МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)"

ФАКУЛЬТЕТ ИННОВАЦИЙ И ВЫСОКИХ ТЕХНОЛОГИЙ
КАФЕДРА АНАЛИЗА ДАННЫХ

Выпускная квалификационная работа по направлению

01.03.02 «Прикладные математика и информатика»

НА ТЕМУ:

ТЕМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ РАЗГОВОРОВ КОНТАКТНОГО ЦЕНТРА

Студент _____ Батаев В.В.

Научный руководитель д.ф-м.н _____ Воронцов К.В.

Зам. зав. кафедрой д.ф-м.н, профессор _____ Бунина Е.И.

МОСКВА, 2017

Содержание

Часть 1

Тематическая сегментация текста

1.1 Постановка задачи

Сформулируем формальную постановку задачи. Пусть дано предложение на естественном языке, которое можно считать последовательностью термов: $w_1, \dots w_n$. В качестве термов могут выступать отдельные слова, n-граммы, коллокации. Для каждого терма требуется поставить метку темы, то есть некоторой обобщающей данное слово сущности в зависимости от контекста. Таким образом решается задача *sequence labelling*, то есть обучения отображения $f : W^k \to T^k$, $k \geq 0$, W, T — конечные множества. Подпоследовательности из одинаковых тем будем называть монотематичными *сегментами*.

1.2 Метрики качества

В предыдущих работах [3][4] для оценки качества сегментации использовалась метрика $WindowDiff$, которая рассчитывается по следующей формуле:

$$WindowDiff(R, C, k) = \frac{1}{N-k} \sum_{i=1}^{N-k} I(|R_{i,i+k} - C_{i,i+k}| > 0),$$

где N — число слов в предложении, k — ширина окна, $R_{i,i+k}$ — эталонное число границ сегментов внутри окна $\{w_i, \dots w_{i+k}\}$, $C_{i,i+k}$ — предполагаемое. Значение данной метрики показывает, сколько в среднем раз алгоритм неверно расставляет сегменты в окне ширины k. Обычно в качестве k берут $\frac{N}{2 \cdot number\_of\_segments}$. Недостатком данной метрики является то, что она никак не учитывает, к каким темам были отнесены получившиеся сегменты.

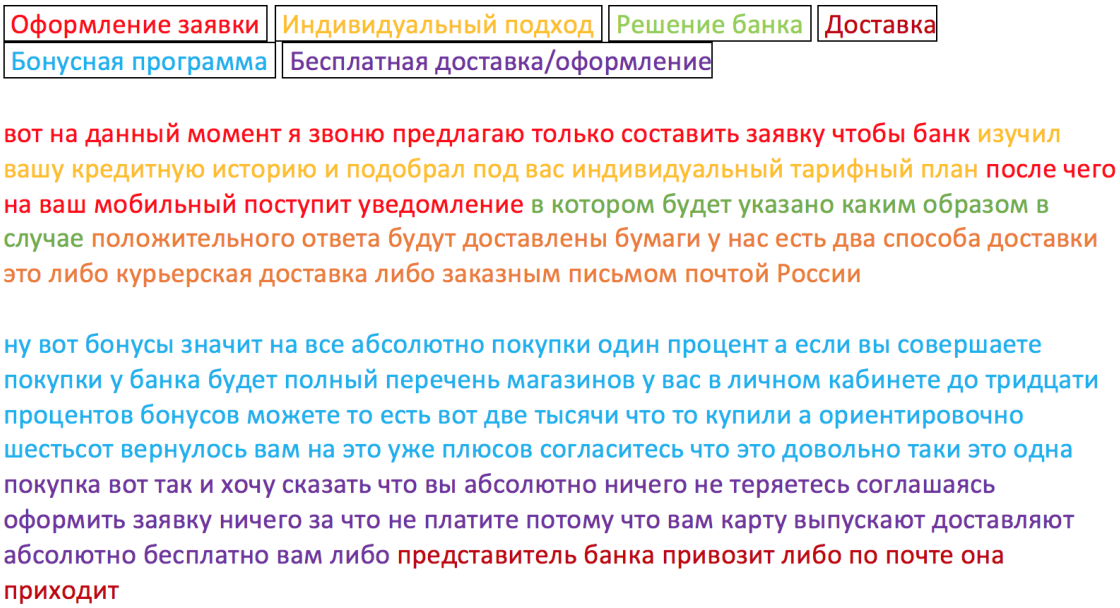

Рис. 1.1: Пример сегментации

Чтобы обойти выше определенную проблему будем мерить усредненный взвешенный $f1 - score$ и $accuracy$ для каждой темы по всем словам из всех предложений корпуса. В таком случае метрика качества сегментации фокусируется на отдельных словах и на точном проведении границ сегментов. Однако, если слово из одной темы окажется в середине сегмента другой темы или ее конце, то метрика оштрафует эти оба случая одинаково.

Если же не важно знать точные границы сегментов, а достаточно получить только порядок следования тем, то метрикой качества может быть редакторское расстояние между последовательностями тем сегментов.

1.3 Обзор существующих методов

При решении задачи тематической сегментации текста существует два принципиально различных подхода. Первый основан на unsupervised learning, не требующей размеченных отсегментированных текстов, в котором основной идеей является предположение, что граница между тематическими сегментами проходит при резкой смене темы локального контекста.

В работе [2] использовался алгоритм, основанный на движении скользящего окна по тексту. Для каждого положения окна строится bag-of-words вектор, сглаженный

с помощью tf-idf по некоторой большой коллекции документов. По построенным векторам строилась последовательность косинусных расстояний между соседними окнами, и в ней точки максимума или точки со значениями, большими некоторого заранее определенного порога, определяются как границы сегментов. Однако же данный метод не позволяет определить темы построенных сегментов.

В работе [4] по корпусу строится вероятностная тематическая LDA модель ([1]), с помощью которой для каждого слова в предложении имеется распределение по темам. Затем векторы распределений тем соседних слов сравниваются с помощью косинусной метрики и производится расстановка сегментов аналогично методу со скользящим окном.

Второй подход основан на supervised learning, использующий разметку текста на сегменты. В работе [8] предлагается решать задачу классификации для каждого предложения — поставить границу сегмента после данного предложения или нет, используя нейросетевую архитектуру, состоящую из рекуррентных и сверточных слоев, которая извлекает векторное представление предложения невысокой размерностью по сравнению с размером словаря. Отметим, что в данной работе также нет прямой возможности узнать, к каким темам относятся построенные сегменты.

Все выше описанные подходы применялись для документов сравнительно большой длины, для которых описания их структуры через входящие в них слова являются достаточно статистически значимыми. В данной же работе основной проблемой является то, что объектами сегментации являются отдельные слова, для которых сложно строить признаки, основанные на большом контексте.

Часть 2

Описание новых предложенных методов

Будем считать, что перед нами стоит supervised задача, поскольку качество ее решения окажется заведомо выше, нежели в unsupervised случае. Первым шагом всех ниже описанных методов будет построение тематической модели по всем документам коллекции с использованием априорных знаний о структуре некоторых тем. Поэтому перейдем к подробному описанию использованной модели.

2.1 Вероятностное тематическое моделирование

Пусть D — множество текстовых документов в коллекции, W — множество слов коллекции, T — множество латентных переменных, именуемых темами.

Для каждого документа $d \in D$ известен вектор частот n_{dw} токена $w \in W$. Необходимо найти условные распределения $p(w|t)$ частоты слова w для темы t и условное распределение $p(t|d)$ частоты темы t в документе d.

Введем следующие обозначения: $\theta_{td} = p(t|d), \phi_{wt} = p(w|t)$.

Для вывода в тематической модели примем гипотезу условной независимости:

$$p(w|d,t) = p(w|t),\ w \in W,$$

то есть токены документа определяются только тематикой.

Коллекция документов рассматривается как случайная и независимая выборка троек $(w_i,\ d_i,\ t_i)$, $i = 1, ..., n$ из дискретного распределения $p(w,d,t)$ на конечном вероятностном пространстве $W \times D \times T$.

В силу данных предпололжений справедливо соотношение:

$$p(w|d) = \sum_{t \in T} p(t|d)p(w|t),\ w \in W, d \in D.$$

Параметры модели образуют матрицы $\Phi = (\phi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$.

Оценки латентных параметров будем искать с помощью EM-алгоритма, максимизируя регуляризованный [6] логарифм правдоподобия:

$$L(\Phi, \Theta) + R(\Phi, \Theta) = \ln \prod_{i=1}^{n} p(w_i|d_i) + R(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(w|d) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

На матрицы Φ, Θ наложены естественные ограничения неотрицательности и нормированности столбцов.

Таким образом для нахождения оценки параметров модели Φ, Θ решается следующая оптимизационная задача:

$$\Phi^*, \Theta^* = \arg\max_{\Phi, \Theta} L(\Phi, \Theta) + R(\Phi, \Theta),$$

$$\sum_{w \in W^m} \phi_{wt} = 1;\ \phi_{wt} \geq 0;$$

$$\sum_{t \in T} \theta_{td} = 1;\ \theta_{td} \geq 0.$$

Регуляризатор $R(\Phi, \Theta)$ отвечает за ограничения, которые могут быть наложены на модель. Такими ограничениями могут быть, например, принадлежность лишь небольшого числа слов к конкретной теме, либо близость векторов тем соседних слов в предложении. Опишем использованные в данной работе регуляризаторы.

Пусть $\alpha_{td} = p(t|d), \beta_{wt} = p(w|t)$ — произвольные распределения.

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \tag{2.1}$$

Регуляризатор 2.1 является при $\beta_0 > 0, \alpha_0 > 0$ *регуляризатором сглаживания* и при $\beta_0 < 0, \alpha_0 < 0$ — *регуляризатором разреживания.*

При $\beta_0 > 0, \alpha_0 > 0$:

$$\sum_{t \in T} KL_w(\beta_{wt}||\phi_{wt}) \to \min_{\Phi}, \sum_{d \in D} KL_t(\alpha_{td}||\theta_{td}) \to \min_{\Theta} \Longrightarrow R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

При $\beta_0 < 0, \alpha_0 < 0$:

$$\sum_{t \in T} KL_w(\beta_{wt}||\phi_{wt}) \to \max_{\Phi}, \sum_{d \in D} KL_t(\alpha_{td}||\theta_{td}) \to \max_{\Theta} \Longrightarrow R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Абсолютные значения коэффициентов α_0 и β_0 отвечают за баланс между разреживанием/сглаживанием распределений модели и максимизацией правдоподобия.

2.2 Частичное обучение тематической модели

Качество тематической модели, которое в первую очередь выражается как интерпретируемость тем, может быть улучшено за счет использования априорных знаний эксперта о структуре тем. В этом случае мы приходим к использованию частичного обучения.

Пусть $D_0 \subset D$ — подмножество размеченных документов, и для каждого документа $d \in D_0$ задано подмножество релевантных тем $T_d \subset T$, к которым он относится, и подмножество нерелевантных тем $\overline{T}_d \subset T$, к которым он не относится.

Инициализируем параметры $\alpha_{td} = \frac{1}{|T_d|}[t \in T_d][d \in D_0]$ с положительным значением параметра α_0, получаем частный случай регуляризатора сглаживания — *регуляризатор частичного обучения по релевантности.*

Если же проинициализировать $\alpha_{td} = \frac{1}{|T_d|}[t \in \overline{T}_d][d \in D_0]$ с отрицательным значением $\overline{\alpha_0}$, то получаем частный случай регуляризатора разреживания — *регуляризатор частичного обучения по нерелевантности.*

При достаточно больших по модулю значениях коэффициентов регуляризации α_0 и $\overline{\alpha_0}$ регуляризаторы частичного обучения обнуляют вероятности нерелевантных тем в документах и токенов в темах.

TODO: какие брались в экспериментах значения α_0 и $\overline{\alpha_0}$, количество итераций EM-алгоритма

2.3 Тематическая моделирование в задаче сегментации

Построим semi-supervised вероятностную тематическую модель, использующую априорные знания о принадлежности некоторых документов к определенным темам. Выход тематической модели задает некоторое приближение для сегментации, поскольку тему слова w_k можно определить как $y_k = \arg\max_{t \in T} p(t|w_k, d)$. Такое приближение на практике оказывается слишком грубым, которое предлагается

уточнять следующими способами:

- Сглаживать векторы тем соседних слов с помощью векторов из окна фиксированной ширины
- Моделировать последовательности тем слов в предложении с помощью дискриминативной модель, а именно *условное случайное поле* (сокращенно CRF)

2.4 Сглаживание векторов тем соседних слов

Напомним, что результатом алгоритма тематического моделирования являются распределения $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$. Из формулы Байеса и гипотезы условной независимости слова при данном документе и теме следует:

$$p(t|w,d) = \frac{p(t,w|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{t\in T}\phi_{wt}\theta_{td}}.$$

Сглаживание будем производить итеративно последовательно по всем словам документа по следующему алгоритму:

$p_0(t|w,d)$ – выход предварительно обученной тематической модели

$$g(t,d,k,i) = f(i)p_s(t|d,w_{k+i})$$

$$p_{s+1}(t|d,w_k) = norm_{t\in T}(\sum_{j=-h,j\neq 0}^{h} g(t,d,k,j))$$

$f(i)$ — функция затухания влияния контекста на распределение текущего слова, h — размер окна, s — номер текущей итерации. С ростом количества итераций увеличивается влияние темы слова на темы остальных слов предложения.

Таким образом данный метод имеет следующие параметры:

- размер окна
- функция затухания контекста
- число итераций сглаживания

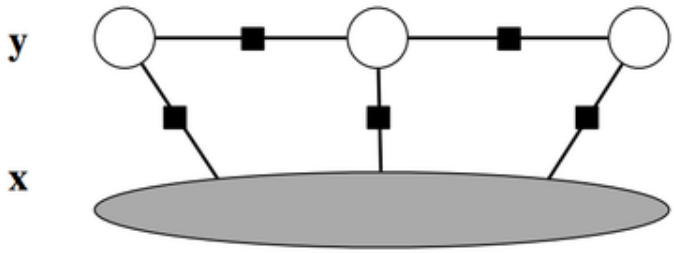

Рис. 2.1: Графическая модель linear chain CRF, в которой переходные факторы зависят от всей обозреваемой последовательности [5]

2.5 Условное случайное поле

В задаче тематической сегментации распределение тем при условии данного слова и документа не учитывает влияния контекста внутри предложения. Для решения данной проблемы предлагается использовать модель условного случайного поля или CRF для моделирования зависимостей между темами разных слов. Преимущество CRF в том, что такая модель является дискриминативной, а следовательно способна более точно моделировать условное распределение $p(\boldsymbol{y}|\boldsymbol{x})$ неизвестной последовательности меток $\boldsymbol{y}$ при условии обозреваемой последовательности слов $\boldsymbol{x}$.

В данной задаче будем использовать linear-chain-CRF, чей граф представляет собой линейную цепь. В таком случае условное распределение можно записать в следующем виде:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{n=1}^{N} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_n, y_{n-1}, \boldsymbol{x})\},$$

где $Z(\boldsymbol{x})$, зависящая от конкретного $\boldsymbol{x}$ функция нормализации

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \prod_{n=1}^{N} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_n, y_{n-1}, \boldsymbol{x})\}$$

$f_k(y_n, y_{n-1}, \boldsymbol{x})$ — произвольные вещественнозначные функции, которые также называются *feature functions*. Они описывают зависимости между соседними метками тем и произвольным подмножеством входных переменных из последовательности. Подчеркиваем, что $f_k(y_n, y_{n-1}, \boldsymbol{x})$ может зависеть от произвольного подмножества $\{x_1, \dots x_N\}$.

В качестве $f_k(y_n, y_{n-1}, \boldsymbol{x})$ будем использовать:

- $I(y_n = y^i, y_{n-1} = y^j)$

- $I(y_n = y^i)p(t = t^j|x_n, d)$
- $I(y_n = y^i)p^*(t = t^j|x_n, d)$ сглаженная версия $p(t|w, d)$ после итеративного алгоритма, описанного в 2.4
- $I(y_n = y^i)(word2vec(x_n))_j$, где $(word2vec(x_n))_j$ — j-ая компонента векторного представление слова x_n из $word2vec$ модели
- $I(y_n = y^i)cosine(word2vec(x_n), word2vec(x_{n+j}))$, где $j \in \{-L, \dots L\}, j \neq 0$, то есть косинусное расстояние от слова в позиции n до слова из окна с центром в позиции n и ширины L

TODO: вывод в linear-chain CRF, обучение параметров θ_k

Часть 3

Результаты экспериментов

3.1 Входные данные

Обучающая выборка состоит из 400 диалогов контактного центра, в которых были оставлены только реплики оператора. В каждой такой реплике проставлены границы сегментов, и каждый из сегментов отнесен к некоторой теме. Число тем было выбрано равным 53 экспертом.

Для построения тематической модели с использованием частичного обучения использовалось open-source библиотека BigArtm [7]. Для каждой темы была размечена небольшая группа до десяти документов, относящихся к данной теме. Регуляризатор частичного обучения запускался на одну итерацию с коэффициентом α_0, равным 100.

TODO: привести примеры тем и сегментов

3.2 Сглаживание векторов тем

Для эксперимента, описанного в 2.4, была взята функция $f(i) = \frac{1}{i^2}$. Тема каждого слова определялась как $\arg\max_{t \in T} p(t|w,d)$. Покажем зависимость качества сегментации от числа итераций и размера окна на графиках 3.1, 3.2, 3.3.

3.3 Метод, основанный на CRF

В качестве feature functions были взяты всевозможные комбинации, описанные в 2.5. Для получения сглаженных версий векторов $p(t|d,w)$ были взяты размер окна 5 и число итераций 3, поскольку они давали наилучшее качество сегментации

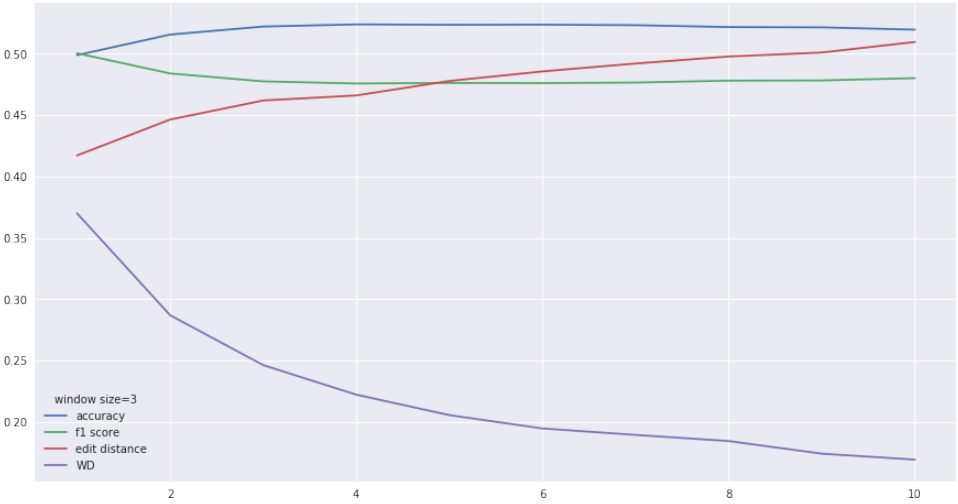

Рис. 3.1: Размер окна 3

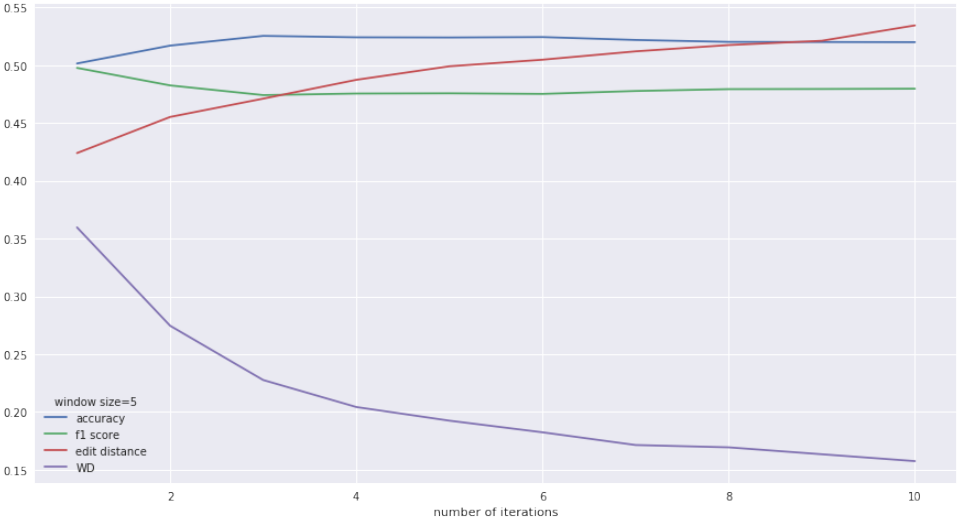

Рис. 3.2: Размер окна 5

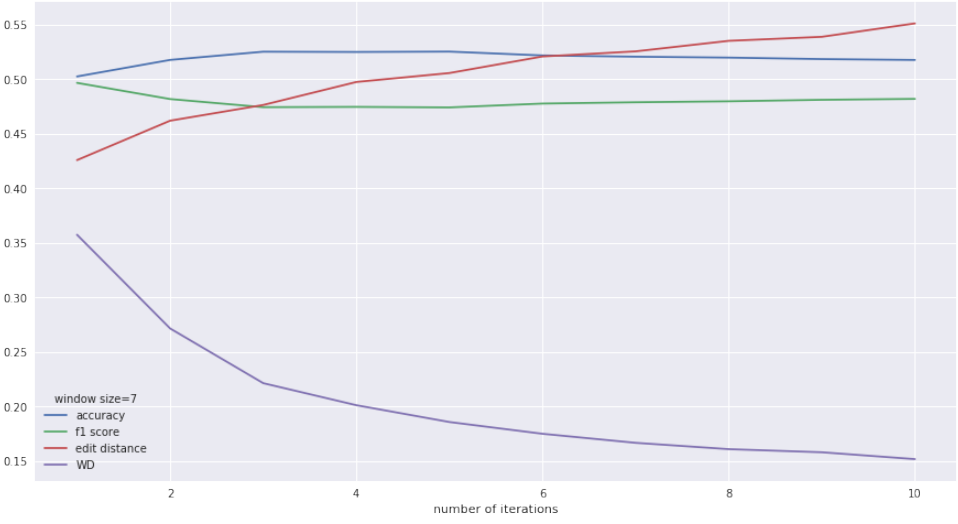

Рис. 3.3: Размер окна 7

Table 3.1: Точность сегментации CRF

Feature functions	accuracy	weighted f1	Edit distance	WD
topics	0.564	0.541	0.444	0.104
w2v features	0.601	0.591	0.381	0.138
topics + w2v features	0.620	0.615	0.358	0.136
smoothed topics	0.521	0.496	0.471	0.117
smoothed topics + w2v features	0.634	0.629	0.361	0.145

по принципу аргмаксимума. Для оценки качества алгоритма использовалась 5-fold кросс-валидация. Результаты представлены в таблице 3.1:

Часть 4

Заключение

TODO: выводы

Литература

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997.

[3] Marti A. Hearst Lev Pevzner. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 2002.

[4] Martin Riedl and Chris Biemann. Text segmentation with topic models.

[5] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[6] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1):303–323, 2015.

[7] Apishev M. Romov P. Dudarenko M. Vorontsov K., Frei O. Bigartm: Open source library for regularized multimodal topic modeling of large collections. *Analysis of Images, Social Networks and Texts.*, 2015.

[8] Liang Wang, Sujian Li, Xinyan Xiao, and Yajuan Lyu. *Topic Segmentation of Web Documents with Automatic Cue Phrase Identification and BLSTM-CNN*, pages 177–188. Springer International Publishing, Cham, 2016.