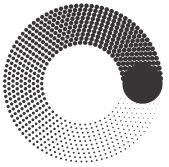
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

****

**МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

***Факультет Информационных технологий***

***Кафедра Информатики и информационных технологий***

**Направление подготовки**

**09.03.02 «Информационные системы и технологии»**

**ЛАБОРАТОРНАЯ РАБОТА № 4**

**Дисциплина:** Инструменты визуализации данных

**Тема:** Визуализация текста: разработка графа знаний по текстовым данным

**Выполнила: студент группы 231-338**

Шаура Илья Максимович

**Дата, подпись \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_\_\_\_\_

(Дата) (Подпись)

**Проверил:** Солонец В.И., к.т.н., доцент **\_\_\_\_\_\_\_\_\_\_**

(Фамилия И.О., степень, звание) (Оценка)

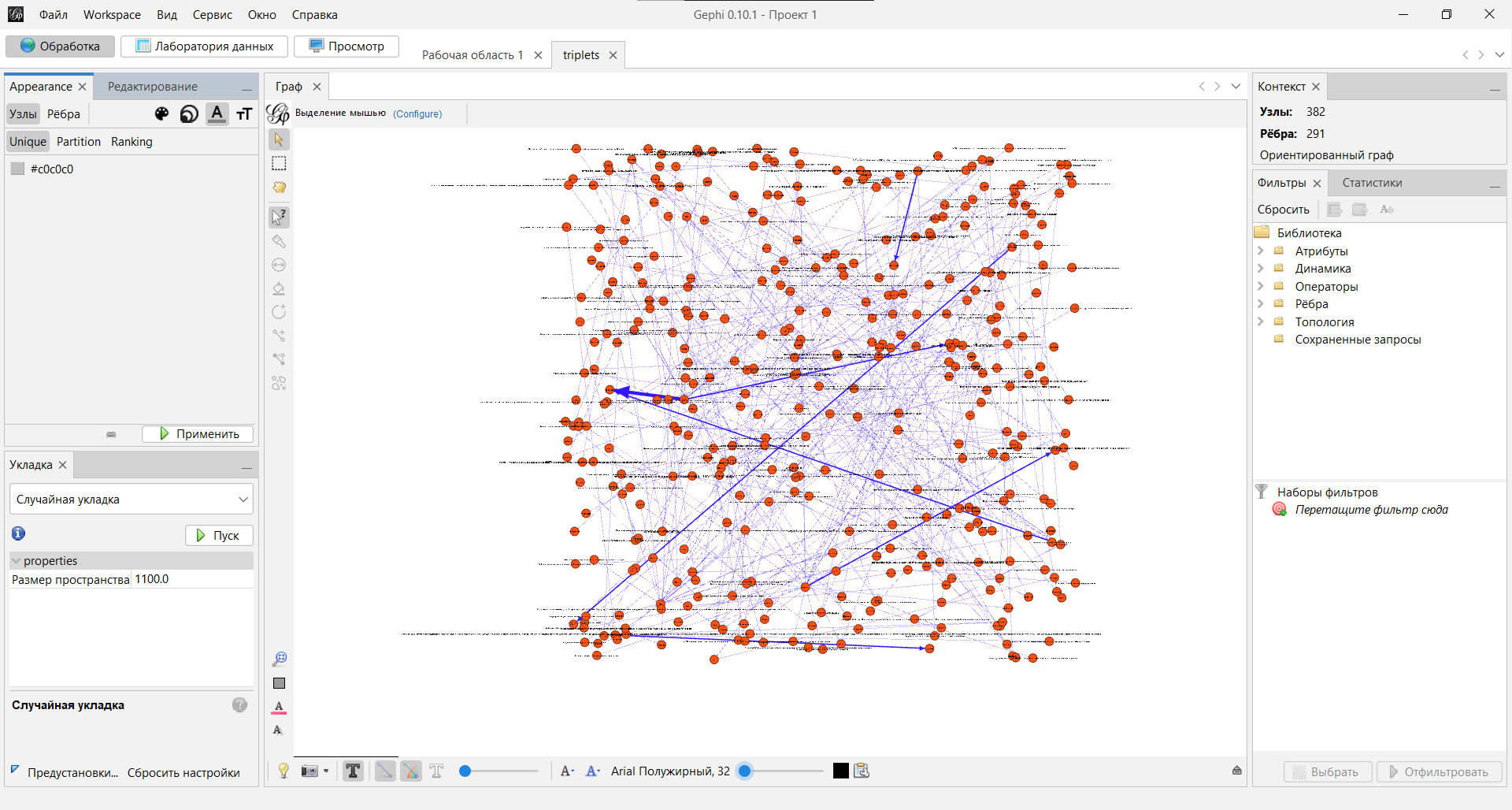
**Дата, подпись \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_\_\_\_\_

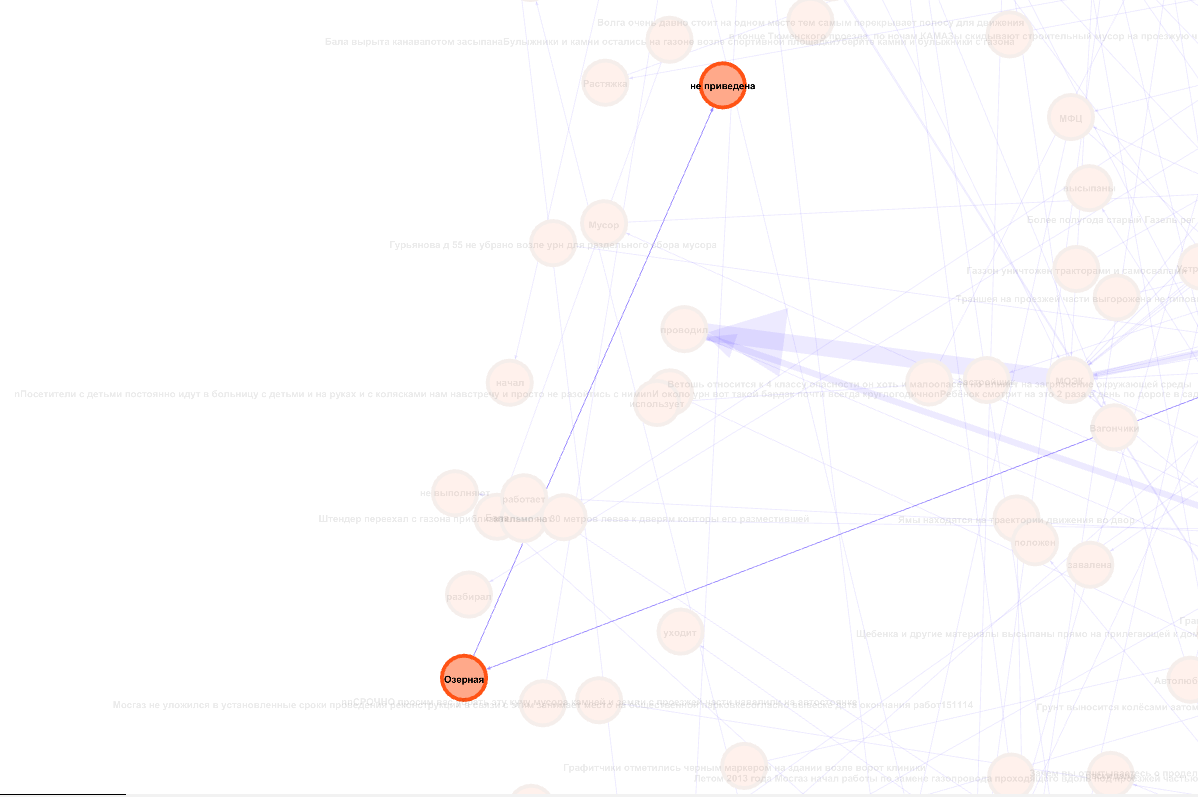
(Дата) (Подпись)

**Замечания:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Москва**

**2024**





import pandas as pd

from tqdm import tqdm

import stanza

import csv

import nltk

from nltk.tokenize import word\_tokenize, sent\_tokenize

df = pd.read\_excel('fill\_info.xlsx')

nltk.download('punkt')

full\_corpus = df["TEXT"].values

sentences = [sent for corp in full\_corpus for sent in sent\_tokenize(corp, language="russian")]

long\_sents = [i for i in sentences if len(i) > 30]

nlp = stanza.Pipeline(lang='ru',

processors='tokenize,pos,lemma,ner,depparse')

triplets = []

for s in tqdm(long\_sents):

    doc = nlp(s)

    for sent in doc.sentences:

        entities = [ent.text for ent in sent.ents]

        res\_d = dict()

        temp\_d = dict()

        for word in sent.words:

            temp\_d[word.text] = {"head": sent.words[word.head-1].text, "dep": word.deprel, "id": word.id}

        for k in temp\_d.keys():

            nmod\_1 = ""

            nmod\_2 = ""

            if (temp\_d[k]["dep"] in ["nsubj", "nsubj:pass"]) & (k in entities):

                res\_d[k] = {"head": temp\_d[k]["head"]}

                for k\_0 in temp\_d.keys():

                    if (temp\_d[k\_0]["dep"] in ["obj", "obl"]) &\

                       (temp\_d[k\_0]["head"] == res\_d[k]["head"]) &\

                        (temp\_d[k\_0]["id"] > temp\_d[res\_d[k]["head"]]["id"]):

                        res\_d[k]["obj"] = k\_0

                        break

                for k\_1 in temp\_d.keys():

                    if (temp\_d[k\_1]["head"] == res\_d[k]["head"]) & (k\_1 == "не"):

                        res\_d[k]["head"] = "не "+res\_d[k]["head"]

                if "obj" in res\_d[k].keys():

                    for k\_4 in temp\_d.keys():

                        if (temp\_d[k\_4]["dep"] =="nmod") &\

                           (temp\_d[k\_4]["head"] == res\_d[k]["obj"]):

                            nmod\_1 = k\_4

                            break

                    for k\_5 in temp\_d.keys():

                        if (temp\_d[k\_5]["dep"] =="nummod") &\

                           (temp\_d[k\_5]["head"] == nmod\_1):

                            nmod\_2 = k\_5

                            break

                    res\_d[k]["obj"] = res\_d[k]["obj"]+" "+nmod\_2+" "+nmod\_1

        if len(res\_d) > 0:

            triplets.append([s, res\_d])

clear\_triplets = []

for tr in triplets:

    for k in tr[1].keys():

        if "obj" in tr[1][k].keys():

            clear\_triplets.append([tr[0], k, tr[1][k]['head'], tr[1][k]['obj']])

file\_path = 'triplets.csv'

with open(file\_path, mode='w', newline='', encoding='utf-8') as file:

    writer = csv.writer(file)

    writer.writerow(["Source", "Target", "Type"])

    for tr in clear\_triplets:

        writer.writerow([tr[0], tr[1], "directed"])

        writer.writerow([tr[1], tr[2], "directed"])

print(f"CSV файл успешно создан: {file\_path}")