ITMO UNIVERSITY

Saint Petersburg, Russia

Social dynamics analysis and modelling using COVID-19 tweets data

4Health

Koshman Varvara

Kshenin Alexander

Shaykina Alevtina

C41131

Saint-Petersburg – 25.06.2020

Project goals and steps

ITMO UNIVERSITY

1. Data acquisition, preprocessing
2. Tweet topic modelling, hashtags clustering
3. Creation of tweet topics popularity time series
4. Cross-correlation analysis

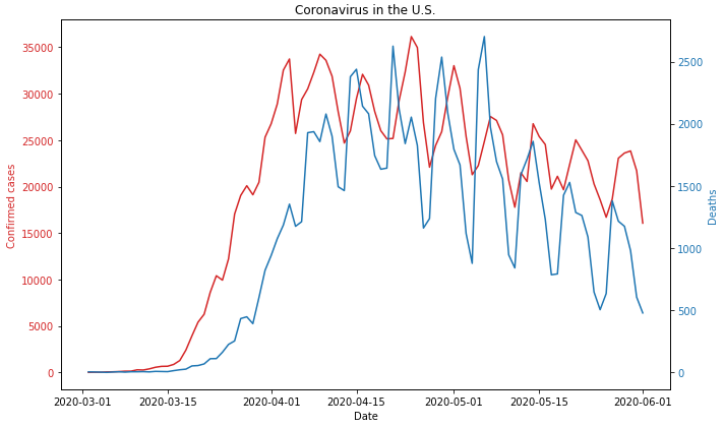
5. Tweet topic popularity prediction ← Goal №1
6. Identification of society states ← Goal №2

ITMO more than a UNIVERSITY

Data

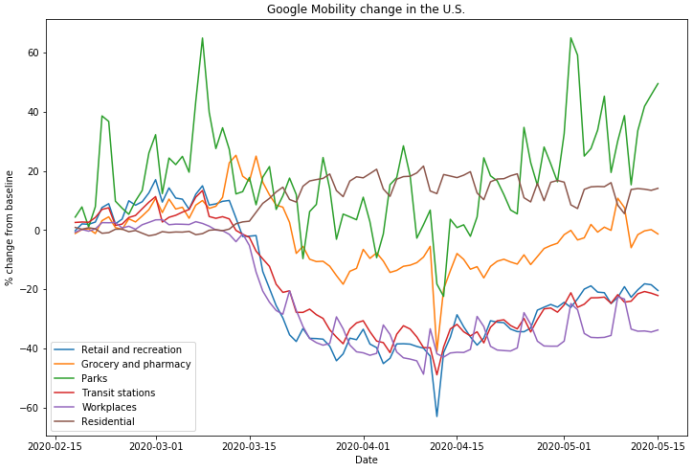
ITMO UNIVERSITY

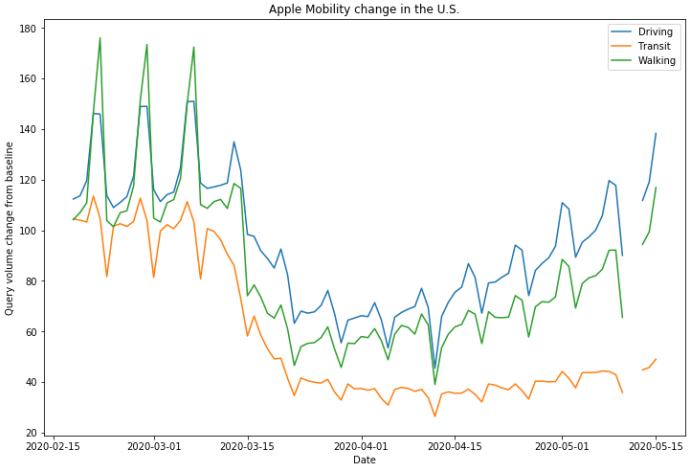
1. CDC Coronavirus Cases in the U.S. (from 21.01.2020 to date)
2. Google Community Mobility Reports (from 15.02.2020 to date)
3. Apple Covid-19 Mobility Trends Reports (from 13.01.2020 to date)



CDC

Google





ITMOre than a UNIVERSITY

Data

ITMO UNIVERSITY

4. Coronavirus (covid19) Tweets
   (12.03 – 31.04.2020)
   - English tweets preprocessing for topic modelling
   - Hashtags extraction, hashtag sequences identification
   - Grouping tweets by country (for analysis of the U.S.)
   - Analysis of metrics for tweet (t) containing topic (p) by author(a)
     (aggregated for topics in time series analysis):

     1) Favorites 2) Retweets 3) Count of tweets with topic

     4) $Popularity(p|t) = (favorites(t) + retweets(t) \times retweet\ weight) \times probability(p|t)$

     5) $Engagement(p|t) = logistic(followers(a)) \times \frac{Popularity(p|t)}{followers(a)}$

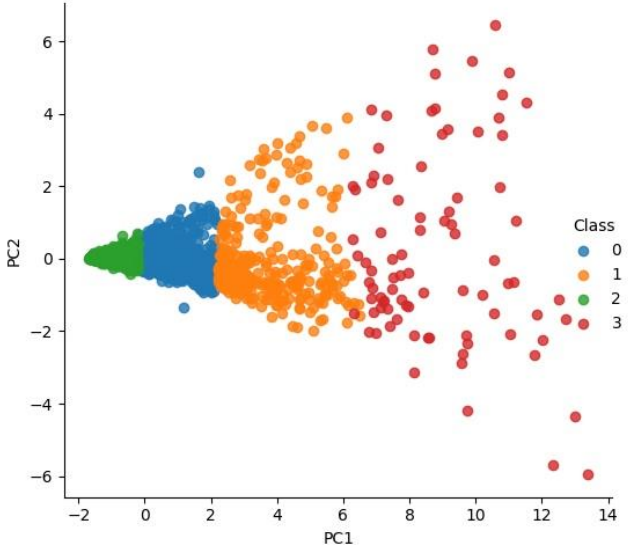
ITMO more than a UNIVERSITY

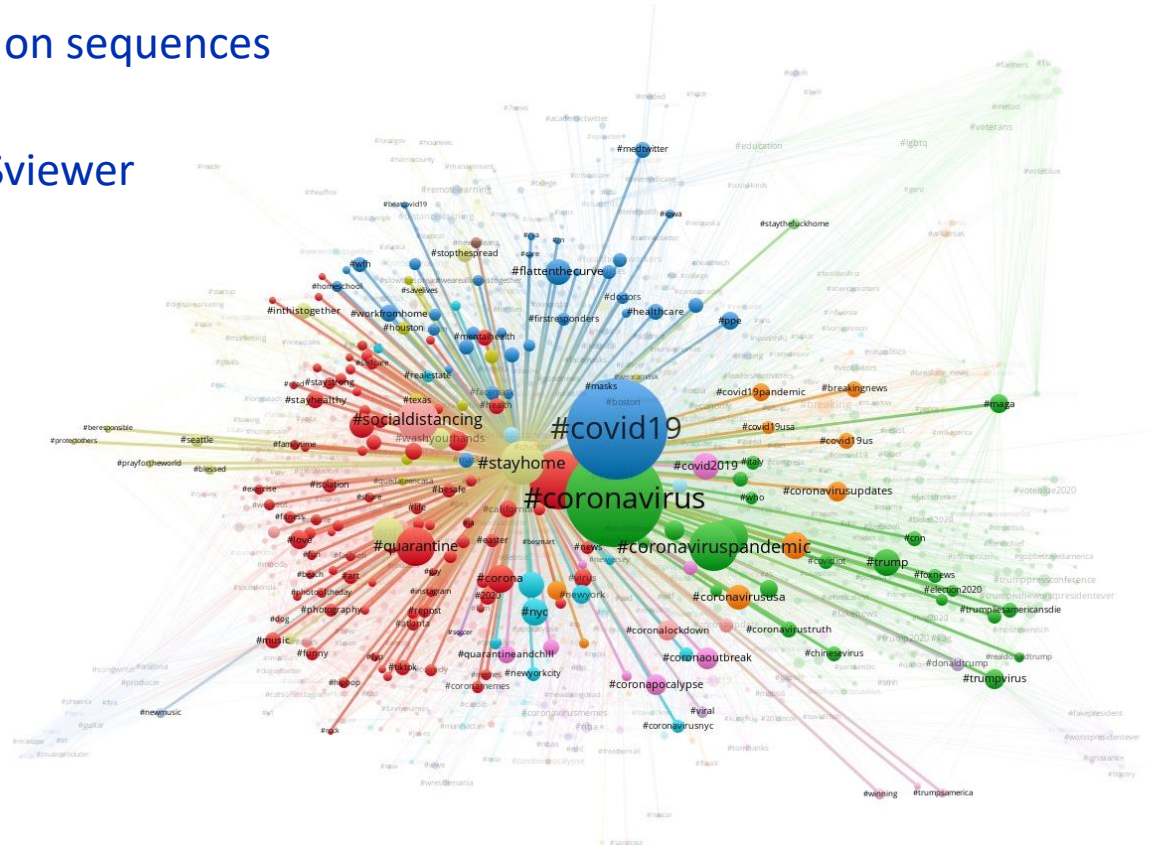
[4] https://www.kaggle.com/smid80/coronavirus-covid19-tweets

Hashtags clustering

ITMO UNIVERSITY

- Trained Word2vec model on sequences of frequent hashtags
- Hashtags co-usage in VOSviewer





ITMOre than a UNIVERSITY

Topic modelling

ITMO UNIVERSITY

20 topics were extracted using BigARTM library:

1. Local government
2. Covid-19 in New York
3. China
4. Music, concerts
5. Covid19 statistics
6. Finance
7. Haircut
8. Student debt stimulus
9. Salary
10. Social distance
11. Free time
12. Memes, getting bored
13. Food delivery
14. Press, conferences
15. Online learning
16. Online chattering
17. Virus diagnostics
18. Personal protective equipment
19. Support Italy
20. Podcasts

+ 7 self-isolation related hashtags topics, based on the results of hashtags clustering

ITMO more than a UNIVERSITY

Cross-correlation matrix

ITMO UNIVERSITY

- Correlations of selected for forecasting topics metrics and Covid-19 + mobility series in matrix
- Was used to distinguish exogeneous data (among Covid-19 + mobility)
- Highlighted values indicate what series were taken as exogeneous

ITMOre than a UNIVERSITY

Topic forecasting

ITMO UNIVERSITY

- We trained ARIMA models for forecasting of different topics popularity and engagement, with the usage of exogeneous data and without it
- Gaussian filter was applied to the data before training
- Parameters for all models were chosen using the Akaike information criterion (aic)
- Exogeneous data for each topic was defined using the analysis presented on the previous slide
- We employed mean absolute percentage error (MAPE) to define quality of models predictions on test data
  - ➢ Model with the usage of exogeneous data showed much better results of forecasting, especially for the first 3 days, however, we experimented with different test duration

ITMO more than a UNIVERSITY

Topic forecasting results

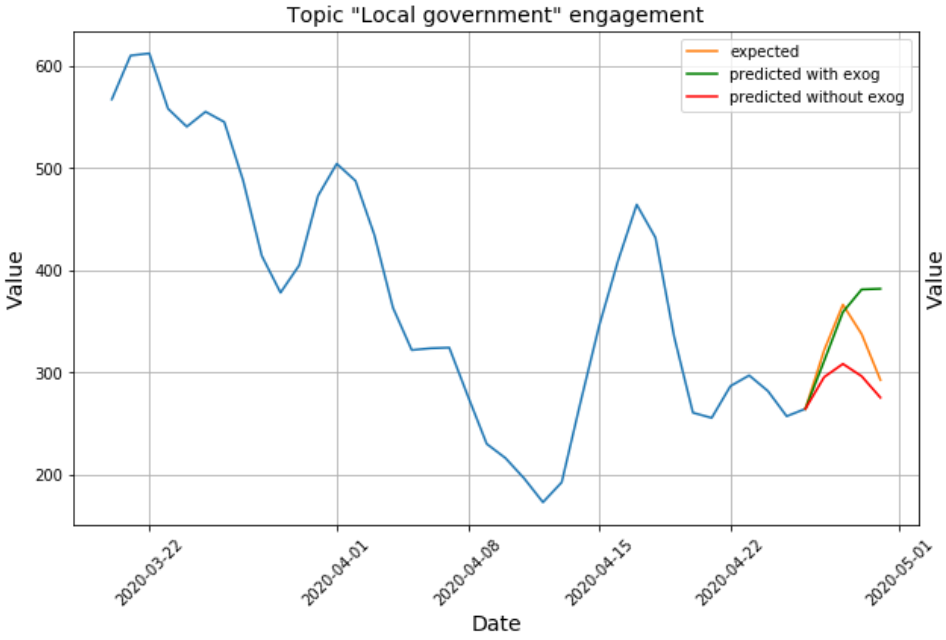
ITMO UNIVERSITY

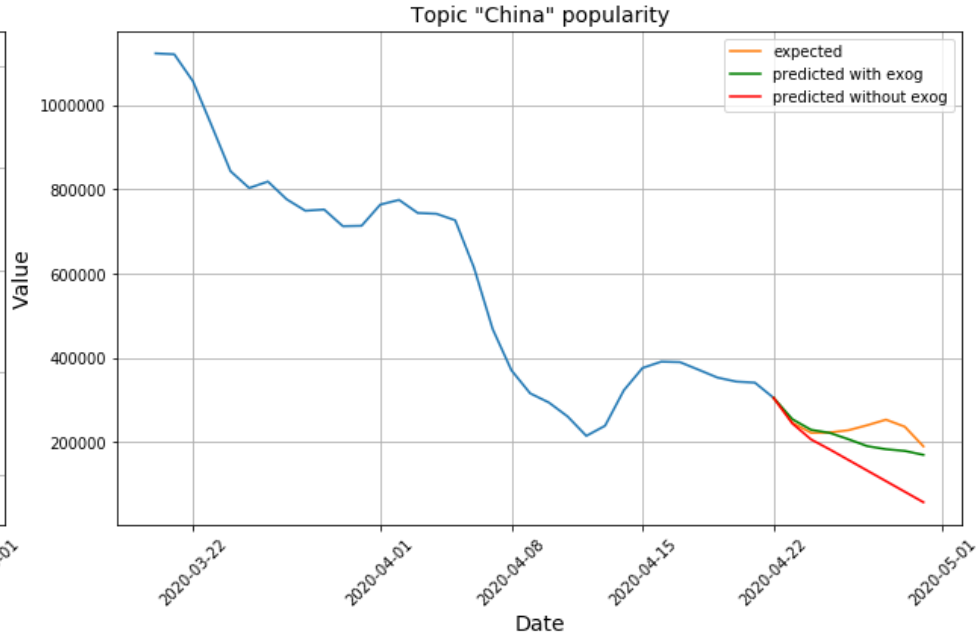
Topic title	metric	MAPE for N days forecasting without and with exogeneous data							
		1 day		2 days		3 days		4 days	
‘Local government’	engagement	8.2	3.3	12.0	2.6	12.1	6.1	10.5	12.2
‘China’	popularity	9.4	2.2	15.8	3.0	20.5	2.7	26.2	4.5
‘Music, concerts’	engagement	7.9	1.0	12.2	5.1	10.6	3.8	9.1	4.2
‘Finance’	engagement	12.8	1.2	28.5	3.5	37.0	5.2	36.0	5.1
‘Salary’	popularity	18.8	5.9	29.3	5.4	33.6	7.5	34.9	18.9
‘Memes, getting bored’	engagement	4.2	1.9	4.9	1.3	5.2	1.1	6.0	7.0
‘Online chattering’	popularity	1.9	2.0	8.5	1.3	17.1	3.7	23.6	6.3
‘Virus diagnostics’	popularity	5.0	1.3	6.8	1.8	10.8	2.4	16.0	5.7
		No exog	Exog	No exog	Exog	No exog	Exog	No exog	Exog

ITMO more than a UNIVERSITY

Topic forecasting results

ITMO UNIVERSITY

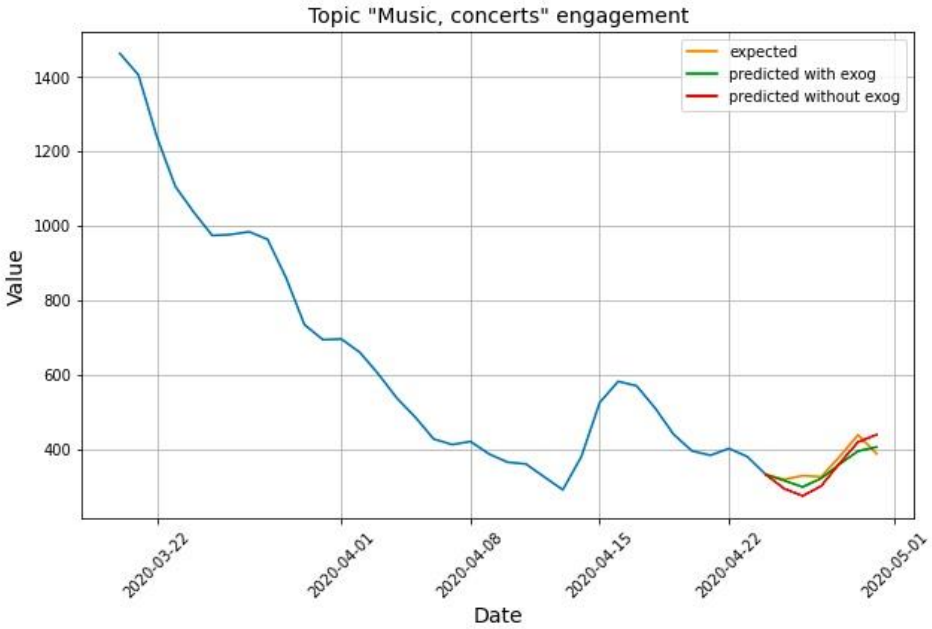


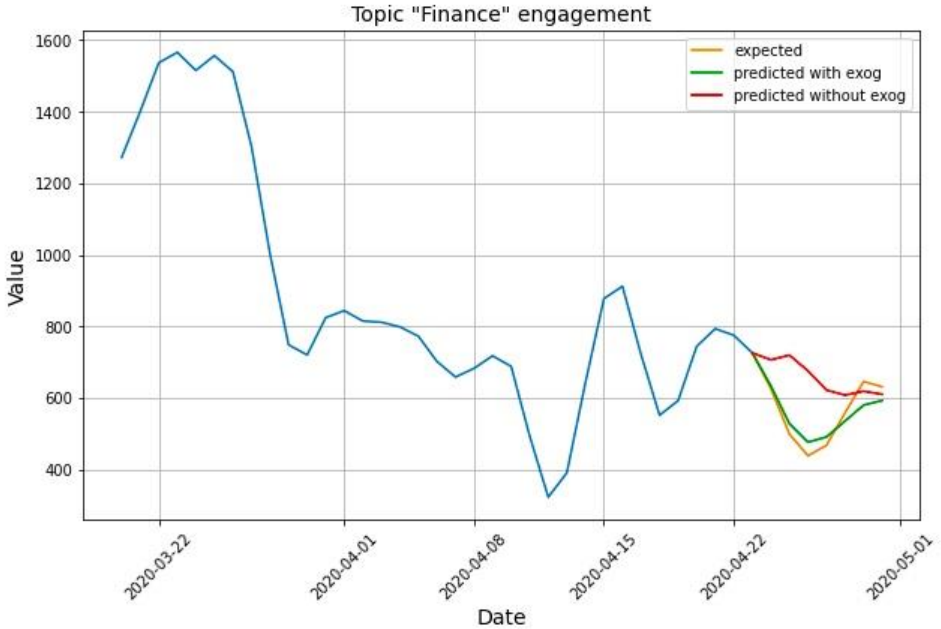


ITMO more than a UNIVERSITY

Topic forecasting results

ITMO UNIVERSITY

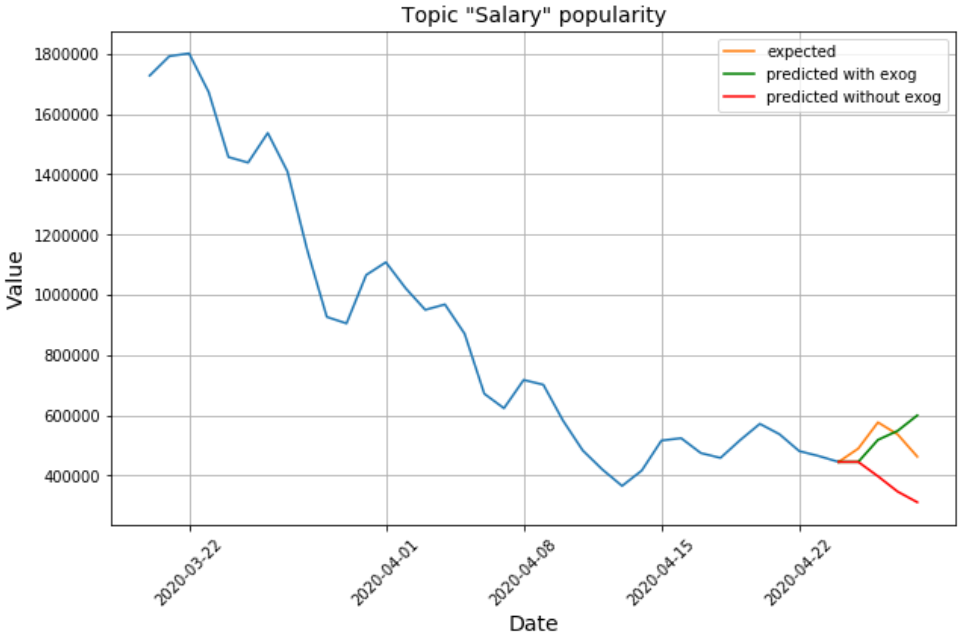


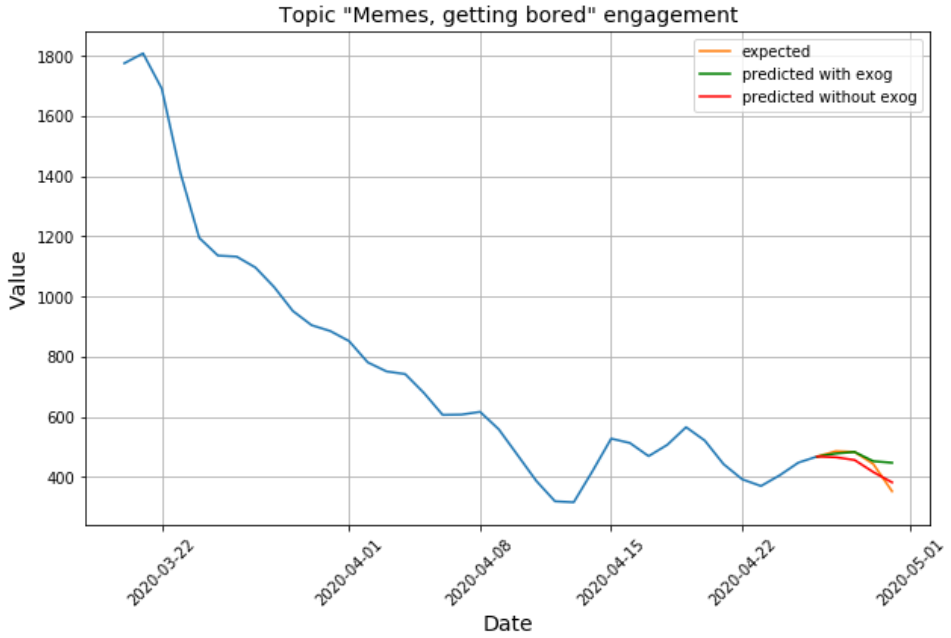


ITMO more than a UNIVERSITY

Topic forecasting results

ITMO UNIVERSITY

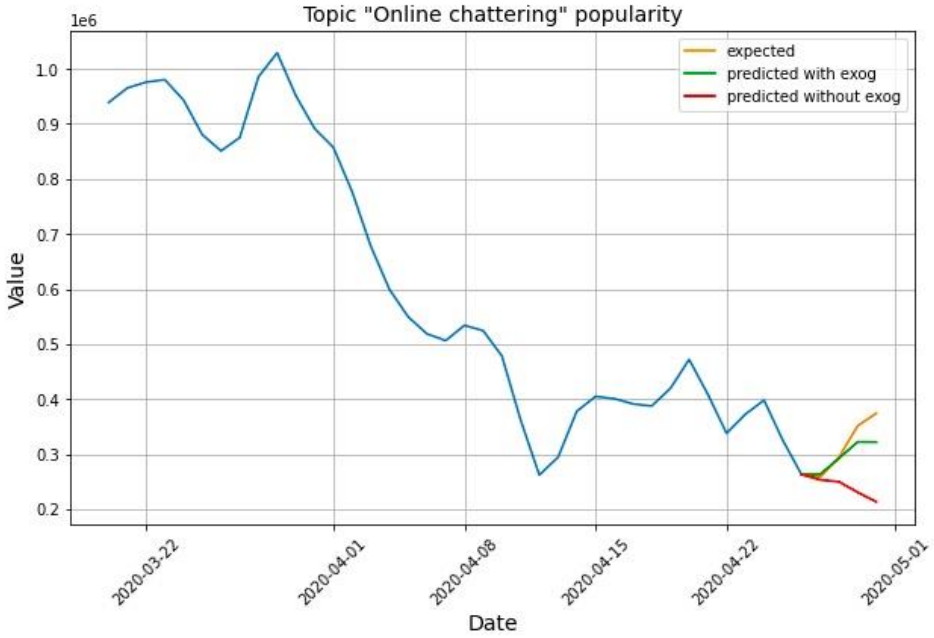


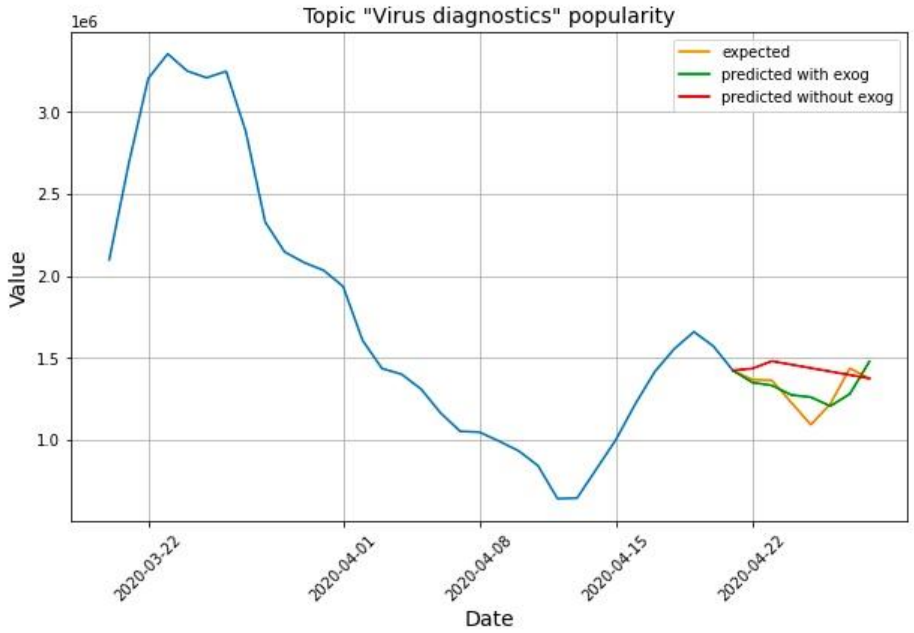


ITMO more than a UNIVERSITY

Topic forecasting results

ITMO UNIVERSITY



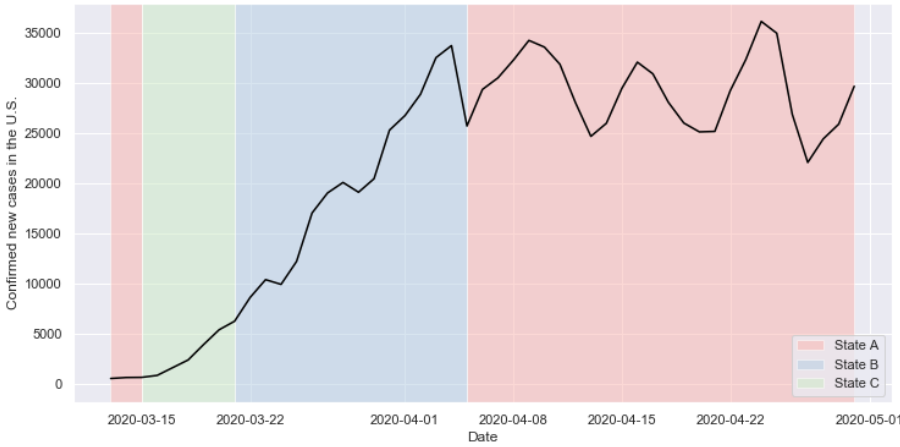


ITMO more than a UNIVERSITY

Social states identification

ITMO UNIVERSITY

- 3 hidden states – states of the society
- Observables (training data) – 7 self-isolation hashtags clusters popularity



ITMOre than a UNIVERSITY

Conclusion

ITMO UNIVERSITY

1. Topic modelling on tweets resulted with 20 well-interpreted topics;
2. The analysis of time series cross-correlations allowed to obtain good variables, exogeneous data for forecasting;
3. ARIMA with selected exogeneous data can be used for short range (around 3 days) forecasting of topics popularity and engagement. It might be useful in promotion of news, recommendations, online services or attracting attention to what is vital in time of pandemic.
4. Shown, that self-isolation related hashtags popularity has a good reflection on mobility and disease data during Covid-19 pandemic.

ITMOre than a UNIVERSITY

Thank you!

Ksh.Al.Dm@gmail.com

ITMOre than a UNIVERSITY