Freelance projects description analysis: Russian freelance market

Ivan Mishalkin

Abstract

The increasing number of freelancers on Russian labour market attracts more and more scientists. The biggest platform for self-employed citizens of the Commonwealth of Independent States is fl.ru. The information from this service may be used for investigating the Russian freelance market. However the researchers may face the problem of lack of the variables available for them to analyze. The main question of this paper is how many responses the offered project will get according to its description. We use two modifications of LDA(Latent Dirichlet Allocation) for supervised learning. The results of this paper are following: it is better to use STM(Structual Topic Model) for this type of the task and the most popular among Russian freelancers are the tasks, which are related to web development and internet shops together. The topics which lead to unpopularity of the project are different for two models.

1 Introduction

In Russian Federation the share of self-employed people is very low compared to the other developed and even developing countries (Strebkov and Shevchuk, 2015). Online market for contract labour grows fast (Agrawal et al., 2013) and this is true for Russia too (Gurova, 2012). As the lack of freelancers is observed than the supply of freelance projects is higher than the demand on it from the freelancers. Some of the projects may seem totally uninteresting for self-employed workers while the other may have some kind of perfect competition between freelancers. The estimation of parameters that lead to interest among Russian freelancers is the goal of this paper. The following questions are also would be take to consideration: what skills have the highest demand on Russian freelance market and what should be taken into account posting a new project.

There are several web-sites, which are used by freelancers to find projects. They are freelance.ru, fl.ru, allfreelancers.su, freelancerbay.com. (Aletdinova, 2016) provides the description of each platform and according and to it the fl.ru is the largest. The platform allows anyone to sign up and to place the freelance project or offer himself as a freelance.

The main problem the freelance researcher may face working with fl.ru is lack of data. Very few projects have fixed reward for job. Often it is written as a whole some for the project, what may mean that the freelancer may do a part of work and he will get proportionally his reward. The other problem is lack of control over the finished projects by fl.ru. So outdated projects may remain being shown as actual endlessly. One of the available and reliable source of data is description of the project in natural language. This paper provides comparison of several ways to deal with text data to extract useful information about Russian freelance market. The methods applied to the solution of this problem are sLDA (the extension of Latent Dirichlet Allocation, that allows the researcher build the model on the relation between target variable and hidden topics distribution over the document), stm (structural topic model) and linear regression with lasso regularization over tf-idf coefficients.

2 Literature background

Traditionally freelance projects are divided into text, design, IT and consulting themes and are researched from certain point of view. In the text category there is a couple of works (Slatkin, 1988) and (Herrmann, 1982).

The first one gives general recommendations for writers and the second is also some kind of article with recommendations for translator. (Koch and Obermaier, 2014) investigates the problem of having two jobs at the same time. The authors try to explain why German freelance journalists have PR as a second job. The methodology is following: the sent about a thousand of emails to freelancers with 6 specific questions. The authors found out that freelancers work as a PRs as it is interesting to them and this is the way they earn money. Nearly a half of the respondents preferred to work as PR if they could afford not to work at the second job.

Very wide area in freelance is IT. (Hsieh and Hsieh, 2013) use content analysis to investigate freelance developers of the mobile applications. The work is quite theoretical and provides readers with some theoretical and managerial implications. More empirical work is (Süß and Kleiner, 2010). The authors investigate freelancers' commitment to firms. They use descriptive analysis and clustering on the answers made by freelancers on special questionnaire. The survey clarifies that freelancer may feel commitment and the degree of commitment is consistent with the other researches. The similar work but again in media and fully theoretical is (Storey et al., 2005).

The lion share of works on freelance is theoretical or based on questionnaires. The good example of empirical work is (Wang et al.,). The authors tried to estimate the period during which the freelance project is done on Amazon Mechanical Turk. This may be actual for researches based on data from platforms with non-closing projects, like fl.ru.In the paper the authors detect topics with LDA and try to implement it. Their model performs rather poor. The other example of using LDA is described in (Kim et al.,). The authors use LDA for freelance data to detect web service abuse. The authors developed the method which strengths are: interpretable clusters of projects and this may reveal the target of the abuse, the method is unsupervised, so there is no need for creating a representative sample with marked target variable. However some mistakes returned from the algorithm can not be corrected unlike in the supervised learning.

3 Analysis and Results

3.1 Data description

The data for the research was collected from Russian freelance website fl.ru. Several times the data was collected with overlapping, so we could track the changes of the projects characteristics. The dates of collecting were 03.11.2016, 25.11.2016, 06.12.2016, 26.12.2016, 21.01.2017, 27.02.2017 and 28.03.2017. The first variable is ”answers” and it is dependent in our model. It shows how many freelancers considered the project interesting.Our goal is to find features, which will maximize ”answers”. The next variable is category. This categories differ from ones at fl.ru. We united some common themes to one, for example, texts and translations were united to one called ”Writing and translation” etc. ”Description” - text in Russian language that describes freelance project. As an id of the project we took the unique web address - the variable ”project_id”. In ”published” we gathered the date of project publication in dd.mm.yyyy format.

As we are going to predict the number of responses to the project it seems obvious that the longer the project is on the platform the more responses it will receive. However, the reality is quite different. We took the datasets collected for the periods which contain the biggest amount of overlapping projects: 25.11.2016, 06.12.2016, 26.12.2016, 21.01.2017. Took the freshest projects, posted at 25.11.2016 and built the graph over 215 projects[Figure 1]. On the y-axis the amount of responses, on the x-axis time points, when the data was collected.

We may notice that the number of responses increases only during the first two weeks and then reaches plato. If we add projects, placed at 25 and 24 of November this conclusion will be even more obvious. This means that we may investigate ”old” projects without any discounting for time it was placed. The first step in sample selection was from each dataset collected at time points choose documents that are a month ”older” than the freshest one in the collection.

Next only documents with more than 20 words were selected. This was done because projects with small description are likely to be fake or contain weblink to more detailed description. Such projects are not the case of this study. On the next step the words should become without variations in number gender,

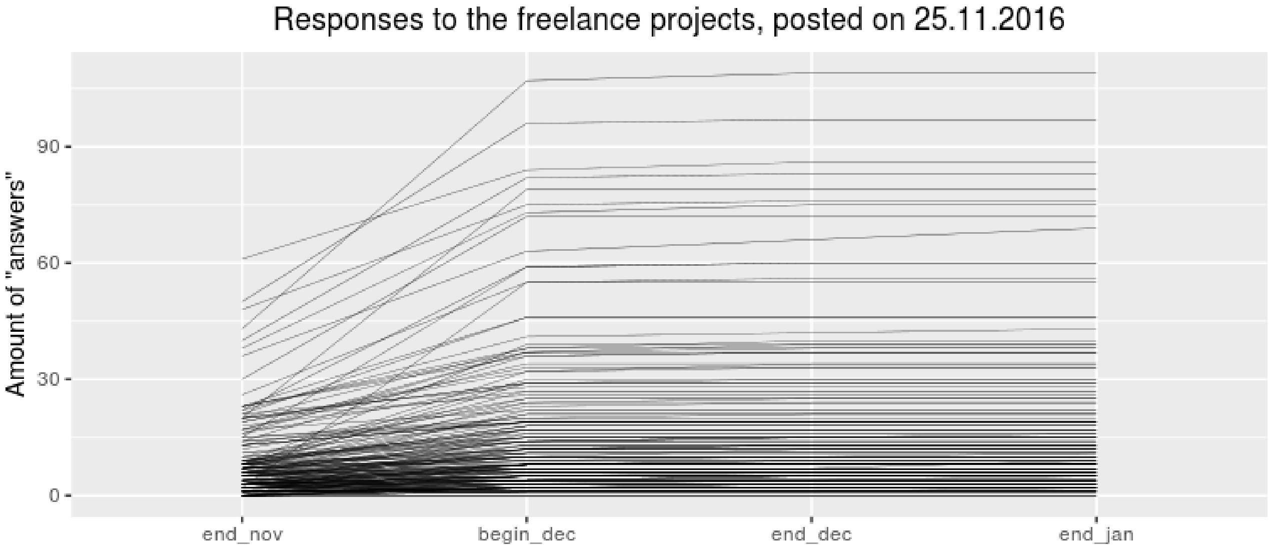

Figure 1: Spaghetti plot of answers on projects over time for projects posted on 25 of November

declension and case. There are two ways to solve this problem: stemming and lemmatization. The first one cuts the variate part of the word leaving only base. Lemmatisation turn the word into the infinitive form. Firstly we tried stemming as it is easier to implement. The texts became hard to read, understand and interpret. So we tried lemmatisation provided by Yandex. Detailed explanation of the algorithm is described in their article (Segalovich, 2003). Lemmatisation performed better and are recommended to use for Russian texts.

After preprocessing 20000 projects remained.

3.2 Models

The problem that researchers face analyzing texts is a huge amount of documents. It seems impossible to analyze manually each document in adequate time. Unsupervised techniques might be a possible solutions for this problem. Latent Dirichlet allocation is a widely used technique.

Another method we use in this work are Structural Topic Models(stm). The main idea of the stm is the ability to include metadata(extra information about the document) into topic model(Roberts et al., 2015). Like many other topic models stm is a generative model of word counts. The key difference of the stm is the idea that topical prevalence and topical content may be a function of documents metadata. The degree the document is associated with the topic is topical prevalence and content refers to the words used within within the topic. Variables explaining prevalence referred to topical covariates, and ones that explain topical content are referred to as topical content covariates. Stm allows to use each of the covariates or none of above. Stm uses Estimation-Maximization algorithm too.

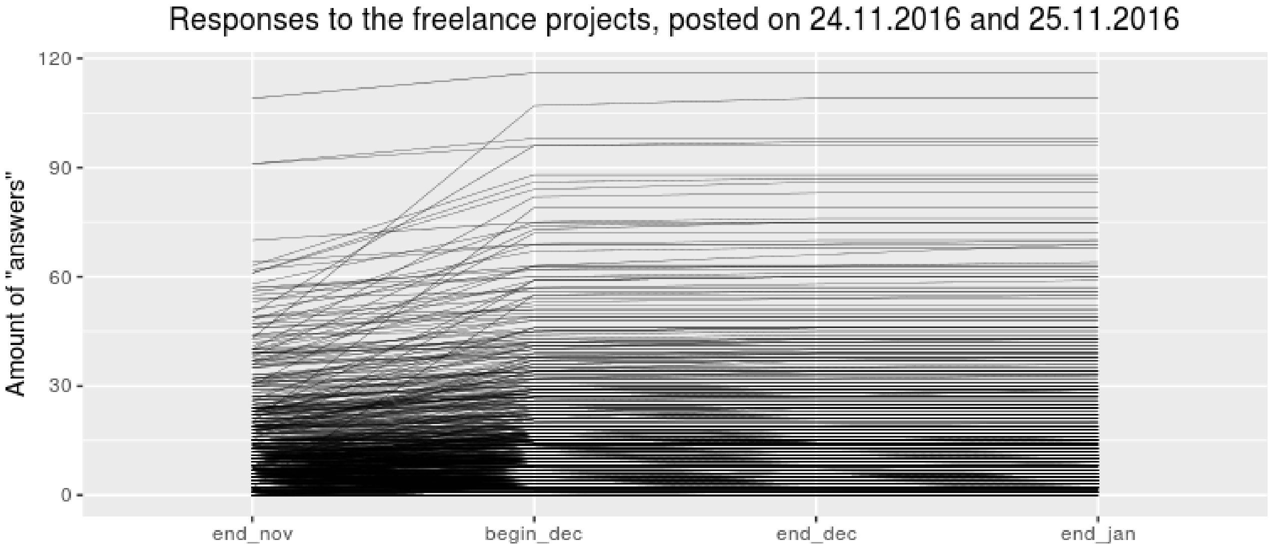

Figure 2: Spaghetti plot of answers on projects over time for projects posted on 24 and 25 of November

The last method we will apply to our data is lasso regression. This method provides both regularization and variable selection.

Before estimating the models, we may want to look at the distribution of the freelancers' responses "answers"[Figure 3].

The most frequent number of responses is 2. The lion share of the projects has from 0 to 13 responses.

We will compare the models with statistic equivalent to the residual sum of squares in OLS regression. The formula is:

$$RSS = \Sigma(y_{observed} - y_{predicted})^2$$

3.2.1 Supervised LDA

At first we estimated the sLDA model with 50 topics. On the plot below we visualize the coefficients[Figure 4]. We tried to interpret generated topics by the most common words in each. All of them are statistically significant except the topic in the middle, which represents some web settings. This topic includes some general words like web, email and tune. It also contains some general English words. This topic is not representative, so shows the insignificance of the coefficient. For Russian speaking people there is a table in appendix, representing top 10 words in the topic and the theme we assigned to this topic[Table 1].

The plot shows the coefficient on the x-axis and the topic on the y-axis. Each document has probability of containing each topic. This plot may help to understand the structure of the supply of labour force on the freelance market. The plot shows that topic, which is connected with full-stack web-site for business development offers the greatest increase in freelancers' responses.

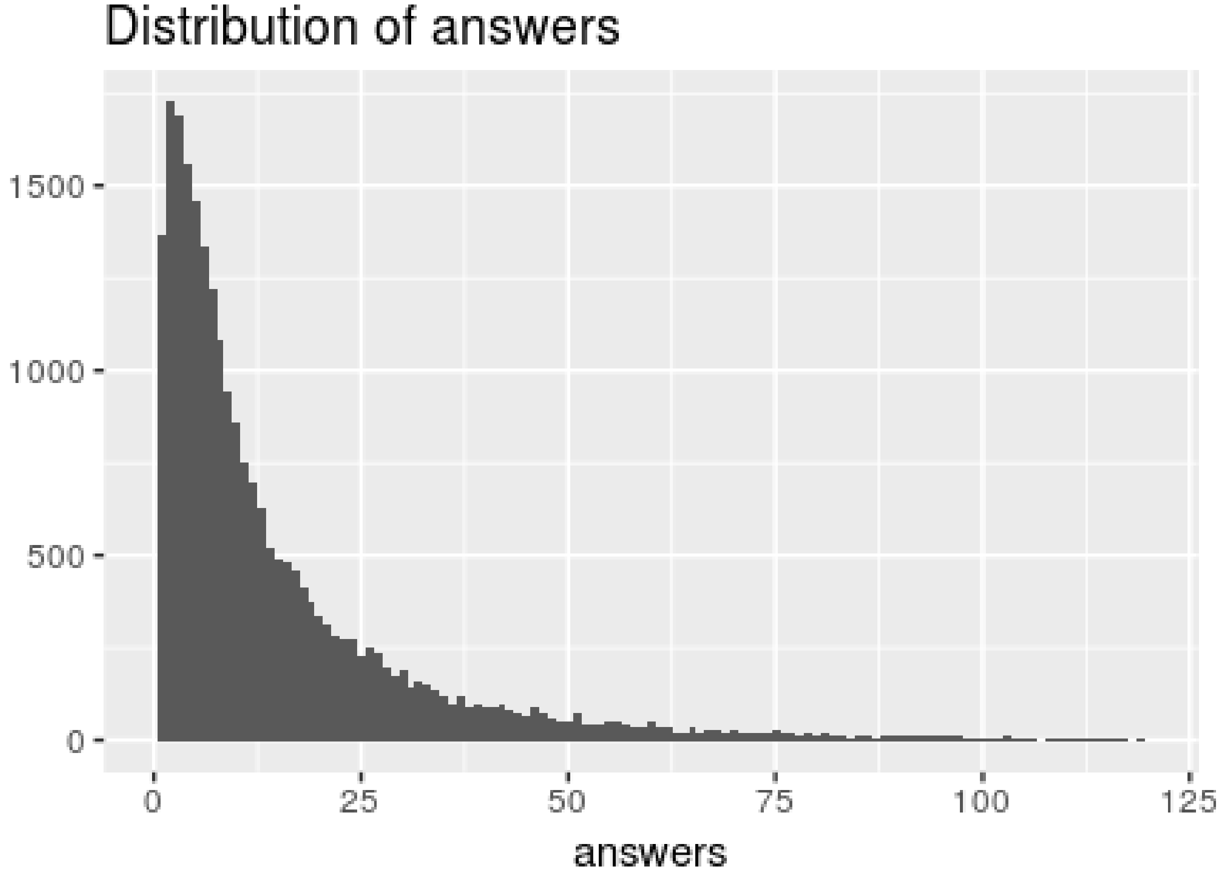

Figure 3: Answers distribution

This may mean that the lion share of freelancers at fl.ru is web-developers. The next after web development goes design. It is interesting to mention that both firm description and description of the demanded worker increase the responses of the freelancers. May be this may be because people do not like uncertainty. Next text related tasks are in demand. There is also an interesting thing - the task on modifying some example adds responses, at the same time tasks related to copying examples are unpopular. Topics related to some specific tasks receive less responses. Probably this is because that there are fewer specialists in databases or in SMM. May be that freelancers at first learn widely used technologies, like php programming language and than spend their time learning some special techniques. An interesting thing is the effect of SMM and SEO. Topic with words related mainly to SEO is the most popular among freelancers. The mixture of SMM and SEO is still interesting for the freelancers, but SMM only decreases the number of responses and this drop is very significant.

After estimating the model we would like to make prediction on our held out sample. As it has been already mentioned above we would like to compare different models, using the squared sum of squares. However, before using this metric we corrected some predictions of the model. Some of the forecasted values estimated to be less than zero and no matter how unpopular the project might be, the number of resposponses can not be negative. So we corrected

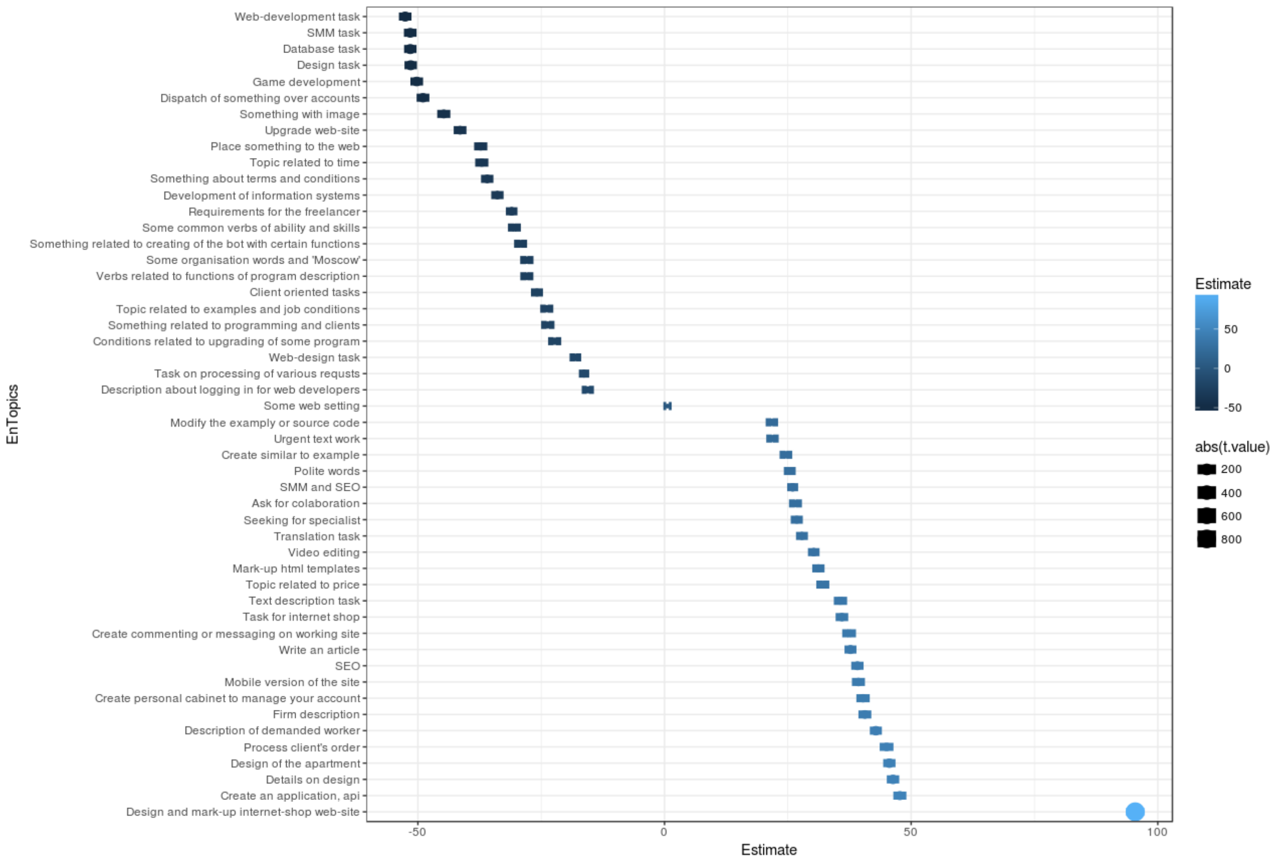

Figure 4

negative values.For sLDA our metric is equal to 3076324.7. If we look at the distribution of the predicted by sLDA values[Figure 5], we notice that a lot of predicted values are negative and correction it to zero decreases the RSS but still the model is rather poor.

3.2.2 Structural topic models

Structural topic model has the great advantage - the ability to include regressors into the topic modelling process. We tried to set answers to prevalence. Topical prevalence captures how much each topic contributes to a document. For topic modelling we chose deterministic initialization as learning method, which uses the spectral algorithm(Roberts et al., 2014). This algorithm is recommended to use on relatively large number of documents as it often performs well. The great advantage of the model is only one neccessar hyperparameter - number of topics. The top words of the topics are shown in the table **??**. The description of the topics is given in the table **??**. This table shows the words, which are associated with each topic. There are 4 ways to find whether the words reflect the topic good or not provided: by their probability to occur in certain topic, by FREX and by score and lift. FREX takes into account both how frequent the word is and how exclusive it is for each topic. Score metric is calculated according to the formula $\beta_{\omega,k}(log\beta_{\omega,k} - 1/K\sum_{k'} log\beta_{\omega,k'}$. Lift is the calculated

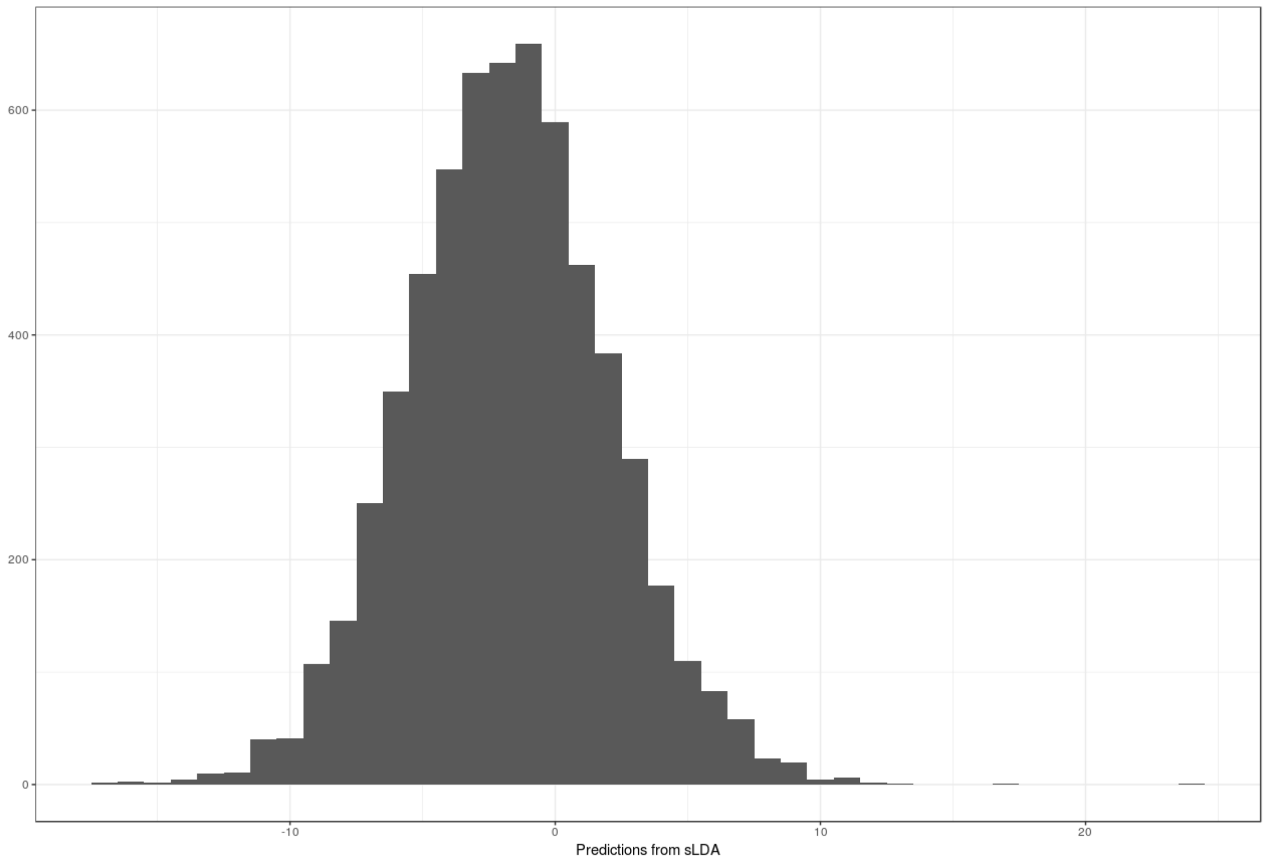

Figure 5

by dividing the topic-word distribution by the empirical word count probability distribution. We may look at the correlation between topics. This might tell us about some topics that are used together.

After we have estimated the topic we would like to predict the responses. We estimate lasso regression over topics. This is similar to what we are going to do in the next chapter. The main difference is in regressors we use:topics or just words. The sum of squared errors is equal to 1999761.26. The result is much better than we have in our previous model. Two histograms on the plot show the destribution of predicted values and real ones[Figure 2]. We may notice that the estimate is biased, unlike sLDA, lasso over stm topics tends to predict more than these is in real life.

It is also interesting to compare estimates for different topics between sLDA and stm. The topic which contributes the most is number 28, which is about web design. The next topic which adds responses is 25. It is about design without web component. Significantly lower effect has topic number 42. It is about SEO(search engine optimization). The next is topic which is a mixture of something about beauty and children. On the fifth place is the topic about time limits and other conditions. The sixth estimator is the topic about internet marketing. The biggest negative effect gives the topic about adding buttons to the web-site and other small tasks. Probably, this may be so because of relatively expensive unpaid time which is spent on explaining the small task.

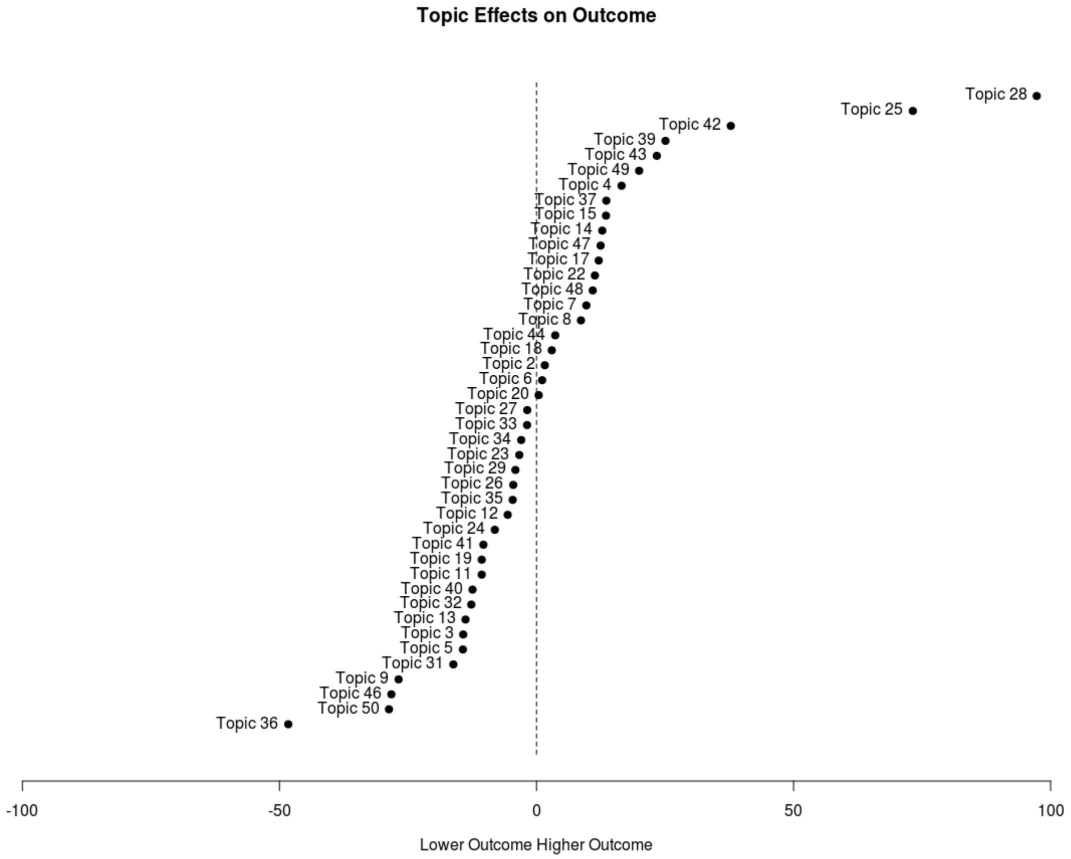

Figure 6: Topics estimates for STM model

The reward is supposed to be small, so experienced freelancers may want to avoid such tasks so as new ones(because of the small reward). Three topics, 50, 46 and 9 has almost the same negative effect. Topic 50 consists of general words. Probably, the freelancers may not like description which is mainly consists of some general words. 46^{th} topic is about rewarding. The negative effect may be explained in the following way: there are some special conditions that the employer has to mention. The topic number 9 is about wordpress and tuning plugins. This may be because of some specificity of this technology.

3.2.3 Lasso regression

Finally we needed very simple algorithm to compare our models with. We have chosen lasso regression over TF-IDF matrix. Columns - TF-IDF coefficients for each word in the vocabulary. Rows represent documents. The main advantages of the lasso regression is its ability to work with very sparse date. It does not have a lot of hyper parameters, and this is very good. We used cross-validation to estimate the optimal lambda. On our hold out sample it has the value of our chosen metric equal to 1900641.

We also would like to know whether our models gain any information or not. To find it out we create a very simple model: we find mean of the amount

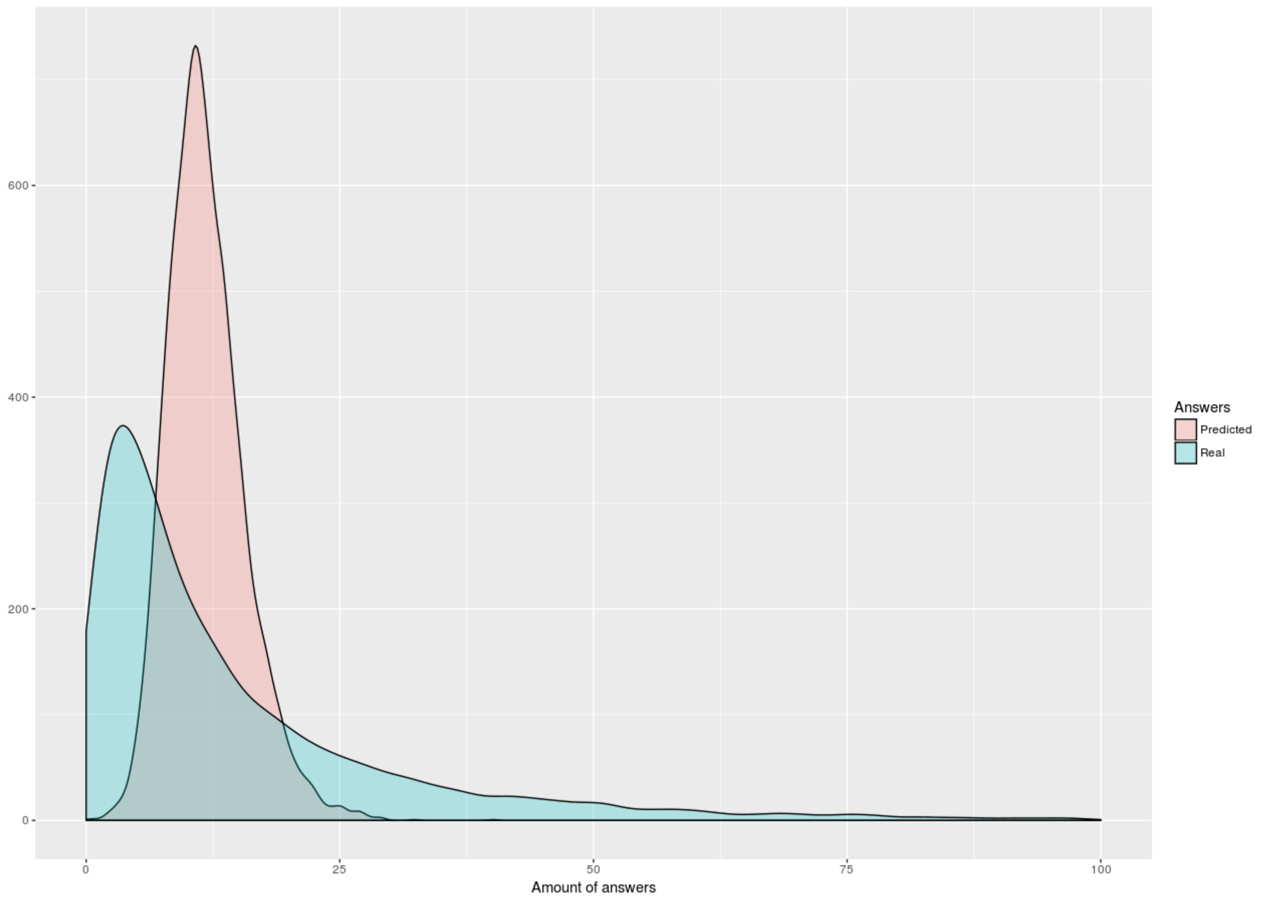

Figure 7: Distributions of predicted and real values

of answers on training dataset and use it to predict values for the test sample. The metric equals 1835396.98. This number is smaller than in lasso regression

4 Conclusion

The goal of this work was to find the best method to work with freelance projects data. Despite the fact that that all of the considered algorithms had sum of squared error higher than the simple model of average value, they may be useful in understanding the structure of freelance market. We understood that there is not much sense in estimating lasso regression over TF-IDF matrix, because the results are uninterpretable, the speed of estimation is nearly the same as sLDA and the prediction is worse than mean.

Stm and sLDA may be useful in understanding the structure of freelance market. Though algorithms used different methods for learning, estimates from both claim that projects in web development and design are more attractive for Russian freelancers than the other areas. The topics from stm seems to be better selected, however the right tune of hyperparameters in sLDA model may also increase its performance. Stm is also good for ability to include various regressors, that may influence on topics distribution.

References

Agrawal, A., Horton, J., Lacetera, N., Lyons, E., and Stern, L. N. (2013). DIGITIZATION AND THE CONTRACT LABOR MARKET: A RESEARCH AGENDA.

Aletdinova, A. A. (2016). Peculiarities of the Russian Freelance Market. *Journal of Siberian Federal University. Humanities & Social Sciences*, 11(10):2734–2741.

Gurova, M. (2012). Факторы, влияющие на выбор фриланса как формы самозанятости(Factors, infuencing on choosing freelance as a form of self-employment). *Теория и практика общественного развития(Theory and practice of public development)*, 7:57–60.

Herrmann, H. H. (1982). Breaking a Lance for the Freelance Translator. *Technical Communication*, 29(4):10–12.

Hsieh, J.-K. and Hsieh, Y.-C. (2013). Appealing to Internet-based freelance developers in smartphone application marketplaces. *International Journal of Information Management*, 33(2):308–317.

Kim, D.-K., Motoyama, M., Voelker, G. M., and Saul, L. K. Topic Modeling of Freelance Job Postings to Monitor Web Service Abuse.

Koch, T. and Obermaier, M. (2014). Blurred lines: German freelance journalists with secondary employment in public relations. *Public Relations Review*, 40(3):473–482.

Roberts, M., Stewart, B., and Tingley, D. (2014). Navigating the Local Modes of Big Data : The Case of Topic Models. *Scholar.Harvard.Edu.*

Roberts, M. E., Stewart, B. M., and Tingley, D. (2015). stm: R Package for Structural Topic Models. *Journal of Statistical Software*, VV(2014).

Segalovich, I. (2003). A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In *MLMTA*, pages 273–280.

Slatkin, E. (1988). The freelance writer and marketing. *Technical Communication*, 35(2):112–117.

Storey, J., Salaman, G., and Platman, K. (2005). Living with enterprise in an enterprise economy: Freelance and contract workers in the media. *Human Relations*, 58(8):1033–1054.

Strebkov, D. and Shevchuk, A. (2015). Трудовые траектории самозанятых профессионалов (фрилансеров)(Labour tracks of self-employed professionals-freelancers). *Мир России. Социология. Этнология (World of Russia. Sociology. Ethnology)*, 1:72–100.

Süß, S. and Kleiner, M. (2010). Commitment and work-related expectations in flexible employment forms: An empirical study of German IT freelancers. *European Management Journal*, 28(1):40–54.

Wang, J., Faridani, S., and Ipeirotis, P. G. Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models.

Appendix

Table 1: SLDA topics description

N	Words	Topics
1	блок верстка ваш весь видео	Web-development task
2	день группа нужно который добрый	SMM task
3	файл данные база таблица добавлять	Database task
4	дизайн товар баннер должный доставка	Design task
5	весь должный картинка игра который	Game development
6	программа наш аккаунт весь нужно	Dispatch of something over accounts
7	весь вариант задание здравствовать предлагать	Something with image
8	сделать добавление возможность выводить api	Upgrade web-site
9	ссылка весь размещать необходимо сумма	Place something to the web
10	свой год который давать вопрос	Topic related to time
11	работа почта объем обязательно оплата	Something about terms and conditions
12	разработка решение система управление информация	Development of information systems
13	опыт работа знание требование команда	Requirements for the freelancer
14	знать мочь работать это делать	Some common verbs of ability and skills
15	это получать ваш мочь курс	Something related to creating of the bot with certain functions
16	срок москва готовый задача небольшой	Some organisation words and 'Moscow'
17	должный иметь задание каждый использовать	Verbs related to functions of program description
18	клиент звонок база услуга договор	Client oriented tasks
19	мочь это просто который например	Topic related to examples and job conditions
20	менеджер время программа каждый номер	Something related to programming and clients

21	бюджет требоваться срок выполнять возможно	Conditions related to upgrading of some program
22	кнопка меню нажатие окно телефон	Web-design task
23	почта присылать обработка настройка документ	Task on processing of various requsts
24	пользователь данные регистрация вводить телефон	Description about logging in for web developers
25	web will page онлайн настраивать	Some web setting
26	сделать нужно примерно пример прикреплять	Modify the examply or source code
27	нужно нужный сразу писать очень	Urgent text work
28	пример нужно это просто чтото	Create similar to example
29	добрый пожалуйста спасибо ответ уважаемый	Polite words
30	сеть рекламный реклама кампания пост	SMM and SEO
31	нужный давать специалист основа требоваться	Ask for colaboration
32	проект кандидат просьба свой обсуждать	Seeking for specialist
33	оплата язык факт день русский	Translation task
34	видео пример ролик минута сценарий	Video editing
35	готовый срок сверстывать дизайн исполнитель	Mark-up html templates
36	цена работа предложение необходимо срок	Topic related to price
37	слово текст ключевой описание заголовок	Text description task
38	товар цена магазин сайт страница	Task for internet shop
39	возможность ссылка должный комментарий сообщение	Create commenting or messaging on working site
40	статья текст тематика тема написание	Write an article
41	продвижение запрос яндекс результат оптимизация	SEO
42	сделать форма мобильный сайт страница	Mobile version of the site
43	личный кабинет свой стоимость функционал	Create personal cabinet to manage your account

44	компания наш год заниматься сфера	Firm description
45	работа опыт проект разработка знание	Description of demanded worker
46	заявка заказ система клиент данные	Process client’s order
47	план дом материал помещение чертеж	Design of the apartment
48	цвет размер стиль фон фото	Details on design
49	приложение сервер данные пользователь файл	Create an application, api
50	сайт работа страница товар нужно	Design and mark-up internet-shop web-site

Table 2: STM topics description

N	Words	Topics
1	онлайн, курс, точка, обучение, школа, итд, тренинг	Education related
2	это, который, мочь, весь, свой, делать, либо	Weakly interpretable general words
3	письмо, рассылка, адрес, отправлять, приходить, отправка, база	E-mail
4	находить, интернет, ремонт, подбирать, оборудование, техника, мастер	Installation and repairing
5	код, скрипт, ошибка, проблема, работать, исправлять, написать	Correct script
6	сторона, окно, листовка, участок, дом, должный, предусматривать	Building architecture
7	страница, главный, блок, меню, форма, должный, ссылка	Web design
8	компания, клиент, услуга, продажа, наш, продукт, менеджер	Task with client
9	шаблон, модуль, cms, wordpress, плагин, устанавливать, настраивать	Wordpress related
10	конструктор, тестирование, ценник, должный, iphone, часы, телефон	Mobile
11	работа, тема, страница, срок, сдача, курсовой, система	Education texts writing

12	должный, вопрос, ответ, скриншот, это, тест, выбирать	web-site related tasks(not web development)
13	отзыв, размещать, ссылка, комментарий, оставлять, размещение, форум	Reviews, recommendations
14	сайт, необходимо, новый, весь, нужно, версия, создавать	Web development
15	товар, заказ, магазин, цена, карточка, корзина, доставка	Internet shop
16	email, skype, web, тело, строчка, связь, prestashop	e-Commerce
17	карта, купить, заказывать, акция, цветок, недвижимость, бесплатный	Orders related
18	категория, каталог, город, фильтр, автомобиль, каждый, производитель	Car related
19	данные, база, поле, таблица, форма, дата, номер	Database
20	весь, вопрос, время, помогать, хотеться, получать, готовый	Noisy topic with clothes
21	приложение, игра, мобильный, разрабатывать, интерфейс, android, ios	Mobile app
22	пользователь, возможность, система, личный, должный, регистрация, кабинет	Private cabinet
23	will, game, can, need, time, english, job	English words
24	сервер, настраивать, доступ, настройка, клиент, crm, подключать	Server options
25	логотип, цвет, фон, название, должный, вариант, белый	General design
26	размер, баннер, изображение, формат, макет, должный, печать	Printing(advertising)
27	раздел, битрикс, этап, задача, оценка, реализация, учет	Documents
28	дизайн, верстка, лендинг, макет, сверстывать, готовый, адаптивный	Web mark up, templates
29	файл, формат, загружать, ссылка, название, загрузка, весь	Files work

30	фото, фотография, описание, пример, необходимо, каждый, фон	Photo
31	программа, сообщение, список, который, аккаунт, должный, время	Parsing
32	заявка, почта, присылать, работа, весь, требоваться, электронный	Rewriting
33	ролик, минута, голос, сценарий, канал, пример, видеоролик	Video
34	группа, сеть, пост, соц, контент, аудитория, подписчик	SMM(social media marketing)
35	работа, опыт, язык, требование, знание, английский, умение	Language skills
36	сделать, нужно, добавлять, кнопка, весь, картинка, это	Small modifications(site)
37	текст, статья, тема, тематика, написать, слово, знак	Copyrighting
38	видео, анимация, презентация, слайд, эффект, сделать, ролик	Presentation
39	детский, год, ребенок, мероприятие, одежда, салон, красота	Beauty/children
40	работа, проект, опыт, программист, знание, задача, разработка	Programming
41	модель, объект, схема, должный, чертеж, деталь, линия	Mechanical drawing
42	сайт, запрос, москва, продвижение, область, оптимизация, seo	SEO(search engine optimisation)
43	срок, цена, ваш, стоимость, писать, работа, предложение	Time limits and conditions
44	проект, искать, специалист, сотрудничество, опыт, давать, который	Ask for collaboration
45	поиск, сервис, позиция, тур, место, билет, определенный	Travelling
46	оплата, работа, скайп, день, задание, рубль, выполнять	Prices
47	дом, план, проект, мебель, материал, интерьер, визуализация	Interior design
48	пример, стиль, картинка, нарисовать, нужно, прикладывать, вложение	Drawing

49	объявление, рекламный, реклама, яндекс, кампания, настройка, гугл	Advertising
50	нужный, задача, возможно, желательно, необходимо, подробный, простой	General words