2025-03-10

Тематическое моделирование

Тематическое моделирование (моделирование топиков, topic modelling) — вид машинного обучения без учителя, основанный на статистическом моделировании.

Применяется для выявления тем в коллекции документов. Принцип тематического моделирования предполагает автоматическую кластеризацию слов, которые часто встречаются в документах в одном контексте (т.е. вместе), с целью выявления групп слов, характеризующих отдельные темы в коллекции документов. Конечная цель тематического моделирования — выявление K основных тем (топиков) в корпусе текстовых данных.

Основные сферы применения тематического моделирования

1. Кластеризация и классификацию документов: распределение документов по группам на основе их содержания (либо выявление этих групп, либо, когда топики уже есть, отнесение документов к какому-либо из них)
2. Информационный поиск: помощь поисковым системам в подборе наиболее релевантных документов
3. Квазиреферирование текста: сжатие больших фрагментов текста в более короткие рефераты с основной информацией
4. Сегментация клиентов: группировка клиентов на основе их отзывов или обзоров
5. Анализ тональности: определение положительного, отрицательного, нейтрального или иного заданного тона для большой коллекции текстов
6. Исследовательский анализ слабоструктурированных данных: обнаружение скрытых закономерностей и тем в большом корпусе текстовых данных

Основные этапы

1. Загрузить текстовые данные
2. Предварительная обработка данных: удаление стоп-слов, знаков препинания, другой нерелевантной информации; лемматизация.

3. Формирование матрицы «документ-термин»: предварительно обработанные текстовые данные преобразуются в матрицу количества слов, где каждая строка — документ, каждый столбец — уникальная лемма.
4. Запуск модели: наиболее популярные подходы включают латентное размещение (аллокацию) Дирихле (Latent Dirichlet Allocation, LDA), неотрицательное матричное разложение (non-negative matrix factorization, NMF), а также методы вроде структурного тематического моделирования (Structural Topic Models, STM). На вход модели подается сформированная ранее матрица.
5. В ходе моделирования формируется матрица «термин-топик»:

	hobby	**pursue**	**ability**	**giving**	...
Topic **4**	0.5	0.3	0.0	0.0	
Topic **5**	0.0	0.0	0.0	0.8	
...					
Topic **88**	0.0	0.2	0.5	0.0	
...					

6. Далее формируется набор наиболее релевантных терминов для каждого топика (обычно по 10):

Topic 4: hobby, rewarding, pursue, pursue hobby, leisure time, leisure, passion, family hobby, time hobby, friend, social
Topic 5: giving, woman, grateful, charitable, giving community, nonprofit, philosophy
Topic 88: ability, contribute, ability travel, time, ability work, pursue, time pursue, family, community, contribute community, contribute society

7. Полученные топики можно далее использовать для визуализации, анализа и решения иных задач на корпусе текстов.

Например, для группировки текстов по темам достаточно составить матрицу «документ-топик», опираясь на статистику встречаемости ключевых терминов топика в каждом конкретном документе (т.е. через определение функции принадлежности) — можно сформировать нечеткое множество для нечеткого отнесения текстов в множестве тем.

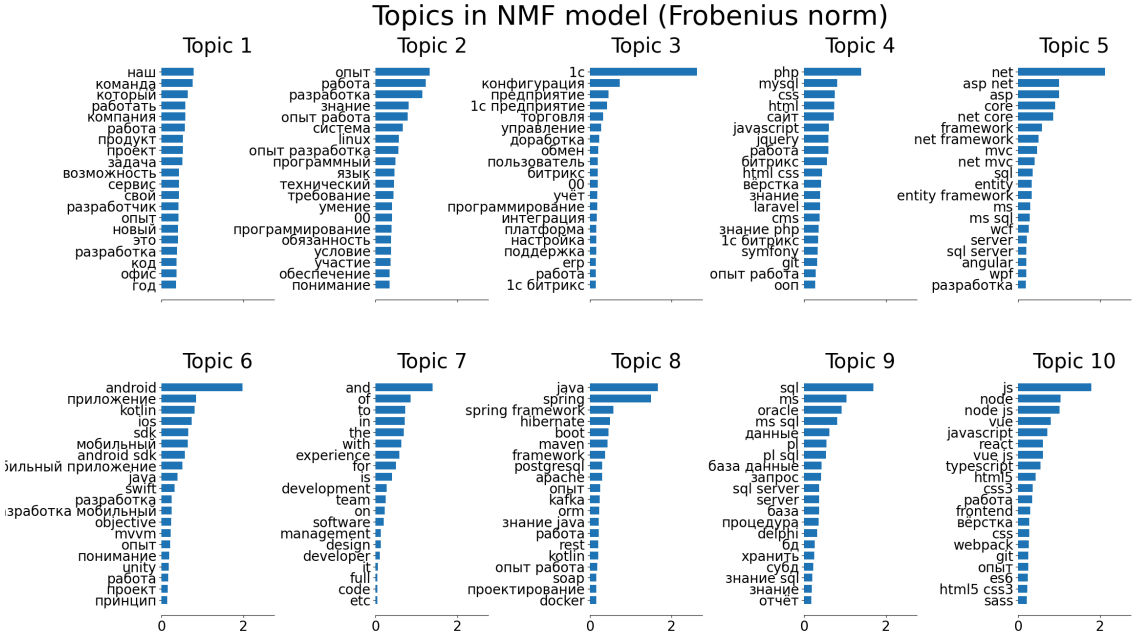
Document	Topic 4	Topic 5	...	Topic 88	...
I have a number of great hobbies and plenty of time to pursue them	0.7	0.3		0.0	
Life is most rewarding when I'm giving back to the community	0.5	0.2		0.8	
I have a great job that gives me the ability to pursue meaningful goals at work	0.3	0.7		0.0	
...					

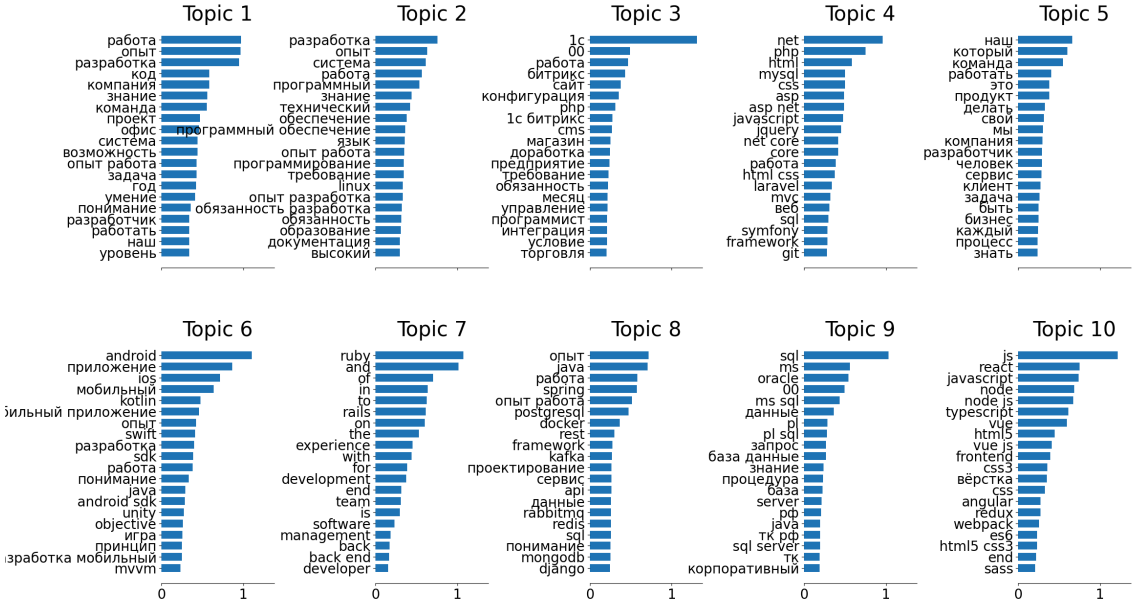
Лабораторная работа №5: Тематическое моделирование

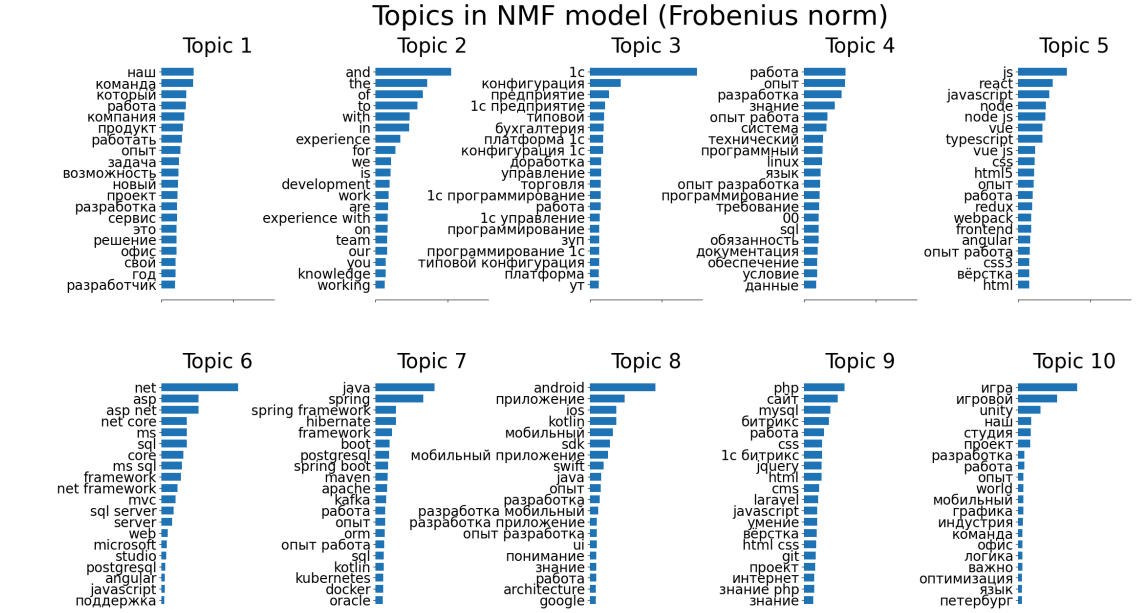
1. Дана коллекция текстовых документов `2021_SPORT`
2. Необходимо провести тематическое моделирование коллекции методами латентного размещения (аллокации) Дирихле и неотрицательного матричного разложения и визуализировать темы с разным количеством топ-слов, реализовав функцию `plot_top_words`.
    1. Для тематического моделирования понадобится `matplotlib`, `sklearn.feature_extraction.text` (`TfidfVectorizer`, `CountVectorizer`), `sklearn.decomposition` (`NMF`, `LatentDirichletAllocation`)
    2. Для препроцессинга нужно воспользоваться материалами прошлой лабораторной (в части стоп-слов, токенизации и лемматизации)
    3. Затем построить через `CountVectorizer`, `TfidfVectorizer` (попробуйте оба) векторное представление каждого текста (для визуализации полезно также сделать что-то вроде `tfidf_feature_names = tfidf_vectorizer.get_feature_names()`)
    4. Полученные вектора передавать уже в модель для тематического моделирования
    5. Количество признаков — начать с 1000, топиков — 10, топ-слов — 20

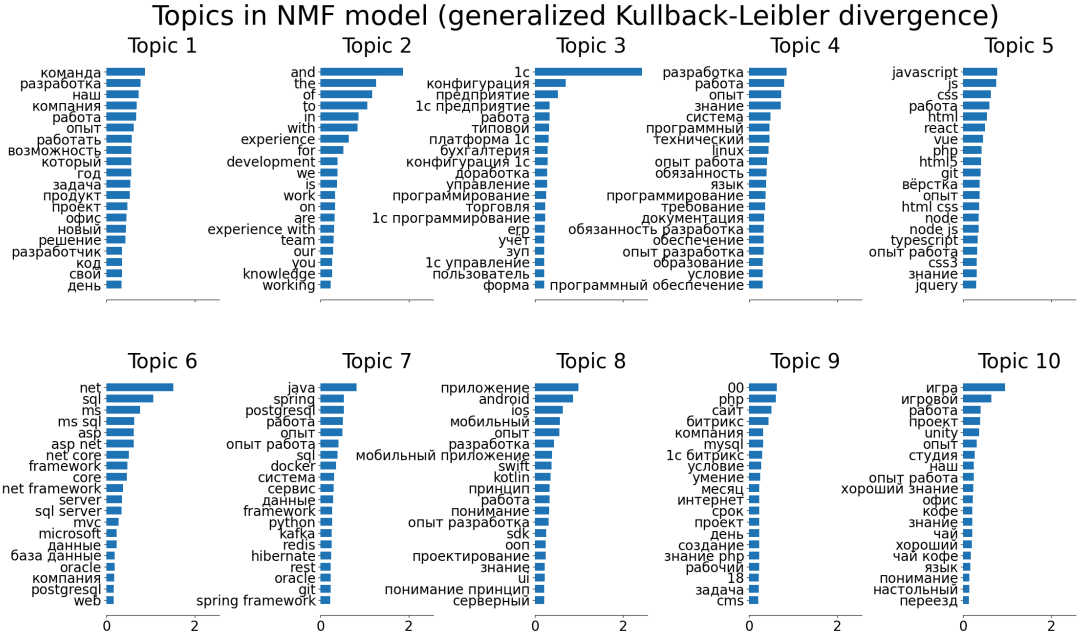
```
# как это может выглядеть (без реализации логики функций)

n_features = 1000
n_components = 10
n_top_words = 20
```

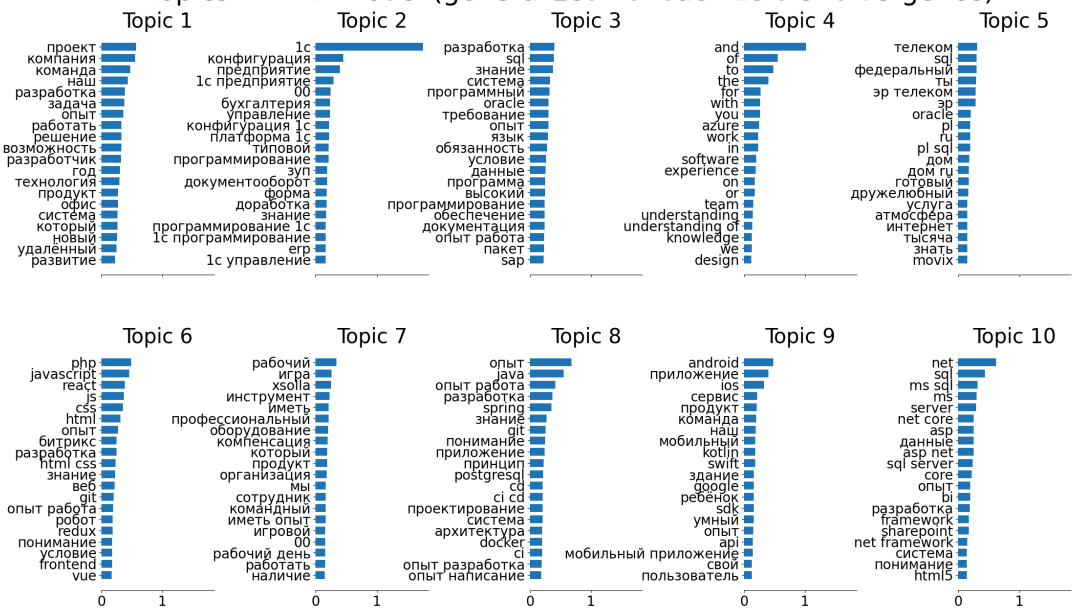
```
# тут все, собственно, происходит

if __name__ == '__main__':
    tfidf_vectorizer, tfidf = create_vectors_tf_idf('./sports')

    nmf = NMF(n_components=n_components, random_state=1, alpha=.1,
l1_ratio=.5).fit(tfidf)
    tfidf_feature_names = tfidf_vectorizer.get_feature_names()
    plot_top_words(nmf, tfidf_feature_names, n_top_words, 'Topics in NMF
model (Frobenius norm)')

    nmf_k = NMF(n_components=n_components, random_state=1,
            beta_loss='kullback-leibler', solver='mu', max_iter=1000,
alpha=.1,
            l1_ratio=.5).fit(tfidf)

    plot_top_words(nmf_k, tfidf_feature_names, n_top_words, 'Topics in
NMF model (generalized Kullback-Leibler divergence)')

    tf_vectorizer, tf = create_vectors_count('./sports')
    lda = LatentDirichletAllocation(n_components=n_components,
max_iter=5,
                                    learning_method='online',
                                    learning_offset=50.,
                                    random_state=0)
    lda.fit(tf)
    tf_feature_names = tf_vectorizer.get_feature_names()
    plot_top_words(lda, tf_feature_names, n_top_words, 'Topics in LDA
model')
```


Topics in NMF model (Frobenius norm)
Topic 1
наш
команда
который
работать
компания
работа
продукт
проект
задача
возможность
сервис
свой
разработчик
опыт
новый
это
разработка
код
офис
год
Topic 2
опыт
работа
разработка
знание
опыт работа
система
linux
опыт разработка
программный
язык
технический
требование
умение
00
программирование
обязанность
условие
участие
обеспечение
понимание
Topic 3
1с
конфигурация
предприятие
1с предприятие
торговля
управление
доработка
обмен
пользователь
битрикс
00
учет
программирование
интеграция
платформа
настройка
поддержка
erp
работа
1с битрикс
Topic 4
php
mysql
css
html
сайт
javascript
jquery
работа
битрикс
html css
вёрстка
знание
laravel
cms
знание php
1с битрикс
symfony
git
опыт работа
ооп
Topic 5
net
asp net
asp
core
net core
framework
net framework
mvc
net mvc
sql
entity
entity framework
ms
ms sql
wcf
server
sql server
angular
wpf
разработка
Topic 6
android
приложение
kotlin
ios
sdk
мобильный
android sdk
ильный приложение
java
swift
разработка
зработка мобильный
objective
mvvm
опыт
понимание
unity
работа
проект
принцип
Topic 7
and
of
to
in
the
with
experience
for
is
development
team
on
software
management
design
developer
it
full
code
etc
Topic 8
java
spring
spring framework
hibernate
boot
maven
framework
postgresql
apache
опыт
kafka
orm
знание java
работа
rest
kotlin
опыт работа
soap
проектирование
docker
Topic 9
sql
ms
oracle
ms sql
данные
pl
pl sql
база данные
запрос
sql server
server
база
процедура
delphi
бд
хранить
субд
знание sql
знание
отчет
Topic 10
js
node
node js
vue
javascript
react
vue js
typescript
html5
css3
работа
frontend
вёрстка
css
webpack
git
опыт
es6
html5 css3
sass


Topics in NMF model (generalized Kullback-Leibler divergence)
Topic 1
работа
опыт
разработка
код
компания
знание
команда
проект
офис
система
возможность
опыт работа
задача
год
умение
понимание
разработчик
работать
наш
уровень
Topic 2
разработка
опыт
система
работа
программный
знание
технический
обеспечение
программный обеспечение
язык
опыт работа
программирование
требование
linux
опыт разработка
обязанность разработка
обязанность
образование
документация
высокий
Topic 3
1с
00
работа
битрикс
сайт
конфигурация
1с битрикс
cms
магазин
доработка
предприятие
требование
обязанность
месяц
управление
программист
интеграция
условие
торговля
Topic 4
net
php
html
mysql
css
asp
asp net
javascript
jquery
net core
core
работа
html css
laravel
mvc
веб
sql
symfony
framework
git
Topic 5
наш
который
команда
работать
это
продукт
делать
свои
мы
компания
разработчик
человек
сервис
клиент
задача
быть
бизнес
каждый
процесс
знать
Topic 6
android
приложение
ios
мобильный
kotlin
ильный приложение
опыт
swift
разработка
sdk
работа
понимание
java
android sdk
unity
objective
игра
принцип
зработка мобильный
mvvm
Topic 7
ruby
and
of
in
to
rails
on
the
experience
with
for
development
end
team
is
software
management
back
back end
developer
Topic 8
опыт
java
работа
spring
опыт работа
postgresql
docker
rest
framework
kafka
проектирование
сервис
api
данные
rabbitmq
redis
sql
понимание
mongodb
django
Topic 9
sql
ms
oracle
ms sql
данные
pl
pl sql
запрос
база данные
знание
процедура
база
server
рф
java
тк рф
sql server
тк
корпоративный
Topic 10
js
react
javascript
node
node js
typescript
vue
html5
vue js
frontend
css3
вёрстка
css
angular
redux
webpack
es6
html5 css3
end
sass


Topics in NMF model (Frobenius norm)
Topic 1
наш
команда
который
работа
компания
продукт
работать
опыт
задача
возможность
новый
проект
разработка
сервис
это
решение
офис
свой
год
разработчик
Topic 2
and
the
of
to
with
in
experience
for
we
is
development
work
are
experience with
on
team
our
you
knowledge
working
Topic 3
1с
конфигурация
предприятие
1с предприятие
типовой
бухгалтерия
платформа 1с
конфигурация 1с
доработка
управление
торговля
1с программирование
работа
1с управление
программирование
зуп
программирование 1с
типовой конфигурация
платформа
ут
Topic 4
работа
опыт
разработка
знание
опыт работа
система
технический
программный
linux
язык
опыт разработка
программирование
требование
00
sql
обязанность
документация
обеспечение
условие
данные
Topic 5
js
react
javascript
node
node js
vue
typescript
vue js
css
html5
опыт
работа
redux
webpack
frontend
angular
опыт работа
css3
вёрстка
html
Topic 6
net
asp
asp net
net core
ms
sql
core
ms sql
framework
net framework
mvc
sql server
server
web
microsoft
studio
postgresql
angular
javascript
поддержка
Topic 7
java
spring
spring framework
hibernate
framework
boot
postgresql
spring boot
maven
apache
kafka
работа
опыт
orm
опыт работа
sql
kotlin
kubernetes
docker
oracle
Topic 8
android
приложение
ios
kotlin
мобильный
sdk
мобильный приложение
swift
java
опыт
разработка
разработка мобильный
разработка приложение
опыт разработка
ui
понимание
знание
работа
architecture
google
Topic 9
php
сайт
mysql
битрикс
работа
css
1с битрикс
jquery
html
cms
laravel
javascript
умение
вёрстка
html css
git
проект
интернет
знание php
знание
Topic 10
игра
игровой
unity
наш
студия
проект
разработка
работа
опыт
world
мобильный
графика
индустрия
команда
офис
логика
важно
оптимизация
язык
петербург


Topics in NMF model (generalized Kullback-Leibler divergence)
Topic 1
команда
разработка
наш
компания
работа
опыт
работать
возможность
который
год
задача
продукт
проект
офис
новый
решение
разработчик
код
свой
день
Topic 2
and
the
of
to
in
with
experience
for
development
we
is
work
on
are
experience with
team
our
you
knowledge
working
Topic 3
1c
конфигурация
предприятие
1с предприятие
работа
типовой
платформа 1с
бухгалтерия
конфигурация 1с
доработка
управление
программирование
торговля
1с программирование
erp
учёт
зуп
1с управление
пользователь
форма
Topic 4
разработка
работа
опыт
знание
система
программный
технический
linux
опыт работа
обязанность
язык
программирование
требование
документация
обязанность разработка
обеспечение
опыт разработка
образование
условие
программный обеспечение
Topic 5
javascript
js
css
работа
html
react
vue
php
html5
git
вёрстка
опыт
html css
node
node js
typescript
опыт работа
css3
знание
jquery
Topic 6
net
sql
ms
ms sql
asp
asp net
net core
framework
core
net framework
server
sql server
mvc
microsoft
данные
база данные
oracle
компания
postgresql
web
0
2
Topic 7
java
spring
postgresql
работа
опыт
опыт работа
sql
docker
система
сервис
данные
framework
python
kafka
redis
hibernate
rest
oracle
git
spring framework
0
2
Topic 8
приложение
android
ios
мобильный
опыт
разработка
мобильный приложение
swift
kotlin
принцип
работа
понимание
опыт разработка
sdk
проектирование
знание
ui
понимание принцип
серверный
0
2
Topic 9
00
php
сайт
битрикс
компания
mysql
1с битрикс
условие
умение
месяц
интернет
срок
проект
день
создание
знание php
рабочий
18
задача
cms
0
2
Topic 10
игра
игровой
работа
проект
unity
опыт
студия
наш
опыт работа
хороший знание
офис
кофе
знание
чай
хороший
чай кофе
язык
понимание
настольный
переезд
0
2
Topics in NMF model (Frobenius norm)
Topic 1
наш
команда
компания
проект
задача
продукт
работать
разработка
опыт
который
разработчик
новый
сервис
решение
приложение
год
свой
возможность
технология
код
Topic 2
1c
конфигурация
предприятие
1с предприятие
бухгалтерия
конфигурация 1с
управление
типовой
платформа 1с
00
программирование
зуп
программирование 1с
1с программирование
erp
1с бухгалтерия
1с управление
документооборот
доработка
форма
Topic 3
разработка
знание
опыт
sql
система
язык
программный
требование
программирование
опыт работа
обязанность
данные
программа
высокий
условие
00
ms
обеспечение
образование
программный обеспечение
Topic 4
and
of
to
the
you
with
for
work
experience
in
or
team
on
we
knowledge
understanding
azure
understanding of
our
development
Topic 5
oracle
sql
pl sql
pl
телеком
эр телеком
ap
ты
федеральный
bi
ru
oracle pl
субд oracle
данные
готовый
дом ru
sql pl
база
интернет
субд
Topic 6
javascript
php
react
css
html
js
опыт
html css
понимание
yii
разработка
опыт работа
git
знание
jquery
веб
написание
css javascript
rest
условие
0
1
Topic 7
xsolla
рабочий
игра
организация
игровой
наличие
командный
наличие собственный
профессиональный
мы
гибкий
спринт
который
профессиональный обучение
добровольный медицинский
добровольный
продукт
сообщество
инструмент
гибкий рабочий
0
1
Topic 8
java
spring
spring framework
framework
hibernate
maven
boot
опыт
spring boot
apache
postgresql
разработка
orm
проект
опыт работа
система
kafka
docker
kubernetes
качество
0
1
Topic 9
android
приложение
ios
kanna
sdk
опыт
мобильный
swift
знание
понимание
разработка
java kotlin
java
мобильный приложение
rest
паттерн
api
серверный
серверный приложение
опыт разработка
0
1
Topic 10
net
net core
core
asp net
asp
framework
net framework
разработка
опыт
система
система контроль
контроль
знание
опыт работа
понимание принцип
solid
mvc
контроль версия
sql
postgresql
0
1


Topics in NMF model (generalized Kullback-Leibler divergence)
Topic 1
проект
компания
команда
наш
разработка
задача
опыт
работать
решение
возможность
разработчик
год
технология
продукт
офис
система
который
новый
удалённый
развитие
Topic 2
1с
конфигурация
предприятие
1с предприятие
00
бухгалтерия
управление
конфигурация 1с
платформа 1с
типовой
программирование
зуп
документооборот
форма
доработка
знание
программирование 1с
1с программирование
erp
1с управление
Topic 3
разработка
sql
знание
система
программный
oracle
требование
опыт
язык
обязанность
условие
данные
программа
высокий
программирование
обеспечение
документация
опыт работа
пакет
sap
Topic 4
and
of
to
the
for
with
you
azure
work
in
software
experience
on
or
team
understanding
understanding of
knowledge
we
design
Topic 5
телеком
sql
федеральный
ты
эр телеком
эр
oracle
pl
ru
pl sql
дом
дом ru
готовый
дружелюбный
услуга
атмосфера
интернет
тысяча
знать
movix
Topic 6
php
javascript
react
js
css
html
опыт
битрикс
разработка
html css
знание
веб
git
опыт работа
робот
redux
понимание
условие
frontend
vue
0
1
Topic 7
рабочий
игра
xsolla
инструмент
иметь
профессиональный
оборудование
компенсация
который
продукт
организация
мы
сотрудник
командный
иметь опыт
игровой
00
рабочий день
работать
наличие
0
1
Topic 8
опыт
java
опыт работа
разработка
spring
знание
git
понимание
приложение
принцип
postgresql
cd
ci cd
проектирование
система
архитектура
docker
ci
опыт разработка
опыт написание
0
1
Topic 9
android
приложение
ios
сервис
продукт
команда
наш
мобильный
kotlin
swift
здание
google
ребенок
sdk
умный
опыт
api
мобильный приложение
свой
пользователь
0
1
Topic 10
net
sql
ms sql
ms
server
net core
asp
данные
asp net
sql server
core
опыт
bi
разработка
framework
sharepoint
net framework
система
понимание
html5
0
1