News Impact

агрегатор найчастіше згаданих новин

Автор: Vasyl Alba

Кеш → Переклад → Ембеддинги → Кластеризація → Дедуп eTLD+1 → Top-N → Резюме

Проблема → Ціль

Проблема: інфошум. Сотні заголовків, мало користі.

Ціль: автоматично виділити **спільні сюжети**, які одночасно висвітлюють різні ЗМІ, і дати короткий нейтральний підсумок з посиланнями.

Що нового (внесок)

- **Багатомовність без болю:** одноразовий LLM-переклад з кешем (sha1) → дешевше, стабільніше.
- **Сюжет ≠ посилання:** семантичні ембеддинги + кластеризація заголовків.
- **Чесна згадка:** дедуплікація **за eTLD+1** (одне посилання на видання `nv.ua`, `bbc.com`, …).
- **Top-N за «перехресним висвітленням»:** ранжування за кількістю унікальних видань.
- **Простий інтерфейс:** Streamlit, JSON-звіт для інтеграцій.

Огляд системи

```
RSS/Atom → Ingest → Titles cache
                  └→ Fetch bodies → Articles cache
Caches → Translate (idempotent: *_en + sha1)
*_en titles → Embeddings → Clustering → Stories
Stories → Dedup by eTLD+1 → Rank Top-N → LLM Summary
                                       ↘️ Streamlit UI + report.json
```

Дані та кеші

- **Ingest:** `feedparser` (з UA-заголовком) + `requests` fallback.
- **Кеші (JSON, дискові):**
  - `out/titles_cache.json` — унікально за URL, поля `title`, `title_en*`, `domain`, `published_at`.
  - `out/articles_cache.json` — `body`, `body_len`, `body_en*`, `fetched_at`.
- **Витяг тексту:** `BeautifulSoup` + евристики + **AMP-fallback**; рефетч коротких статей.

Переклад (idempotent)

- Викликаємо OpenAI **один раз** на унікальний текст.
- Зберігаємо: `*_en`, `*_en_sha1`, `translator_model`, `translated_at`.
- Якщо sha1 не змінився → **не** платимо вдруге.
- Ембеддинги будуються по `title_en` (fallback `title`).

Кластеризація та дедуп

- **Ембеддинги:** `all-MiniLM-L6-v2`.
- **Кластеризація:** косинус, жадібно; поріг налаштовується.
- **Дедуп у сюжеті:** одна ланка на **реєстрабельний домен**: `sport.nv.ua`, `techno.nv.ua` → **`nv.ua`**.

Ранжування та резюме

- **Ранжування Top-N:**
  i. кількість унікальних видань у сюжеті,
  ii. кількість згадок,
  iii. рання поява.
- **Резюме:** OpenAI (`gpt-4.1-mini` , `temperature=0`), 2–3 речення, нейтрально.
- В UI: **[ДОМЕН] Назва статті** (назва — клікабельна).

Controls

RSS feeds

feeds.bbci.co.uk ×
cnbc.com ×
feeds.finance.ya… ×
investing.com ×

Max items (ingest)

1600

Max age (hours)

48

Similarity threshold

0.72

Top-N stories

10

OpenAI model

gpt-4.1-mini

Translate titles to English

Translate article bodies (costly)

Article timeout (s)

10

Update cache

Summarize

News Impact — Top Most-Mentioned Stories (distinct outlets)

Top stories | All titles

Top 10 most-mentioned stories (by distinct outlets)

Bail has been posted for MP Kuznetsov, who is suspected of corruption in the procurement of drones and electronic warfare equipment.

Bail has been posted for MP Oleksiy Kuznetsov, who is suspected of corruption related to the procurement of drones and electronic warfare equipment. The High Anti-Corruption Court set his bail at eight million hryvnias, lower than the prosecution's demand of 30 million. Kuznetsov, a member of the Servant of the People party, was among four individuals detained in an investigation revealing a scheme involving up to 30% kickbacks in state contracts.

outlets: 4 · first seen: 2025-09-02T02:47:58.301155+00:00

- **[CENSOR.NET]** Corruption in drone procurement: an 8 million hryvnia bail was posted for "Servant of the People" member Kuznetsov
- **[NV.UA]** Bail has been posted for MP Kuznetsov, who is suspected of corruption in the procurement of drones and electronic warfare equipment.
- **[PRAVDA.COM.UA]** Corruption in the procurement of electronic warfare systems and drones: a bail of 8 million was posted for MP Kuznetsov
- **[TSN.UA]** Corruption in the procurement of drones and electronic warfare systems: bail posted for MP Kuznetsov

Ukrainian soldiers have liberated Novoekonomichne in the Donetsk region from Russian forces (video).

Ukrainian forces have liberated the village of Novoekonomichne in the Donetsk region from Russian control, as reported by the General Staff of the Armed Forces of Ukraine. The operation, carried out by the 425th Separate Assault Regiment "Skelya," took about two weeks and involved raising the Ukrainian flag in the settlement. This action is part of ongoing efforts by Ukrainian troops to regain control over temporarily occupied areas in Donetsk.

outlets: 4 · first seen: 2025-09-02T03:11:15.512902+00:00

- **[CENSOR.NET]** The Armed Forces of Ukraine have liberated the village of Novoekonomichne in the Donetsk region and raised the Ukrainian flag there, according to the General Staff. VIDEO
- **[RBC.UA]** Ukrainian soldiers have liberated Novoekonomichne in the Donetsk region from Russian forces (video).
- **[TSN.UA]** In Donetsk region, the Armed Forces of Ukraine have liberated the settlement of Novoekonomichne from the occupiers.
- **[UKRINFORM.NET]** Ukrainian forces liberate Novoekonomichne in Donetsk region

Ukraine has likely struck a Russian target with the "Flamingo" missile for the first time. What is known about this?

Ukraine likely conducted its first strike using the domestically produced "Flamingo" cruise missile on a Russian target in annexed Crimea near Armyansk on August 30. The missile reportedly has a range of up to 3,000 kilometers and can carry a warhead weighing over a ton, potentially allowing Ukraine to target deep into Russian territory. While video footage and satellite images suggest damage to a Russian FSB border post, independent confirmation and detailed damage assessments remain unavailable.

Демо (UI)

Top stories: по одному лінку на домен у межах сюжету. All titles: ієрархія «домен → піддомени». Якщо піддомен один — показуємо один рівень.

Метрики (що рахуємо)

- **Coverage:** `items_ingested`, кількість кластерів.
- **Cross-outlet spread:** середня/медіана унікальних доменів у Top-N.
- **Extraction rate:** частка `body_len>0`.
- **Translation rate:** частка записів із `title_en`.
- **Вартість/затримка:** # **нових** перекладів/резюме × тарифи API.

Обмеження → Як обходимо

- JS-рендер/пейволи → AMP-fallback, більший таймаут, рефетч «коротких».
- Відсутність RSS у деяких ЗМІ → альтернативні фіди, м’який пропуск.
- Помилкова близькість у кластерах → тюнінг порога; (план) агломеративна кластеризація.
- Вартість LLM → кеш за sha1, «translate-once».

План далі (roadmap)

- Агломеративна кластеризація (average linkage).
- Headless-рендер для JS-сайтів.
- Тематична класифікація/теги, історичні тренди.
- Зважування видань (якість/аудиторія).

Запуск

Вимоги: Python 3.11, `.env` з `OPENAI_API_KEY`.

Встановлення

```
mamba env create -f requirements.yml
mamba activate news-impact
```

UI

```
streamlit run streamline.py
```

- **Update cache** → збір/витяг/переклад/кеш.
- **Summarize** → кластери/Top-N/резюме → `out/report.json`.

CLI (опційно)

```
python -m src.main --out out/report.json --max-age-hours 48 --sim-thr 0.72 --topk 5 --oa-model gpt-4.1-mini
```

Репозиторій

```
.
├── src/main.py          # пайплайн і логіка
├── streamline.py        # Streamlit UI
├── requirements.yml     # середовище
├── .env                 # OPENAI_API_KEY
└── out/
    ├── titles_cache.json
    ├── articles_cache.json
    └── report.json
```