

# Text Mining for Journalism

Matthew Shardlow

[m.shardlow@mmu.ac.uk](mailto:m.shardlow@mmu.ac.uk)

# Text Mining for Journalism

## The need

- A journalist must be a temporary expert in a wide variety of topics

### Air pollution plans to tackle wood burners

🕒 22 May 2018

| Science & Environment



#### Pump it down

How to turn carbon dioxide into rock



### Agency warns of water deficits for England

Enough water to meet the needs of 20 million people is lost through leakage every day, the report says.

🕒 1 hour ago | Science & Environment | 🗨️ 323

### 2 hospitals to offer proton beam therapy for cancer

The Straits Times - 9 May 2018

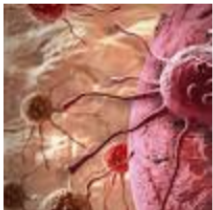
Proton beams are not better at killing cancer cells than conventional radiation, but they do it with far less damage to surrounding healthy tissues ...



### Complex medical imaging systems to be installing led in new NHS ...

News-Medical.net - 9 May 2018

Proton beam therapy uses very high energy beams of protons and though more expensive to deliver than conventional x-rays ...



### UK's clean car goal 'not ambitious enough'

🕒 21 May 2018 | Business | 🗨️ 590

# Text Mining for Journalism Approach

- Focus on scientific literature
- The Five w's
  - What
  - Where
  - When
  - Who
  - Why

# Text Mining for Journalism Approach

- Focus on scientific literature
- The Five w's
  - **What**
  - Where
  - When
  - Who
  - Why



air pollution uk

"PROTON BEAM THERAPY"

SEARCH

QUERY: "Proton BEAM Therapy" x

# Text Mining for Journalism Approach

- Focus on scientific literature
- The Five w's
  - What
  - **Where**
  - When
  - Who
  - Why



# Text Mining for Journalism Approach

- Focus on scientific literature
- The Five w's
  - What
  - Where
  - **When**
  - Who
  - Why



# Text Mining for Journalism Approach

- Focus on scientific literature

- The Five w's

- What
- Where
- When
- **Who**
- Why



# Text Mining for Journalism Approach

- Focus on scientific literature
- The Five w's
  - What
  - Where
  - When
  - Who
  - **Why**





# Text Mining for Journalism Implementation

- Choice of framework – Apache UIMA
  - Reusability of components
  - Access components from other frameworks
  - Oasis Standard
  - Apache support



# Text Mining for Journalism Implementation

- Metadata descriptors
  - Automatically generated using java annotations
  - Documentation is kept close to code

```
@Component(OperationType.ANNOTATOR)
@ResourceInput(
    type = ProcessingResourceType.CORPUS,
    encoding = CharacterEncoding.UTF_8,
    dataFormat = @DataFormat(dataFormat = DataFormatType.BINARY_CAS))
@ResourceOutput(
    type = ProcessingResourceType.DOCUMENT,
    encoding = CharacterEncoding.UTF_8,
    language = @Language(languageId="en"),
    dataFormat = @DataFormat(dataFormat = DataFormatType.TEXT))
@LanguageCapability("en")
@TypeCapability(
    inputs = {},
    outputs = {})
```

```
public class TextMiningForJournalismApplication extends JCasAnnotator_ImplBase
{
```

# Text Mining for Journalism Implementation

- Choice of distribution – Maven Central
- Publicly accessible
- Open source

GroupId	ArtifactId	Latest Version	Updated	Download
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">JournalismTypeSystem</a>	<a href="#">1.1.0</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">StanfordNLPTagger</a>	<a href="#">1.1.0</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">TextMiningForJournalismApplication</a>	<a href="#">0.0.2</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">WhenAnnotator</a>	<a href="#">1.1.0</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">jar-with-dependencies.jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">WhoAnnotator</a>	<a href="#">1.1.0</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">jar-with-dependencies.jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>
<a href="#">uk.ac.mmu.tdmlab.journalism</a>	<a href="#">LocationAnnotator</a>	<a href="#">1.1.0</a> <a href="#">all (2)</a>	11-May-2018	<a href="#">pom</a> <a href="#">jar</a> <a href="#">jar-with-dependencies.jar</a> <a href="#">javadoc.jar</a> <a href="#">sources.jar</a>

# Text Mining for Journalism Implementation

- Upload to platform
  - Via maven coordinates
  - Simple easy to use interface



*\* We support UIMA and GATE  
components that are registered in  
Maven*

MAVEN COORDINATES

OMTD EDITOR

UPLOAD XML

Provide maven coordinates

uk.ac.mmu.tdmlab.journalism

TextMiningForJournalismApplciation

0.0.2

RESOLVE

# Text Mining for Journalism Implementation

- Workflow creation via platform

## OpenMinTeD Workflow Editor

Tools

Workflow Canvas | 0931732053353301-53286924-730f-4a51-aaac-77a1440e333d

- CERMINE PDF Reader Collection search tools  
CERMINE <https://github.com/CeON/CERMINE>.
- Variable Disambiguator Assign variable IDs to sentences based on calculating the similarity between the sentencetext and the description of the variable.
- [test test contact](#)
- [Text Mining For Journalism Component](#) A UIMA component to annotate entities relevant to journalism. The named entities include: "Person", "Organisation", "Location" and "Time/Date".

```
graph LR; omtDImporter[omtdImporter] -- output --> PdfReader[PdfReader]; PdfReader -- output --> TextMining[Text Mining For Journalism Component];
```

# Text Mining for Journalism

## Running the application

# Video Demonstration