Web-Intelligence WS 25/26

Name: Irmak Damla Özdemir **Matrikel-Nr.:** 5123127

Identifizieren Sie geeignete Datenquellen und Daten im Web für Ihren Anwendungsfall. Wenden Sie ein Analyse-Tool an, um verschiedene Daten für Ihren Anwendungsfall zu vergleichen.

Trending Health News: Data Sources in Web

In this project, my focus is on analyzing health-related news articles published by reliable online sources. The main goal is to identify which health topics are currently most discussed in the media. It is essential to select appropriate data sources that provide continuously updated and structured information. RSS feeds from health news providers are particularly suitable, as they offer data that can be easily processed with analytical tools.

Data Sources for Health News Analysis

For this use case, I use five trustworthy health-related sources.

Spiegel Gesundheit: A German health news section publishing articles about medical research, nutrition, and mental health.

Tagesschau Gesundheit: A national news source offering daily updates on current health topics, healthcare reforms, and social debates.

World Health Organization (WHO): Official RSS feed with global health updates, disease reports, and public health initiatives.

Deutsches Ärzteblatt: Medical professional news covering health policy, research, and system-related discussions.

These sources were chosen because they cover both global and local health perspectives. Moreover, they provide consistent data formats, which allows automated data retrieval through Python scripts without violating website policies.

Analyze Tools

First, for processing and comparing the collected data, Python is selected as the main analysis tool because of its flexibility. *pandas* library is used to clean the data, convert publication dates into a uniform format, and count the frequency of specific words and topics across all sources.

Some example keywords for topic hints to be categorized:

Topic	**Keywords**
Covid	covid, coronavirus
Vaccination	vaccine, vaccination, immunization, Impfung, Impfstoff, impfen
Pandemic	pandemic, epidemic, Ausbruch, Pandemie, Epidemie
Measles	measles
Influenza	influenza, coughing, Erkältung, schnupfen, nasenschwellung, fieber, nasal congestion, fever
Depression	depression, depressed, anxiety, angst, depressiv, psychisch
Addiction	addiction, alcohol, smoking, drugs, Sucht, Rauchen, Alkohol, Drogen
Nutrition and lifestyle	nutrition, diet, obesity, exercise, Yoga, Meditation
Policy and system	healthcare, hospital, reform, insurance, Krankenhaus, Krankenversicherung

Tabelle 1: Topic categories

```
def detect_topic(text):
    text = str(text).lower()
    for topic, keywords in TOPIC_KEYWORDS.items():
        if any(word in text for word in keywords):
            return topic
    return "Other"
```

Listing 1: Topic detection

```
def fetch_sources():
    rows = []
    for name, url in SOURCES.items():
        print(f"Fetching: {name} -> {url}")
        feed = feedparser.parse(url)
        print(f"{name}: {len(feed.entries)} entries")
        for e in feed.entries:
            rows.append({
                "source": name,
                "title": e.get("title", ""),
                "summary": e.get("summary", ""),
                "link": e.get("link", ""),
                "published": e.get("published", e.get("updated", "")),
            })
    return pd.DataFrame(rows)
```

Listing 2: Fetching

```
def save_csv(df: pd.DataFrame, out_dir: str = "."):
    os.makedirs(out_dir, exist_ok=True)
    fn = f"categorized_healthnews_{datetime.now(UTC).strftime('%Y%m%d')
      }.csv"
    out_path = os.path.join(out_dir, fn)
    columns_to_save = ["source", "published", "topic", "title", "summary
      ", "link"]
    df.to_csv(
        out_path,
        index=False,
        encoding="utf-8-sig",
        lineterminator="\n",
        sep=",",
        columns=columns_to_save )
```

Listing 3: Saving as a CSV File

By using simple data frames and aggregation functions in *pandas*, it becomes possible to visualize which topics dominate the media over a period. These insights are visualized in a low-code environment **AppSheet**:

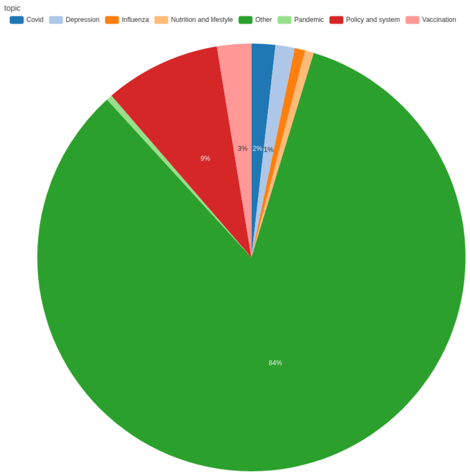

Abbildung 1: Frequency of topics

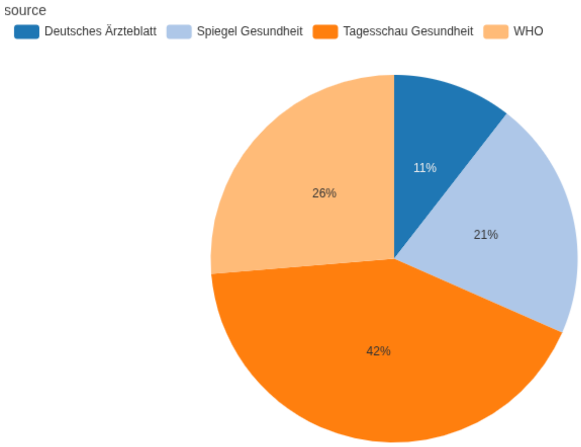

Abbildung 2: Percentage of sources

Comparison Dimensions for Data Sources

Tagesschau publishes the most frequent updates, followed by Spiegel and WHO with regular but less frequent posts. Ärzteblatt releases content based on official or scheduled publications.

Spiegel and WHO cover the widest range of topics, from infectious diseases to mental health and lifestyle. Ärzteblatt is more specialized, focusing mainly on policy and healthcare systems.

WHO provides a global outlook on health issues, while the other sources primarily emphasize national perspectives. Spiegel and Tagesschau occasionally include international topics but remain Germany-centered.

Most sources publish in German, whereas WHO uses English. This linguistic difference made defining and matching keywords across sources more challenging for me.

Overall, the comparison showed some differences between the sources in terms of frequency, topic range, and regional scope. I found it interesting to see how national sources focused more on local health policies and the language differences also made me realize how keyword design can affect analysis accuracy.