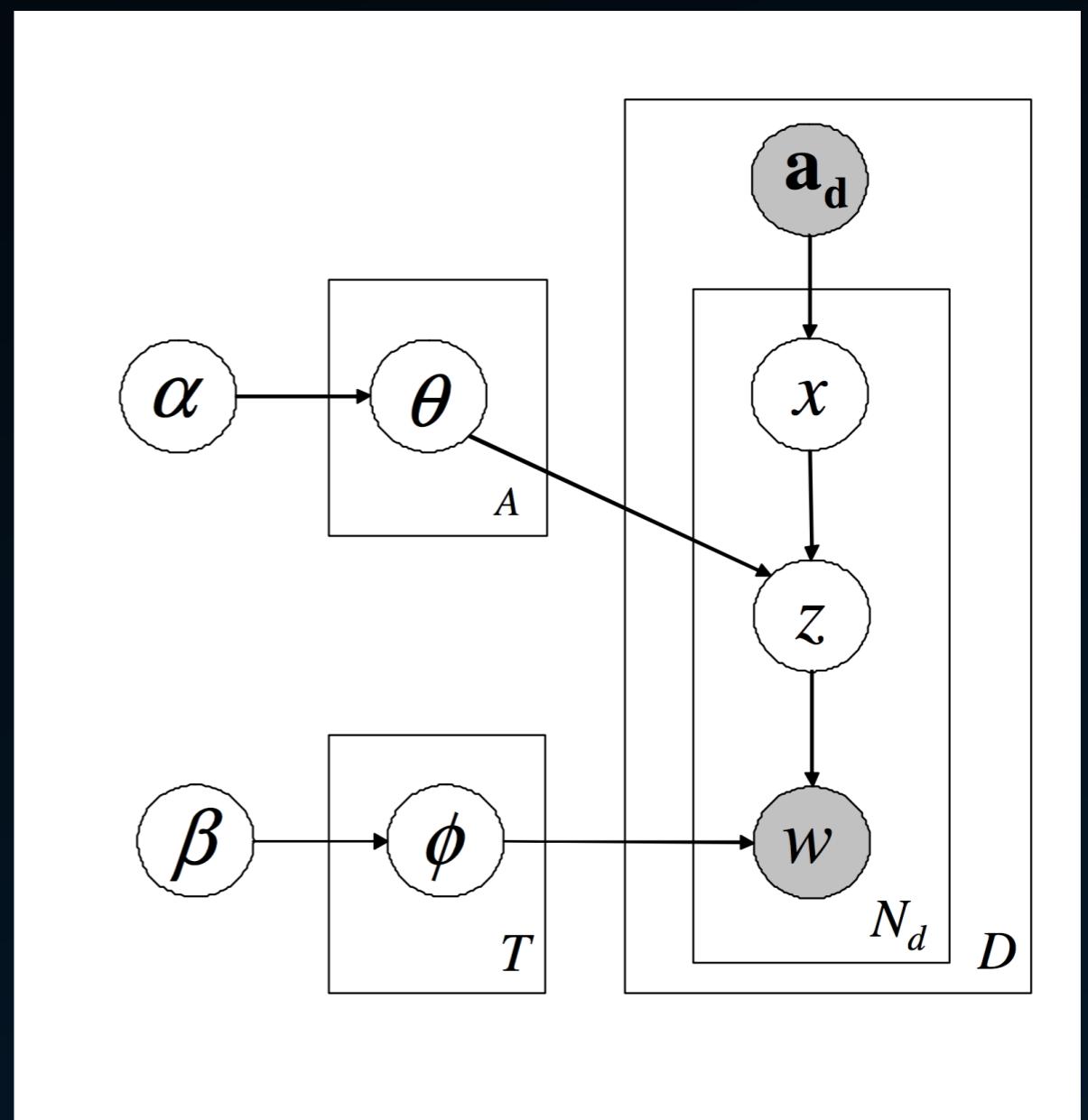


Rideró

Жизнь после LDA...

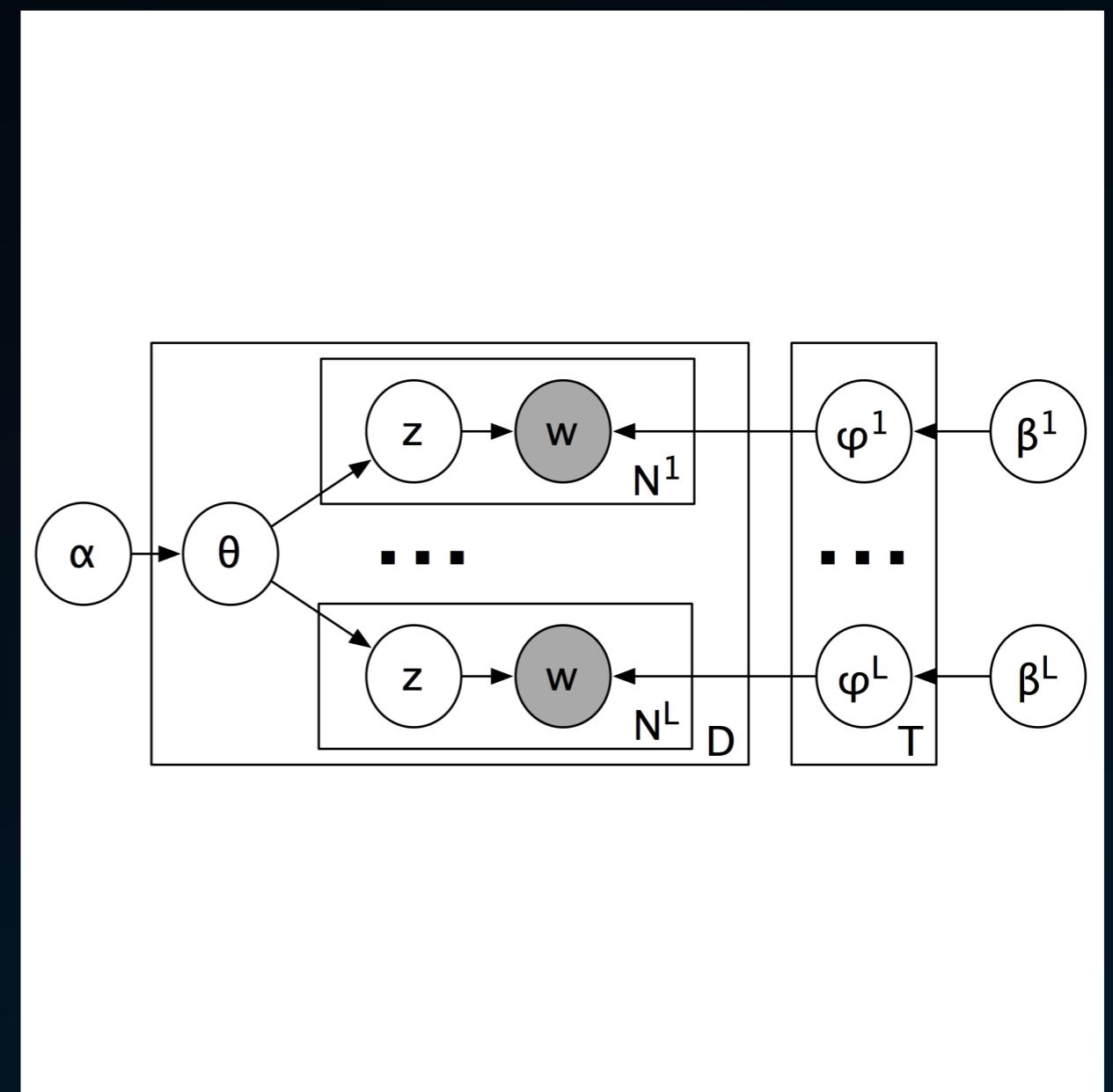
ATM – Author topic models

[https://mimno.infosci.cornell.edu/
info6150/readings/398.pdf](https://mimno.infosci.cornell.edu/info6150/readings/398.pdf)



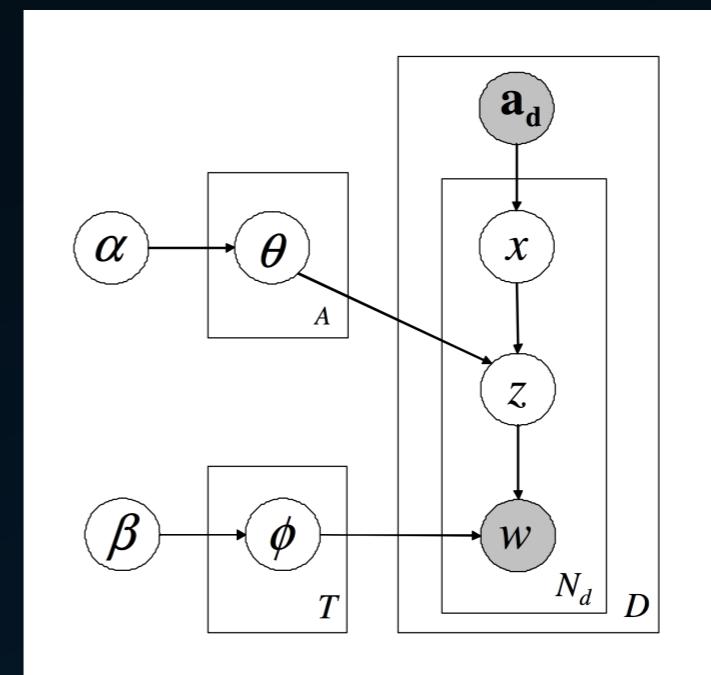
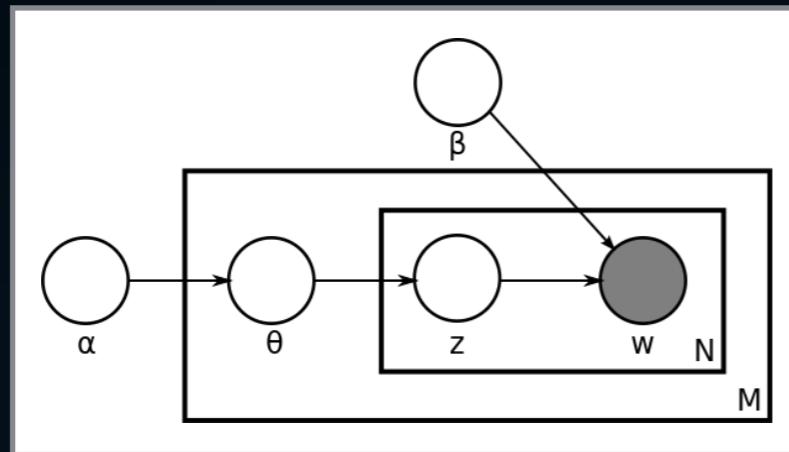
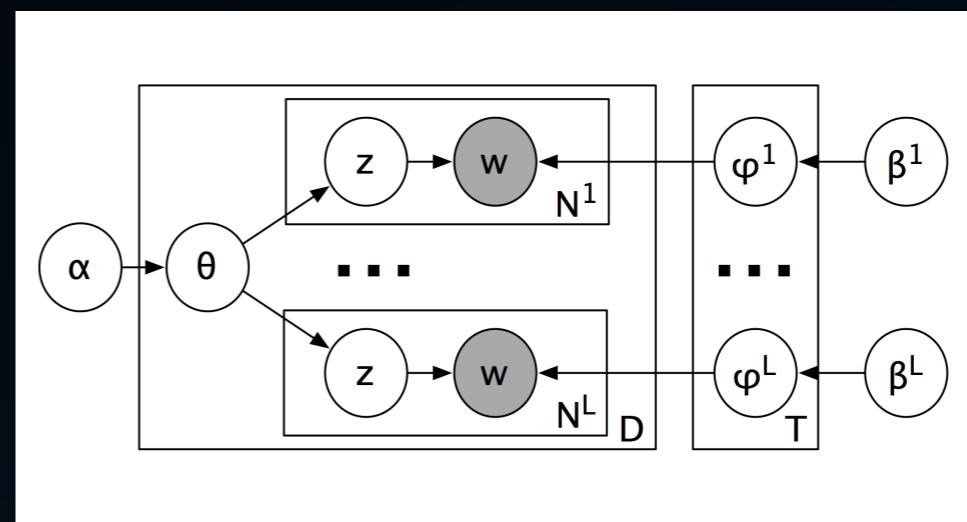
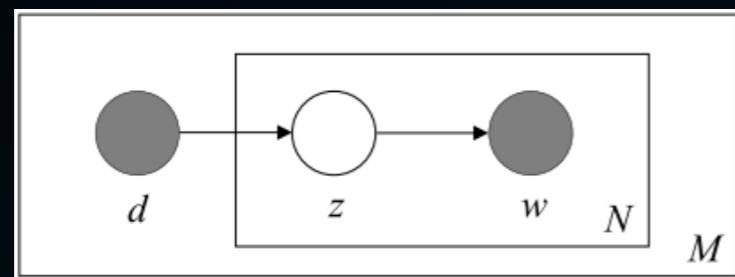
Жизнь после LDA...

ML-LDA –
мультиязычная модель



[http://www.aclweb.org/anthology/
D09-1092](http://www.aclweb.org/anthology/D09-1092)

Жизнь после LDA...



Все очень круто, но...

...сложно.

Свой вывод для каждой модели.

Библиотек с **готовыми реализациями** для простых людей нет. Хорошо, если есть отдельные реализации на C++.

Но есть одна библиотека, дающая простой **интерфейс для построения** достаточно сложных моделей.

Riderо

BigARTM



BigARTM

Модификации модели реализуются через регуляризаторы
– Описывайте модели через ограничения.

Различные регуляризаторы просто складываются друг с другом – Вы просто передаёте их в качестве параметров.

Можно гибко реализовывать различные модели несколькими модальностями (авторы, комментарии)

Простой интерфейс позволяет не сильно задумываться о математике, лежащей в основе.

RiderÓ

BigARTM быстрее



Gensim.LdaMulticore



BigARTM

Библиотека	Число процессоров	Время обучения модели ¹
BigARTM	1	35 минут
LdaModel	1	369 минут
VW.LDA	1	73 минуты
BigARTM	4	9 минут
LdaMulticore	4	60 минут
BigARTM	8	4.5 минуты
LdaMulticore	8	57 минут

¹ Использовалась коллекция документов английской Википедии,
 $|D| \approx 3.7 \times 10^6$.

http://www.machinelearning.ru/wiki/images/a/ae/MelLain_parallel_slides.pdf

RiderÓ

Как это было

Моделирование темы пользователя.

1-ая модальность текст книги

2-ая модальность краткая аннотация книги

3-ая модальность биография пользователя

Что получилось

1. человек, время, ребёнок, __МУЖСКОЕ_ИМЯ__,
слово, вопрос, момент, новое, сила, отношение
2. работа, развитие, родитель, сила, ответ, день,
проблема, техника, практика, психология
3. психолог, стаж, психология, специалист,
__число_между_3_и_10__, категория, НПЦ, тренинг,
мельников, патология

Что получилось

1. __МУЖСКОЕ_ИМЯ__, год, война,
__число_между_10_и_20__, два, часть, начальник,
дом, идти, армия
2. война, событие, россия, век, судьба, страна, ссср,
воспоминание, приключение, союз
3. конец, военный, начало, омск, служилый, училище,
ссср, граница, сибирь

Что получилось

1. время, девушка, голова, человек, голос, рука, взгляд, глаз, __ЖЕНСКОЕ_ИМЯ__
2. происходить, век, девушка, день, путешествие, судьба, реальность, дело, новое, отношение
3. фэнтези, поэт, декабрь, театр, решить, найти, первое, городок, герой, писательство

Что получилось

1. время, энергия, тело, вода, человек, организм, вид, земля, планета
2. женщина, знание, болезнь, рецепт, предложить, тело, заболевание, планета, природа
3. знание, награда, педагог, питание, сообщение, монография, биография, аудитория

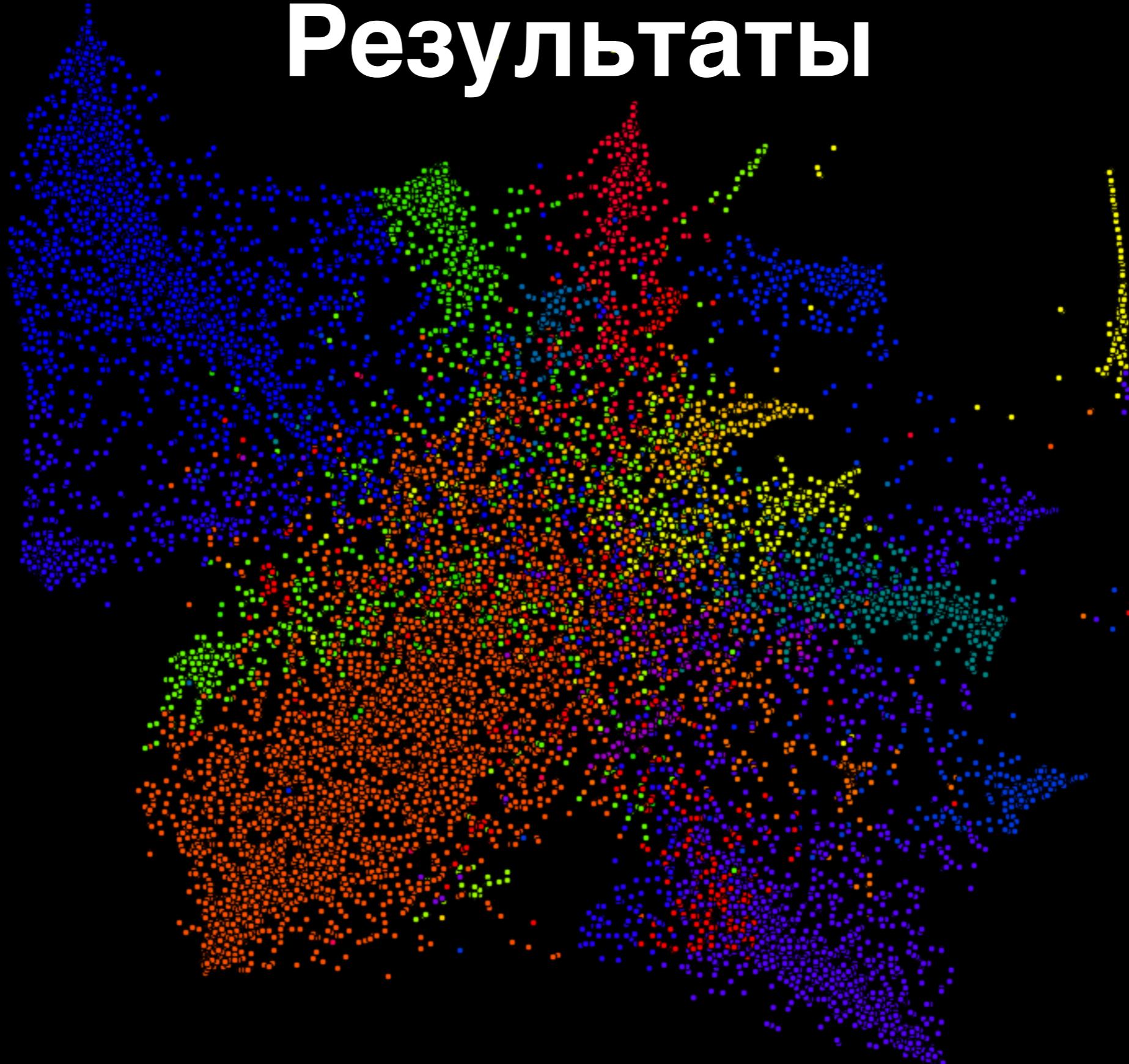
Что получилось

1. работа, время, дело, должный, жизнь, система, год, находиться
2. работа, стратегия, узнать, ответ, пример, прочитать, опыт, помочь, материал, форма
3. менеджмент, компания, участие, консультант, топменеджер, новое, управление, номинация, технология

Результаты

Книга	Аннотация	Биография
человек, время, ребёнок, __МУЖСКОЕ_ИМЯ__, слово, вопрос, момент, новое, сила, отношение	работа, развитие, родитель, сила, ответ, день, проблема, техника, практика, психология	психолог, стаж, психология, специалист, __число_между_3_и_10__, категория, нпц, тренинг, мельников, патология
__МУЖСКОЕ_ИМЯ__, год, война, __число_между_10_и_20__, два, часть, начальник, дом, идти, армия	война, событие, россия, век, судьба, страна, ссср, воспоминание, приключение, союз	конец, военный, начало, омск, служилый, училище, ссср, граница, сибирь
время, девушка, голова, человек, голос, рука, тогда, взгляд, глаз, __ЖЕНСКОЕ_ИМЯ__	происходить, век, девушка, день, путешествие, судьба, реальность, дело, новое, отношение	фэнтези, поэт, декабрь, театр, решить, найти, первое, городок, герой, писательство
время, энергия, тело, вода, человек, организм, вид, земля, планета	женщина, знание, болезнь, рецепт, два, предложить, тело, заболевание, планета, природа	знание, награда, педагог, питание, сообщение, монография, биография, аудитория
работа, время, дело, должный, жизнь, система, год, находится	работа, стратегия, узнать, ответ, пример, прочитать, опыт, помочь, материал, форма	менеджмент, компания, участие, консультант, топменеджер, новое, управление, номинация, технология

Результаты



Rideró

Другие матрицы

Пользователь – Предмет / Страница

Пользователь – Пользователь

Документ – Документ

Термин – Термин

Итого

Для «мягкой кластеризации» данных можно использовать тематическое моделирование.

Легко интерпретировать результат.

Следуя данной методологии, можно делать модели, включающие в себя несколько способов описания объектов.

BigARTM – хороший инструмент для построения таких моделей.

Riderо

Больше полезных ссылок

[http://www.machinelearning.ru/wiki/index.php?
title=BigARTM](http://www.machinelearning.ru/wiki/index.php?title=BigARTM)

[http://www.machinelearning.ru/wiki/index.php?
title=Аддитивная регуляризация тематических моделей](http://www.machinelearning.ru/wiki/index.php?title=Аддитивная_регуляризация_тематических_моделей)

https://www.youtube.com/watch?v=frLW8UVp_lk&t=1725s

Больше ссылок в бонусных слайдах

Спасибо

Вопросы?



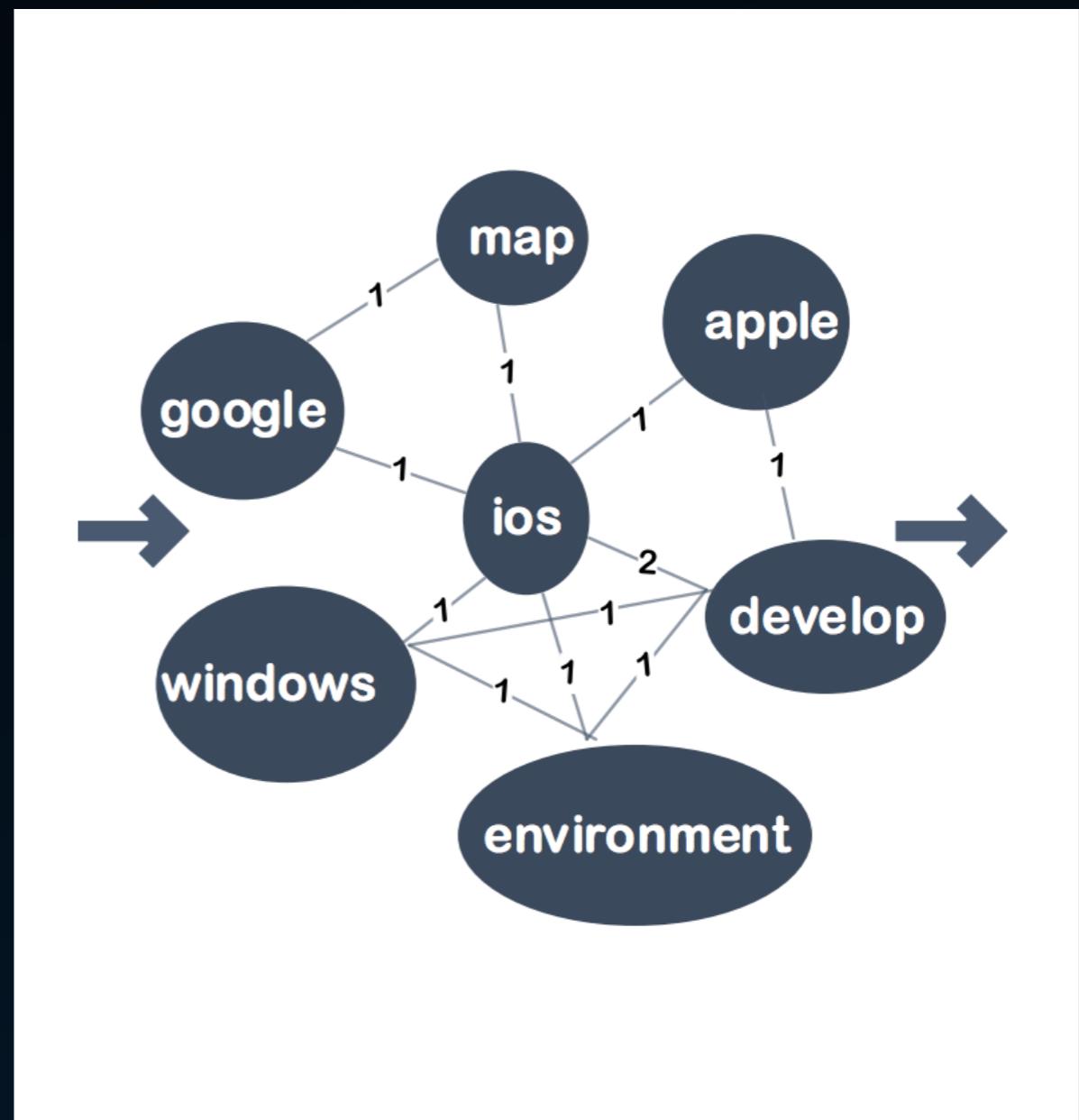
fb: /ruslan.talipov.75

vk, telegram: /roosh_roosh

Жизнь после LDA...

WNTM – word network
topic model

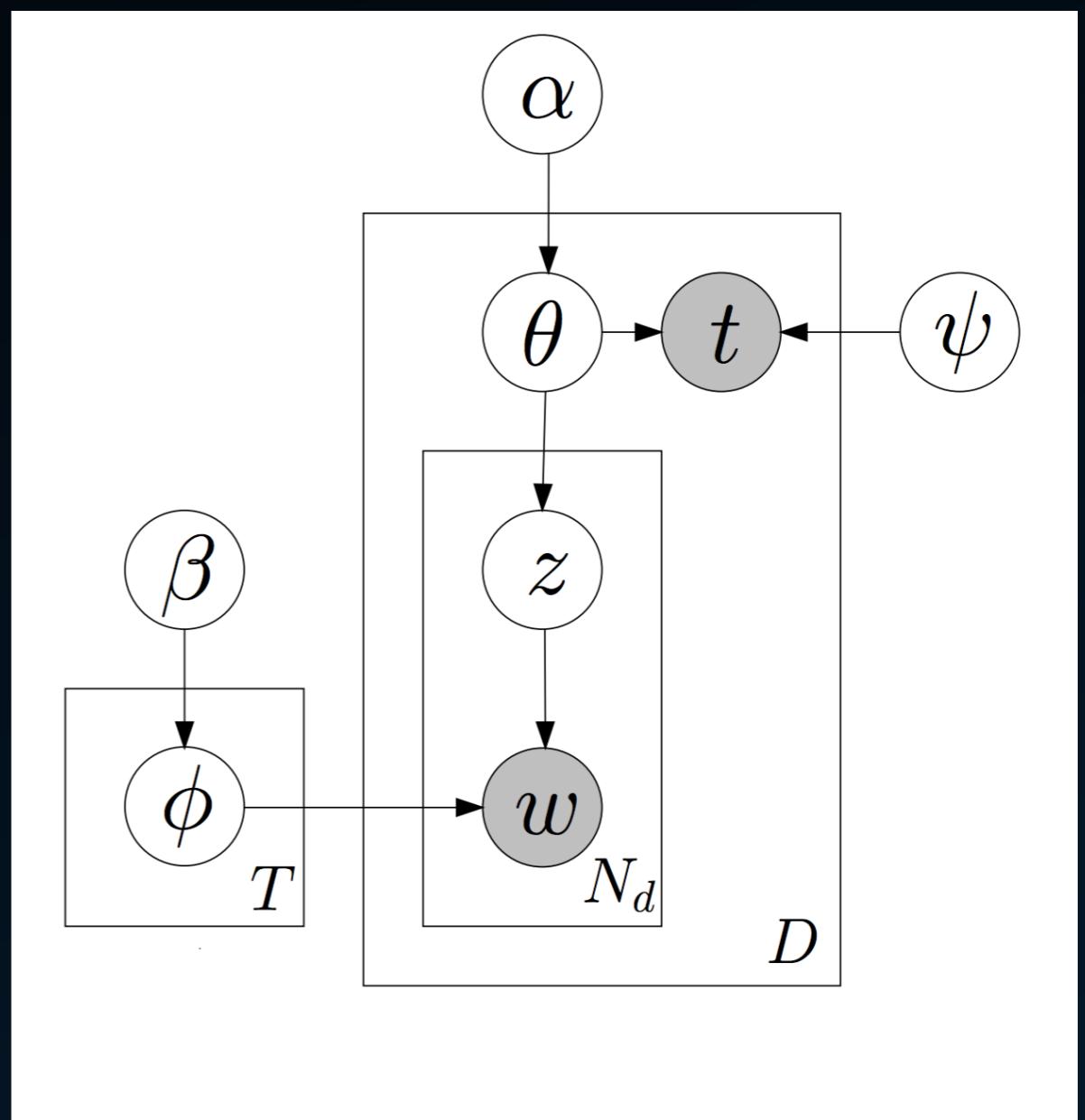
<https://arxiv.org/abs/1412.5404>



Жизнь после LDA...

TOT – topic over time

<https://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf>



Жизнь после LDA...

ssLDA – частично обученная

[http://www.cis.pku.edu.cn/faculty/vision/zlin/Publications/
2013-Neucom-RSSLDA.pdf](http://www.cis.pku.edu.cn/faculty/vision/zlin/Publications/2013-Neucom-RSSLDA.pdf)

Жизнь после LDA...

BitermTM

[https://pdfs.semanticscholar.org/
f499/5dc2a4eb901594578e3780a6f33dee02dad1.pdf](https://pdfs.semanticscholar.org/f499/5dc2a4eb901594578e3780a6f33dee02dad1.pdf)

Жизнь после LDA...

mLDA – MULTI-CONTEXTUAL TOPIC MODELS

<http://www-personal.umich.edu/~qmei/pub/kdd2013-Tang.pdf>

Жизнь после LDA...

HDP – Hierarchical Dirichlet Processes

[https://people.eecs.berkeley.edu/~jordan/papers/
hierarchical-dp.pdf](https://people.eecs.berkeley.edu/~jordan/papers/hierarchical-dp.pdf)

Жизнь после LDA...

TNG – Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval

<https://people.cs.umass.edu/~mccallum/papers/tng-icdm07.pdf>

Жизнь после LDA...

CTM – A CORRELATED TOPIC MODEL OF SCIENCE

[http://www.cs.columbia.edu/~blei/papers/
BleiLafferty2007.pdf](http://www.cs.columbia.edu/~blei/papers/BleiLafferty2007.pdf)

Жизнь после LDA...

NetPLSA – Topic Modeling with Network Regularization

<http://wwwconference.org/www2008/papers/pdf/p101-meia.pdf>