

## ОТЧЕТ О СЕМИНАРЕ ПО МАШИННОМУ ОБУЧЕНИЮ НА CL120

### Метаданные

Полное название: «Машинное обучение и параллельные вычисления».

Форма проведения: Открытый семинар,  
кафедральный пилот будущего курса.

Место проведения: Аудитория №120.

Время проведения: С 5 по 26 марта 2026 г.

Периодичность: 1 раз в неделю.

Всего проведено: 4 пары.

### Техническая база

Вычислительный кластер CL120 в составе:

Списанный и восстановленный сервер.

ПК компьютерного класса а.120, в качестве вычислительных нод.

Инфраструктура лаборатории ЭВТ им. Павла Чеповского.

### Участники

Студенты 3-4 курсов МГИМО.

Практиканты лаборатории, в качестве ассистентов.

Инженеры лаборатории, как соавторы курса:

М.Звягинцев, А.Слюсарь, Ю.Акулин

**Автор курса, зав. лаборатории:** А.Немченко

Куратор проекта, заведующий кафедрой Н.В. Артамонов

### Гости:

Проректор по цифровизации Е.В. Погребняк

(выполнил часть одного из заданий).

Преподаватель каф. МЭИТ Иван Бондарев

(прослушал весь материал и выполнил все задачи).

### Статистика посещаемости и формальные результаты

Записалось: 28 чел.				
	I (05.03)	II (12.03)	III (20.03)	IV (26.03)
Присутствовали	17	9	5	5
Попытались, но не справились с задачей	2	3	2	0
Попытались, и справились с задачей	13	5	3	5

### Ссылки на материалы к семинару

имеется гостевой доступ, пароль: 1q2w3e

Опорный курс семинара

«Практикум по машинному обучению. Часть 2. Нейросети.»

Ссылка: <https://lms3.mgimo.ru/course/view.php?id=293>

Основной практический пример из курса

"Проект «Классификация новостей»"

Ссылка: <https://lms3.mgimo.ru/mod/book/view.php?id=10998>

Курсы, которые желательно предварительно изучить:

«Корпоративные информационные системы»

Ссылка: <https://lms3.mgimo.ru/course/view.php?id=153>

«Практикум по машинному обучению. Часть 1. Классические модели.»

Ссылка: <https://lms3.mgimo.ru/course/view.php?id=177>

## Сценарии занятий

### 05.03.2026 16:00

1. 5 мин. Представление, рассадка, и введение.

2. 5 мин. Инструменты (терминал и пр.)

3. 30 мин. Задача 0-2. БД (постгрес и пр.)

<https://lms3.mgimo.ru/mod/book/view.php?id=10998&chapterid=184>

ассистенты помогают студентам.

4. 30 мин. Задача 1\_1. Первый скрипт на кластере:

<https://lms3.mgimo.ru/mod/book/view.php?id=10998&chapterid=189>

5. 10 мин. Вопросы. Репозиторий.

Домашнее задание:

Из репозитория скачать `1_1_all_articles_parquet.py`

- изучить и понять логику;

- с помощью любого ИИ откомментировать код.

### 12.03.2026 16:00

1. 5 мин. Проверка домашки. Повторение пройденного.

Кто прокомментировал `1_1_all_articles_parquet.py` с ИИ?

Повторить в БД:

```
select * from f_create_train_test(80000+НОМЕР_ПК, 20000+НОМЕР_ПК);
```

2.0. 30 мин. ТРАДИЦИОННАЯ МЕТОДИКА

Ручная редакция ЛОГИН/ПАРОЛЬ для каждого скрипта.

Подготовка данных

`1_2_articles_lemmatization.py`

`1_3_train_test_lem.py`

`1_4_key_weight.py`

Фильтр и Кластеризации

1\_5\_filter\_news.py

2\_1\_lda\_multi\_labeling.py

Обучение (3 модели)

классика: 3\_1\_train\_final\_model.py

нейросеть на CPU: 3\_2\_train\_nn\_model.py

нейросеть на GPU: 3\_3\_train\_gpu\_pytorch.py

ПРИМЕНЕНИЕ ОБУЧЕННОЙ МОДЕЛИ

Массовая разметка новостей с помощью модели

4\_1\_apply\_model\_to\_all\_v2.py

5\_1\_export\_to\_db.py

## 2.1. 20 мин. АЛЬТЕРНАТИВНАЯ МЕТОДИКА

1) Авто-редактура ЛОГИН/ПАРОЛЬ для всех скриптов сразу

```
time ./0.py.sh
```

обязательно редактировать файл перед запуском:

ввести свои ЛОГИН/ПАРОЛЬ

2) ПРОГОН ВСЕХ ПРОГРАММ.py одним скриптом.sh

```
time ./1-5.py.sh
```

- этот shell-скрипт (подобен bat-файлам в windows) последовательно запустит все py-скрипты проекта, выводя по ходу работы на экран сообщения, что он делает;
- среднее время работы 10-15 минут;
- рекомендуется перед запуском изучить код.

Пока скрипт работает 5-7 мин. у доски:

о месте нашей методики работы с BigData в индустрии IT.  
IT на плоскости, далее в 3D и многомерной диаграмме.

3. 20 мин. Анализ результата автоматической разметки новостной базы.

Работа с дигитлаб. Терминал, postgres.

Большой запрос.

4. 5 мин. Вопросы. Домашка.

Продолжить изучение результатов используя большой запрос.

Собрать статистику правильной/ошибочной классификации новостей моделью.

5. 10 мин. Инструменты: дигитлаб, юпитер, терминал. (Андрей)

Удаленная работа с кластером из любого класса кафедры в т.ч. из Windows.

- работа с Postgres не требует включения нод;

- изучение кода скриптов не требует включения нод;
- запуск скриптов для параллельной обработки данных требует включения нод!

**20.03.2026**

1. 10 мин. Проверка домашки. Повторение пройденного.  
Кто поработал с ИИ, кодом, базой, в т.ч.  
через терминал Jupiter`a  
(<http://192.168.121.170:8000>)  
Обсуждение кода предложенного студентами.

Вспоминаем, что Цель - это анализ исторического периода  
конца 20 - начала 21 века.

Средства:

- а) Достоверные данные крупнейших мировых новостных источников (только английский язык), для выявления тенденций и критических точек периода и их возможных причин.
  - б) Построение моделей для глубокого анализа текстов новостей и метаданных (даты, источники, etc.)
  - в) Собственно анализ, с использованием SQL-запросов и их визуализация из результатов.
2. 20 мин. Исправление ошибки допущенной в Задаче 5.  
Оптимизация подготовки данных.
  3. 30 мин. Анализ результата разметки новостной базы.  
Терминал, postgres: Запросы.  
Дигитлаб, юпитер: Скрипты визуализации.
  4. 10 мин. а) Место нашего подхода к разработке  
в общем поле Софта.  
  
б) Другие подходы:
    - интегрированные среды с компилятором и пр.;
    - сначала тест кода, потом основной код;
    - ИИ, плюсы и минусы: древняя мудрость «Чтобы правильно задать вопрос - нужно знать большую часть ответа» заиграла новыми красками.  
в) Связь между нашим и «не нашими» подходами.
  5. 10 мин. Вопросы. Домашка.  
  
а) Продолжить изучение результатов, используя дигитлаб, юпитер и визуализацию.

- б) Предложить свою оригинальную идею для обучения модели, указать Цель и источник данных.

**26.03.2026**

1. 10 мин. Проверка домашки. Повторение пройденного.
  - а) Кто готов разобрать по косточкам `3_2_train_nn_model_v3.py`?
  - б) Кто принес идею, план или готовый проект ML?
  
2. 50 мин. Удаленная работа с сервером
  - 2.1. Подключение к серверу
    - а) К postgres-базе
    - б) Виртуальная среда python
    - в) Вызов py-Скриптов
  
  - 2.2. Визуализация результатов
    - а) `7_2_visualize_query.py`
    - б) `7_3_visualize_big_query.py`  
модернизация для использования аргументов
    - в) `read_parquet.py`  
просмотр содержимого паркетных данных
  
  - 2.3. Интерпретация графиков/диаграмм
    - а) Политика
    - б) Экономика
  
3. 20 мин. Сохранение проекта и разворачивание его из архива на другой площадке
  - 3.1. Бекап
    - а) Postgres базы данных
    - б) Виртуальной среды Python
    - в) PY-Скриптов
  
  - 3.2. Восстановление
    - а) Postgres базы данных
    - б) Виртуальной среды Python
    - в) PY-Скриптов

**Визуальные результаты, которых достигли все, успешно завершившие все задачи.**

Базы у всех участников незначительно отличаются, кроме того обучение моделей на больших данных имеет вероятностный характер, поэтому графики у всех не совпадают, однако тенденции которые они отражают мало чем отличаются. Ниже приводятся два варианта каждого графика, выбранные случайно.

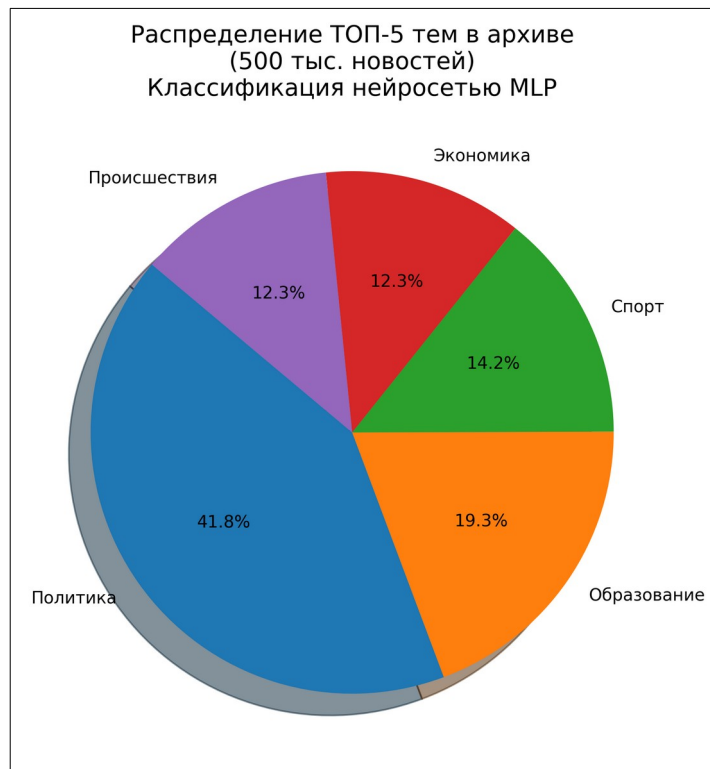


Рис. 1.1

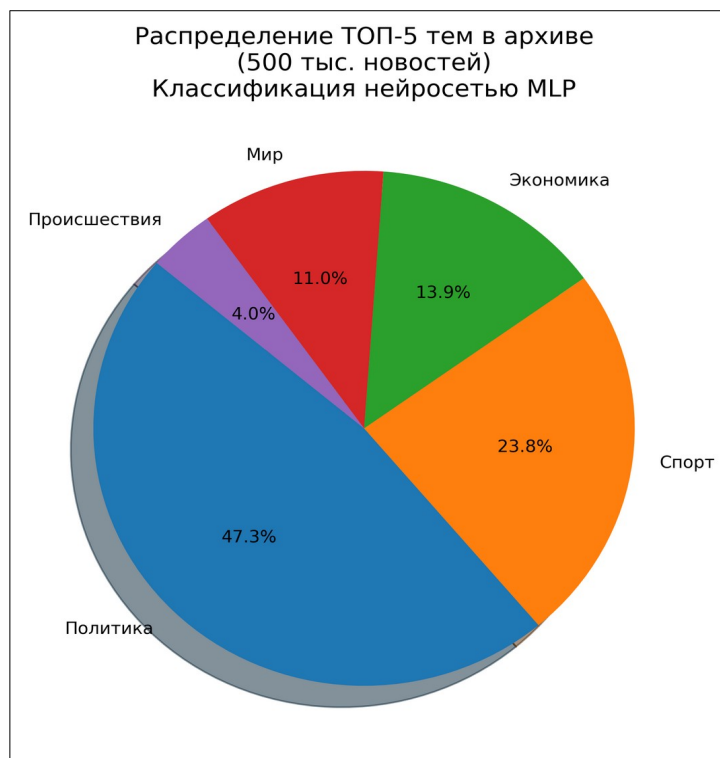


Рис. 1.2

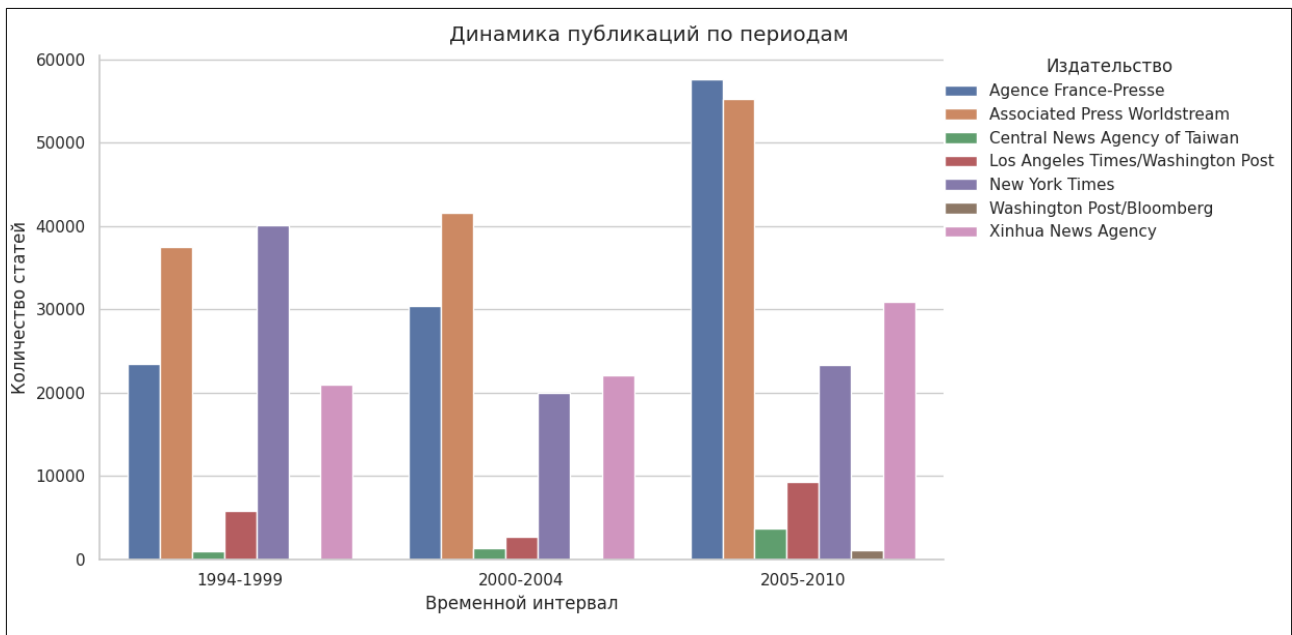


Рис. 2.1.

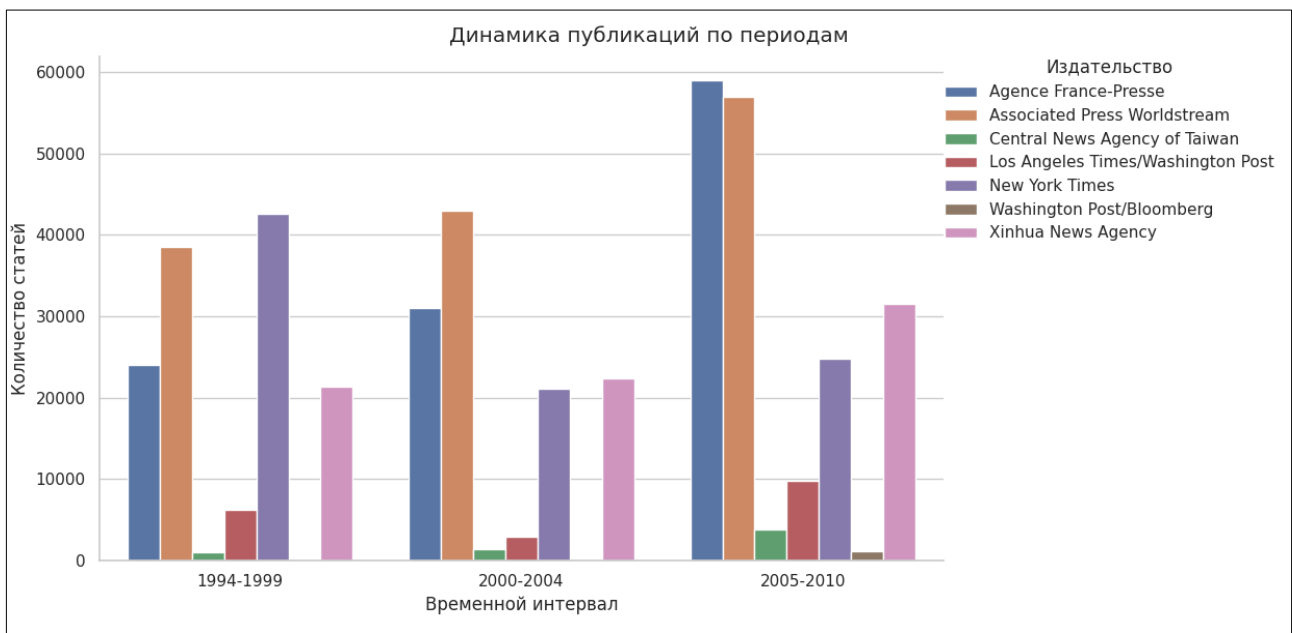


Рис. 2.2

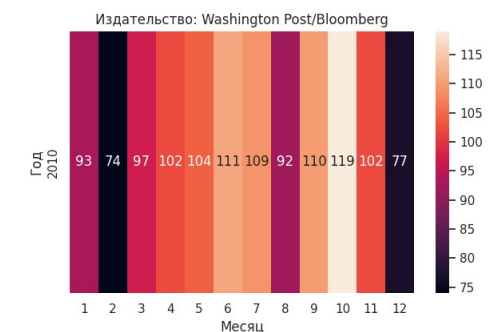
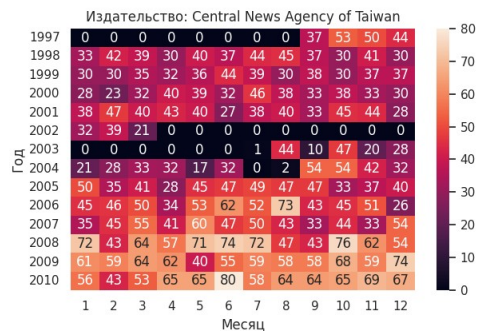
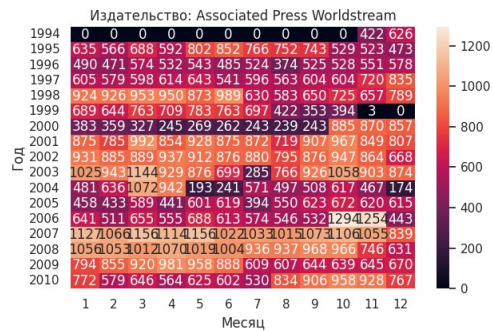
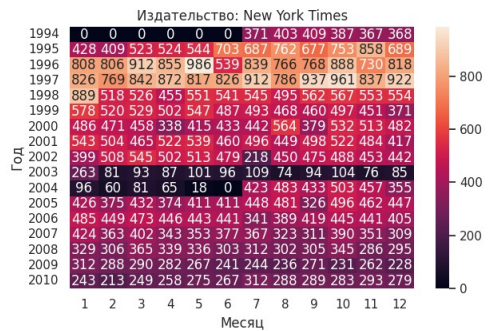
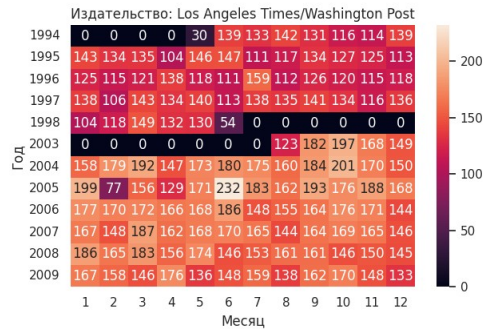
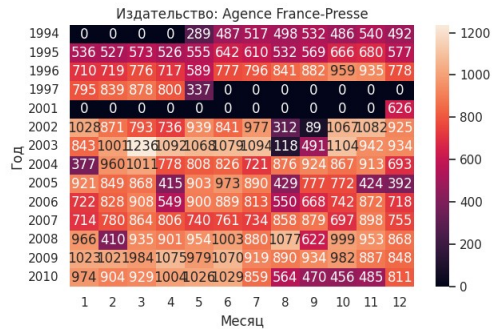


Рис. 3.1.

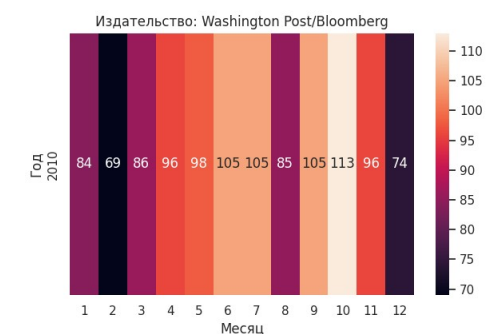
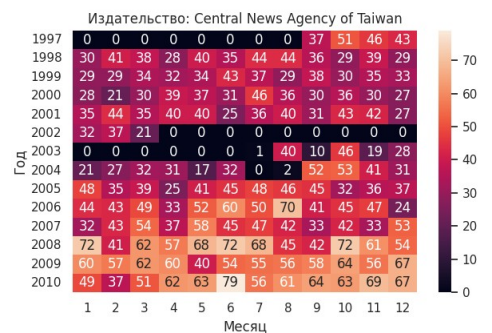
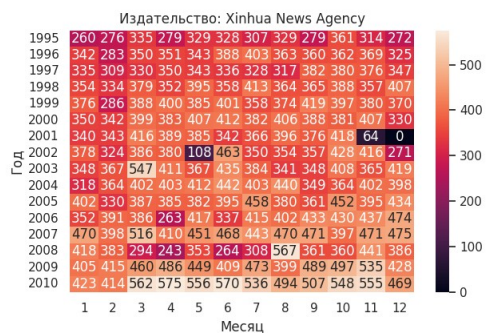
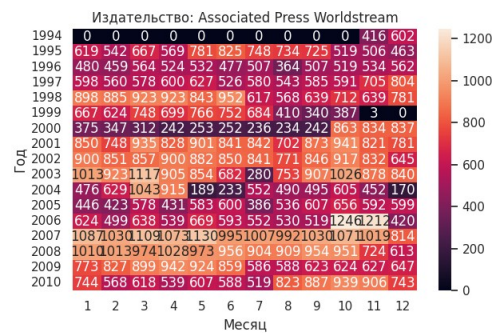
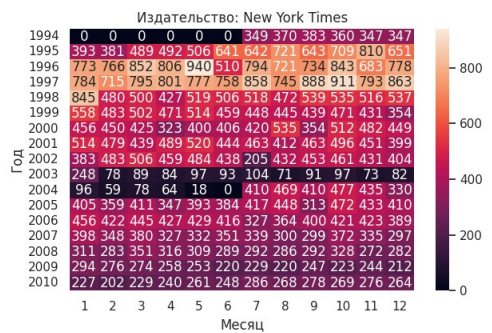
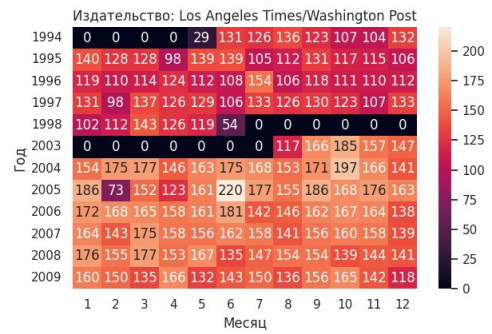
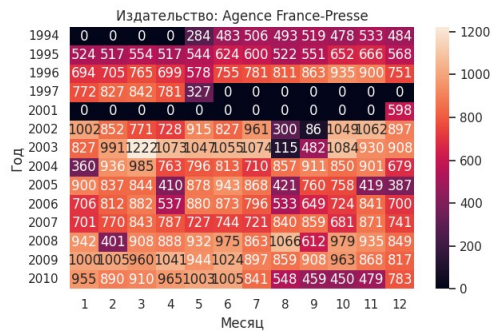


Рис. 3.2

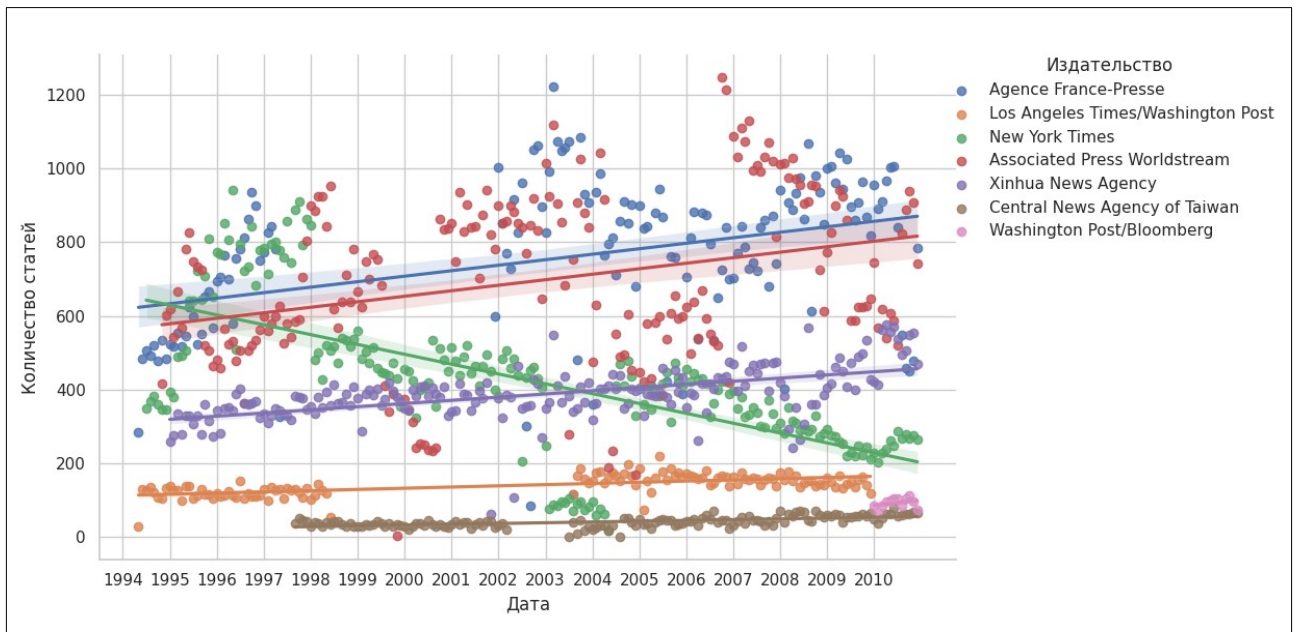


Рис. 4.1.

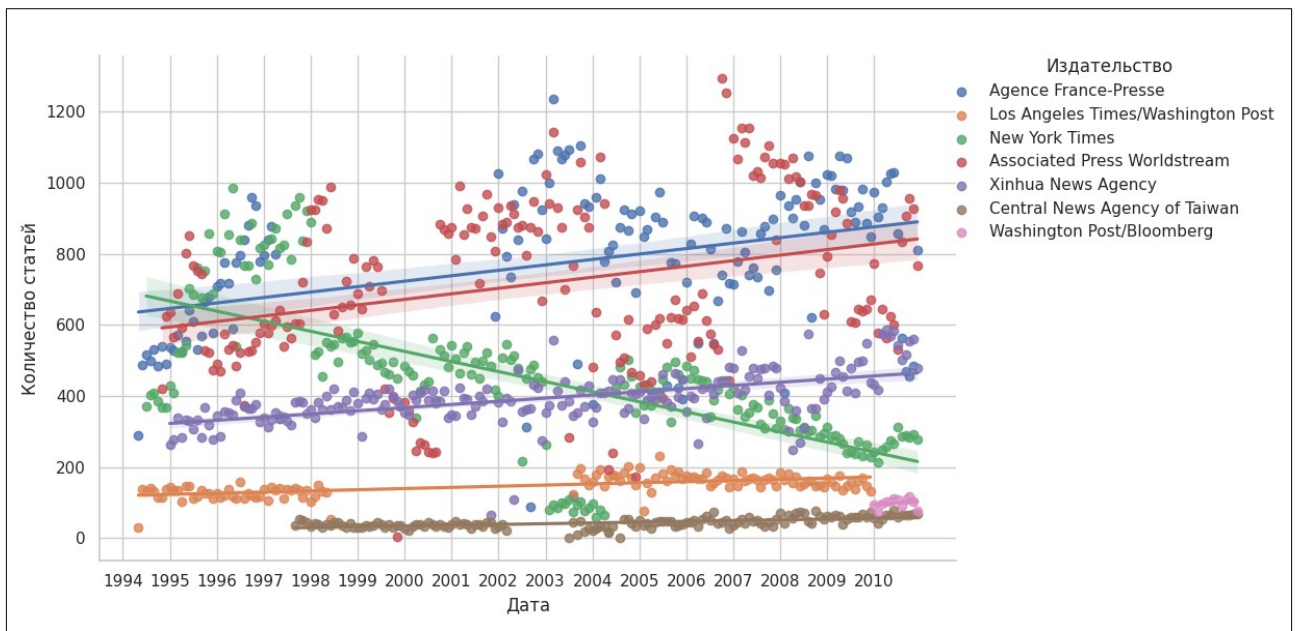


Рис. 4.2.

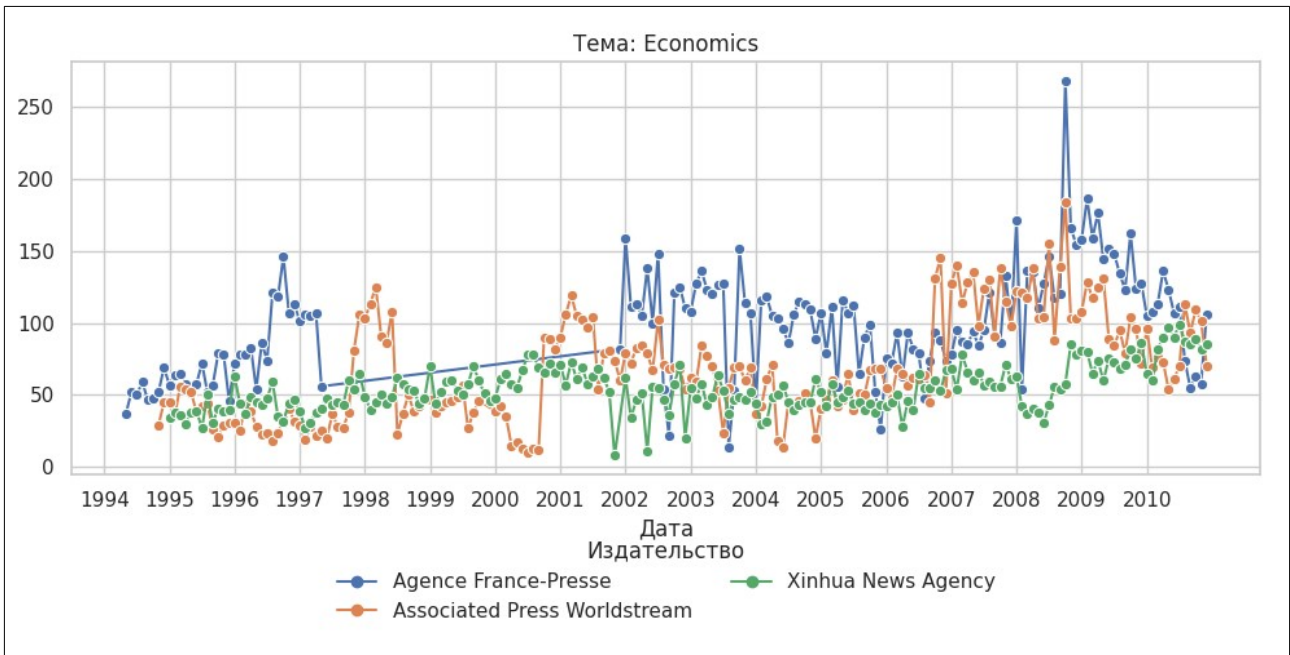


Рис. 5.1.

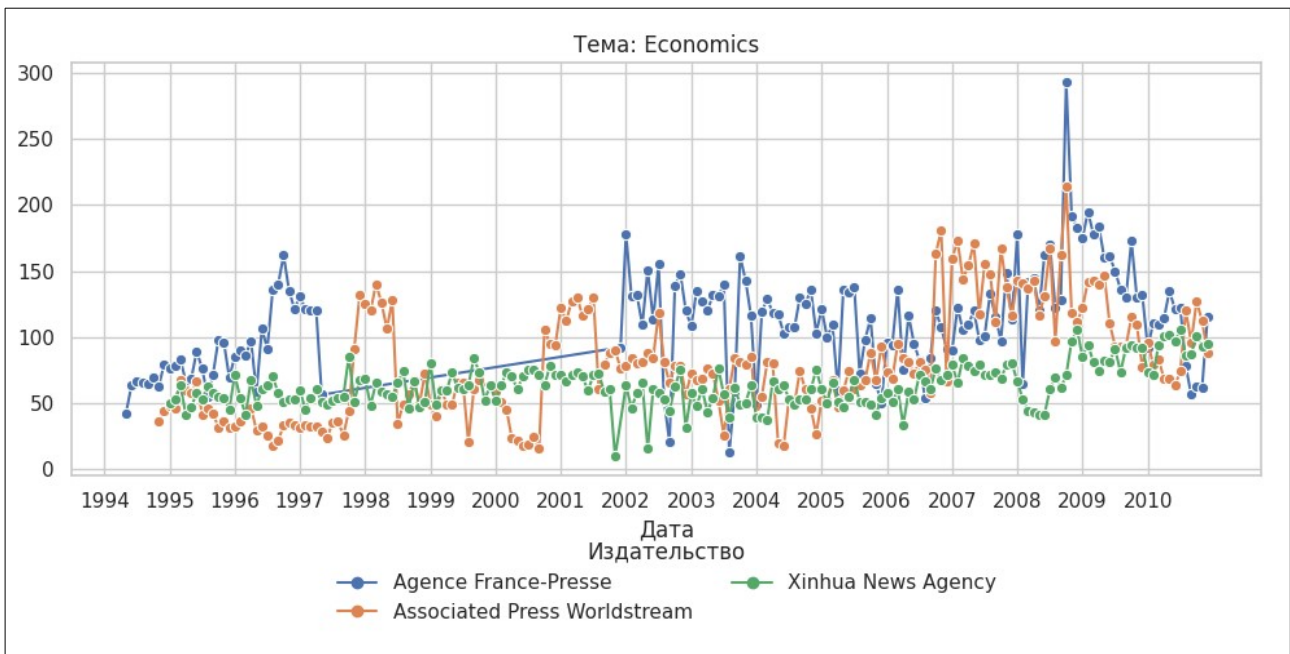


Рис. 5.2.

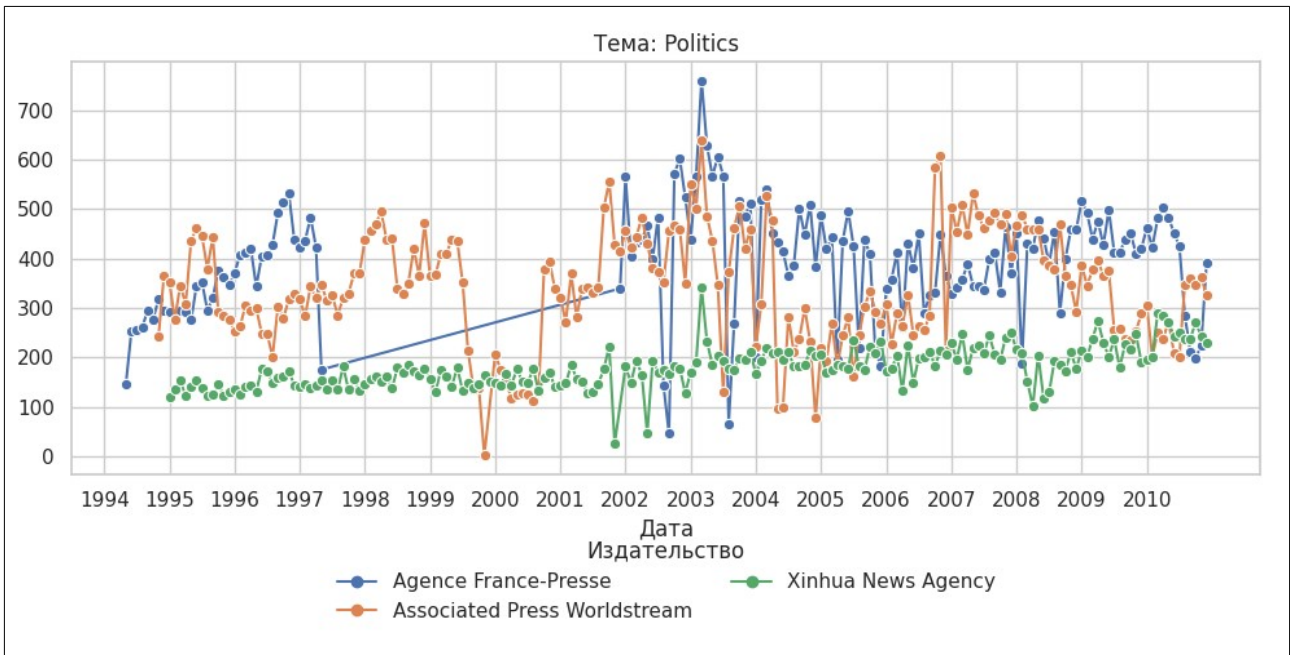


Рис. 6.1.

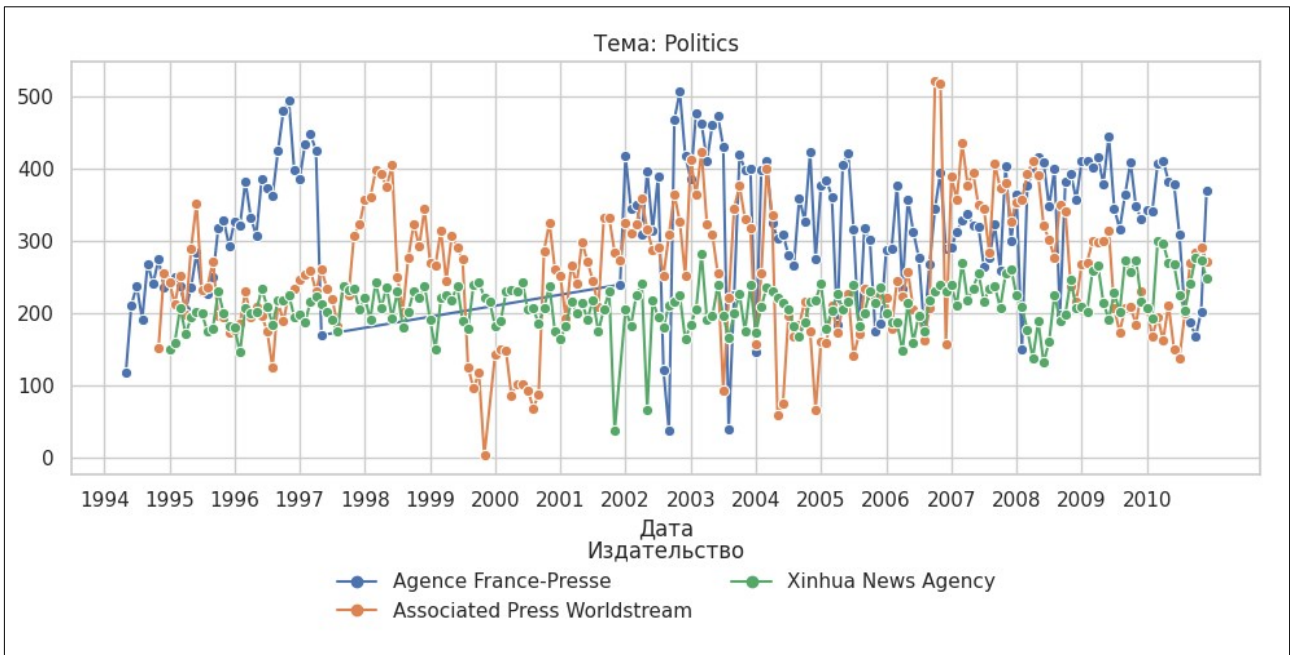


Рис. 6.2.

По ходу семинара велась видеозапись и запись демонстрации экрана. Материалы обрабатываются и по мере готовности выкладываются здесь: [https://disk.yandex.ru/d/ldCY\\_q\\_jd2BdFg](https://disk.yandex.ru/d/ldCY_q_jd2BdFg)

На момент написания данного текста, там выложены 4 коротких ролика, это фрагменты будущего большого видеоотчета.

### **Оценка результатов данного семинара, как эксперимента.**

17% из записавшихся,  
30% из принявших участие в первом занятии,  
100% из принявших участие в последнем занятии,  
освоили предложенный им материал.

Мы предполагали подобную динамику, так как материал, сложный по сути, совершенно новый в технологическом (кластер, параллельные вычисления) и инструментальном (Linux, командная строка, консоль Postgres) отношениях.

Поэтому мы оцениваем этот результат как, в целом, положительный.

Заведующий лабораторией ЭВТ  
им. Павла Чеповского

Немченко А.Б.

Контакты для связи:  
мобильный: +7 926 932 3001  
внутренний: 16-79  
телеграм: @an2kus  
e-mail: an2k@mail.ru