

## ▼ Topic Modeling and Gibbs Sampling

Задача: описать текст через распределение весов по некоторому фиксированному набору тем. Например, для набора тегов Политика, Военные сражения, Спорт, Интернет, Драма представить роман как вектор (0.3, 0.2, 0, 0, 0.5), а статью в газете про допинг в велоспорте как вектор (0.1, 0, 0.7, 0, 0.2).

Для чего, например, это нужно: имея векторное представление для текстов, тексты можно кластеризовать, находить похожие.

Условие: даны только набор текстов и количество тем.

## ▼ Немного теории

Будем представлять текст как неупорядоченный набор слов (Bag-of-words model). Предположим, для каждого тега выбрано распределение  $\phi_k$  над списком всевозможных слов (словарем). Вектор  $\phi_k$  - это вектор длины  $N$  из неотрицательных величин, в сумме дающих 1. Вектора  $\phi_k$  не зависят от текста, они заданы заранее. Будем считать, что  $\phi_k$  заданы распределением Дирихле  $Dir(\beta)$ . Теперь, чтобы собрать текст  $d$  из  $n$  слов, будем действовать следующим образом:

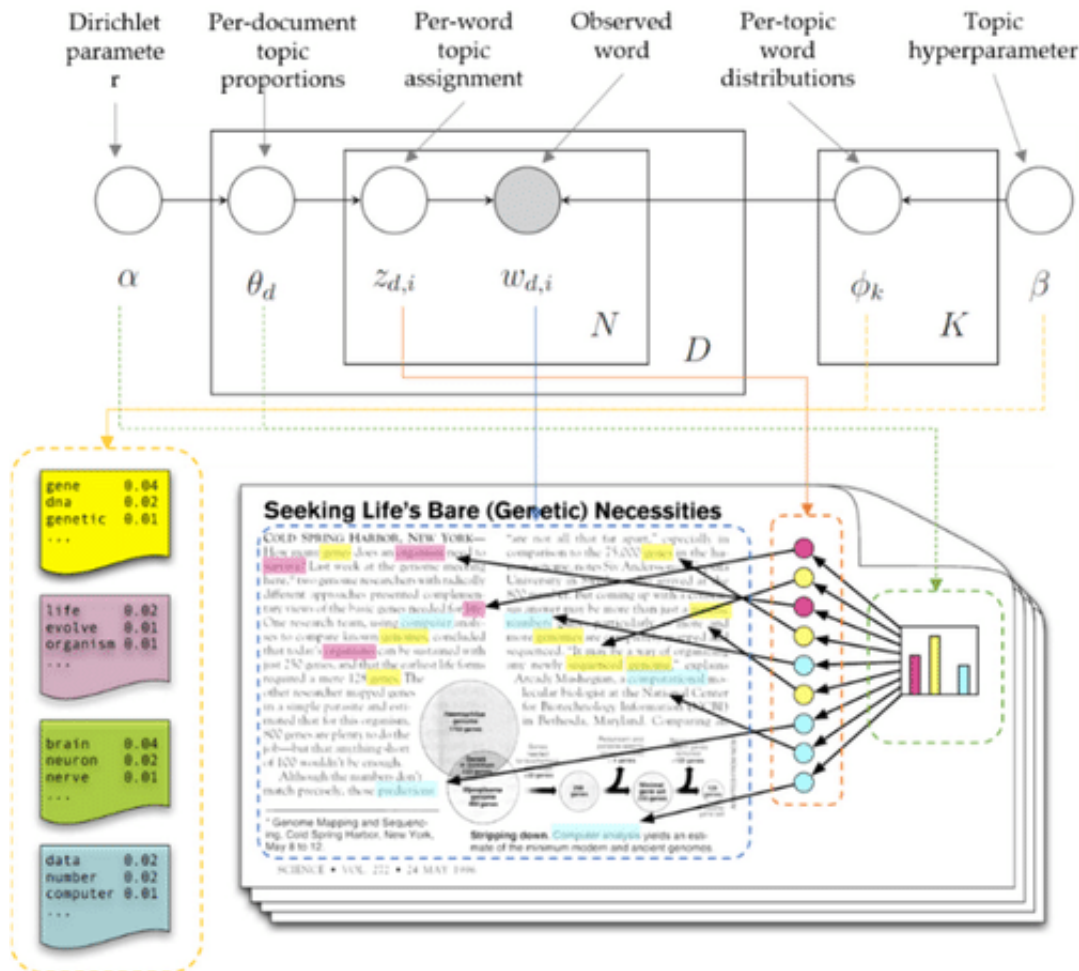
- выберем распределение для тегов  $\theta_d$ . Вновь,  $\theta_d$  - это вектор длины  $K$  из неотрицательных величин, в сумме дающих 1. Поэтому естественно брать  $\theta_d \sim Dir(\theta|\alpha)$
- Для  $i$  от 1 до  $n$ :
  - выберем тег  $z_i$  согласно распределению  $\theta_d$
  - выберем слово  $w_i$  из распределения для данного тега, т.е.  $w_i \sim \phi_{z_i}$
  - добавляем слово  $w_i$  в текст.

Полученная модель называется моделью LDA (Latent Dirichlet Allocation). Описанная схема генерации текста задает совместное распределение скрытых и наблюдаемых параметров по всем текстам корпуса размера  $N$ :

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi | \alpha, \beta) = Dir(\theta | \alpha) Dir(\phi | \beta) \prod_i Cat(z_i | \theta) Cat(w_i | \phi_{z_i}).$$

Здесь  $\mathbf{w}$  и  $\mathbf{z}$  обозначают вектора слов и тегов по всем текстам,  $\theta$  - набор из  $\theta_d$  для каждого текста,  $\phi$  - набор из  $\phi_k$  для каждого тега (матрица  $K \times N$ ).

(a) LDA document generation process



(b) An illustrative example of LDA document generation process

(c) Two outputs of LDA

(c-1) Per-document top

	Topic 1	Topic
Doc 1	0.20	0.50
Doc 2	0.50	0.02
Doc 3	0.05	0.12
...	...	...
Doc N	0.14	0.25

(c-2) Per-topic word di

	Topic 1	Topic
word 1	0.01	0.05
word 2	0.02	0.02
word 3	0.05	0.12
...	...	...
word N	0.04	0.01

Наша задача - восстановить распределение  $p(\mathbf{z}, \theta, \phi | \mathbf{w}, \alpha, \beta)$ .

Немного упростим жизнь, и поставим себе задачей восстановить распределение  $p(\mathbf{z} | \mathbf{w}, \alpha, \beta) = \int \int p(\mathbf{z}, \theta, \phi | \mathbf{w}, \alpha, \beta) d\theta d\phi$ .

В этот момент на помощь приходит алгоритм Gibbs Sampling. Напомним, для оценки на  $\mathbf{z} = (z_1, z_2, \dots, z_m)$  используется схема:

$$z_i^{(t)} \sim p(z_i^{(t)} | z_1 = z_1^t, \dots, z_{i-1} = z_{i-1}^t, z_{i+1} = z_{i+1}^{t-1}, z_m = z_m^{t-1}).$$

Условные распределения выводятся так. Сначала замечаем, что

$$p(z_i | \mathbf{z}_{\hat{i}}, \mathbf{w}, \alpha, \beta) = \frac{p(z_i, \mathbf{z}_{\hat{i}}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{z}_{\hat{i}}, \mathbf{w} | \alpha, \beta)} = \frac{p(\mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{z}_{\hat{i}}, \mathbf{w} | \alpha, \beta)}.$$

Здесь  $\mathbf{z}_{\hat{i}}$  - вектор без  $i$ -ой компоненты.

Далее расписываем:

$$p(\mathbf{z}, \mathbf{w} | \alpha, \beta) = \int \int p(\mathbf{z}, \mathbf{w}, \theta, \phi | \alpha, \beta) d\theta d\phi = \int \int \text{Dir}(\theta | \alpha) \text{Dir}(\phi | \beta) \text{Cat}(\mathbf{z} | \theta) \text{Cat}(\mathbf{w} | \phi) d\theta d\phi$$

и обнаруживаем, что оба интеграла в последнем выражении вычисляются аналитически

$$\int \text{Dir}(\theta | \alpha) \text{Cat}(\mathbf{z} | \theta) d\theta = \prod_d \int \text{Dir}(\theta_d | \alpha) \text{Cat}(\mathbf{z}_d | \theta_d) d\theta_d = \prod_d \int \frac{1}{B(\alpha)} \prod_k \theta_{d,k}^{\alpha-1} \prod_i \theta_{d,z_i} d\theta_d$$

$$\prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(\alpha)}$$

Здесь  $n_{d,k}$  - количество тэгов  $k$  в тексте  $d$ ,  $n_{d,\cdot}$  - вектор длины  $K$  из этих величин.

$$\text{Аналогично, второй интеграл } \int \text{Dir}(\phi | \beta) \text{Cat}(\mathbf{w} | \phi) d\phi = \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)},$$

где  $n_{k,\cdot}$  - вектор длины  $N$  встречаемости слов внутри тэга  $k$ .

Получаем:

$$p(\mathbf{z}, \mathbf{w} | \alpha, \beta) = \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(\alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)}.$$

$$\text{Теперь } p(z_i | \mathbf{z}_{\hat{i}}, \mathbf{w}, \alpha, \beta) = \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(\hat{n}_{d,\cdot} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\hat{n}_{k,\cdot} + \beta)}.$$

Выражение упрощается дальше, расписывая бета-функцию через гамма-функции. Напом

$B(x_1, \dots, x_k) = \frac{\Gamma(x_1) \dots \Gamma(x_k)}{\Gamma(x_1 + \dots + x_k)}$ , а также  $\Gamma(n) = (n-1)\Gamma(n-1)$ . Получим:

$$p(z_i = s | \mathbf{z}_{\hat{i}}, \mathbf{w}, \alpha, \beta) \propto (n_{d,s} + \alpha_s - 1) \frac{n_{s,w_i} + \beta_s - 1}{\sum_w (n_{s,w} + \beta_w) - 1}.$$

Знак  $\propto$  означает пропорциональность с точностью до коэффициентов, независящих от  $i$

С этого места можно полностью собрать алгоритм моделирования плотности  $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ , количество слов, отнесенных к тегу  $k$ ,  $W$  - общее количество слов в корпусе,  $\bar{\beta} = \sum_w \beta_w$

Алгоритм:

- заведем счетчики  $n_{k,w}, n_{d,k}, n_k$
- случайным образом расставим теги словам, обновим счетчики  $n_{k,w}, n_{d,k}, n_k$
- пока не сойдемся к стационарному режиму:
  - для каждого  $i$  от 1 до  $W$ :
    - для каждого  $k$  от 1 до  $K$ :
      - вычисляем  $p(z_i = k | \dots) = (n_{d,k} + \alpha_k - 1) \frac{n_{k,w_i} + \beta_k - 1}{n_k + \bar{\beta} - 1}$
      - сэмплим новый  $z_i$  из полученного распределения  $p(z_i = k | \dots)$

Восстановив распределение для  $\mathbf{z}$ , можем оценить  $\theta$  и  $\phi$ , о которых мы ненадолго забыли через матожидание по апостериорным распределениям. Получите формулы самостоятельно.

Литература:

<http://u.cs.biu.ac.il/~89-680/darling-lda.pdf>

<https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture20-topic-models.pdf>


Перейдем к практике.

## ▼ Датасет

Возьмем популярный датасет [20 Newsgroups](#), встроенный в пакет `sklearn`. Датасет состоит из классифицированных на 20 категорий. Датасет разбит на `train` и `test`. Для загрузки используйте `fetch_20newsgroups`, в параметрах указать, что мета-информацию о тексте загружать не нужно.

```
import numpy as np
from sklearn.datasets import fetch_20newsgroups

newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers', 'footers'))
```

 Downloading 20news dataset. This may take a few minutes.  
 Downloading dataset from <https://ndownloader.figshare.com/files/5975967> (14 MB)

Выведем список категорий текстов:

```
newsgroups_train.target_names
```

```
[ 'alt.atheism',  
  'comp.graphics',  
  'comp.os.ms-windows.misc',  
  'comp.sys.ibm.pc.hardware',  
  'comp.sys.mac.hardware',  
  'comp.windows.x',  
  'misc.forsale',  
  'rec.autos',  
  'rec.motorcycles',  
  'rec.sport.baseball',  
  'rec.sport.hockey',  
  'sci.crypt',  
  'sci.electronics',  
  'sci.med',  
  'sci.space',  
  'soc.religion.christian',  
  'talk.politics.guns',  
  'talk.politics.mideast',  
  'talk.politics.misc',  
  'talk.religion.misc']
```


Атрибут `target` хранит номера категорий для текстов из обучающей выборки:

```
newsgroups_train.target[:10]
```

```
array([ 7,  4,  4,  1, 14, 16, 13,  3,  2,  4])
```

Доступ к самим текстам через атрибут `data`. Выведем текст и категорию случайного при

```
n = 854
print('Topic = {0}\n'.format(newsgroups_train.target_names[newsgroups_train.target_names.index(newsgroups_train.data[n])])
print(newsgroups_train.data[n])
```

 Topic = rec.motorcycles


```
hey... I'm pretty new to the wonderful world of motorcycles... I just
bought
a used 81 Kaw KZ650 CSR from a friend.... I was just wondering what kind of
saddle bags I could get for it (since I know nothing about them) are there
bags for the gas tank? how much would some cost, and how much do they
hold?
thanks for your advice!!! I may be new to riding, but I love it
already!!!!
:)
```

## ▼ Векторное представление текста

Представим текст как вектор индикаторов вхождения слов из некоторого словаря в текст. Сформируем словарь на основе нашего набора текстов. Для этого используем модуль `CountVectorizer` из библиотеки `sklearn`.


```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.stop_words import ENGLISH_STOP_WORDS

vectorizer = CountVectorizer(lowercase=True, stop_words=ENGLISH_STOP_WORDS,
                             analyzer='word', binary=True)
vectorizer.fit(newsgroups_train.data)
```

 `CountVectorizer(analyzer='word', binary=True, decode_error='strict', dtype=<class 'numpy.int64'>, encoding='utf-8', input='count', lowercase=True, max_df=1.0, max_features=None, min_df=1, ngram_range=(1, 1), preprocessor=None, stop_words=frozenset({'a', 'about', 'above', 'across', 'after', 'afterwards', 'again', 'against', 'all', 'almost', 'alone', 'along', 'already', 'also', 'although', 'always', 'am', 'among', 'amongst', 'amoungst', 'amour', 'an', 'and', 'another', 'any', 'anyhow', 'anyone', 'anything', 'anyway', 'anywhere', ...}), strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b', tokenizer=None, vocabulary=None)`


Количество проиндексированных слов:

```
len(vectorizer.vocabulary_)
```

 101322

Проиндексированные слова и их индексы:

```
vectorizer.vocabulary_
```

 { 'wondering': 96879,  
 'enlighten': 37256,  
 'car': 25717,  
 'saw': 80420,  
 'day': 31927,  
 'door': 34741,  
 'sports': 84312,  
 'looked': 57247,  
 'late': 55606,  
 '60s': 9843,  
 'early': 35902,  
 '70s': 11174,  
 'called': 25437,  
 'bricklin': 24108,  
 'doors': 34742,  
 'really': 76269,  
 'small': 83208,  
 'addition': 16806,  
 'bumper': 24583,  
 'separate': 81450,  
 'rest': 77676,  
 'body': 23430,  
 'know': 54493,  
 'tellme': 87913,  
 'model': 62594,  
 'engine': 37208,  
 'specs': 84050,  
 'years': 99608,  
 'production': 73174,  
 'history': 46690,  
 'info': 49800,  
 'funky': 41874,  
 'looking': 57250,  
 'mail': 59071,  
 'fair': 39296,  
 'number': 66680,  
 'brave': 23973,  
 'souls': 83779,  
 'upgraded': 92389,  
 'si': 82337,  
 'clock': 27889,  
 'oscillator': 68519,  
 'shared': 81848,

```
'experiences': 38637,  
'poll': 72039,  
'send': 81378,  
'brief': 24125,  
'message': 60923,  
'detailing': 33127,  
'procedure': 73122,  
'speed': 84088,  
'attained': 20236,  
'cpu': 30233,  
'rated': 75904,  
'add': 16791,  
'cards': 25769,  
'adapters': 16776,  
'heat': 45997,  
'sinks': 82666,  
'hour': 47310,  
'usage': 92588,  
'floppy': 40727,  
'disk': 34011,  
'functionality': 41836,  
'800': 12266,  
'floppies': 40725,  
'especially': 37867,  
'requested': 77460,  
'summarizing': 86076,  
'days': 31942,  
'network': 65377,  
'knowledge': 54504,  
'base': 21688,  
'upgrade': 92386,  
'haven': 45775,  
'answered': 18757,  
'thanks': 88277,  
'folks': 40950,  
'mac': 58777,  
'plus': 71778,  
'finally': 40284,  
'gave': 42680,  
'ghost': 43239,  
'weekend': 95865,  
'starting': 84921,  
'life': 56530,  
'512k': 8677,  
'way': 95674,  
'1985': 3483,  
'sooo': 83714,  
'market': 59552,  
'new': 65475,  
'machine': 58818,  
'bit': 22913,  
'sooner': 83713,  
'intended': 50302,  
'picking': 71139,
```



'powerbook': 72399,  
'160': 2940,  
'maybe': 59935,  
'180': 3248,  
'bunch': 24591,  
'questions': 75095,  
'hopefully': 47151,  
'somebody': 83660,  
'answer': 18755,  
'does': 34617,  
'anybody': 18883,  
'dirt': 33760,  
'round': 79037,  
'introductions': 50597,  
'expected': 38604,  
'heard': 45975,  
'185c': 3301,  
'supposed': 86308,  
'make': 59153,  
'appearence': 19072,  
'summer': 86079,  
'anymore': 18889,  
'don': 34696,  
'access': 16417,  
'macleak': 58864,  
'rumors': 79373,  
'price': 72952,  
'drops': 35113,  
'line': 56676,  
'like': 56591,  
'ones': 67929,  
'duo': 35385,  
'just': 52861,  
'went': 95966,  
'recently': 76397,  
'impression': 49251,  
'display': 34113,  
'probably': 73091,  
'swing': 86648,  
'got': 43914,  
'80mb': 12356,  
'120': 2223,  
'feel': 39787,  
'better': 22475,  
'yea': 99602,  
'looks': 57254,  
'great': 44241,  
'store': 85349,  
'wow': 97051,  
'good': 43852,  
'solicit': 83610,  
'opinions': 68088,  
'people': 70436,  
'use': 92600.

```
-----,
'worth': 97019,
'taking': 87383,
'size': 82759,
'money': 62794,
'hit': 46692,
'active': 16690,
'realize': 76261,
'real': 76241,
'subjective': 85740,
'question': 75083,
've': 93598,
'played': 71620,
'machines': 58826,
'computer': 28881,
'breifly': 24044,
'figured': 40178,
'actually': 16715,
'uses': 92624,
'daily': 31623,
'prove': 73551,
'helpful': 46163,
'hellcats': 46134,
'perform': 70494,
'advance': 17003,
'email': 36809,
'll': 56988,
'post': 72291,
'summary': 86077,
'news': 65517,
'reading': 76218,
'time': 88835,
'premium': 72761,
'finals': 40285,
'corner': 29852,
'tom': 89184,
'willis': 96403,
'twillis': 90490,
'ecn': 36082,
'purdue': 73974,
'edu': 36234,
'electrical': 36612,
'engineering': 37212,
'weitek': 95921,
'address': 16818,
'phone': 70985,
'information': 49822,
'chip': 26997,
'article': 19689,
'c5owcb': 25113,
'n3p': 64543,
'world': 96983,
'std': 85008,
'com': 28405,
'tombaker': 89195.
```

```
'baker': 21413,  
'understanding': 91640,  
'errors': 37739,  
'basically': 21725,  
'known': 54509,  
'bugs': 24497,  
'warning': 95535,  
'software': 83557,  
'things': 88484,  
'checked': 26809,  
'right': 78195,  
'values': 93325,  
'aren': 19421,  
'set': 81587,  
'till': 88819,  
'launch': 55653,  
'suchlike': 85935,  
'fix': 40440,  
'code': 28151,  
'possibly': 72290,  
'introduce': 50592,  
'tell': 87908,  
'crew': 30414,  
'ok': 67681,  
'213': 4520,  
'liftoff': 56550,  
'ignore': 48697,  
'course': 30091,  
'term': 88042,  
'rigidly': 78222,  
'defined': 32452,  
'doubt': 34809,  
'using': 92637,  
'quote': 75177,  
'allegedly': 17943,  
'read': 76207,  
'presenting': 72823,  
'argument': 19462,  
'weapons': 95811,  
'mass': 59702,  
'destruction': 33112,  
'commonly': 28575,  
'understood': 91650,  
'switching': 86666,  
'topics': 89281,  
'point': 71950,  
'evidently': 38232,  
'allowed': 18006,  
'later': 55612,  
'analysis': 18442,  
'given': 43387,  
'consider': 29305,  
'class': 27692,  
'responded': 77648
```

```

    responded': 77410,
    'request': 77459,
    'treatment': 89867,
    'astrocytomas': 20036,
    'couldn': 30015,
    'thank': 88272,
    'directly': 33745,
    'bouncing': 23749,
    'probs': 73111,
    'sean': 81078,
    'debra': 32138,
    'sharon': 81860,
    'thought': 88567,
    'publicly': 73811,
    'sure': 86350,
    'glad': 43476,
    'accidentally': 16434,
    'rn': 78616,
    'instead': 50188,
    'rm': 78551,
    'trying': 90158,
    'delete': 32588,
    'file': 40217,
    'september': 81478,
    'hmmm': 46823,
    'shows': 82222,
    'scsi': 80945,
    'controler': 29587,
    'range': 75816,
    '5mb': 9463,
    'controller': 29593,
    '6mb': 10776,
    '10mb': 1883,
    'burst': 24698,
    'note': 66349,
    'increase': 49478,
    'quadra': 74958,
    'version': 93839,
    'exist': 38534,
    'pc': 70140,
    'mode': 62589,
    '16': 2939,
    'wide': 96303,
    'fast': 39508,
    '12mb': 2402,
    '20mb': 4463,
    '32': 6203,
    '15': 2777,
    '40mb': 7521,
    'data': 31820,
    'twice': 90481,
    'esdi': 37829,
    'correct': 29884,
    'reach': 76185,
    'on': 4306

```

```
20 : 4500,  
'faster': 39515,  
'ide': 48521,  
'96': 13781,  
'facts': 39243,  
'posted': 72303,  
'newsgroup': 65530,  
'ibm': 48358,  
'sheet': 81934,  
'available': 20583,  
'ftp': 41728,  
'sumex': 86062,  
'aim': 17514,  
'stanford': 84859,  
'36': 6576,  
'44': 7699,  
'report': 77367,  
'compare': 28642,  
'txt': 90546,  
'173': 3156,  
'161': 2956,  
'problem': 73103,  
'inconsiant': 49447,  
'documented': 34589,  
'apple': 19094,  
'salesperson': 80105,  
'said': 80055,  
'maximum': 59921,  
'synchronous': 86803,  
'ansynchronous': 18760,  
'slower': 83137,  
'interface': 50378,  
'think': 88487,  
'driven': 35083,  
'true': 90100,  
'win': 96428,  
'downloaded': 34856,  
'icons': 48451,  
'bmp': 23319,  
'figure': 40177,  
'change': 26651,  
'wallpaper': 95442,  
'help': 46158,  
'appreciated': 19162,  
'thanx': 88284,  
'brando': 23947,  
'board': 23375,  
'year': 99605,  
'work': 96948,  
'diskdoubler': 34014,  
'autodoubler': 20502,  
'licensing': 56494,  
'stac': 84760,  
'technologies': 87783,  
'camera': 60024
```

```
owners': 69024,  
'compression': 28828,  
'technology': 87784,  
'writing': 97166,  
'memory': 60760,  
'lost': 57331,  
'reference': 76700,  
'wrong': 97178,  
'problems': 73108,  
'hard': 45557,  
'say': 80430,  
'fault': 39560,  
'decompress': 32266,  
'troubled': 90075,  
'recompress': 76488,  
'icon': 48441,  
'usually': 92683,  
'reappears': 76295,  
'mentioned': 60812,  
'freeware': 41493,  
'expansion': 38592,  
'utility': 92736,  
'dd': 32033,  
'expand': 38582,  
'compressed': 28823,  
'unless': 91980,  
'installed': 50172,  
'product': 73172,  
'unlikely': 91985,  
'holes': 46964,  
'related': 76993,  
'fixed': 40445,  
'sad': 79982,  
'makes': 59163,  
'reluctant': 77108,  
'buy': 24796,  
'stinky': 85238,  
'hey': 46359,  
'competition': 28723,  
'ducati': 35282,  
'900gts': 13404,  
'1978': 3473,  
'17k': 3230,  
'runs': 79398,  
'paint': 69449,  
'bronze': 24239,  
'brown': 24271,  
'orange': 68235,  
'faded': 39254,  
'leaks': 55875,  
'oil': 67645,  
'pops': 72167,  
'1st': 4107,  
'accel': 16382,  
'shock': 82127
```

```
shop : 82121,  
'trans': 89639,  
'leak': 55870,  
'sold': 83591,  
'bike': 22729,  
'owner': 69023,  
'want': 95488,  
'3495': 6425,  
'thinking': 88493,  
'3k': 7148,  
'nice': 65711,  
'stable': 84757,  
'mate': 59762,  
'beemer': 22088,  
'jap': 51774,  
'axis': 20810,  
'motors': 63120,  
'tuba': 90259,  
'irwin': 50998,  
'honk': 47105,  
'computrac': 28894,  
'richardson': 78120,  
'tx': 90525,  
'cmptrc': 28018,  
'lonestar': 57217,  
'org': 68325,  
'dod': 34594,  
'0826': 784,  
'r75': 75486,  
'yep': 99646,  
'pretty': 72907,  
'jew': 52036,  
'understand': 91636,  
'jewish': 52041,  
'jews': 52045,  
'believe': 22223,  
'covenant': 30119,  
'yhwh': 99729,  
'patriarchs': 70006,  
'abraham': 16258,  
'moses': 63039,  
'case': 25910,  
'establishes': 37910,  
'moral': 62940,  
'follow': 40954,  
'mankind': 59362,  
'decide': 32200,  
'boundaries': 23751,  
'fall': 39341,  
'sadducees': 79994,  
'believed': 22224,  
'torah': 89307,  
'required': 77469,  
'pharisees': 70881,  
'-----': 10007
```

'ancestors': 18487,  
'modern': 62628,  
'judaism': 52725,  
'interpretation': 50470,  
'lead': 55850,  
'morality': 62944,  
'nuances': 66647,  
'talmud': 87426,  
'essence': 37892,  
'biblical': 22653,  
'man': 59263,  
'christian': 27148,  
'necessarily': 65141,  
'indicate': 49564,  
'outside': 68760,  
'relationship': 77001,  
'speculate': 84073,  
'trouble': 90074,  
'created': 30359,  
'image': 48959,  
'means': 60461,  
'different': 33540,  
'come': 28443,  
'conclusion': 28979,  
'upsets': 92440,  
'cart': 25870,  
'wants': 95493,  
'script': 80907,  
'shaky': 81798,  
'foundation': 41256,  
'mix': 62034,  
'metaphors': 60982,  
'unashamedly': 91409,  
'living': 56885,  
'christ': 27138,  
'example': 38371,  
'little': 56849,  
'jesus': 52019,  
'person': 70636,  
'recorded': 76520,  
'utterances': 92777,  
'narratives': 64835,  
'followers': 40957,  
'references': 76702,  
'comtemporany': 28905,  
'historians': 46683,  
'revelation': 77874,  
'aside': 19815,  
'second': 81132,  
'hand': 45418,  
'worse': 97006,  
'attempt': 20247,  
'debunk': 32151,  
'christianity': 27152,  
.....



```
'initially': 49945,  
'bible': 22649,  
'interpret': 50469,  
'humanity': 47648,  
'guess': 44639,  
'faith': 39317,  
'relevation': 77041,  
'comes': 28449,  
'inherent': 49906,  
'subjectiveness': 85742,  
'absolute': 16289,  
'undoubtedly': 91694,  
'multiple': 63839,  
'codes': 28164,  
'founded': 41258,  
'parent': 69723,  
'child': 26952,  
'swear': 86584,  
'assume': 19982,  
'swears': 86586,  
'simply': 82583,  
'told': 89166,  
'trooper': 90063,  
'pub': 73796,  
'bar': 21560,  
'children': 26962,  
'wrongness': 97183,  
'disobeys': 34057,  
'inappropriate': 49351,  
'quite': 75154,  
'happy': 45533,  
'animals': 18610,  
'analogy': 18434,  
'hold': 46950,  
'water': 95614,  
'knows': 54510,  
'type': 90580,  
'gist': 43376,  
'incidentally': 49396,  
'young': 99907,  
'considers': 29314,  
'directive': 33743,  
'gets': 43138,  
'older': 67763,  
'piaget': 71115,  
'learns': 55898,  
'david': 31891,  
'religion': 77072,  
'description': 32989,  
'external': 38785,  
'tank': 87474,  
'option': 68165,  
'ssf': 84651,  
'redesign': 76600,  
- - - - -
```

'deleted': 32589,  
'yo': 99869,  
'ken': 53721,  
'let': 56169,  
'wingless': 96495,  
'orbiter': 68251,  
'options': 68169,  
'list': 56790,  
'today': 89130,  
'23': 4670,  
'edition': 36199,  
'york': 99893,  
'times': 88854,  
'reports': 77376,  
'connor': 29238,  
'panel': 69561,  
'proposals': 73425,  
'dropped': 35111,  
'giant': 43260,  
'fuel': 41778,  
'tanks': 87481,  
'used': 92603,  
'launching': 55658,  
'space': 83871,  
'shuttles': 82293,  
'building': 24510,  
'station': 84965,  
'existing': 38545,  
'shuttle': 82292,  
'wings': 96499,  
'tail': 87346,  
'removed': 77197,  
'currently': 30989,  
'considered': 29312,  
'presented': 72819,  
'advisory': 17057,  
'meeting': 60605,  
'yesterday': 99662,  
'reported': 77371,  
'low': 57387,  
'cost': 29982,  
'modular': 62660,  
'approach': 19181,  
'studied': 85618,  
'team': 87738,  
'msfc': 63573,  
'teams': 87743,  
'jsc': 52656,  
'larc': 55541,  
'supporting': 86303,  
'srt': 84592,  
'crystal': 30707,  
'city': 27499,  
'lerc': 56138,

'reston': 77692,  
'site': 82719,  
'locations': 57127,  
'helping': 46166,  
'respective': 77639,  
'activities': 16697,  
'key': 53808,  
'features': 39747,  
'bus': 24710,  
'developed': 33229,  
'lockheed': 57133,  
'qualified': 74982,  
'sts': 85595,  
'elv': 36790,  
'provides': 73569,  
'propulsion': 73452,  
'gn': 43666,  
'communications': 28598,  
'management': 59274,  
'air': 17534,  
'force': 41042,  
'power': 72393,  
'capability': 25648,  
'obtained': 67294,  
'flights': 40669,  
'solar': 83583,  
'arrays': 19620,  
'provide': 73563,  
'kw': 54972,  
'vehicle': 93654,  
'flies': 40667,  
'arrow': 19652,  
'optimize': 68158,  
'microgravity': 61430,  
'environment': 37419,  
'spacelab': 83884,  
'missions': 61947,  
'utilize': 92738,  
'vehilce': 93657,  
'source': 83802,  
'30': 6044,  
'human': 47639,  
'tended': 87969,  
'opposed': 68115,  
'old': 67759,  
'sexist': 81666,  
'achieved': 16570,  
'common': 28571,  
'module': 62665,  
'modified': 62646,  
'lab': 55284,  
'docking': 34562,  
'ports': 72229,  
'added': 16795,

```
'international': 50437,  
'partners': 69858,  
'labs': 55311,  
'place': 71514,  
'nodes': 66110,  
'docked': 34559,  
'60': 9762,  
'habitability': 45187,  
'eva': 38126,  
'nasda': 64854,  
'esa': 37783,  
'modules': 62666,  
'permanent': 70552,  
'presence': 72812,  
'3rd': 7312,  
'habitation': 45191,  
'acrv': 16670,  
'assured': 19994,  
'return': 77833,  
'freedom': 41474,  
'derived': 32940,  
'based': 21698,  
'mike': 61561,  
'griffin': 44345,  
'alot': 18051,  
'design': 33018,  
'love': 57373,  
'assumes': 19984,  
'lightweight': 56581,  
'assembly': 19916,  
'computed': 28880,  
'51': 8668,  
'inclination': 49405,  
'orbit': 68247,  
'build': 24505,  
'occurs': 67354,  
'phases': 70894,  
'initial': 49935,  
'research': 77504,  
'reached': 76187,  
'transferred': 89685,  
'visits': 94232,  
'adopted': 16951,  
'non': 66154,  
'language': 55481,  
'deployed': 32863,  
'10': 1469,  
'keeping': 53675,  
'orbiters': 68252,  
'logistics': 57191,  
'supply': 86295,  
'tolerance': 89169,  
'14': 2593,  
'2nd': 5688,
```

'thermal': 88410,  
'control': 29584,  
'radiator': 75631,  
'internationals': 50442,  
'finish': 40324,  
'24': 4749,  
'systems': 86883,  
'exception': 38399,  
'major': 59141,  
'changes': 26659,  
'reduced': 76636,  
'prices': 72956,  
'forsale': 41194,  
'behalf': 22143,  
'brother': 24260,  
'moving': 63196,  
'moved': 63183,  
'offer': 67520,  
'black': 23022,  
'decker': 32228,  
'duster': 35435,  
'portable': 72208,  
'vaccum': 93242,  
'purchased': 73968,  
'12': 2222,  
'sr': 84531,  
'1000': 1471,  
'dual': 35259,  
'cassette': 25936,  
'player': 71621,  
'fm': 40810,  
'band': 21496,  
'graphics': 44152,  
'equalizer': 37586,  
'high': 46527,  
'dubing': 35273,  
'tape': 87499,  
'deck': 32225,  
'treble': 89873,  
'sound': 83782,  
'bet': 22453,  
'fixable': 40441,  
'80': 12265,  
'25': 4899,  
'monolux': 62839,  
'zoom': 100977,  
'microscope': 61457,  
'1200x': 2230,  
'magnification': 59017,  
'japan': 51775,  
'includes': 49413,  
'accessories': 16428,  
'50': 8581,  
'sunbeam': 86108,

'1400': 2595,  
'hair': 45268,  
'dryer': 35154,  
'head': 45919,  
'salons': 80131,  
'ask': 19822,  
'bro': 24188,  
'everylast': 38211,  
'bag': 21361,  
'leather': 55908,  
'brand': 23943,  
'osterizer': 68572,  
'pusle': 74040,  
'matic': 59800,  
'blender': 23128,  
'speeds': 84102,  
'cookbook': 29688,  
'binolux': 22820,  
'binoculars': 22819,  
'7x35': 12225,  
'extra': 38803,  
'angle': 18580,  
'525ft': 8740,  
'1000yds': 1536,  
'proctor': 73154,  
'silex': 82499,  
'spray': 84355,  
'steam': 85032,  
'dry': 35151,  
'iron': 50927,  
'contact': 29439,  
'reply': 77358,  
'expeditiously': 38614,  
'included': 49410,  
'lastly': 55597,  
'reasonable': 76308,  
'look': 57244,  
'happened': 45523,  
'japanese': 51777,  
'citizens': 27490,  
'war': 95496,  
'ii': 48746,  
'prepared': 72773,  
'stick': 85191,  
'concentration': 28939,  
'camp': 25509,  
'trial': 89919,  
'short': 82145,  
'step': 85097,  
'gassing': 42622,  
'nazis': 64974,  
'originally': 68386,  
'imprison': 49259,  
'final': 40279,

'solution': 83641,  
'dreamt': 35029,  
'partly': 69855,  
'afford': 17210,  
'run': 79380,  
'camps': 25528,  
'devastation': 33222,  
'caused': 26052,  
'goering': 43761,  
'total': 89373,  
'weren': 95971,  
'gassed': 42618,  
'generally': 42923,  
'died': 33500,  
'malnutrition': 59232,  
'disease': 33957,  
'certainly': 26431,  
'tiff': 88778,  
'works': 96972,  
'philosophically': 70960,  
'complexity': 28771,  
'led': 55941,  
'programs': 73244,  
'poor': 72142,  
'writers': 97163,  
'making': 59171,  
'bizarre': 22959,  
'files': 40237,  
'inability': 49329,  
'load': 57077,  
'images': 48972,  
'save': 80403,  
'general': 42913,  
'interchange': 50346,  
'environments': 37426,  
'despite': 33074,  
'fact': 39225,  
'saying': 80434,  
'goes': 43763,  
'worried': 97001,  
'abuse': 16331,  
'chalk': 26599,  
'immense': 49051,  
'unnecessary': 92029,  
'format': 41154,  
'words': 96938,  
'spec': 84005,  
'appendix': 19086,  
'page': 69414,  
'capitalized': 25674,  
'emphasis': 36958,  
'sort': 83759,  
'success': 85911,  
'designed': 33024,

'powerful': 72403,  
'flexible': 40646,  
'expense': 38628,  
'simplicity': 82576,  
'takes': 87380,  
'effort': 36334,  
'handle': 45454,  
'specification': 84031,  
'application': 19116,  
'complete': 28755,  
'job': 52351,  
'able': 16214,  
'import': 49217,  
'generating': 42930,  
'applications': 19119,  
'program': 73220,  
'won': 96869,  
'deal': 32085,  
'trapped': 89800,  
'agrees': 17378,  
'reasoning': 76314,  
'asking': 19831,  
'kind': 54125,  
'rates': 75906,  
'single': 82647,  
'male': 59191,  
'drivers': 35086,  
'yrs': 99983,  
'paying': 70081,  
'performance': 70498,  
'cars': 25863,  
'replies': 77355,  
'received': 76389,  
'27': 5071,  
'close': 27905,  
'1992': 3495,  
'dodge': 34601,  
'stealth': 85030,  
'rt': 79238,  
'twin': 90491,  
'turbo': 90333,  
'300hp': 6061,  
'tickets': 88752,  
'accidents': 16436,  
'house': 47314,  
'taken': 87375,  
'defensive': 32428,  
'driving': 35093,  
'airbag': 17539,  
'abs': 16277,  
'security': 81179,  
'alarm': 17757,  
'1500': 2779,  
'500': 8582,



```
'decut': 32319,  
'state': 84941,  
'farm': 39461,  
'insurance': 50247,  
'additional': 16807,  
'100': 1470,  
'000': 1,  
'umbrella': 91316,  
'policy': 72011,  
'standard': 84835,  
'300': 6045,  
'bought': 23738,  
'company': 28634,  
'accident': 16432,  
'ticket': 88750,  
'11': 1912,  
'quoted': 75178,  
'hope': 47148,  
'helps': 46169,  
'steve': 85158,  
'flynn': 40808,  
'university': 91945,  
'delaware': 32572,  
'45': 7764,  
'kevin': 53793,  
'remembered': 77150,  
'correctly': 29892,  
'asked': 19826,  
'similar': 82546,  
'situation': 82731,  
'inquiry': 50071,  
'age': 17301,  
'eagle': 35883,  
'talon': 87427,  
'tsi': 90186,  
'awd': 20738,  
'record': 76519,  
'clean': 27737,  
'illinois': 48906,  
'820': 12437,  
...}
```

Индекс, например, для слова anyone:

```
vectorizer.vocabulary_.get('car')
```


 25717

А теперь преобразуем строку в вектор:

```
text = 'I was wondering if anyone out there could enlighten me on this car I saw'
x = vectorizer.transform([text])
```

Какой тип имеет объект, на который указывает `x`?

```
type(x)
```


```
 scipy.sparse.csr.csr_matrix
```

Разреженная матрица!

## ▼ Отступление про разреженные матрицы


Список ненулевых элементов матрицы:

```
x.data
```

```
 array([1, 1, 1, 1])
```


Индексы строк и столбцов для ненулевых элементов:

```
x.nonzero()
```

```
 (array([0, 0, 0, 0], dtype=int32),
    array([25717, 37256, 80420, 96879], dtype=int32))
```


Преобразование к объекту `ndarray` (именно после приведения к такому виду разреженнь в функции, например, библиотеки Numpy):

```
x.toarray()
```

```
 array([[0, 0, 0, ..., 0, 0, 0]])
```

Вернемся к словарю. Раскодируем вектор `x` в список слов:

```
vectorizer.inverse_transform(x)
```

```
 [array(['car', 'enlighten', 'saw', 'wondering'], dtype='<U81')]
```

Пропало слово `1`. Но дело в том, что по умолчанию `CountVectorizer` отбрасывает последние символы. На это указывает параметр `token_pattern='(?u)\b\w+\b'`.

Переведем весь набор текстов обучающего датасета в набор векторов, получим матрицу

```
X_train = vectorizer.fit_transform(newsgroups_train.data)
X_train.shape
```



О пользе разреженных матриц. Отношение числа ненулевых элементов ко всем элементам

```
X_train.nnz / np.prod(X_train.shape)
```



Задача: запустить модель LDA и Gibbs Sampling с числом тегов 20. Вывести топ-10 слов и полученных тегов с тегами из датасета, сделать выводы.