КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

МЕТОД ОПРЕДЕЛЕНИЯ ТЕМАТИКИ МАТЕМАТИЧЕСКИХ ДОКУМЕНТОВ НА ОСНОВЕ ВЕРОЯТНОСТНОЙ МОДЕЛИ СКРЫТОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ

Выполнил студент гр. 05-603 **Альмухаметов Дамир Альбертович**

Специальность: математика и компьютерные науки

Кафедра компьютерной математики и информатики

Научный руководитель: **Липачёв Евгений Константинович**

К. ф.-м. н., доц. каф. КМиИ

КАЗАНЬ - 2020

Постановка задачи

- Дана коллекция из D документов, обозначаемая как $D = \{d_1, d_2, \dots, d_D\}$, и словарь терминов $V = \{v_1, v_2, \dots, v_V\}$.

Заранее задано количество тем K.

Найти: параметры тематической модели $p(d|v) = \sum_{k \in K} \vartheta_{dk} \varphi_{kv}$

- $\vartheta_{dk} = p(d|k)$ – вероятность документа d соответствовать темам k. $\Theta = \{\bar{\vartheta}_d\}_{d=1}^{D} - D \times K$-матрица;
- $\varphi_{kv} = p(k|v)$ – вероятность темы k соответствовать терминам v. $\Phi = \{\bar{\varphi}_k\}_{k=1}^{K} - K \times V$-матрица.

Цель работы

- Описать вероятностную модель скрытого размещения Дирихле;
- Выбрать наиболее подходящий метод оценки матриц Θ и Φ для модели скрытого размещения Дирихле;
- Реализовать выбранный метод на высокоуровневом языке программирования Python;
- Протестировать программную реализацию на коллекции русскоязычных математических документов, хранящихся в pdf формате.

Модель скрытого размещения Дирихле

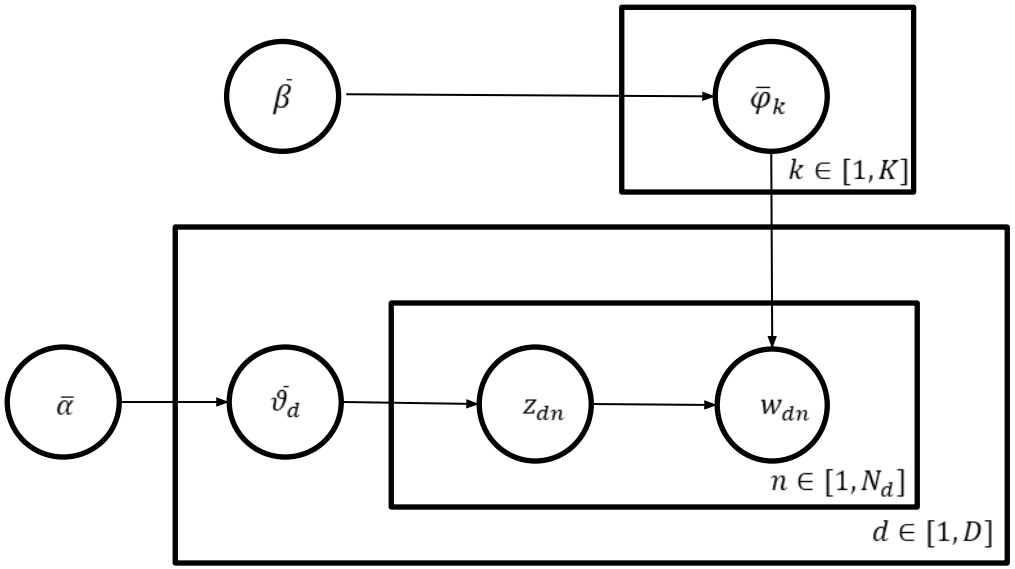

Рис. 1: Байесовская сеть для модели скрытого размещения Дирихле. Внешний блок представляет документы, а внутренний – набор тем и слов в документе.

Наблюдаемые переменные

- N_d - количество слов в документе d
- w_{dn} - слово в документе d

Скрытые переменные

- z_{dn} - значение темы для слова w_{dn}
- $\bar{\vartheta}_d$ - вероятность встретить тему в документе d
- $\bar{\varphi}_k$ - вероятность встретить слово в теме k

Метод оценки матриц Θ и Φ

Модифицированный вариационный вывод Байеса

EM-алгоритм – итерационный процесс , в котором каждая итерация состоит из двух шагов – E (expectation) и M (maximization).

E-шаг:

$$p(k|d,v) = \frac{p(v|k)p(k|d)}{p(v|d)} = \frac{\varphi_{vk}\vartheta_{kd}}{\sum_{s\in K}\varphi_{vs}\vartheta_{sd}}$$

M-шаг:

$$\vartheta_{kd} = \frac{n_{dk} + \alpha_k}{n_d + \alpha_0}$$

$$\varphi_{vk} = \frac{\rho n_{vk} + \beta_v}{\rho n_k + \beta_0}, \rho \in (0,1]$$

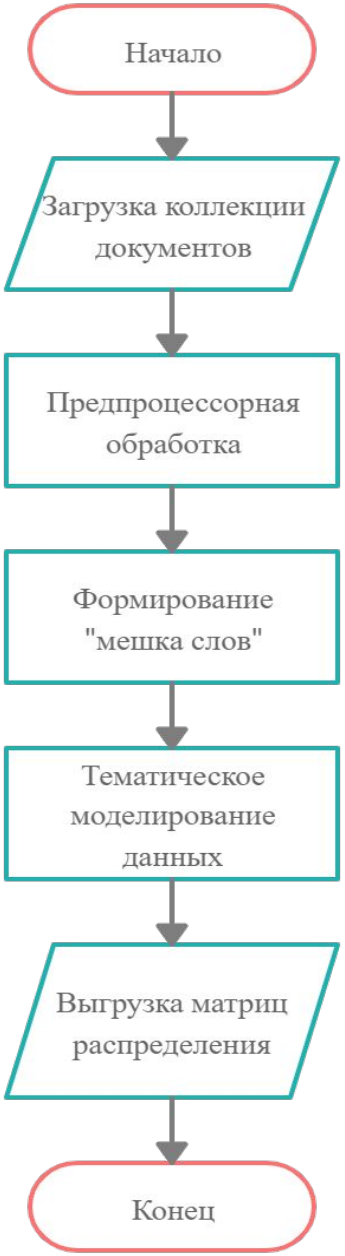

Схема работы программы

Загрузка коллекции документов и предпроцессорная обработка

```
def get_docs_set():
    directory = path_to_docs
    files = os.listdir(directory)
    docs_set = list()
    docs_set_name = list()
    for f in files:
        print(f)
        text = text_cleaner(extract_text(directory + f))
        docs_set.append(text)
        docs_set_name.append(f)
    return (docs_set, docs_set_name)
```

Извлечение текста из PDF документа

```
def extract_text(path_to_file):
    resource_manager = PDFResourceManager()
    file_extract = io.StringIO()
    converter = TextConverter(resource_manager, file_extract)
    interpreter = PDFPageInterpreter(resource_manager, converter)
    with open(path_to_file, 'rb') as file:
        for page in PDFPage.get_pages(file, caching=True, check_extractable=True):
            interpreter.process_page(page)
        text = file_extract.getvalue()
    converter.close()
    file_extract.close()
    return text
```

- Удаление пунктуации;
- Лемматизация слов («формам изгибных колебаний» → «форма изгибный колебание»);
- Удаление стоп слов ({'ну', 'за', 'было', 'тем', 'мне', 'что', 'эти', …}).

```
def text_cleaner(text):
    text = re.sub(r'[^а-яА-ЯёЁ -]+', ' ', text)
    text = re.sub(r'-', '', text)
    text = re.sub(r'\s+', ' ', text)
    text = text.lower()
    lemmat_words = list()
    words = text.split()
    for w in words:
        if not (len(w) == 1 or len(w) > 20):
            p = morph.parse(w)[0]
            lemmat_words.append(p.normal_form)
    text = ' '.join([lw for lw in lemmat_words if lw not in(ru_stop_words)])
    return text
```

Формирование «мешка слов»

Порядок слов в документе не важен, документ – «мешок слов».

```
def parse_doc_set(docs, vocab):
    D_len = len(docs)
    word_ids = list()
    word_cts = list()
    for d in range(D_len):
        words = str.split(docs[d])
        d_dict = dict()
        for w in words:
            if (w in vocab):
                word_token = vocab[w]
                if (not word_token in d_dict):
                    d_dict[word_token] = 0
                d_dict[word_token] += 1
        word_ids.append(list(d_dict.keys()))
        word_cts.append(list(d_dict.values()))
    return (word_ids, word_cts)
```

Процесс тематического моделирования

```
def update_lambda_docs(self, docs):
    rho_t = pow(self._tau_0 + self._update_ct, -self._kappa)
    self._rho_t = rho_t
    (gamma, s_stats) = self.do_e_step_corpus(docs)
    self._lambda = self._lambda * (1 - rho_t) + \
                   rho_t * (self._beta + self._D * s_stats / len(docs))
    self._Elogbeta = dirichlet_expectation(self._lambda)
    self._expElogbeta = numpy.exp(self._Elogbeta)
    return gamma
```

```python
def do_e_step(self, word_ids, word_cts):
    part_D = len(word_ids)
    gamma = 1 * numpy.random.gamma(100., 1. / 100., (part_D, self._K))
    Elogtheta = dirichlet_expectation(gamma)
    expElogtheta = numpy.exp(Elogtheta)
    s_stats = numpy.zeros(self._lambda.shape)
    for d in range(part_D):
        ids = word_ids[d]
        cts = word_cts[d]
        gamma_d = gamma[d, :]
        expElogtheta_d = expElogtheta[d, :]
        expElogbeta_d = self._expElogbeta[:, ids]
        phi_norm = numpy.dot(expElogtheta_d, expElogbeta_d) + 1e-100
        while True:
            last_gamma = gamma_d
            gamma_d = self._alpha + expElogtheta_d * \
                      numpy.dot(cts / phi_norm, expElogbeta_d.T)
            Elogtheta_d = dirichlet_expectation(gamma_d)
            expElogtheta_d = numpy.exp(Elogtheta_d)
            phi_norm = numpy.dot(expElogtheta_d, expElogbeta_d) + 1e-100
            meanchange = numpy.mean(abs(gamma_d - last_gamma))
            if meanchange < mean_change_thresh:
                break
        gamma[d, :] = gamma_d
        s_stats[:, ids] += numpy.outer(expElogtheta_d.T, cts / phi_norm)
    s_stats = s_stats * self._expElogbeta
    return (gamma, s_stats)
```

Тестирование работы программы

В качестве коллекции русскоязычных математических документов были использованы труды математического центра имени Н.И. Лобачевского. Коллекция состояла из 20 томов.

В каждом томе содержались статьи различной направленности и длины, от одной страницы до 34 страниц.

Слов в коллекции: 313992

Уникальных слов: 31082

Результаты обработки коллекции

```
Кол-во слов в докуметне 1 --- 21059
Кол-во слов в докуметне 2 --- 2522
Кол-во слов в докуметне 3 --- 21957
Кол-во слов в докуметне 4 --- 15560
Кол-во слов в докуметне 5 --- 3300
Кол-во слов в докуметне 6 --- 18795
Кол-во слов в докуметне 7 --- 3653
Кол-во слов в докуметне 8 --- 2437
Кол-во слов в докуметне 9 --- 21086
Кол-во слов в докуметне 10 --- 179
Кол-во слов в докуметне 11 --- 147
Кол-во слов в докуметне 12 --- 150
Кол-во слов в докуметне 13 --- 17006
Кол-во слов в докуметне 14 --- 11863
Кол-во слов в докуметне 15 --- 4162
Кол-во слов в докуметне 16 --- 48592
Кол-во слов в докуметне 17 --- 45880
Кол-во слов в докуметне 18 --- 32280
Кол-во слов в докуметне 19 --- 22300
Кол-во слов в докуметне 20 --- 21064
```

```
Тема 118:
        функция    ---    0.01451141
         задача    ---    0.01386783
      уравнение    ---    0.01161890
        решение    ---    0.01070110
        теорема    ---    0.00709562
        условие    ---    0.00689531
        система    ---    0.00678978
          пусть    ---    0.00649745
  пространство     ---    0.00627105
         метод     ---    0.00621360

Тема 139:
          пусть    ---    0.00556454
   литература      ---    0.00517794
  ключевыеслов     ---    0.00376323
        теорема    ---    0.00320365
         матема    ---    0.00263575
   такимобраз      ---    0.00237961
     например      ---    0.00234695
      длявсех      ---    0.00206908
   вчастность      ---    0.00183186
 обозначимчерез    ---    0.00162887
```

Результаты обработки одной статьи из коллекции

УДК 534.121.1

ПАРАМЕТРИЧЕСКИЙ РЕЗОНАНС ПРИ ИЗГИБНО-КРУТИЛЬНЫХ КОЛЕБАНИЯХ ПЛАСТИНЫ

Б. Аффане[1]

1 boudkhil.affane@gmail.com; Казанский (Приволжский) федеральный университет, институт математики и механики им. Н.И. Лобачевского

Статья посвящена исследованию параметрического резонанса крутильных колебаний возникающего при изгибных колебаниях консольных пластин. В работе проводится поиск окон параметрического резонанса крутильных колебаний. Для первых двух мод изгибных колебаний построены окна параметрического резонанса крутильных колебаний.

Ключевые слова: изгибные колебания, крутильные колебания, параметрический резонанс

Введение

Консольно закрепленные пластины стержневого типа, у которых толщина h, ширина b и длина L таковы, что $h \ll b \ll L$ используются, в частности при определении логарифмического декремента колебаний (ЛДК) материала на основе исследования затухающих изгибных колебаний тест-образцов по первой основной моде с учетом аэродинамического демпфирования [2, 3]. Однако при определении ЛДК материалов на основе исследования вынужденных колебаний тест-образцов по второй формам изгибных колебаний, наряду с фиксируемыми изгибными колебаниями, наблюдаются также и высокочастотные крутильные колебания. Для корректного описания указанных механических эффектов в основу базовых уравнений движения положим уточненные геометрически нелинейные уравнения колебаний удлиненных пластин, предложенные в работе [1]. Для простоты, уравнения [1] анализируются в пренебрежении демпфированием (внутренним и внешним). Модель предсказывает возбуждение крутильных колебаний (КК) пластины за счет параметрического резонанса при наличии высокоамплитудных изгибных колебаний (ИК). В данной работе ставится задача теоретического изучения этого явления с целью построения окон параметрического резонанса КК. Задача сводится к решению бесконечной цепочке обыкновенных дифференциальных уравнений с периодическими коэффициентами, эта цепочка усекается и определяются мультипликаторы усеченной системы. Для первых двух мод ИК построены окна параметрического резонанса крутильных колебаний. Показано, что при типичных значениях параметров для основной моды ИК окна резонанса чрезвычайно узки, в силу чего фиксация роста КК не представляется возможной на имеющемся экспериментальном оборудовании. Иначе обстоит ситуация для второй моды. Здесь окна параметрического резонанса проявляются более отчетливо и может ставится задача их экспериментального обнаружения

```
Вероятность темы 1     ---   0.00448833
Вероятность темы 2     ---   0.00448833
Вероятность темы 3     ---   0.00448833
Вероятность темы 4     ---   0.00448833
Вероятность темы 5     ---   0.00448833
Вероятность темы 6     ---   0.00448833
Вероятность темы 7     ---   0.00448833
Вероятность темы 8     ---   0.00448833
Вероятность темы 9     ---   0.00448833
Вероятность темы 10    ---   0.00448833
Вероятность темы 11    ---   0.00448833
Вероятность темы 12    ---   0.00448833
Вероятность темы 13    ---   0.00448833
Вероятность темы 14    ---   0.00448833
Вероятность темы 15    ---   0.00448833
Вероятность темы 16    ---   0.00448833
Вероятность темы 17    ---   0.00448833
Вероятность темы 18    ---   0.00448833
Вероятность темы 19    ---   0.91472172
Вероятность темы 20    ---   0.00448833
```

```
Тема 19:
        колебание    ---   0.04512674
             окно    ---   0.02551494
         резонанс    ---   0.02355376
         пластина    ---   0.02355373
параметрический      ---   0.02159261
         изгибный    ---   0.01963145
      крутильный     ---   0.01767029
       уравнение     ---   0.01767022
             мода    ---   0.01570909
           задача    ---   0.01374789
         значение    ---   0.01178674
           случай    ---   0.00982561
  мультипликатор     ---   0.00982558
         частота     ---   0.00982557
     проводиться     ---   0.00982554
```

Перспективы развития работы

- Использование методов оптического распознавания символов (Optical Character Recognition, OCR)
- Создание русскоязычного словаря математических терминов на основе советской математической энциклопедии в пяти томах под редакцией Виноградова И. М.

Заключение

Основные результаты работы:

- Исследованы модель скрытого размещения Дирихле и методы оценки матриц Θ и Φ для модели скрытого размещения Дирихле;
- Разработана программа для автоматической обработки коллекций документов и выявления скрытых тем в ней;
- Программа протестирована на коллекции русскоязычных математических документов, состоящей трудов математического центра имени Н.И. Лобачевского. Коллекция состояла из 20 томов;
- Предложены дальнейшие перспективы развития работы.

Спасибо за внимание!