Тематическое моделирование отзывов Контур.Отель

Байтенова Асыл Житков Алексей

При чем тут Контур.Отель?

Простой способ управлять отелем

С ним можно:
управлять номерным фондом
получать брони с сайта отеля и площадок бронирования
представлять сведения в МВД о гостях через интернет

Тематическое моделирование

метод извлечения тем из текста

Что сделали?

Подготовили данные

привели буквы к нижнему регистру

удалили все лишние знаки

токенизация

лемматизация

удалили стоп-слова

стемминг

биграммы и триграммы

корпус и словарь

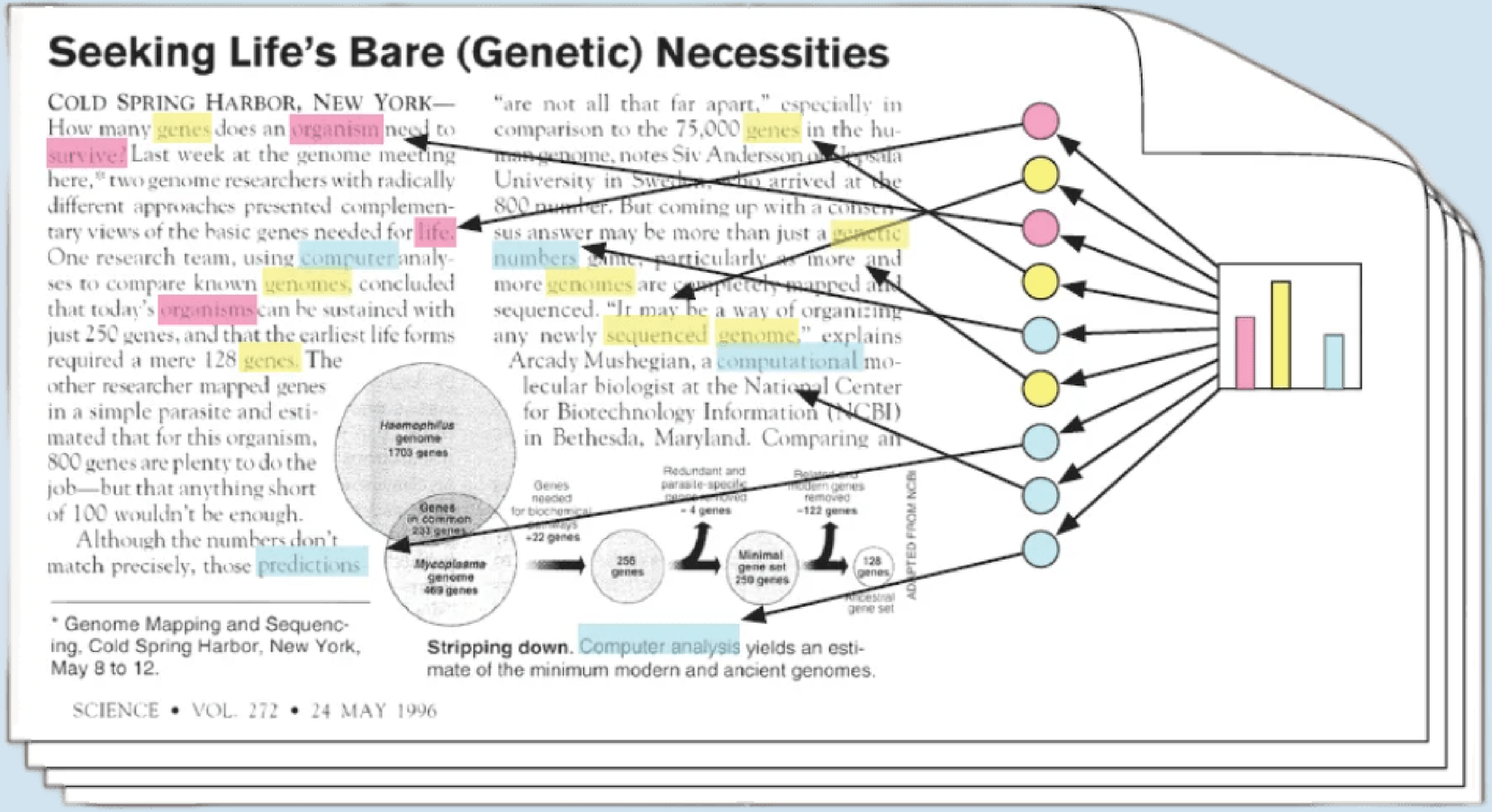
LDA

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
. . .	



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 "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

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

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Selected Topic: 0 Previous Topic Next Topic Clear Topic

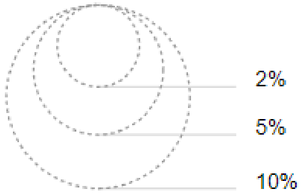
Slide to adjust relevance metric:(2)

λ = 1

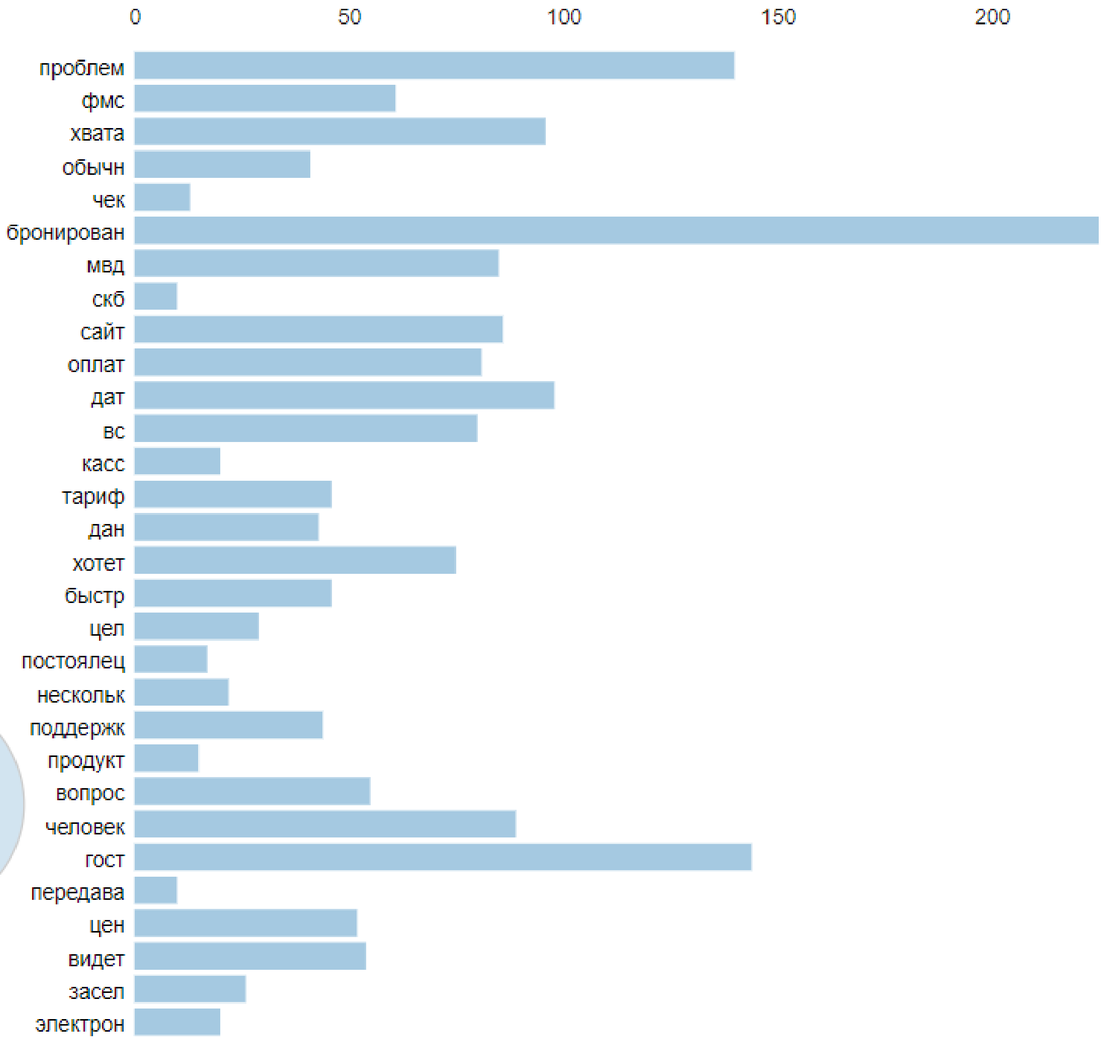
0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)

Marginal topic distribution



Top-30 Most Salient Terms[1]



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Selected Topic: 1 Previous Topic Next Topic Clear Topic

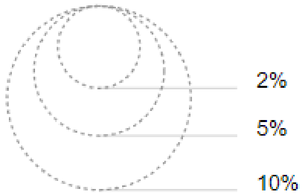
Slide to adjust relevance metric:(2)

λ = 1

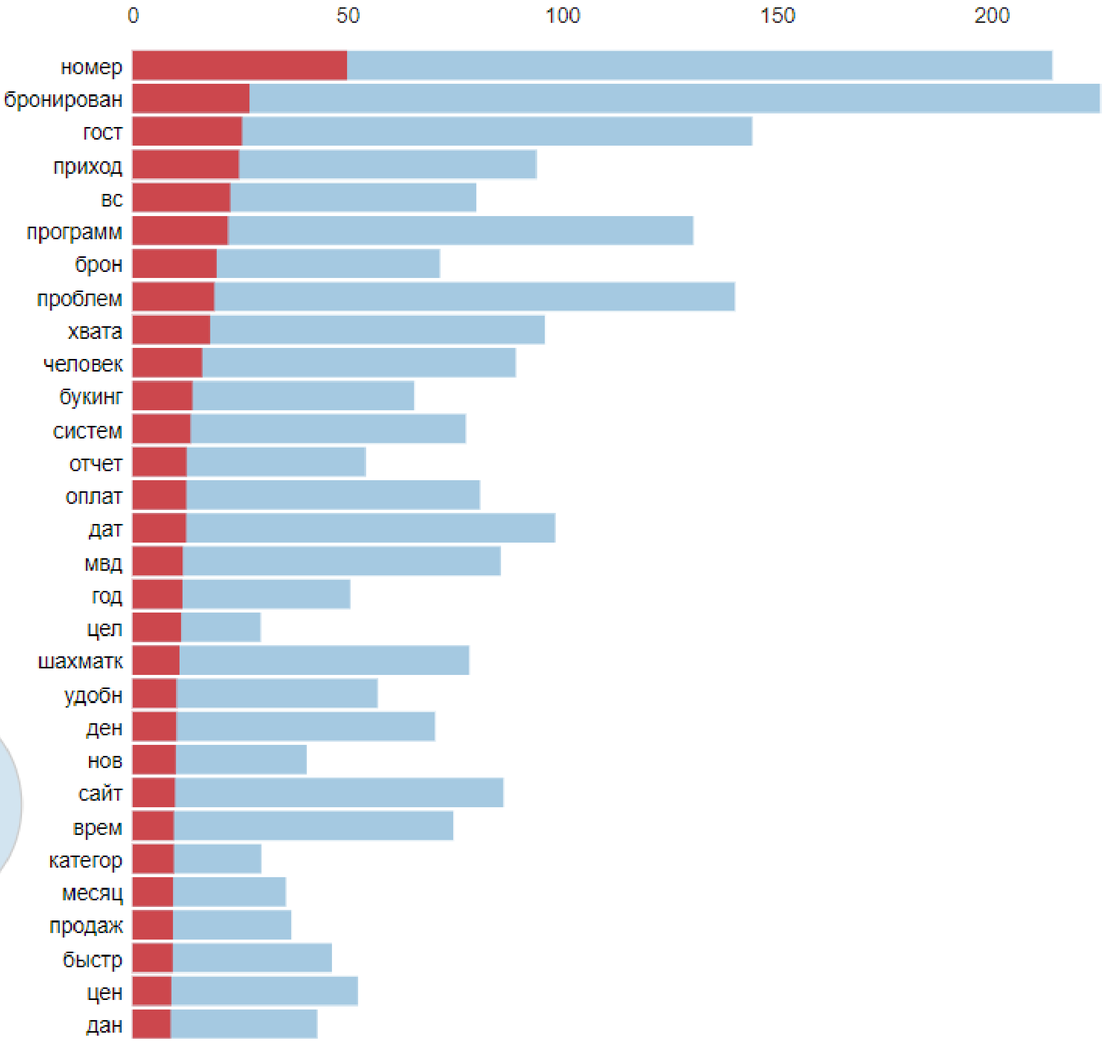
0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)

Marginal topic distribution

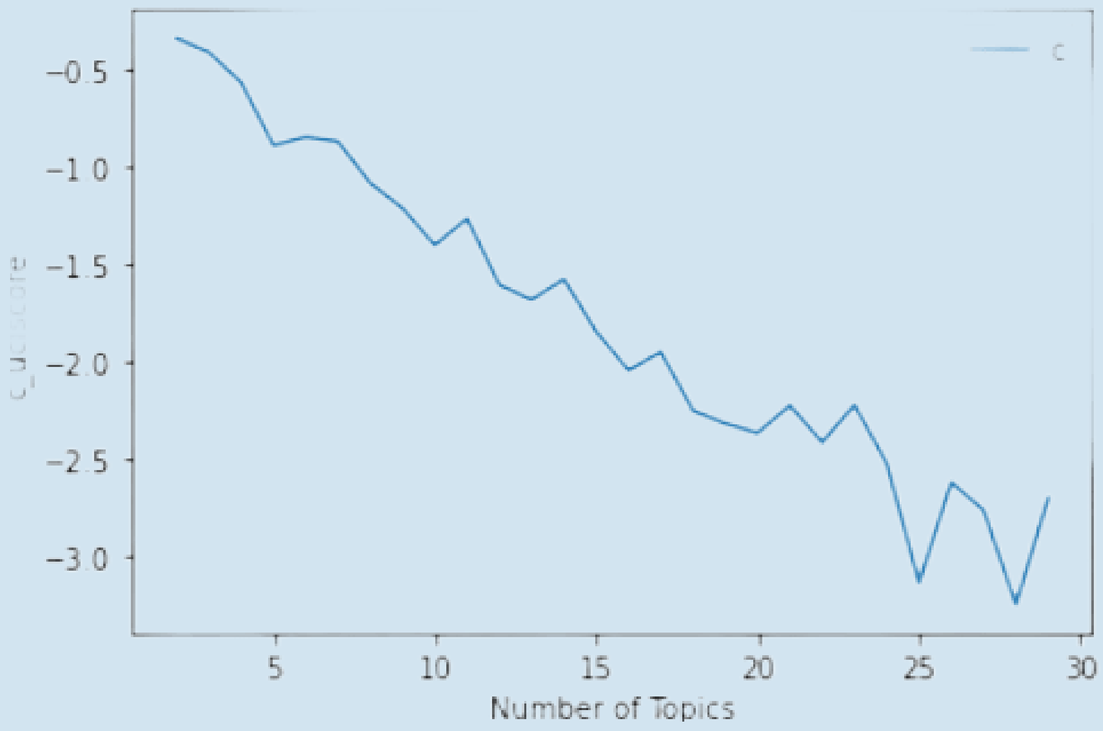


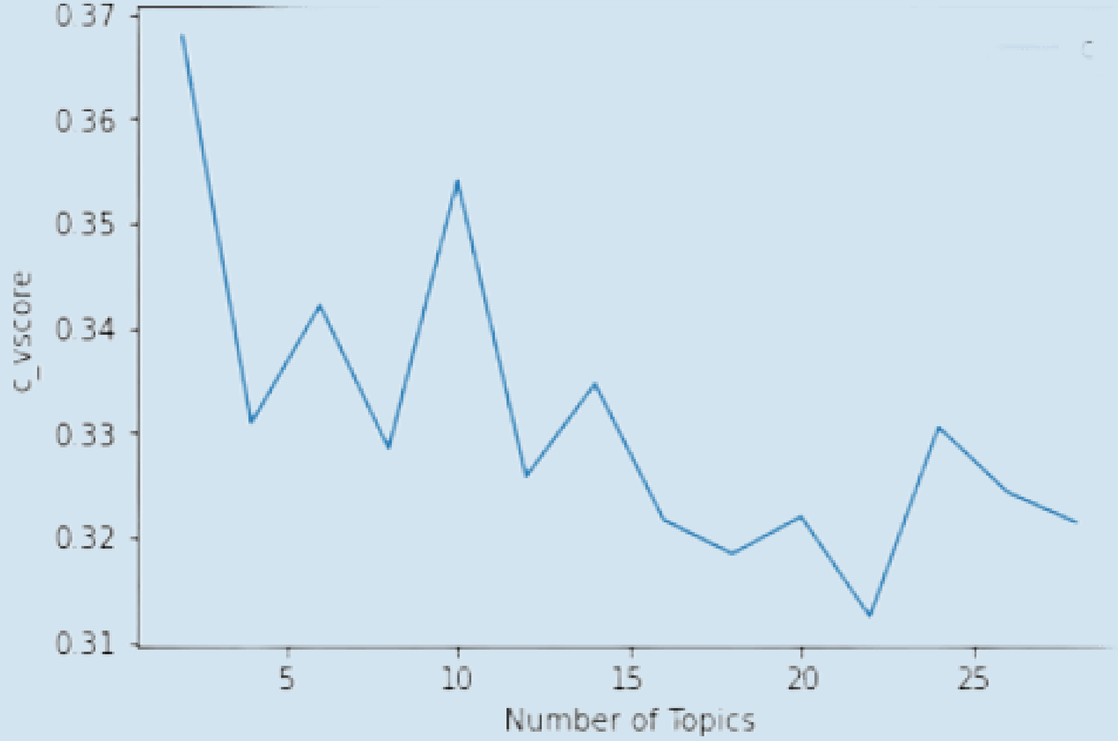
Top-30 Most Relevant Terms for Topic 1 (14.9% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Недоумевание: -7.797711406326151
Согласованность: 0.33515353226516625





```
[(0,
  '0.020*"проблем" + 0.016*"номер" + 0.011*"хвата" + 0.010*"программ" + '
  '0.009*"бронирован" + 0.009*"сайт" + 0.009*"человек" + 0.008*"прост" + '
  '0.007*"видет" + 0.007*"гост"'),
 (1,
  '0.021*"номер" + 0.011*"бронирован" + 0.011*"гост" + 0.010*"приход" + '
  '0.010*"вс" + 0.009*"программ" + 0.008*"брон" + 0.008*"проблем" + '
  '0.008*"хвата" + 0.007*"человек"'),
 (2,
  '0.019*"бронирован" + 0.011*"гост" + 0.009*"номер" + 0.008*"вс" + '
  '0.008*"систем" + 0.007*"цен" + 0.007*"тариф" + 0.006*"человек" + '
  '0.006*"модул" + 0.006*"отчет"'),
 (3,
  '0.020*"фмс" + 0.012*"номер" + 0.009*"дат" + 0.007*"дан" + 0.007*"вопрос" + '
  '0.007*"программ" + 0.006*"врем" + 0.006*"бронирован" + 0.006*"гост" + '
  '0.006*"скб"'),
 (4,
  '0.020*"бронирован" + 0.018*"номер" + 0.013*"гост" + 0.009*"дат" + '
  '0.008*"букинг" + 0.008*"программ" + 0.007*"возможн" + 0.007*"шахматк" + '
  '0.007*"хотет" + 0.006*"человек"'),
 (5,
  '0.014*"бронирован" + 0.013*"оплат" + 0.013*"номер" + 0.011*"хотет" + '
  '0.011*"программ" + 0.009*"дат" + 0.008*"менеджер" + 0.007*"гост" + '
  '0.007*"проблем" + 0.006*"дела"'),
 (6,
  '0.012*"проблем" + 0.009*"человек" + 0.009*"мвд" + 0.007*"вс" + '
  '0.007*"программ" + 0.007*"дат" + 0.007*"сайт" + 0.007*"чек" + '
  '0.007*"приход" + 0.006*"касс"'),
 (7,
  '0.026*"бронирован" + 0.012*"номер" + 0.011*"программ" + 0.010*"приход" + '
  '0.010*"мвд" + 0.009*"шахматк" + 0.008*"проблем" + 0.008*"гост" + '
  '0.006*"сайт" + 0.005*"ден"'),
 (8,
  '0.018*"хвата" + 0.016*"проблем" + 0.011*"гост" + 0.009*"обычн" + '
  '0.009*"сайт" + 0.007*"номер" + 0.006*"мвд" + 0.006*"дат" + 0.006*"сво" + '
  '0.005*"постоя"')]
```

0. Проблема бронирования номерного фонда в программе
1. Бронирование номера гостем (оффлайн)
2. Проблема сопоставления тарифа в программе
3. Корректность данных отправляемых в ФМС
4. Бронирование номера гостем с букинга
5. Проблема оплаты бронирования
6. Проблема введения кассовой дисциплины
7. Бронирование номера в программе, расчет суток
8. Проблема постановки на учет в МВД

Было ли что-то интересное?

КОНЕЧНО

Хотели оптимизировать

Проработали стоп-слова

И у них получилось!

Хотели оптимизировать

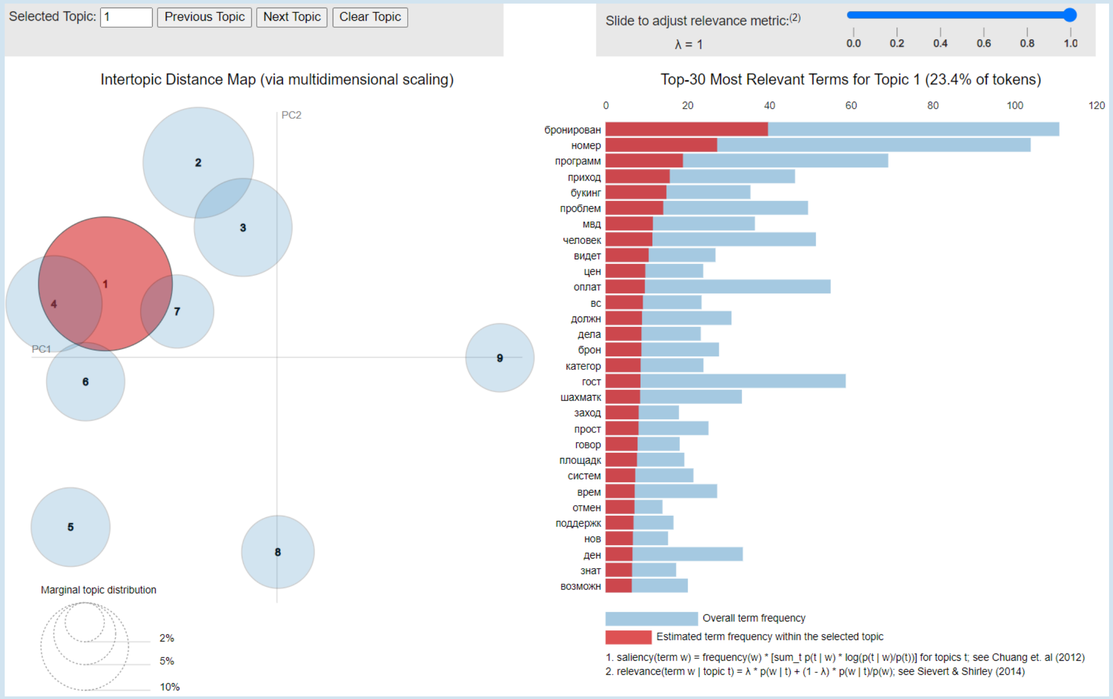
Warning!

gensim v.4.2

больше не поддерживает необходимые методы

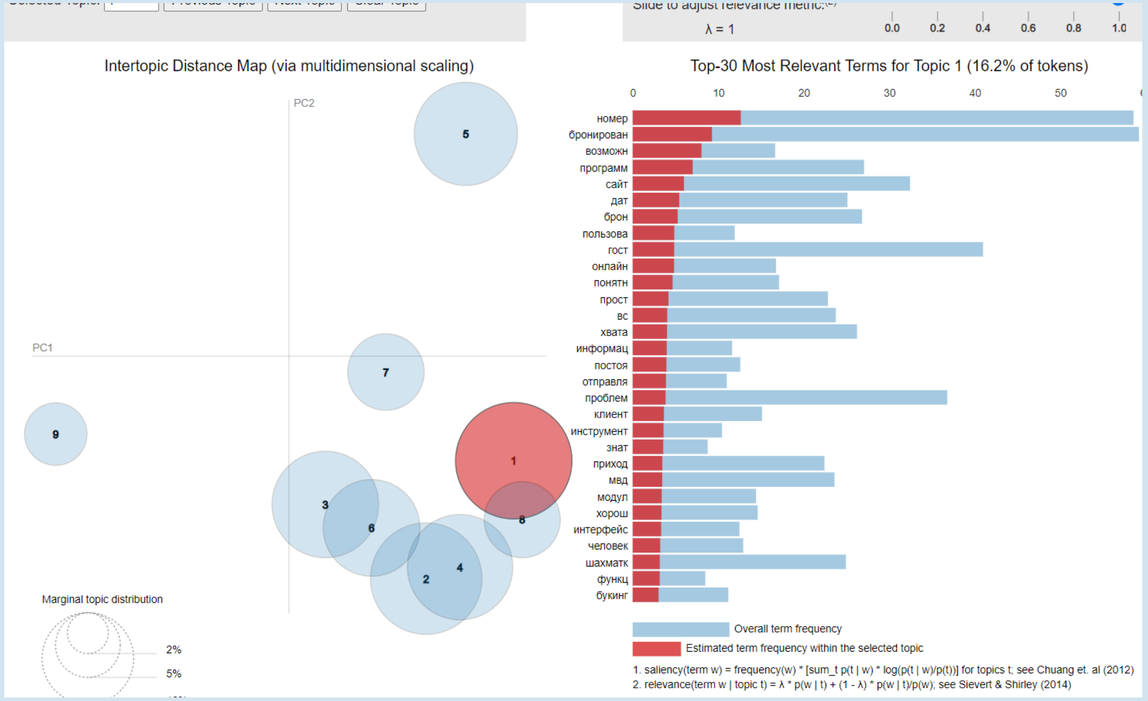
Разделение на DataFrames

Критики



Недоумевание: -7.791838288677627
Согласованность: 0.347034202177979

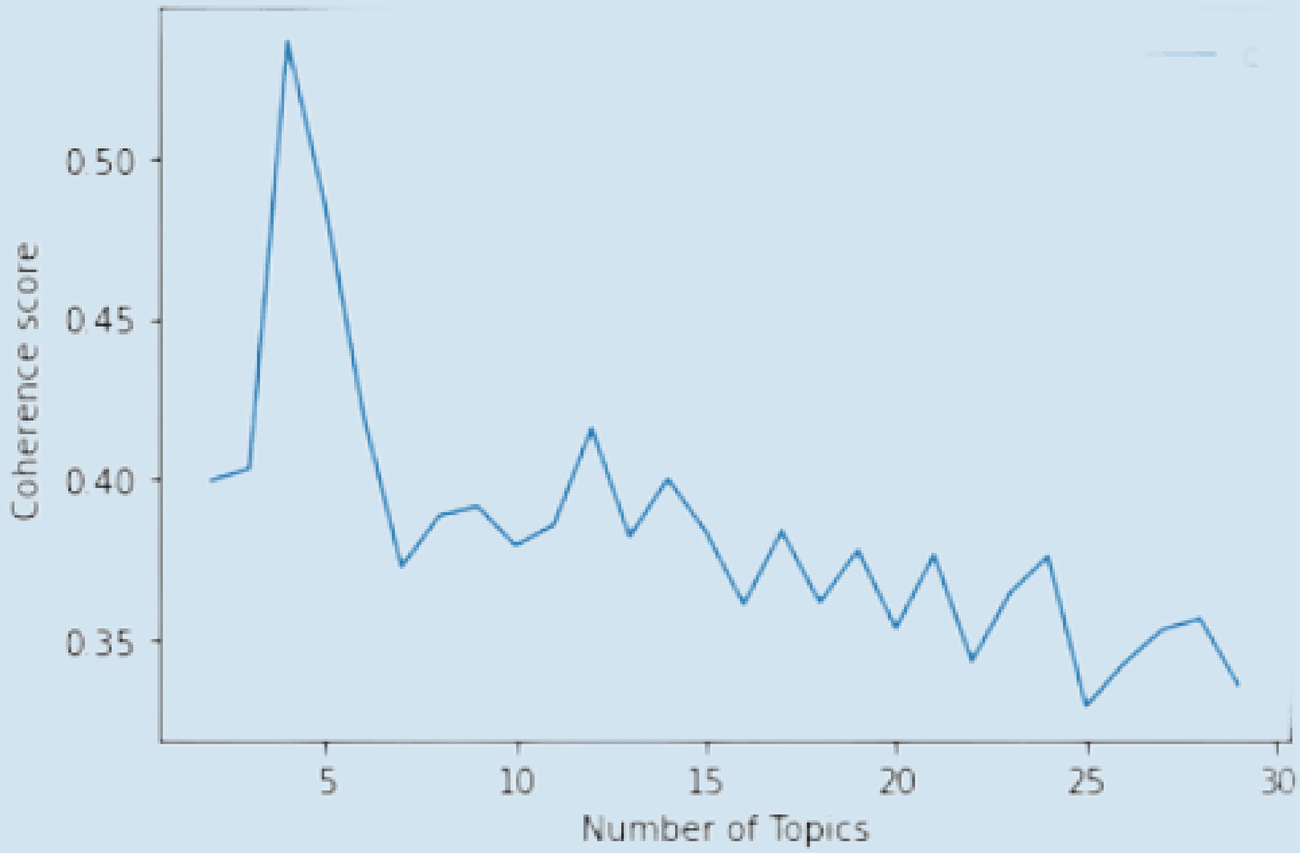
Промоутеры



Недоумевание: -7.765739013919867
Согласованность: 0.290426013309044 9

LSA

Согласованность: 0.4626934457180264



LSA

```
[(0,
  '0.534*"номер" + 0.289*"бронирован" + 0.217*"человек" + 0.186*"программ" + '
  '0.153*"оплат" + 0.149*"гост" + 0.140*"ден" + 0.139*"сайт" + 0.124*"категор'
  '+ 0.124*"видет"'),
 (1,
  '0.413*"номер" + -0.356*"бронирован" + 0.165*"строк" + 0.156*"страниц" + '
  '-0.150*"приход" + 0.142*"категор" + 0.142*"заезжа" + 0.140*"сайт" + '
  '-0.133*"букинг" + -0.132*"вс"'),
 (2,
  '0.494*"бронирован" + 0.332*"букинг" + -0.226*"фмс" + -0.215*"дат" + '
  '-0.215*"программ" + -0.173*"гост" + 0.166*"номер" + -0.134*"проблем" + '
  '0.133*"отмен" + -0.119*"сутк"'),
 (3,
  '-0.324*"систем" + -0.324*"поздн" + -0.298*"выезд" + -0.296*"ден" + '
  '0.218*"бронирован" + -0.216*"след" + -0.162*"человек" + -0.151*"продаж" + '
  '-0.143*"постав" + 0.136*"шахматк"'),
 (4,
  '-0.338*"фмс" + 0.243*"должн" + -0.232*"номер" + 0.228*"шахматк" + '
  '0.205*"видет" + 0.201*"сумм" + -0.199*"букинг" + -0.188*"гост" + '
  '0.178*"клиент" + -0.129*"дат"')]
```

0) Корректное бронирование категории номера с сайта
1)Бронирование с букнига соответствие информации категории номера
2)Бронирование номера с букинга гостем проблема отмены бронирования
3)Учет системой позднего выезда/ раннего заезда из шахматки
4)Учет суммы долга в шахматке при оплате пост.счет

А есть ли практическая польза?

Да, для диплома

Спасибо за внимание!

Тематическое моделирование отзывов Контур.Отель

Байтенова Асыл

Житков Алексей