Text Mining for Journalism

Matthew Shardlow
m.shardlow@mmu.ac.uk

Text Mining for Journalism
The need

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

Pump it down

How to turn carbon dioxide into rock

Air pollution plans to tackle wood burners

22 May 2018 | Science & Environment

Agency warns of water deficits for England

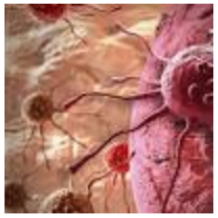
Enough water to meet the needs of 20 million people is lost through leakage every day, the report says.

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 prot and though more expensive to deliver than conventional x-r

UK's clean car goal 'not ambitious enough'

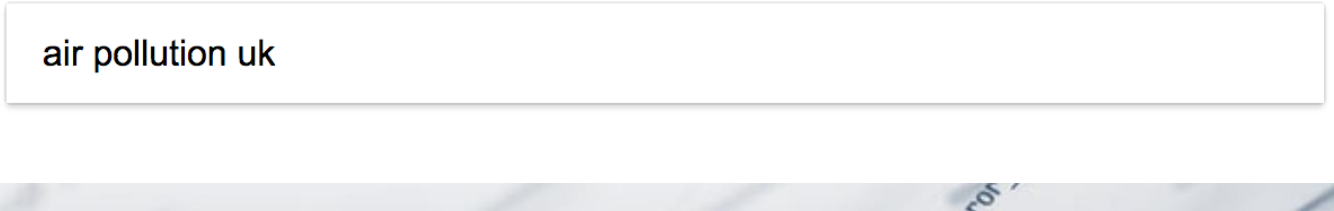
21 May 2018 | Business | 590

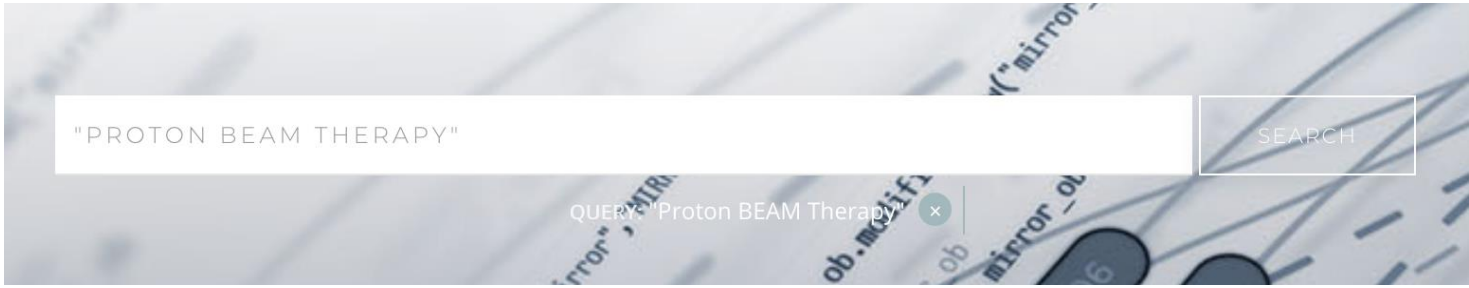
Text Mining for Journalism Approach

- 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





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
{
```

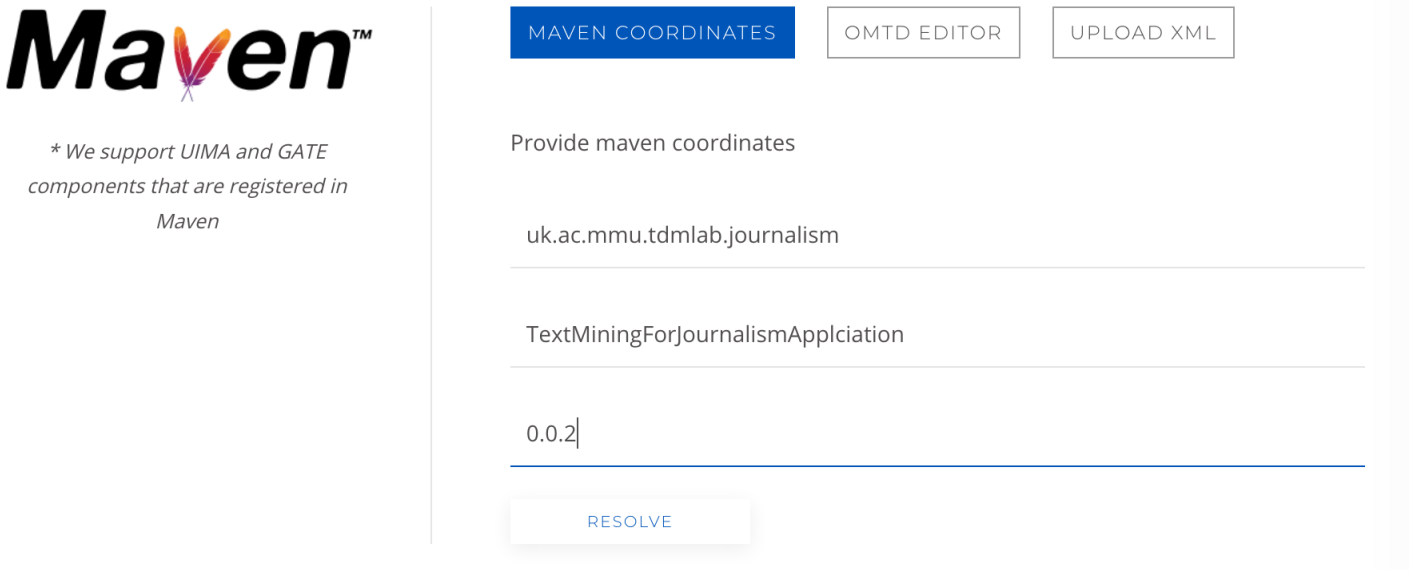
Text Mining for Journalism Implementation

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

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

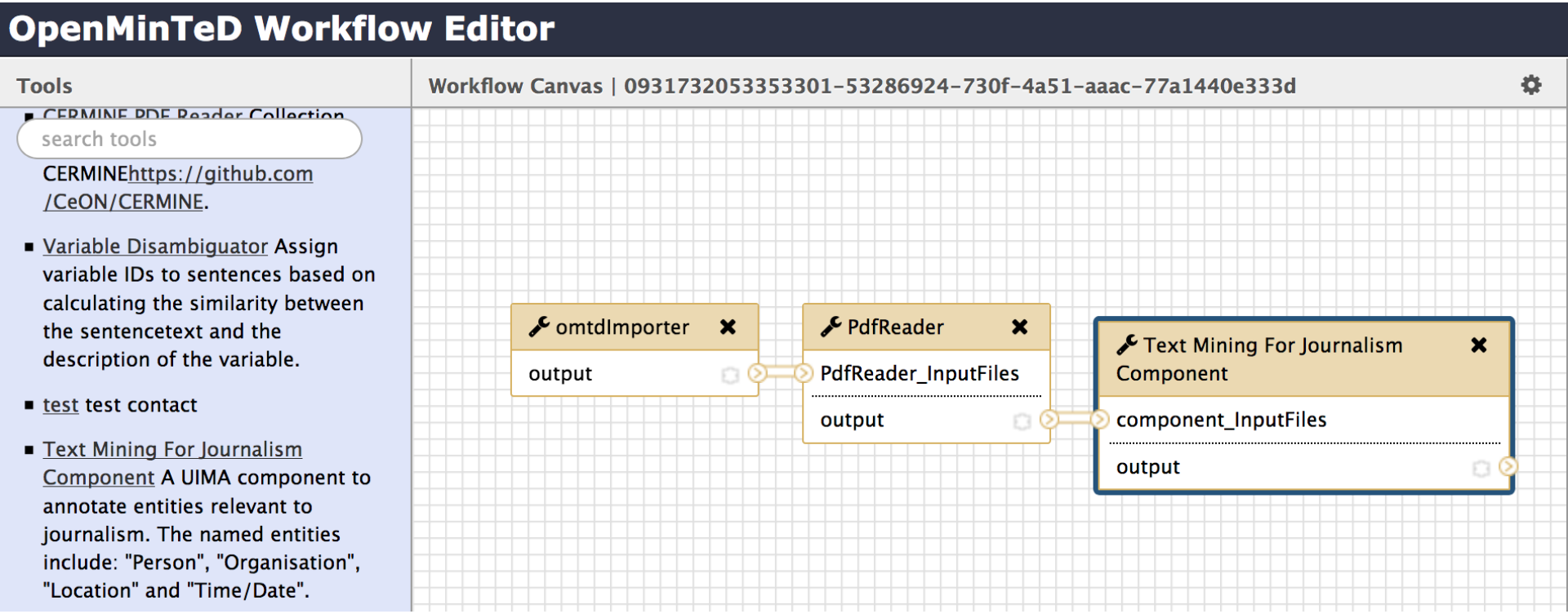
Text Mining for Journalism Implementation

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



Text Mining for Journalism Implementation

- Workflow creation via platform



Text Mining for Journalism
Running the application

Video Demonstration