



DST-OFF

Задача 01.

Сервис проверки поддельных новостей (fake news) в
сфере технологий и инноваций



НИКОЛАЙ ГАНИБАЕВ
Капитан команды

Telegram: [nganibaev](#)

Email: ganibaev@gmail.com

Тел. +7 903 851 5919

Функции: аналитика (фреймворки fasttext, natasha), формирование базы данных новостей



АЛЕКСАНДР ГАНИБАЕВ
Аналитик, разработчик

Telegram: [aganibaev](#)

Email: acleriot@gmail.com

Тел. +7 960 111 4192

Функции: программирование api-сервиса (фреймворк fastapi) и работа с github, автоматическое разворачивание в прод



АЛЕКСАНДР ГАНИБАЕВ
Аналитик, разработчик

Telegram: [valthazari](#)

Email: vasif.faradzhov@yandex.ru

Тел. +7 906 062 9089

Функции: описание концепции, презентация и дизайн

Подход к решению задачи*

1

СБОРЩИК НОВОСТЕЙ ИЗ
АВТОРИТЕТНЫХ
ИСТОЧНИКОВ

Работает постоянно в фоне. Требуется список авторитетных источников и алгоритм разбора страниц для сбора текста новостей. В целях хакатона база данных новостей будет собрана вручную с mos.ru.

2

API

Работает по запросу. На входе текст, на выходе оценка достоверности новости (1-100 баллов)

3

Интерфейс

для демонстрации работы API

*Базовый стек: Python, Natasha, FastText, FastApi

```
print(f"Загружается список новостей mos.ru  
end = '\r', flush = 'True')

Загружается список новостей mos.ru с 01.09.2021

[82]: urls_to_load = set("https://www.mos.ru/news/ite

[*]: Path('news_dir').mkdir(exist_ok=True)
path_news_dir = Path('news_dir/mos_ru')
path_news_dir.mkdir(exist_ok=True)

dl_urls = []
err_io_stream = io.StringIO()

def beauty_parse(text):
    text_news = reduce_content(text)
    t_result = re.findall('""datetime="(.*?)".*'
    t_date = t_result[0]
    t_title = text_beautifier(t_result[1])
    t_text = text_beautifier(t_result[2].replac
    return t_date+'\n'+t_title+'\n'+t_text

def news_parser(text, url, url_source=None):
    url_source = url_source or url
    file_b = beauty_parse(text).encode()
    filename = url_source.split('/')[-2]
    with open(path_news_dir / filename, 'wb') as
        shutil.copyfileobj(io.BytesIO(file_b),
        dl_urls |= url_source

read_urls_contents(urls_to_load, session=curren
if len(urls_to_load) == len(dl_urls):
    print('Все новости загружены')
```

File Edit View Run Kernel Git Snippets Tabs Settings Help

NewsLoad.ipynb x DSphere.ipynb x 101171073 x

/ news_dir / mos_ru /

Name

100031073

100035073

100044073

100088073

100146073

100150073

100163073

100235073

100252073

100313073

100384073

100425073

100470073

100485073

100535073

100554073

100580073

100618073

100626073

100642073

100661073

100704073

100711073

100725073

100759073

```

1 2022-01-11 07:07:00
2 В акселерационных программах Московского экспортного центра приняло участие более 200 компа
3 В 2021 году более 200 компаний прошли акселерационные программы, подготовленные Московским
4 хотя выйти на новые зарубежные рынки. Об этом рассказал руководитель Департамента предприн
5 «За 2021 год было проведено 11 потоков акселерации. Более 200 компаний получили финансиро
6 количество участников акселераторов увеличилось втрое», – отметил глава ведомства. В прошед
7 интернет-инициатив, платформой по развитию корпоративных инноваций GenerationS, международн
8 Российского экспортного центра и другими организациями.
9 Совместно с Московской школой управления «Сколково» реализуется программа «Экспортеры 2.0».
10 после акселерации начала продажи инфракрасных обогревателей в Нидерландах. Другое предприят
11 Еще одна востребованная программа – проект GoGlobal для ИТ-компаний, который организован со
12 компаний.
13 Выпускники проекта GoGlobal заключили более сотни экспортных контрактов. В этом году некото
14 на использование виджета виртуальной примерки с партнерами из Великобритании.
15 Часть акселераторов в этом году начали работать впервые. К примеру, программа Tech Export t
16 Они смогли проконсультироваться с экспертами, протестировать гипотезы и сформировать страте
17 Также совместно с платформой GenerationS была организована программа Global Partners Progra
18 и логистики. Некоторые участники обучения вошли в топ-100 мировых стартапов на конкурсе Sl
19 Экспорт в один клик: город помогает предпринимателям развивать бизнес онлайн по всему миру.
20 Московский экспортный центр – это проект Департамента предпринимательства и инновационного

```

[133]: df_news = {}

df_news[0] = """Москва заняла первое место среди европейских городов в рейтинге инноваций, помогающих в борьбе с COVID-19. В мире Москва занимает третье место, уступая лишь Нью-Йорку и Сан-Франциско.

Москва признана первой среди европейских городов в рейтинге инноваций, помогающих в формировании устойчивости коронавируса.

Среди мировых мегаполисов российская столица занимает третью строчку – после Сан-Франциско и Нью-Йорка. Пятерку замыкают Бостон и Лондон.

Добиться высоких показателей Москве помогло применение искусственного интеллекта. Это методика уже помогла рентгенологам проанализировать снимки легких.

Среди них алгоритмы компьютерного зрения на основе искусственного интеллекта, которые уже помогли рентгенологам проанализировать снимки легких.

Еще одно инновационное решение – облачная платформа, которая объединяет пациентов, врачей, медицинские организации, страховые компании.

Способствовали высоким результатам и технологии, которые помогают адаптировать жизнь горожан во время пандемии. Это проекты по созданию «умных» городов.

Эксперты агентства StartupBlink оценивали принятые в Москве меры с точки зрения эпидемиологических показателей и влияния на экономику.

Московский опыт

В борьбе с коронавирусом Москва отказалась от жестких мер, которые могут нанести ущерб экономике.

Московская система здравоохранения за время пандемии показала высокую эффективность.

Столица поддерживает бизнес, выделяя субсидии на аренду помещений.

Как составляется рейтинг

Рейтинг составляется на базе глобальной карты инновационных решений по борьбе с коронавирусом.

Алгоритм рейтинга учитывает количество и тип инновационных решений.

df_news[1] = """Москва заняла первое место среди европейских городов в рейтинге инноваций, помогающих в борьбе с COVID-19.

В мире Москва занимает третье место, уступая лишь Нью-Йорку и Сан-Франциско.

Москва признана первой среди европейских городов в рейтинге инноваций, помогающих в формировании устойчивости коронавируса.

Среди мировых мегаполисов российская столица занимает третью строчку – после Сан-Франциско и Нью-Йорка. Пятерку замыкают Бостон и Лондон.

Добиться высоких показателей Москве помогло применение искусственного интеллекта. Это методика уже помогла рентгенологам проанализировать снимки легких.

Среди них алгоритмы компьютерного зрения на основе искусственного интеллекта, которые уже помогли рентгенологам проанализировать снимки легких.

Еще одно инновационное решение – облачная платформа, которая объединяет пациентов, врачей, медицинские организации, страховые компании.

Способствовали высоким результатам и технологии, которые помогают адаптировать жизнь горожан во время пандемии. Это проекты по созданию «умных» городов.

Эксперты агентства StartupBlink оценивали принятые в Москве меры с точки зрения эпидемиологических показателей и влияния на экономику.

Московский опыт

В борьбе с коронавирусом Москва отказалась от жестких мер, которые могут нанести ущерб экономике.

Московская система здравоохранения за время пандемии показала высокую эффективность.

Столица поддерживает бизнес, выделяя субсидии на аренду помещений.

Как составляется рейтинг

Рейтинг составляется на базе глобальной карты инновационных решений по борьбе с коронавирусом.

Алгоритм рейтинга учитывает количество и тип инновационных решений.

df_news[2] = """Москва стала лидером в Европе в рейтинге инноваций, помогающих в борьбе с COVID-19.

В мире российская столица заняла третье место, обогнав Лондон и Барселону.

Москва заняла первое место среди европейских городов в рейтинге инноваций, помогающих в борьбе с COVID-19, опередив Лондон и Барселону.

В российской столице применяются почти 160 передовых решений для борьбы с распространением коронавируса. Среди них алгоритмы компьютерного зрения на основе искусственного интеллекта, которые уже помогли рентгенологам проанализировать снимки легких.

Рейтинг составляется на базе глобальной карты инновационных решений по борьбе с коронавирусом и оценивает около 100 ведущих городов мира.

df_news[3] = """Москва заняла первое место в Европе по инновациям в борьбе с COVID-19.

Москва обошла европейские столицы в рейтинге инноваций по устойчивости к COVID-19, опередив Лондон и Барселону, сообщается в рейтинге инноваций.

Российская столица также заняла третье место среди мировых мегаполисов. В пятерку лидеров вошли Бостон и Лондон.

Занять лидирующие позиции в рейтинге Москве помогли около 50 передовых решений, которые применяются для борьбы с распространением коронавируса.

Одно из таких решений – алгоритмы компьютерного зрения на основе искусственного интеллекта, которые уже помогли рентгенологам проанализировать снимки легких.

Также высоким результатам способствовали технологии, помогающие адаптировать жизнь москвичей во время пандемии. Среди них проекты по созданию «умных» городов.

```
from transformers import pipeline
p = pipeline(
    task='zero-shot-classification',
    model='cointegrated/rubert-base-cased-nli-threeway'
)
p(
    sequences=df_news[8],
    candidate_labels="Хорошо, Плохо",
    hypothesis_template="{.}"
)
```

```
{'sequence': 'Проектный офис ФЭСН РАНХиГС завершает 30 проектов. И начинает новые.\nЕжегодно в мае-июне Проектный офис Факультета экономических и социальных наук проводит конкурс заявок на участие в проектах.\nСреди компаний-заказчиков были крупные российские компании и организации: Сбер, РЖД, Газпром, Альфа-Банк, ВТБ, Дом.РФ, Очаково, Сегежа-Групп а также представительства зарубежных компаний: BMW, Delonghi, L'Oreal, Pfizer.\nФЭСН является крупнейшим университетским проектным центром не только по количеству и сложности проектов, и не только по количеству студентов-участников. Так, еще пять лет назад ФЭСН вовлек в проектную деятельность университеты пяти стран Европы и стал разрабатывать с ними полугодовые проекты для студентов.\nВ рамках проекта ФЭСН работали вместе с магистрами из Германии и Бельгии. А в одном из проектов участвовали команды из Германии, Бельгии, Италии, 4 проекта.\n', 'labels': ['Хорошо', 'Плохо'], 'scores': [0.7546549439430237, 0.2453451007604599]}
```



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

```
from transformers import pipeline
p = pipeline(
    task='zero-shot-classification',
    model='cointegrated/rubert-base-cased-nli-threeway'
)
p(
    sequences=df_news[8],
    candidate_labels="Хорошо, Плохо",
    hypothesis_template="{ }."
)
```

```
{'sequence': 'Проектный офис ФЭСН РАНХиГС завершает 30 проектов. И начинает новые.\nЕжегодно в мае-июне Проектный офис Факультета экономических и  
зультаты заказчиком – российским и международным компаниям.\nСреди компаний-заказчиков были крупные российские компании и организации: Сбер, РЖД,  
ого, холдинг САВООБ ФУДС, Альфа-Банк, ВТБ, Дом РФ, Очаково, Сегежа-Групп а также представительства зарубежных компаний: BMW, DeLonghi, L'Oreal, Pf  
ФЭСН является крупнейшим университетским проектным центром не только по количеству и сложности проектов, и не только по количеству студентов-участ  
ы. Так, еще пять лет назад ФЭСН вовлек в проектную деятельность университеты пяти стран Европы и стал разрабатывать с ними полугодовые проекты дл  
делались усилиями проектных групп нескольких стран; например, в проекте от генерального директора BMW Russia по «Разработке маркетинговой стратеги  
ии» бакалавры ФЭСН работали вместе с магистрами из Германии и Бельгии. А в одном из проектов участвовали команды из Германии, Бельгии, Италии, 4 п  
, 'labels': ['Хорошо', 'Плохо'],  
, 'scores': [0.7546549439430237, 0.2453451007604599]}
```

```
[236]: from sentence_transformers import SentenceTransformer, util
        model = SentenceTransformer('sentence-transformers/LaBSE')

[241]: import numpy as np
        # Two lists of sentences
        sentences1 = list(df_news.values())

        sentences2 = list(df_news.values())

        #Compute embedding for both lists
        embeddings1 = model.encode(sentences1, convert_to_tensor=True)
        embeddings2 = model.encode(sentences2, convert_to_tensor=True)

        #Compute cosine-similarities
        cosine_scores = util.cos_sim(embeddings1, embeddings2).numpy().mean()
        cosine_scores

0.7916351

[244]: cosine_scores = util.cos_sim(embeddings1, embeddings2)
        #Output the pairs with their score
        for i in range(1):
            for j in range(len(sentences2)):
                print("{} \t\t {} \t\t Score: {:.4f}".format(i, j, cosine_scores[i][j]))

0           0           Score: 1.0000
0           1           Score: 1.0000
0           2           Score: 0.9246
0           3           Score: 0.9011
0           4           Score: 0.8411
0           5           Score: 0.9015
0           6           Score: 0.9513
0           7           Score: 0.7405
0           8           Score: 0.5089
```

0	0	Score: 1.0000
0	1	Score: 0.9433
0	2	Score: 0.8958
0	3	Score: 0.8647
0	4	Score: 0.8255
0	5	Score: 0.9211
0	6	Score: 0.9210
0	7	Score: 0.6971
0	8	Score: 0.5612

Ввод [11]: 1 df_news[1].splitlines()[1]

executed in 22ms, finished 21:08:48 2022-06-11

Out[11]: 'В мире Москва занимает третье место, уступая лишь Нью-Йорку и Сан-Франциско.'

```
Ввод [28]: 1 import pandas as pd
2 from dostoevsky.tokenization import RegexTokenizer
3 from dostoevsky.models import FastTextSocialNetworkModel
4 import fasttext
5
6 fasttext.FastText.eprint = lambda x: None
7 tokenizer = RegexTokenizer()
8
9 model = FastTextSocialNetworkModel(tokenizer=tokenizer)
10
11 messages = df_news.values()
12 # messages = df_topics
13 # messages = [x.splitlines()[1] for x in df_news.values()]
14
15 results = model.predict(messages)
16
17 p = pd.DataFrame(messages, columns=['text'])
18 p['positive']=""
19 p['negative']=""
20 p['result']=""
21
22 for id, sentiment in enumerate(results):
23     p['positive'][id] = f"{sentiment['positive']:.2f}"
24     p['negative'][id] = f"{-sentiment['negative']:.2f}"
25     p['result'][id] = f"{sentiment['positive']-sentiment['negative']:.2f}"
26
27 p
```

executed in 851ms, finished 21:25:32 2022-06-11

Out[28]:

	text	positive	negative	result
0	Москва заняла первое место среди европейских г...	0.06	-0.19	-0.12
1	Москва заняла первое место среди европейских г...	0.08	-0.18	-0.10
2	Москва стала лидером в Европе в рейтинге иннов...	0.09	-0.17	-0.08
3	Москва заняла первое место в Европе по инновац...	0.08	-0.15	-0.07
4	Москва заняла первое место в Европе по инновац...	0.07	-0.17	-0.10
5	Москва стала первой в Европе среди городов с и...	0.08	-0.22	-0.14
6	Москва заняла восьмое место среди европейских ...	0.08	-0.16	-0.08
7	Бостон занял первое место среди европейских го...	0.08	-0.22	-0.14
8	Проектный офис ФГСН РАНХиГС завершает 30 проек...	0.07	-0.20	-0.13

Ввод [29]:

```
1 from autocorrect import Speller
2 spell = Speller(lang='ru', fast=True, only_replacements=True)
```

executed in 8.94s, finished 21:25:42 2022-06-11

Ввод [30]:

```
1 # Проходим spell-чеккером по всем текстам
2 p["spell_correct"] = p['text'].apply(lambda x: spell(str(x)))
3 p
```

executed in 119ms, finished 21:25:42 2022-06-11

...

Ввод [36]:

1 p

executed in 44ms, finished 21:25:53 2022-06-11

Out[36]:

	text	positive	negative	result	spell_correct	err_count
0	Москва заняла первое место среди европейских г...	0.06	-0.19	-0.12	Москва заняла первое место среди европейских г...	1
1	Москва заняла первое место среди европейских г...	0.08	-0.18	-0.10	Москва заняла первое место среди европейских г...	0
2	Москва стала лидером в Европе в рейтинге иннов...	0.09	-0.17	-0.08	Москва стала лидером в Европе в рейтинге иннов...	0
3	Москва заняла первое место в Европе по инновац...	0.08	-0.15	-0.07	Москва заняла первое место в Европе по инновац...	0
4	Москва заняла первое место в Европе по инновац...	0.07	-0.17	-0.10	Москва заняла первое место в Европе по инновац...	0
5	Москва стала первой в Европе среди городов с и...	0.08	-0.22	-0.14	Москва стала первой в Европе среди городов с и...	0
6	Москва заняла восьмое место среди европейских ...	0.08	-0.16	-0.08	Москва заняла восьмое место среди европейских ...	0
7	Бостон занял первое место среди европейских го...	0.08	-0.22	-0.14	Бостон занял первое место среди европейских го...	0
8	Проектный офис ФЭСН РАНХиГС завершает 30 проек...	0.07	-0.20	-0.13	Проектный офис ФЭСН РАНХиГС завершает 30 проек...	1

Ввод [31]:

1 from spellchecker import SpellChecker

```
Ввод [102]: 1 from natasha import (
2     Segmenter,
3     MorphVocab,
4     NewsEmbedding,
5     NewsMorphTagger,
6     NewsSyntaxParser,
7     NewsNERTagger,
8     Doc,
9 )
10
11 emb = NewsEmbedding()
12 segmenter = Segmenter()
13 morph_vocab = MorphVocab()
14 ner_tagger = NewsNERTagger(emb)
15 # morph_tagger = NewsMorphTagger(emb)
16 # syntax_parser = NewsSyntaxParser(emb)
17
```

executed in 2.44s, finished 05:47:48 2022-08-12

```
Ввод [114]: 1 # text = ''.join(m.Lemmatize(df_news[0]))
2 text = df_news[0]
3 doc = Doc(text)
4 doc.segment(segmenter)
5 doc.tag_ner(ner_tagger)
6 # doc.tag_morph(morph_tagger)
7 # doc.parse_syntax(syntax_parser)
```

executed in 104ms, finished 05:58:08 2022-08-12

```
Ввод [115]: 1 for span in doc.spans:
2     span.normalize(morph_vocab)
3
4 NER_extract = dict(pd.DataFrame([(x.type, x.normal) for x in doc.spans], columns=['type', 'title'])
5     .groupby('type').title.unique())
6
7 if 'ORG' in NER_extract:
8     print('Найдены организации: ', ', '.join(NER_extract['ORG']))
9
10 if 'LOC' in NER_extract:
11     print('\nНайдены места: ', ', '.join(NER_extract['LOC']))
12
13 if 'PER' in NER_extract:
14     print('\nНайдены люди: ', ', '.join(NER_extract['PER']))
```

executed in 22ms, finished 05:58:12 2022-08-12

Найдены организации: startupBlink

Найдены места: Москва, Нью-Йорку, Сан-Франциско, Лондон, Барселону, Нью-Йорка, Бостон, Москве











Заголовок

- кликбейт/желтизна заголовка
- тональность заголовка
- близость семантики заголовка содержанию новости



Новости

- тональность новости
- Охват /распространенность новости в авторитетных источниках
- проверка на опечатки

Сравнение с первоисточником:



Формируем для каждой новости набор метрик/оценок:

семантическая близость
заголовков

01



семантическая близость
первых абзацев (обычно
суть новости)

02



семантическая близость
текстов без первого абзаца

03



04

близость по тональности
заголовков



05

близость по тональности
первых абзацев



06

близость по тональности
текстов без первого абзаца



Охват/распространенность новости в авторитетных источниках

- учитываются новости с семантической схожестью более _80_% за _2_ дня



Тональность новости

- стандартная новость нейтральна



Дополнительные метрики*



Отсутствие лишних упоминаний по сравнению с источником

Полнота всех упоминаний по сравнению с источником

Соответствие связей между сущностями в первоисточнике и новости

*рассчитываются после выявления в тексте новости и первоисточника именованных сущностей (NER) - персоналий, организаций, мест, числительных









