news_signals: An NLP Library for Text and Time Series

Example Workflow

Creating and exploring news signals

Creating a Signal for *OpenAI*

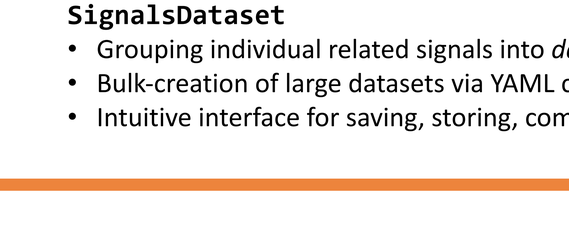
```python
from news_signals import signals, wikidata_utils

entity_name = 'OpenAI'
entity =
wikidata_utils.search_wikidata(entity_name)[0]

# let's create a signal
signal = signals.AylienSignal(
    name=entity['label'],
    params={"entity_ids": [entity['id']]}
)
```

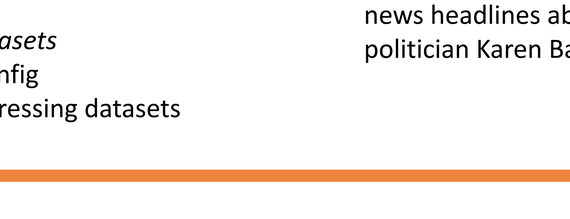
Specify time-range and plot news volume

```python
# let's instantiate our signal for the time period we
care about
start = '2023-01-01'
end = '2023-11-28'
timeseries_signal = signal(start, end)
timeseries_signal.plot(figsize=(10,3), title='News
Volume – OpenAI')
```



Plot anomalies in news coverage

```python
# did the signal have any unexpected spikes?
anomaly_signal = signal.anomaly_signal()
anomaly_signal.anomalies.plot(color='red',
figsize=(10,3), title s='Anomalies – Openai',
ylable='anomaly score')
```



What are the news on the biggest spike?

```python
# let's have a look at the biggest anomaly
highest_anomaly_day = signal.anomalies.idxmax()
highest_anomaly_day

# what was going on that day?
signal = signal.sample_stories_in_window(
    start=highest_anomaly_day,
    end=highest_anomaly_day
+  datetime.timedelta(days=1)
)

for s in signal.feeds_df.stories[0]:
   print(s['title'])
```

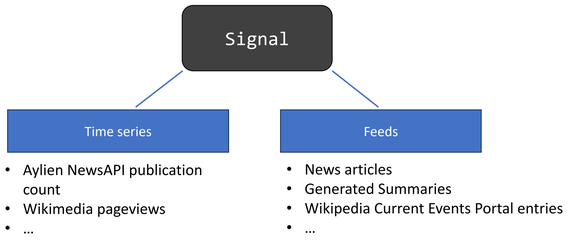
```
• What to know about OpenAI's new interim CEO Emmett Shear
• First on CNBC: CNBC Transcript: Microsoft CEO Satya Nadella Speaks
  with CNBC's Jon Fortt on "Fast Money" Today
• OpenAI's wild 60 hours: Here's what happened
• Exclusive–OpenAI investors considering suing the board after CEO's
  abrupt firing – sources
• Microsoft CEO says OpenAI governance needs to change no matter where
  Sam Altman ends up
• OpenAI Investors, Led by Thrive, Angle to Bring Back Altman
• Sam Altman, OpenAI drama playing out like 'telenovela': Expert
• OpenAI staff threaten mass exodus to join ex-CEO
...
```

TLDR

- Open-source Python library for creating and exploring datasets with **text + timeseries**
  - Inputs: buckets of text
  - Outputs: timeseries
- We provide 3 example datasets: US-politicians, Nasdaq-100, S&P-500, covering millions of news articles and thousands of entities.

Motivation & Vision

- Exploring connection between NLP and time series research, e.g. time series forecasting from textual data
- Domains: financial data analysis, healthcare, social sciences



3rd party API integrations

WIKIDATA · quantexa · AYLIEN

Aylien NewsAPI (now Quantexa News Intelligence)

- Retrieving news articles for custom filters (e.g. Wikidata entity, text query)
- Retrieving time series of publication volume

Wikipedia / Wikidata

- Retrieving page views time series for Wikipedia articles (= entities)
- Retrieving mentions of entities in Wikipedia Current Events Portal

More integrations planned!

`Signals & SignalDataset API`

`Signals`

- Stores and explores data about a real-world entity
- Pandas-like interface – a Signal's data is stored in dataframes

`SignalsDataset`

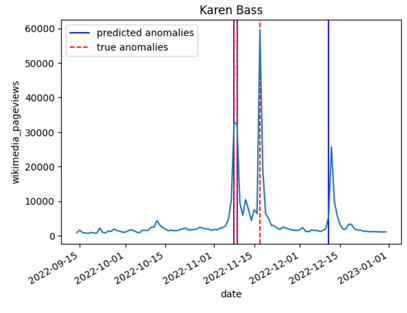
- Grouping individual related signals into *datasets*
- Bulk-creation of large datasets via YAML config
- Intuitive interface for saving, storing, compressing datasets

Features

- Saving and loading signals datasets via Google Drive & Google Cloud storage
- Docker container & Kubernetes job configuration for repeated dataset creation
- Dataset enrichment/transformations:
  - Time series anomaly detection
  - Abstractive multi-doc summarization
  - Wikimedia pageviews time series

Anomaly Classification Experiments

Example predictions by random forest trained on lexical features to predict (binary) anomalies in Wikimedia pageviews from news headlines about US-politician Karen Bass.



Github

https://github.com/AYLIEN/news-signals-datasets

Paper

→ Example Jupyter notebooks, walk-through videos, datasets!

Chris Hokamp, Demian Gholipour Ghalandari, Parsa Ghaffari

{chrishokamp, demiangholipour, parsaghaffari}@quantexa.com

quantexa