**Хакатон, команда 2**

**Ощепков Дмитрий Владимирович РУДН**

**Даниил Шлык**

**Евгений Ибатуллин**

**Колосова Екатерина Дмитриевна Дубна**

**1. Методология исследования**

**1.1 Цели исследования**

Цель данного исследования — провести анализ текстовых данных, содержащихся в наборах данных о публикациях научного контента. Мы будем использовать собранный нами датасет, модели для обработки естественного языка и линейную комбинацию результатов этих моделей для поиска интересующей нас информации.

**1.2 Данные**

Данные были получены из CSV файлов. Каждый файл содержит колонки, которые включают заголовок документа, аннотацию, название источника, издателя, ссылку, ключевые слова и тд.

Датасет, помогающий нам оценивать текст в колонках, был собран c arXive и частично вручную с сайта ОИЯИ.

**1.3 Обработка данных**

Для обработки текстовых данных использовались следующие шаги:

1. **Загрузка данных**: Считывание CSV файлов с помощью библиотеки pandas.
2. **Предобработка**: Приведение всех данных к строковому формату для обеспечения однородности, выделели только существенные на наш взгляд признаки.
3. **Инициализация новых колонок**: Создание новых колонок для хранения оценок, а также общей оценки (counter).
4. **Применение моделей**: Для каждого документа применяются функции обработки текста, которые возвращают оценки для разных аспектов документа.

**1.4 Оценка**

Оценка каждой строки происходит путем вычисления взвешенной суммы из других оценок различных колонок. Веса настраивались, исходя из здравого смысла.

Какова главная идея? Собираем псевдо - обучающую выборку примерно от 100 до 2000 значений на признак. Признаки: 'Title', 'Abstract', 'Source title', 'Publisher', 'Link', 'Author Keywords'

Например, учим TfidVectorizer на 2000 названиях статей близких по тематике на условие Хакатона. Соответственно, когда мы применяем его к колонке с названием, содержащим похожие слова, векторайзер оценивает его близко к 1, или, напротив, близко к 0.

Еще мы перевели весь текст в инфинитивные формы и убрали предлоги. Таким образом, применяем 10 условно разных векторайзеров на 10 разных признаков и собираем некоторое множество значений от 0 до 1. Умножаем их на веса и складываем.

Вследствие чего, получаем наиболее похожие на описание нас строки датасета с наибольшими значениями сумм.

**1.5 Диапазоны значений**

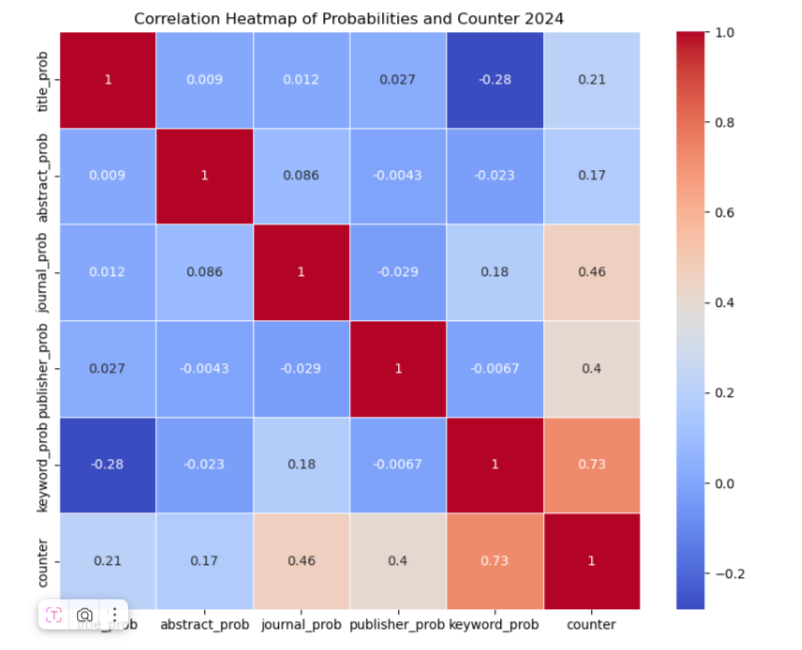
Для 2024 года топ 100 наибольших целевых коэффициентов оказались в пределах от 16.4 до 16.9

Для 2020 года топ 100 наибольших целевых коэффициентов оказались в пределах от 16.4 до 19 (то есть здесь более выгодные для нас организации)

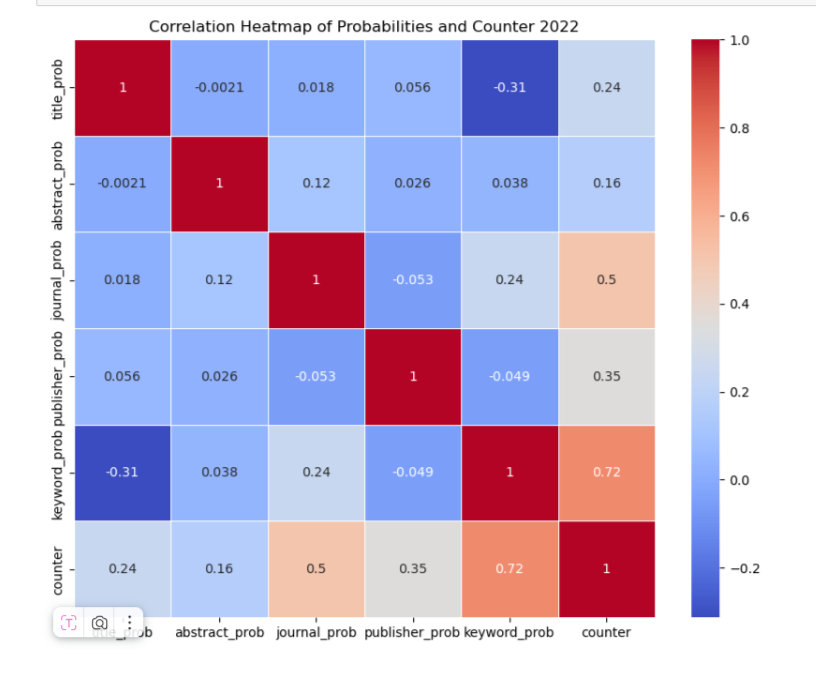
**1.6 Графики корреляции признаков и целевой переменной**

**Самый информативный график для понимая картины в целом**

**Матрица за 2024**



**Матрица за 2020**

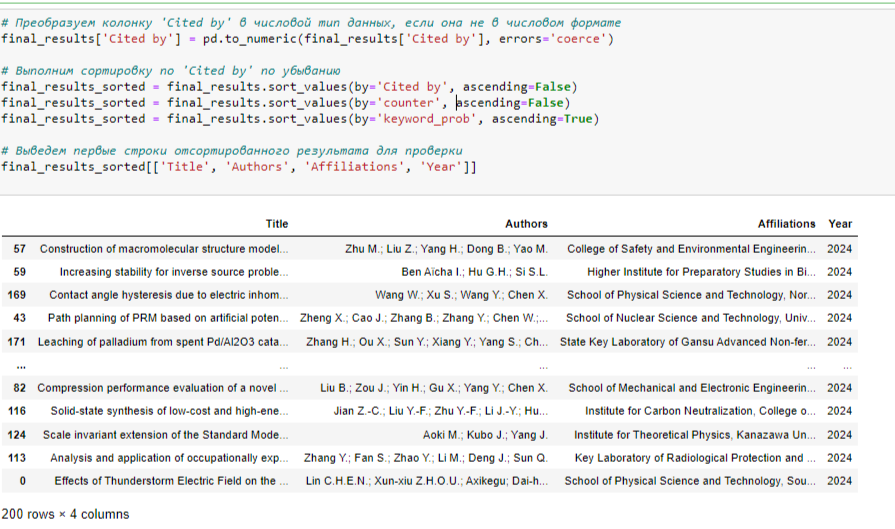


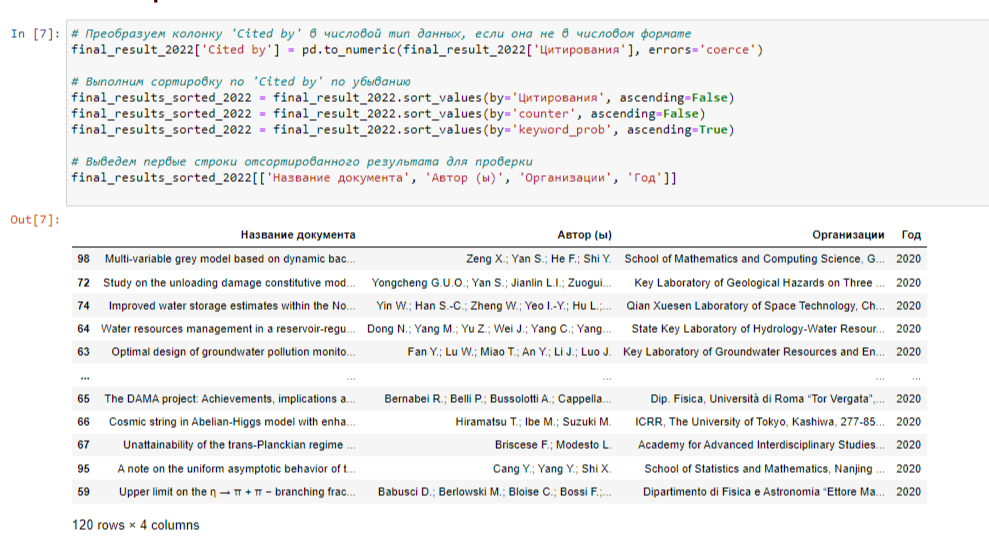
Соответственно, почти такие же матрицы для 2022, 2021, 2023 годов

**1.7 Статистика**

Сортируем значения по количеству цитировний автора, по counter, а потом с наименьшим keyword (так как у него слишком высокая корреляция с ответом получилась из-за немного неадекватно подобранного веса, а хочется в первую очередь видеть варианты с равноценным вкладом всех признаков)

Перед нами по 200 статей за каждый год, максимально похожих на ОИЯИ по всем признакам (так как изначально мы брали обучающий датасет по ОИЯИ или из подобных организаций)





Файлы полных финальных таблиц и рекомендаций будут приложены вместе с отчетом и программным кодом. https://github.com/Doshchepkov/Hackha

Маленькая часть за 2024 год:

1. Author: Zhu M.; Liu Z.; Yang H.; Dong B.; Yao M., Affiliations: College of Safety and Environmental Engineering, Shandong University of Science and Technology, 579 Qianwangang Rd, Huangdao District, Qingdao, 266590, China; State Key Laboratory of Mining Disaster Prevention and Control Co-founded by Shandong Province and the Ministry of Science and Technology, Shandong University of Science and Technology, Qingdao, 266590, China

2. Author: Ben Aïcha I.; Hu G.H.; Si S.L., Affiliations: Higher Institute for Preparatory Studies in Biology-Geology (ISEP-BG), University of Carthage, La Soukra, Tunis, 2036, Tunisia; LAMSIN, National Engineering School of Tunis, B.P. 37, Tunis, 1002, Tunisia; School of Mathematical Sciences, Nankai University, Tianjin, 300071, China; School of Mathematics and statistics, Shandong University of Technology, Shangdong, 255049, China

3. Author: Wang W.; Xu S.; Wang Y.; Chen X., Affiliations: School of Physical Science and Technology, Northwestern Polytechnical University, Shaanxi Province, Xi'an, 710129, China; Computer Science Department, Georgia State University, Atlanta, 30301, GA, United States

4. Author: Zheng X.; Cao J.; Zhang B.; Zhang Y.; Chen W.; Dai Y.; Zhao J., Affiliations: School of Nuclear Science and Technology, University of South China, Hengyang, 421001, China

5. Author: Zhang H.; Ou X.; Sun Y.; Xiang Y.; Yang S.; Chen Z., Affiliations: State Key Laboratory of Gansu Advanced Non-ferrous Metal Materials, Lanzhou University of Technology, Gansu, Lanzhou, 730050, China; School of Material Science and Engineering, Lanzhou University of Technology, Gansu, Lanzhou, 730050, China; State Key Laboratory of Nickel and Cobalt Resources Comprehensive Utilization, Gansu, Jinchang, 737100, China; Lanzhou Petrochemical Research Center, Petrochemical Research Institute, PetroChina, Gansu, Lanzhou, 730050, China

6. Author: Quan C.; Ji S.; Yao R.; Liu W.; Yang J.; Li X., Affiliations: Key Laboratory for Organic Electronics and Information Displays & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications (NJUPT), Nanjing, 210023, China; Tongda College, Nanjing University of Posts & Telecommunications, Yangzhou, 225100, China; School of Science, Nanjing University of Posts & Telecommunications (NJUPT), Nanjing, 210023, China

7. Author: Chen M.; Batlle C.; Escachx B.; Costa-Castelló R.; Na J., Affiliations: Institut de Robòtica i Informàtica Industrial, CSIC-UPC. C/ Llorens i Artigas 4-6, Barcelona, 08028, Spain; Departament de Matemàtiques, Institut d'Organització i Control, EPSEVG, UPC, Vilanova i la Geltru, 08800, Spain; Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, 650500, China; Yunnan Key Laboratory of Intelligent Control and Application, Kunming, 650500, China; Universitat Politècnica de Catalunya (UPC). Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (ETSEIB)., Av. Diagonal 647, 2, Barcelona, 08028, Spain

8. Author: Chen G.; Su S.; Xu Q.; Lv H.; Zhao Y.; Xia L.; Zhang G.; Hu K., Affiliations: Institute of Safety Science & Engineering, South China University of Technology, Guangzhou, 510640, China; Guangdong Institute of Special Equipment Inspection and Research, Guangdong, Foshan, 528251, China

9. Author: Yang R.; Cao K.; Zhao W.; Wang Q.B.; Lu C.M.; Zhang Y., Affiliations: School of Physical Education, China University of Mining and Technology, Jiangsu Province, Xuzhou, 221116, China; Logistics University of Chinese People’s Armed Police Forces, Tianjin, 300309, China; Department of Physical Education, Guangxi University of Chinese Medicine, Guangxi Zhuang Autonomous Region, Nanning, 530021, China

10. Author: You H.; Luo W.; Chammingkwan P.; Taniike T., Affiliations: Guangxi Key Laboratory of Advanced Structural Materials and Carbon Neutralization, Guangxi Colleges and Universities Key Laboratory of Environmental-friendly Materials and New Technology for Carbon Neutralization, School of Materials and Environment, Guangxi Minzu University, Nanning, 530105, China; Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1, -, 1 Asahidai, Ishikawa, Nomi, 923-1292, Japan; Guangxi Colleges and Universities Key Laboratory of Natural and Biomedical Polymer Materials, Guilin University of Technology, Guilin, 541004, China

Маленькая часть за 2020 год:

1. Author: Zeng X.; Yan S.; He F.; Shi Y., Affiliations: School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, No. 1, Jinji Road, Guilin, 541004, Guangxi, China; School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang, 471023, China; School of Sciences, Southwest Petroleum University, Chengdu, 610500, China

2. Author: Yongcheng G.U.O.; Yan S.; Jianlin L.I.; Zuogui C.; Qinglin X.; Bin Y.A.N., Affiliations: Key Laboratory of Geological Hazards on Three Gorges Reservoir Area, Ministry of Education, China Three Gorges University, Yichang, 443002, China; College of Civil Engineering and Architecture, China Three Gorges University, Yichang, 443002, China; Hubei Key Laboratory of Disaster Prevention and Mitigation, China Three Gorges University, Yichang, 443002, China

3. Author: Yin W.; Han S.-C.; Zheng W.; Yeo I.-Y.; Hu L.; Tangdamrongsub N.; Ghobadi-Far K., Affiliations: Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, 104 Youyi Road, Beijing, 100094, China; School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia; College of Water Sciences, Beijing Normal University, Beijing, 100875, China; Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, United States; Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, United States

4. Author: Dong N.; Yang M.; Yu Z.; Wei J.; Yang C.; Yang Q.; Liu X.; Lei X.; Wang H.; Kunstmann H., Affiliations: State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China; College of Hydrology and Water Resources, Hohai University, Nanjing, China; Department of Water Resources, China Institute of Water Resources and Hydropower Research, Beijing, China; Institute of Meteorology and Climate Research (IMK-IFU), Karlsruhe Institute of Technology, Campus Alpine, Garmisch-Partenkirchen, Germany; College of Civil Engineering, Tianjin University, Tianjin, China; Institute of Geography, University of Augsburg, Augsburg, Germany

5. Author: Fan Y.; Lu W.; Miao T.; An Y.; Li J.; Luo J., Affiliations: Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun, 130021, China; Jilin Provincial Key Laboratory of Water Resources and Environment, Jilin University, Changchun, 130021, China; College of New Energy and Environment, Jilin University, Changchun, 130021, China

6. Author: Zhao J.; Li C.; Yang T.; Tang Y.; Yin Y.; Luan X.; Sun S., Affiliations: Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling, 712100, Shaanxi, China; Institute of Water Saving Agriculture in Arid Regions of China, Northwest A&F University, Yangling, 712100, Shaanxi, China

7. Author: Dan Li; Meng L.-Z.; Liu X.-L.; Guo Y.-F.; Deng T.-L.; Liu Y.-H.; Yang L., Affiliations: School of Materials Science and Engineering, Linyi University, Linyi, 276000, China; Tianjin Key Laboratory of Marine Resources and Chemistry, Tianjin University of Science and Technology, Tianjin, 300457, China

8. Author: Tang X.; Li J., Affiliations: School of Mathematics and Physics, Jinggangshan University, Ji'an, Jiangxi, 343009, China; School of Electronics and Information Engineering, Jinggangshan University, Ji'an, Jiangxi, 343009, China

9. Author: Cui Y.; Feng S.; Chen P.; Ye Y.; Wu Y.; Li C.; Yang R.; Wang H., Affiliations: School of Resource Environment and Safety Engineering, University of South China, Hengyang, 421001, Hunan, China; Hunan Province Engineering Technology Research Center of Uranium Tailings Treatment, Hengyang, 421001, China; Hunan Province Engineering Research Center of Radioactive Control Technology in Uranium Mining and Metallurgy, Hengyang, 421001, China

10. Author: He C.; Huang G.; Liu L.; Li Y.; Zhang X.; Xu X., Affiliations: State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Beijing Normal University, Beijing, 100875, China; Center for Energy, Environment and Ecology Research, UR-BNU, School of Environment, Beijing Normal University, Beijing, 100875, China; Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, S4S 0A2, Saskatchewan, Canada; Centre for Environment & Sustainability, University of Surrey, Guildford, GU2 7XH, United Kingdom

Как итог: выявлены адекватные рекомендации, учитывающие многие признаки. По названиям организаций и статей уже хорошо заметно, что алгоритм работает адекватно. Если подбирать веса более разумно (сделать влияние ключевых слов чуть ниже, например), то можно получить более точную оценку.

К огромному сожалению, 2022 год ночью во время обработки сломался из-за проблем с компом, а 2020 и 2021 год почти обработались, но я не успеваю пару часов в дедлайн (только до марта успел обработать). Суть точно такая же, только другие ответы.