ShiftRy

СеМаНтИчЕсКиЕ сДвИгИ в РуССкиХ нОвОсТяХ

Настя Чижикова и Женя Егорова

Проект ShiftRy

- ShiftRy 2020: анализ диахронических изменений в употреблении слов на российских новостях;
- период с 2010 по 2020;
- новостные веб-сайты газет с разными политическими взглядами: лояльные, оппозиционные, смешанные

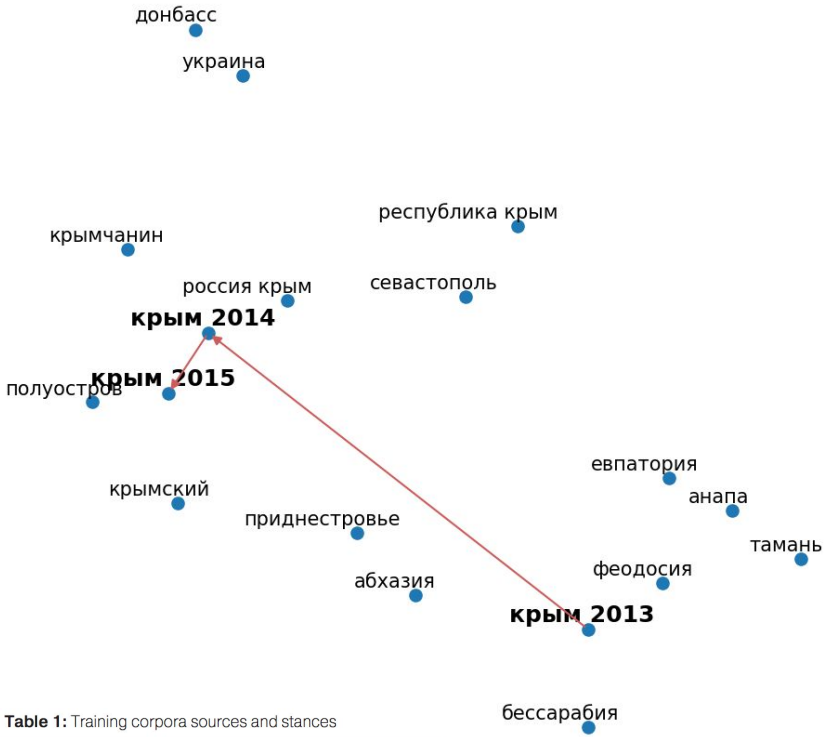


Table 1: Training corpora sources and stances

Nr.	Title	URL	Stance
1	Fontanka.ru	https://www.fontanka.ru/	Opposition
2	Gazeta.ru	https://www.gazeta.ru/	Loyal
3	Interfax	https://www.interfax.ru/	Neutral
4	Izvestia	https://iz.ru/	Loyal
5	KP	https://www.kp.ru/	Loyal
6	Lenta.ru	https://lenta.ru/	Mixed
7	Novaya Gazeta	https://novayagazeta.ru/	Opposition
8	N + 1	https://nplus1.ru/	Scientific
9	RBC	https://www.rbc.ru/	Neutral
10	The Village	https://www.the-village.ru/	Opposition

Наша идея

- Рассмотреть отдельно провластные и оппозиционные СМИ, сохранив разделение по годам, и сравнить расположения векторов слов в пространстве;

- Попробовать получать ближайших соседей при помощи маскирующих языковых моделей и сравнить результаты по годам и политической направленности.

Данные

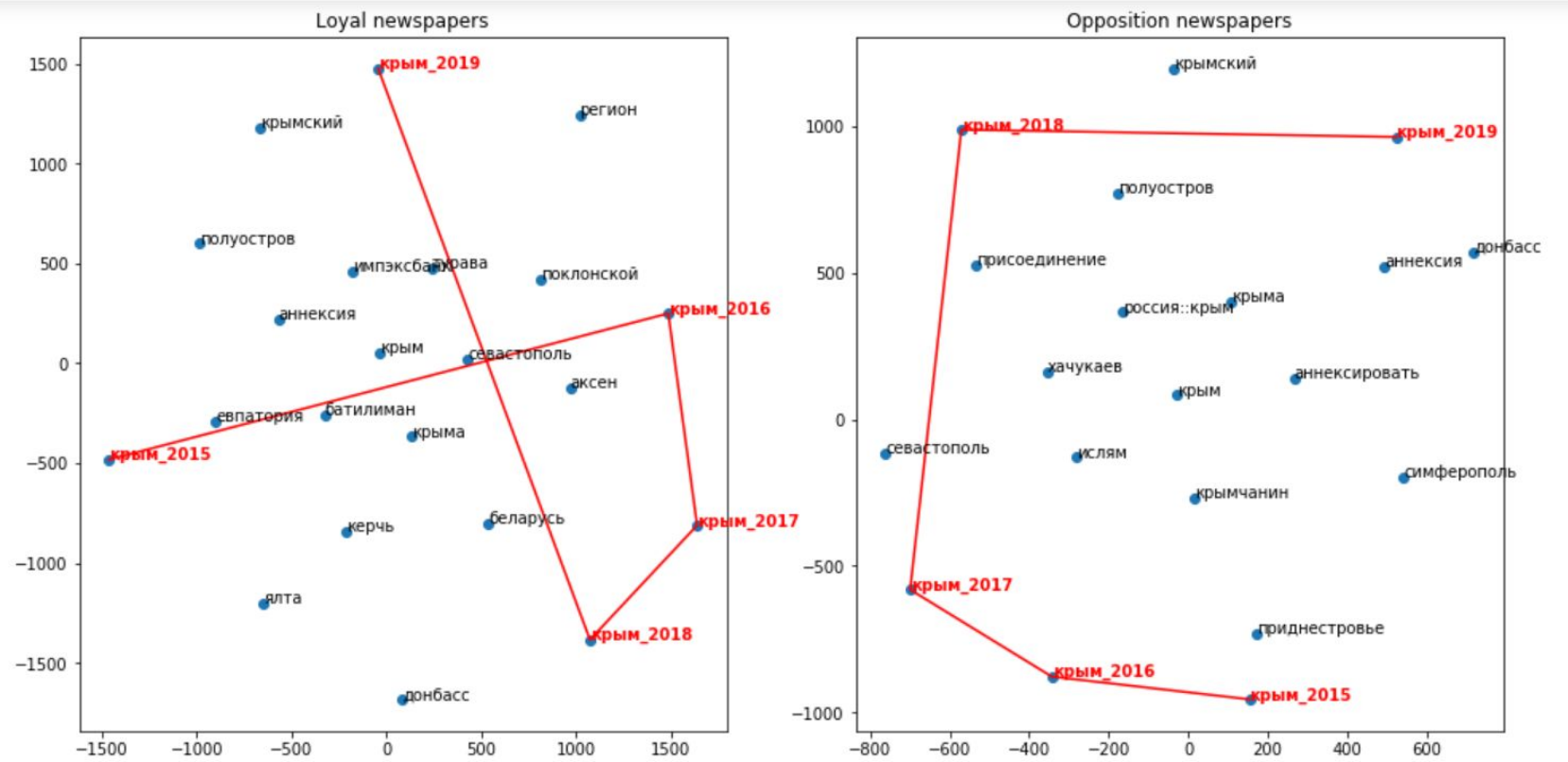
- данные с архива RusVectores с 2015 по 2019 год

- для каждого года - отдельный .txt файл для оппозиции и лоял (прошедший препроцессинг: токенизация (razdel), удаление пунктуации и стоп-слов, лемматизация и pos-tagging (nltk))

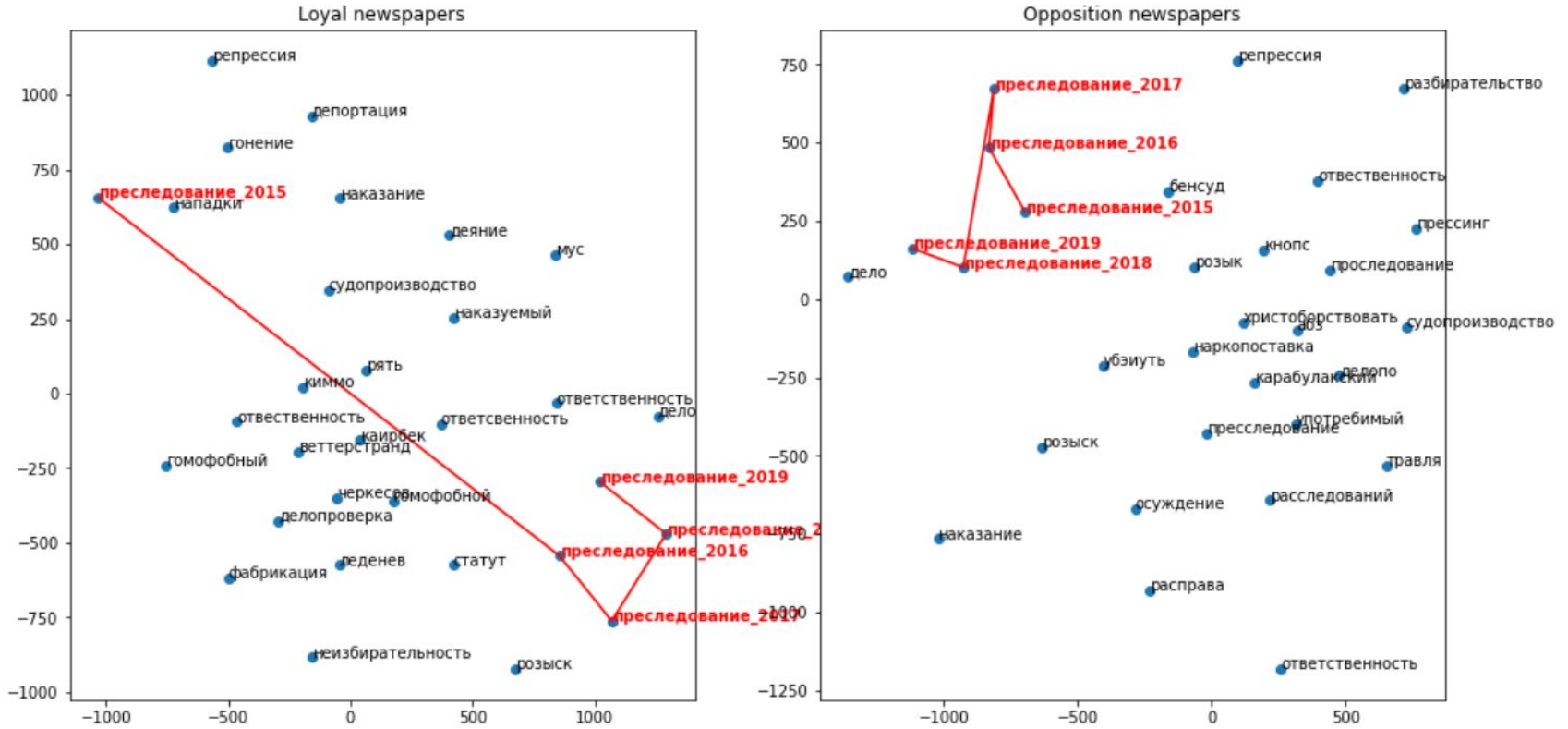
Реализация

- за основу бралась модель с RusVectores ‘taiga_upos_skipgram_300_2_2018’;
- данная модель дообучалась на новостных текстах каждого года, разделенных по политической направленности, полученные модели сохранялись;
- для построения векторов слов в одном пространстве модели выравнивались при помощи Procrustes transformation: алгоритм приближения матрицы X к матрице Y путем набора трансформаций
- при помощи t-SNE строились векторы заданного слова и 6 его ближайших соседей

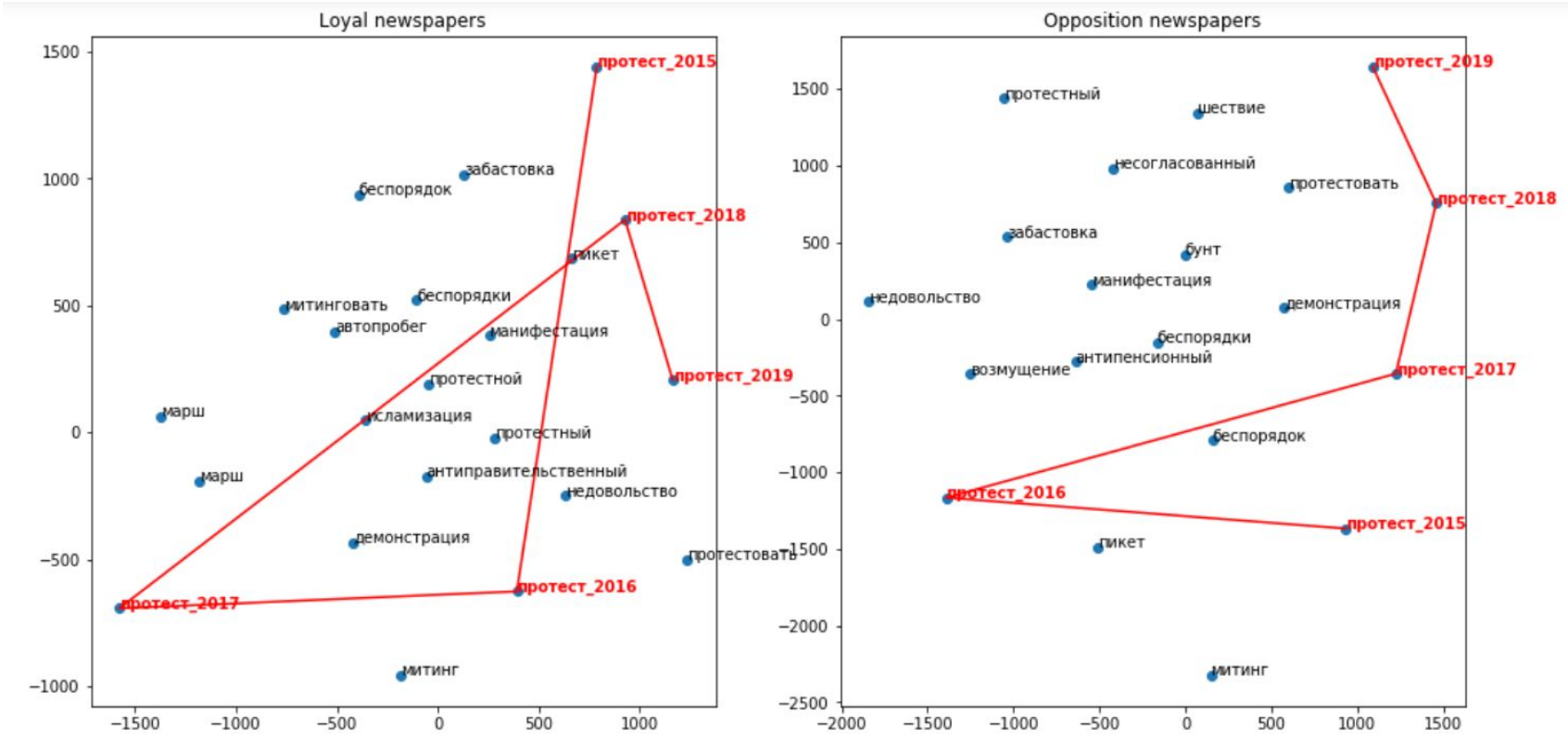
Результаты


Loyal newspapers
крым_2019
крымский
регион
полуостров
поклонской
аннексия
крым_2016
крым
севастополь
аксен
евпатория
батилиман
крыма
крым_2015
керчь
беларусь
крым_2017
ялта
крым_2018
донбасс
Opposition newspapers
крымский
крым_2018
крым_2019
полуостров
присоединение
аннексия
донбасс
крыма
россия::крым
хачукаев
аннексировать
крым
севастополь
ислям
симферополь
крымчанин
крым_2017
приднестровье
крым_2016
крым_2015

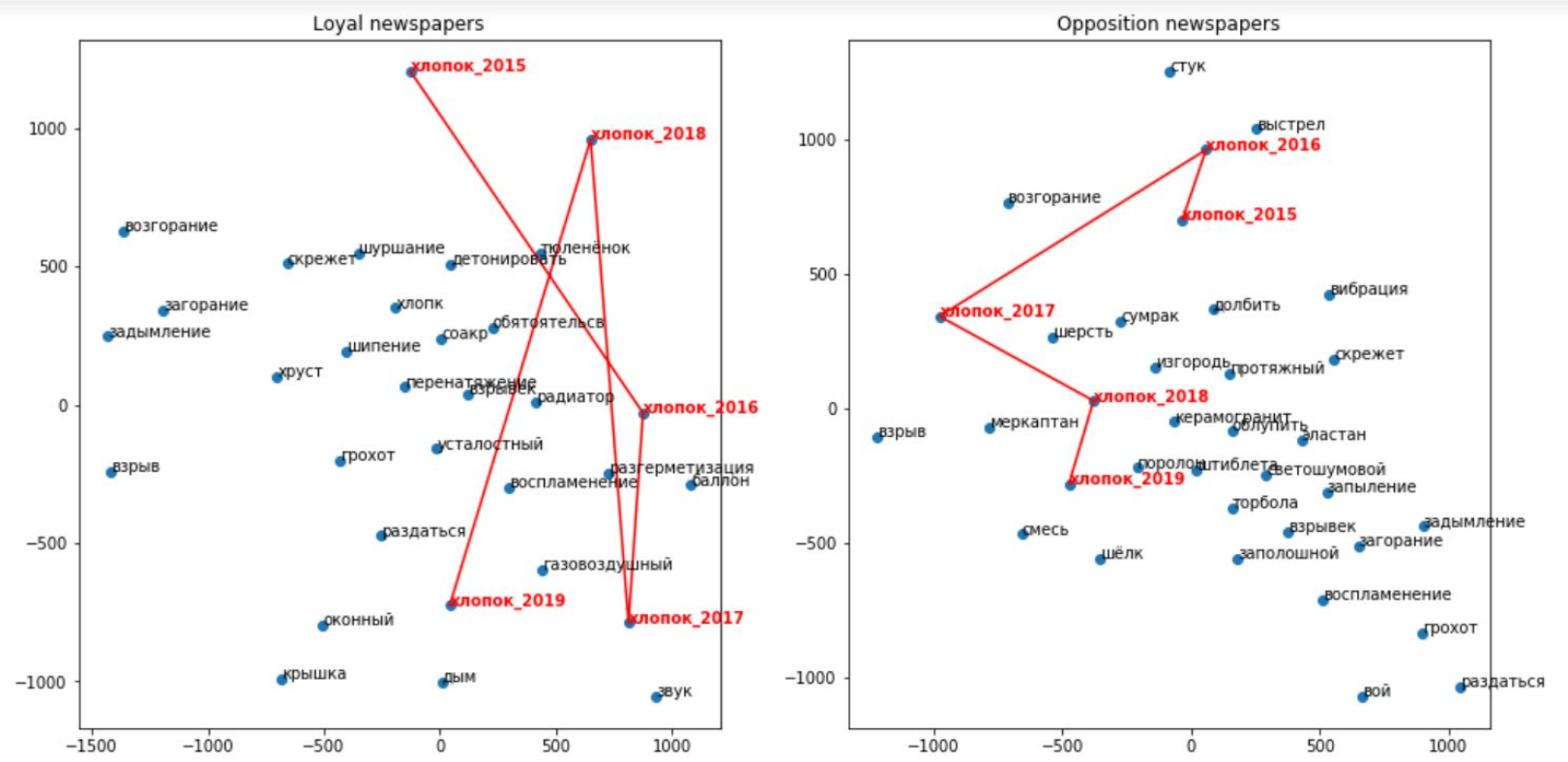
Результаты


Loyal newspapers
репрессия
депортация
гонение
преследование_2015
нападки
наказание
деяние
мус
судопроизводство
наказуемый
рять
киммо
ответственность
ответственность
ответсвенность
дело
каирбек
веттерстранд
гомофобный
преследование_2019
черкесев
гомофобной
делопроверка
преследование_2016
фабрикация
реденев
статут
преследование_2017
неизбирательность
розыск
Opposition newspapers
репрессия
преследование_2017
разбирательство
преследование_2016
ответственность
бенсуд
преследование_2015
прессинг
преследование_2019
кнопс
преследование_2018
розык
дело
проследование
христобоствовать
судопроизводство
наркопоставка
убэиуть
карабулакский
употребимый
преследование
розыск
травля
осуждение
расследований
наказание
расправа
ответственность

Результаты


Loyal newspapers
протест_2015
забастовка
беспорядок
протест_2018
пикет
митинговать
беспорядки
автопробег
манифестация
протест_2019
протестной
марш
исламизация
протестный
марш
антиправительственный
недовольство
демонстрация
протестовать
протест_2017
протест_2016
митинг
Opposition newspapers
протест_2019
протестный
шествие
несогласованный
протестовать
протест_2018
забастовка
бунт
манифестация
недовольство
демонстрация
беспорядки
возмущение
антипенсионный
протест_2017
беспорядок
протест_2016
протест_2015
пикет
митинг

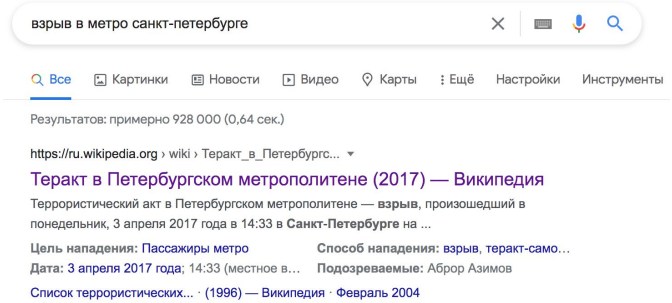
Результаты


Loyal newspapers
хлопок_2015
хлопок_2018
хлопок_2016
хлопок_2019
хлопок_2017
возгорание
загорание
задымление
взрыв
скрежет
шуршание
детонировать
тюленёнок
хлопк
обятоятельсв
соакр
шипение
хруст
перенатяжение
взрывек
радиатор
усталостный
грохот
разгерметизация
воспламенение
баллон
раздаться
газовоздушный
оконный
крышка
дым
звук
Opposition newspapers
стук
выстрел
хлопок_2016
возгорание
хлопок_2015
вибрация
хлопок_2017
сумрак
долбить
шерсть
изгородь
протяжный
скрежет
хлопок_2018
керамогранит
облупить
взрыв
меркаптан
эластан
хлопок_2019
светошумовой
запыление
торбола
взрывек
задымление
смесь
загорание
шёлк
заполошной
воспламенение
грохот
вой
раздаться

Языковые модели?

- за основу брался русскоязычный BERT: ‘rubert_cased_L-12_H-768_A-12_pt’;
- в качестве данных были взяты те же тексты с убранными pos-тегами;
- модели дообучалась на восстановление маскированных токенов.

Языковые модели



"взрыв в [MASK] в Санкт-Петербурге"

2015	2016	2017	2018
['понедельник', 'год', 'четверг', 'день', 'воскресение', 'период']	['случиться', '##ити', '##еть', '[PAD]', '##еться', 'метро']	['**метро**', 'порту', 'понедельник', 'четверг', 'дом', 'спорткомплекс']	['четверг', 'час', 'понедельник', 'случиться', '##ити', 'пропасть']

"Присоединение [MASK] к Российской Федерации"

2015	2016	2017	2018
['**крым**', 'полуостров', 'регион', 'россия', 'республика', 'территория']	['россия', 'страна', 'регион', 'это', 'федерация', 'государство']	['россия', 'крым', 'регион', 'страна', 'полуостров', 'киев']	['россия', 'страна', 'киев', 'крым', 'присоединение', 'республика']

Языковые модели

"[MASK] выиграл выборы в америке"

2015	2016	2017	2018
['год', 'также', 'политик', 'это', 'однако', 'спортсмен']	['миллиардер', 'год', **'трамп'**, 'это', 'накануне', 'ранее']	['также', 'однако', 'затем', 'кроме', 'это', **'трамп'**]	['год', 'политик', 'октябрь', 'мужчина', **'трамп'**, 'ноябрь']

"В России продолжаются преследования [MASK]."

Провластные			Оппозиционные		
2015	2016	2017	2015	2016	2017
['человек', 'ребенок', 'преступник', **'медведь'**, 'автомобиль', 'россиянин']	['человек', 'преступник', 'мужчина', 'ребенок', 'подозревать', 'женщина']	['подозревать', 'коррупция', 'террорист', 'боевик', 'человек', 'россия']	['р', 'россия', 'ст', 'обвинять', 'подозревать', 'задержать']	['страна', 'человек', 'россия', **'иностранец'**, 'воина', 'запад']	['человек', **'геев'**, 'коррупция', 'гражданин', 'год', 'задержать']

Возможные улучшения

- дополнить коллекцию текстов по годам до 2014 и после 2020 и обучить на них модели;
- в языковых моделях использовать нелемматезированные тексты:

"Против [MASK] завели уголовное дело."

The Village 2019 (оппозиционно)	Lenta.ru 2019 (провластно)
['него', 'политика', 'студента', 'них', 'художника', 'ученого']	['него', 'нее', 'мужчины', 'девушки', 'злоумышленника', 'задержанного']

"В России продолжаются преследования [MASK]."

The Village 2019 (оппозиционно)	Lenta.ru 2019 (провластно)
['оппозиционеров', 'россиян', 'политзаключенных', 'активистов', 'протестующих', 'заключенных']	['животных', 'биатлонистов', 'россиян', 'пилотов', 'геев', 'спортсменов']

СПАСИБО ЗА ВНИМАНИЕ!