# Topic modeling analysis #

One of the most important decisions that needs to be made before conducting a topic modeling analysis is to decide an optimal number of topics. Most researchers suggest experimenting with different numbers to evaluate the quality of topic outputs. On one hand, analyzing larger data sets with a low number of topics can result it topics that are too general, missing out on the granularity of the data. On the other hand, using too many topics will generate outputs with topics too rare and non-representative.

For the current analysis, I experimented with 10, 30, 50 and 100 topics. The outputs generated using 10 and 30 topics were more interesting than those of 50 and 100. For this reason, the below analysis, will include excerpts from using 10 and 30 topics.

## How to interact with the visuals? ##

All visuals included in the current folder are interactive. However, due to technical difficulty of embedding these visuals directly on GitHub, I opted to include static screen shots followed by URLs. By clicking the URLs, the user of the project can engage with the visual directly.

### Actions for interactivity: ###

* Use “Slide to adjust relevance metric:” 1.0 is the most relevant words in each topic, 0.0 is least relevant, terms that may have appeared only in one or two documents
* Use the top left “Selected Topic” box to view topics clustered together.
  + *For example, in the 10-topic visual only 6 topics are visible on the left chart, 7-10 are clustered together under the number 10, formatted in bold.* *To view topic 8, type that number in the “Selected Topic” box, which make the terms of that topic appear on the right side of the screen.*
* The circles represent topics while the sizes correlate with their prevenance
* The space between the circles show the similarity of topics

|  |  |  |
| --- | --- | --- |
| Topics | Terms | Topic title |
| Topic 1 | Женский  Население  Девушка  Положение  Муж  Мужчина  отношение | женщина в обществе |
| Topic 2 | Население  Муж  Ребёнок  Мужчина  Положение | роль женщины в семье |
| Topic 3 | Женский  Ребёнок  область |  |
| Topic 4 | Область  Ребёнок  Революция | предстоящие перемены |
| Topic 5 | Население  Семья  Ребёнок | Демография |
| Topic 6 | Край  Женский  Население  Экономический, труд | женский труд в экономике |
| Topic 7 | Хозяйство  борьба | Борьба с предрассудками |
| Topic 8 |  |  |
| Topic 9 |  |  |
| Topic 10 |  |  |

[(0, '0.004\*"калмыцкий" + 0.004\*"женский" + 0.004\*"область" + 0.004\*"турецкий" + 0.003\*"узбекистан" + 0.003\*"ребёнок" + 0.003\*"кибитка" + 0.003\*"калмык" + 0.003\*"край" + 0.003\*"положение"'),

(1, '0.000\*"население" + 0.000\*"мужчина" + 0.000\*"положение" + 0.000\*"область" + 0.000\*"семья" + 0.000\*"ребёнок" + 0.000\*"муж" + 0.000\*"больший" + 0.000\*"хозяйство" + 0.000\*"борьба"'),

(2, '0.000\*"чувашский" + 0.000\*"чувашка" + 0.000\*"население" + 0.000\*"женский" + 0.000\*"муж" + 0.000\*"чуваш" + 0.000\*"мужчина" + 0.000\*"ребёнок" + 0.000\*"отношение" + 0.000\*"образ"'),

(3, '0.008\*"край" + 0.006\*"женский" + 0.006\*"население" + 0.006\*"зырянский" + 0.005\*"зырянка" + 0.004\*"экономический" + 0.004\*"труд" + 0.004\*"область" + 0.004\*"коми" + 0.003\*"зырянин"'),

(4, '0.000\*"население" + 0.000\*"женский" + 0.000\*"больший" + 0.000\*"ребёнок" + 0.000\*"положение" + 0.000\*"мужчина" + 0.000\*"семья" + 0.000\*"область" + 0.000\*"отношение" + 0.000\*"власть"'),

(5, '0.010\*"курд" + 0.005\*"население" + 0.004\*"курдский" + 0.004\*"таджикистан" + 0.004\*"семья" + 0.004\*"ребёнок" + 0.004\*"таджичка" + 0.003\*"положение" + 0.003\*"башкирский" + 0.003\*"муж"'),

(6, '0.006\*"татарский" + 0.006\*"китайский" + 0.006\*"китай" + 0.005\*"марийка" + 0.004\*"чувашка" + 0.004\*"область" + 0.004\*"ребёнок" + 0.004\*"революция" + 0.004\*"чувашский" + 0.004\*"мужчина"'),

(7, '0.006\*"персидский" + 0.005\*"женский" + 0.005\*"население" + 0.004\*"мордва" + 0.004\*"персия" + 0.004\*"девушка" + 0.004\*"положение" + 0.004\*"муж" + 0.004\*"мужчина" + 0.003\*"отношение"'),

(8, '0.000\*"население" + 0.000\*"женский" + 0.000\*"муж" + 0.000\*"девушка" + 0.000\*"положение" + 0.000\*"отношение" + 0.000\*"ребёнок" + 0.000\*"область" + 0.000\*"образ" + 0.000\*"мужчина"'),

(9, '0.006\*"население" + 0.006\*"муж" + 0.005\*"ребёнок" + 0.004\*"мужчина" + 0.004\*"положение" + 0.004\*"афганистан"

* A finding is that this didn’t work to locate the religious minorities I hope to find
* It is talking about talking about the Muslim women in the same way as everyone else, so they are not racialized
* How do these women get propagandized in other in another corpus – future work – next step and a possible inclination

What are the main topics that are shared among brochures? Do they match to the results of the close reading?

How are the Muslim women, who represent 13 out of 30 ethnic groups accounted for in the brochures, described in comparison to their non-Muslim counterparts? How does the topic modeling demonstrate these discrepancies? Are there any discrepancies that came up in the topic modeling outputs?

What decisions are made to determine which stopwords to keep and which to remove in the pre-processing? How do these decisions impact the topic modeling output?

* Considering that topic modeling is designed to analyze representative or thematic samples of large data sets, what the benefits of utilizing it with a small corpus?