LDA тематическое моделирование с использованием pyro

План

- Задача
- Тренировочный и тестовый сет
- LDA
- Pyro
- Схема процесса
- Примеры
- Результаты

Задача

- Подготовить тренировочный и тестовый наборы данных из webis-ambient15 - https://webis.de/data/webis-ambient-15.html
- Вычислить LDA-модель корпуса webis-ambient15
- Проецировать обычные текстовые документы в пространство LDA
- Проецировать темы в пространство LDA
- Реализовать в pyro - https://docs.pyro.ai/en/stable/

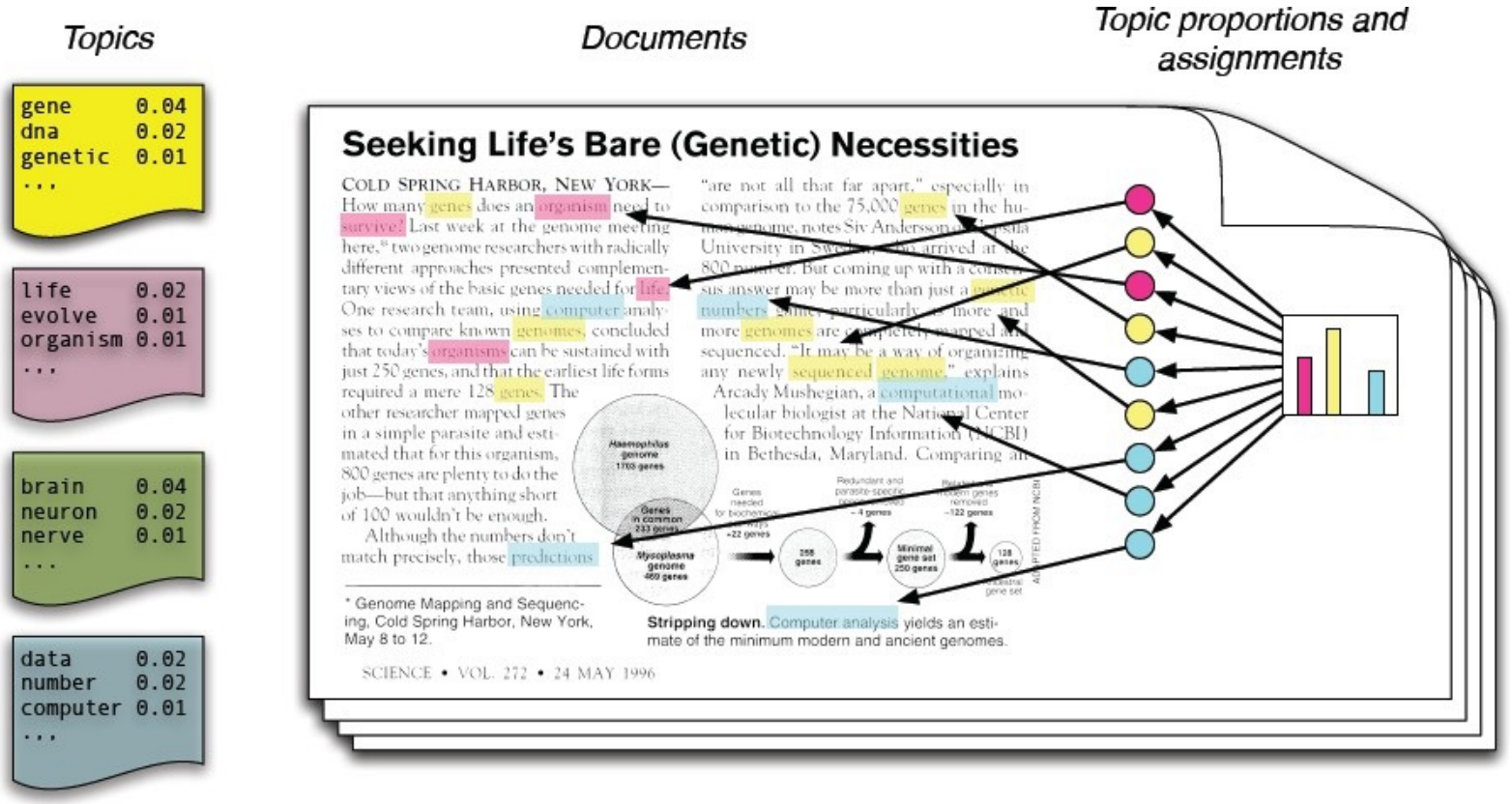
Webis-Ambient-15: Документы

- 3.100 http://www.dog-tracker.com/breeds/beagle.html Beagle, puppies, dog, breeder, stud service, standard Beagle breed info and standard, with free ads for puppies for sale , stud service and a lot more. ... Known as the "singing Beagle," he has a sweet hunting ...
- 4.1 http://en.wikipedia.org/wiki/Bronx The Bronx - Wikipedia, the free encyclopedia ... five boroughs were independent cities, the Bronx would rank as the ninth most ... The name refers to the Bronx River, and rivers are commonly referred to with the ...
- 5.99 http://www.amazon.ca/s?ie=UTF8&keywords=cain&index=books&page=1 Amazon.ca: cain: Books ... Decision Making, conflict by Jim Cain and Barry Jolliff (Paperback - Nov 1997) ... From America's #1 Food Magazine by Anne Chappell Cain (Hardcover - Sep 15 2006) ...
- 6.93 http://www.camelsaust.com.au/ Camels Australia Export - Central Australia - Northern Territory - Australia ... Camel Industry ... sustainable development of the camel industry through the use, ... estimated the present feral camel population in the Northern ...
- 7.65 http://www.coralseavillastobago.com/ Coral Sea Villas Tobago Coral Sea Villas, Bon Accord, Tobago. Caribbean villas, 5 minutes from the airport and close to beaches, shops, restaurants
- …

Webis-Ambient-15: Темы

- 3 Beagle
- 4 Bronx
- 5 Cain
- 6 Camel
- 7 Coral Sea
- 8 Cube
- 9 Eos
- 10 Excalibur
- 32 Purple Haze
- 33 Raam
- 34 Rhea
- 35 Scorpion
- 36 The Little Mermaid
- 37 Tortuga
- 38 Urania
- 39 Wink
- 40 Xanadu
- …

LDA: процесс


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
Topic proportions and assignments
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
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical tasks +22 genes
256 genes
Redundant and parasite-specific genes removed ~4 genes
Minimal gene set 250 genes
Recent modern genes removed ~122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
SCIENCE • VOL. 272 • 24 MAY 1996

LDA: модель

1. Маша любит зеленые и синие вещи.
2. Петя нравятся красные и розовые вещи.
3. Даша любит прямоугольники и треугольники.
4. Дима нравятся круги.
5. Паша любит зеленые треугольники.

Распределение по предложениям:

Предложение 1 и 2: 100% цветов темы

Предложение 3 и 4: 100% тематических форм

Предложение 5: 50% цвета темы, 50% формы темы

Topic "Colors": { 20% зеленый, 20% синий, 20% красный, 20% розовый, ... }

Тема "Формы": { 30% кругов, 30% прямоугольников, 30% треугольников, ... }

LDA: процедура

Семплирование Гиббса:

Набор документов с предполагаемым количеством тем T

Случайно присваиваем тему каждому слову в каждом документе

Для каждого документа d Для каждого слова w

Вычислить p(t | d) // долю слов в документе d, которые относятся к теме t

Вычислить p(w | t) // доля назначений на тему t среди всех документов, которые приходятся на данное слово w.

Присвоить w новую тему t с вероятностью p (тема t | документ d) * p (слово w | тема t).

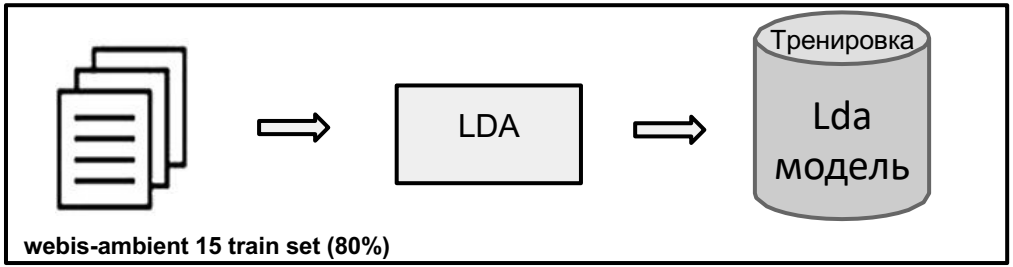
Pyro: модель

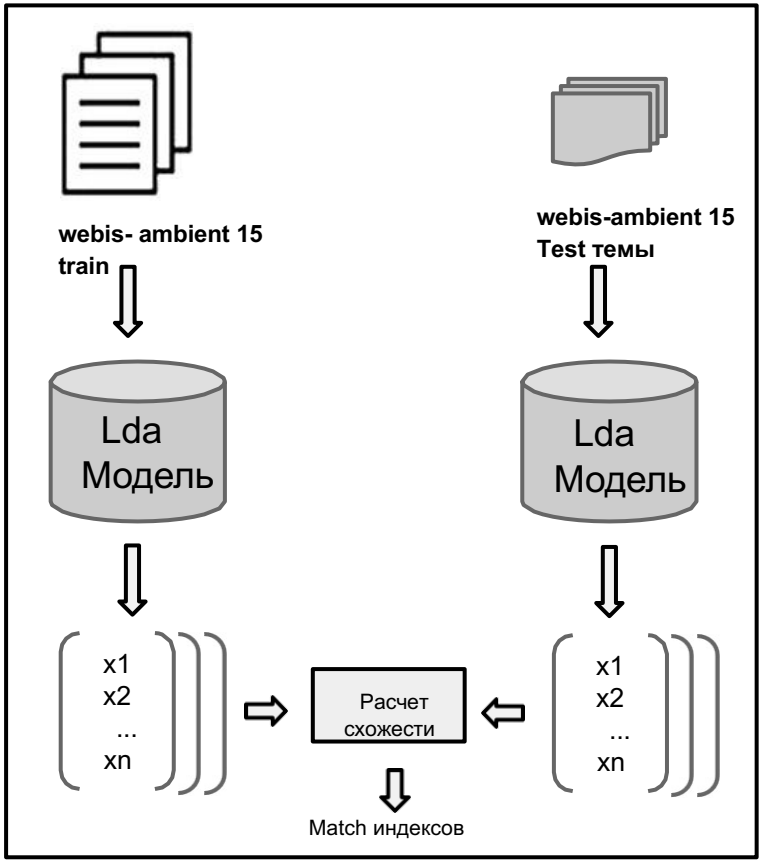
- Библиотека тематического моделирования, реализованная на языке программирования Python
- Использует фреймворк pytorch
- Подаем на вход данные как массив идентификаторов слов в форме [num_words_per_doc, num_documents]
- Используем вывод локальных тематических переменных, достигаемый с помощью многослойного перцептрона
- Используем распределения Гамма и Дирихле
- Задействуем Adam оптимизатор и gradient clipping

Pyro: параметры модели

- num-topics
- num-words
- num-docs
- num-words-per-doc
- num-steps
- layer-sizes
- learning-rate
- batch-size

Схема процесса





Примеры тем

Пример 1

20.5*query + 10.6*queries + 3.1*index + 1.3*database + 1.0*answer + 0.9 *indexing + 0.7*databases + 0.6*answers + 0.6*indexes + 0.6*querying

Пример 2

13.0*server + 8.5*client + 5.3*request + 4.1*requests + 3.5*clients + 2.0 *session + 1.7*proxy + 1.5*remote + 1.3*response + 1.0*service

Пример 3

19.5*pp + 10.3*fig + 8.7*proc + 5.2*vol + 3.5*ii + 2.3*conf + 1.9*iii + 1.8*he + 1.6*unit + 1.5*int

Результаты

Topic x1 - 82.61%

0.042*translation + 0.035*word + 0.023*english + 0.015*alignment + 0.013*phrase + 0.013*chinese + 0.011*sentence + 0.009*mt + 0.009*corpus + 0.009*target

Topic x2 - 78.05%

0.020*word + 0.013*corpus + 0.013*sentence + 0.010*lexical + 0.010*verb + 0.009*noun + 0.009
*semantic + 0.009*sentences + 0.009*syntactic + 0.008*linguistics

Topic x3 – 80.11%

27 : 0.068*items + 0.058*item + 0.026*trust + 0.024*recommendation + 0.021*ratings + 0.019*rating + 0.018*profile + 0.015*recommendations + 0.014*recommender + 0.014*collaborative