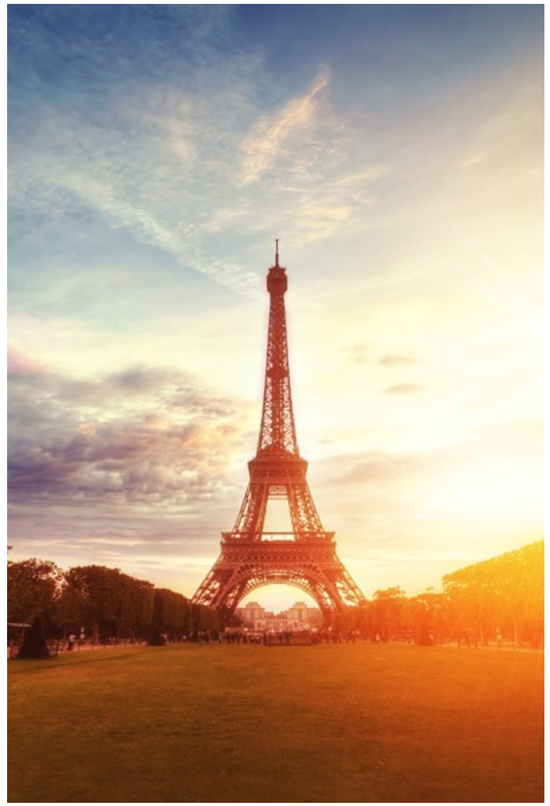
Document Clustering

With the explosion of texts available for machine processing, document clusterization offers an automated way of organising texts, extracting topics and information retrieval.

Reporting On Terrorism

This case study aims to find similarities between articles about terrorist attacks as reported by The Guardian.



Terrorism Related Articles

The goal is to identify common expressions and language used to report on the attacks and find out if there are differences in how different attacks are being covered.

The initial assumptions is that it should be possible to distinguish meaningful groups of articles based on the place of an attack as mentioned in the Paris, Beirut, and the Language Used to Describe Terrorism article.

Document Clusterization In Python

1

Split Articles Into Words

We collect all the words from all the articles. For each article, we count the occurrences of each word. Resulting matrix enables us to asses similarity of different documents.

2

Document Clusterization

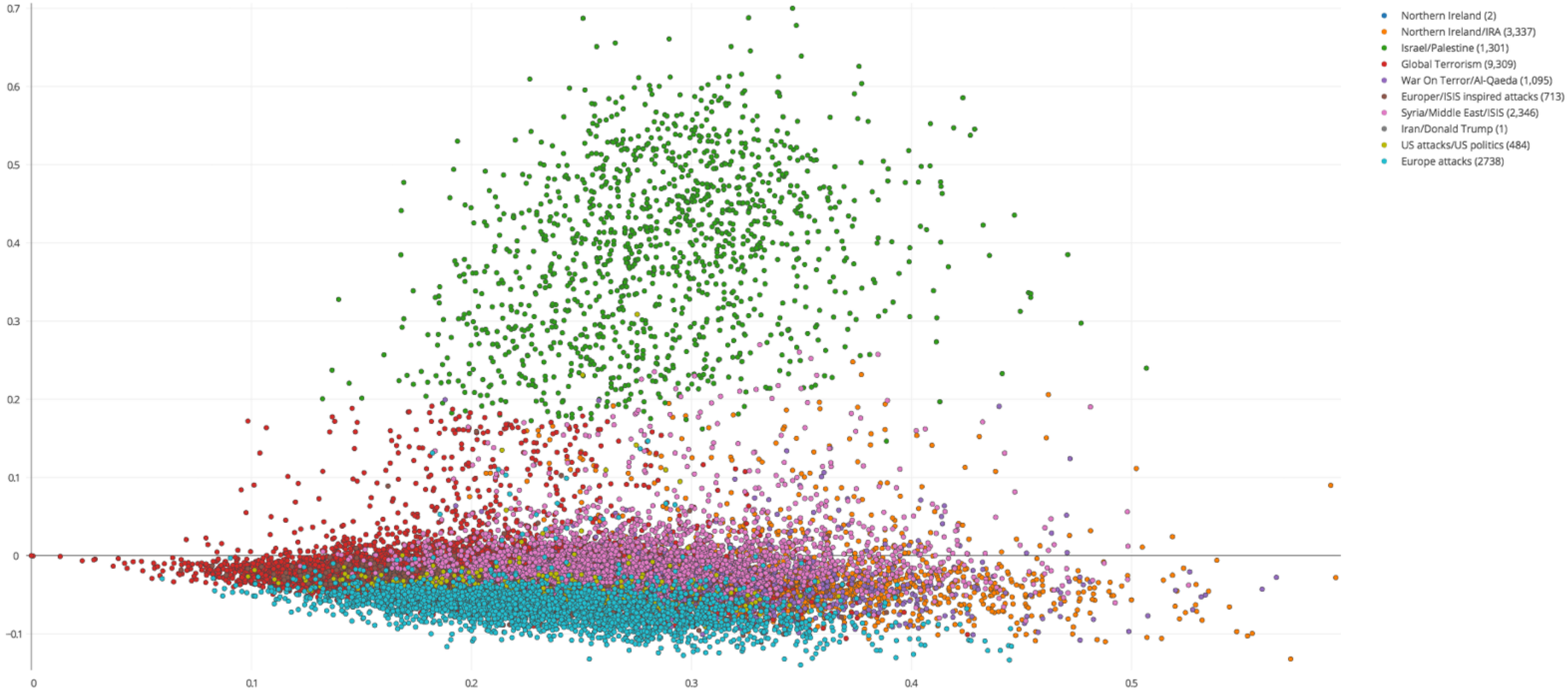
Based on similarity of words in articles, we try to group articles into 10 groups. Articles within each group will share same words.

3

Decomposition

To be able to visualise the clustered results, we reduce the amount of dimensions within our dataset to 2 and 3.

Clusterization Results



Interactive version at https://capstone-visualisation.herokuapp.com/2d-scatter.html

Clusterization Results

Interactive version at https://capstone-visualisation.herokuapp.com/

Identified Groups

01. N. Ireland / Orange Group

Contains only two articles about the Orange Volunteers, a terrorist group targeting catholics in N. Ireland in the 90s.

02. N. Ireland / I.R.A

3,337 articles mostly concerned with N. Ireland and IRA related activity.

03. Israel / Palestine

1,301 articles mostly concerned with the Israeli-Palestinian conflict.

04. Global Terrorism

The largest group of 9,309 articles with no distinguished topics other then terrorist attacks across the globe.

05. War On Terror / Al-Qaeda

1,905 articles mostly published before 2012, mostly reporting on Al-Qaeda inspired attacks and War on Terror.

06. Europe / Isis Inspired Attacks

713 articles mostly reporting on attacks from 2016 and 2017 with emphasis on the ISIS connection.

07. Syria / Middle East / Isis

2,346 articles mostly reporting on Syrian war, continuing war on terror in Afghanistan and Pakistan and rise of Isis.

08. Afghanistan

Single pre-9/11 of the US administration article asking Taliban to stop funding Bin-Laden.

09. U.S. Attacks / U.S. Politics

484 articles mostly covering attacks happening in the US and US politics.

10. Europe Attacks

2,738 articles mostly covering attacks in the UK and across Europe.

Topical Analysis

1

Create Dictionary Of Texts

Collect all the words from all the articles and create a dictionary. corpus where each article is represented as a set of words, disregarding the order of words or grammar.

2

Create Document Generation Model

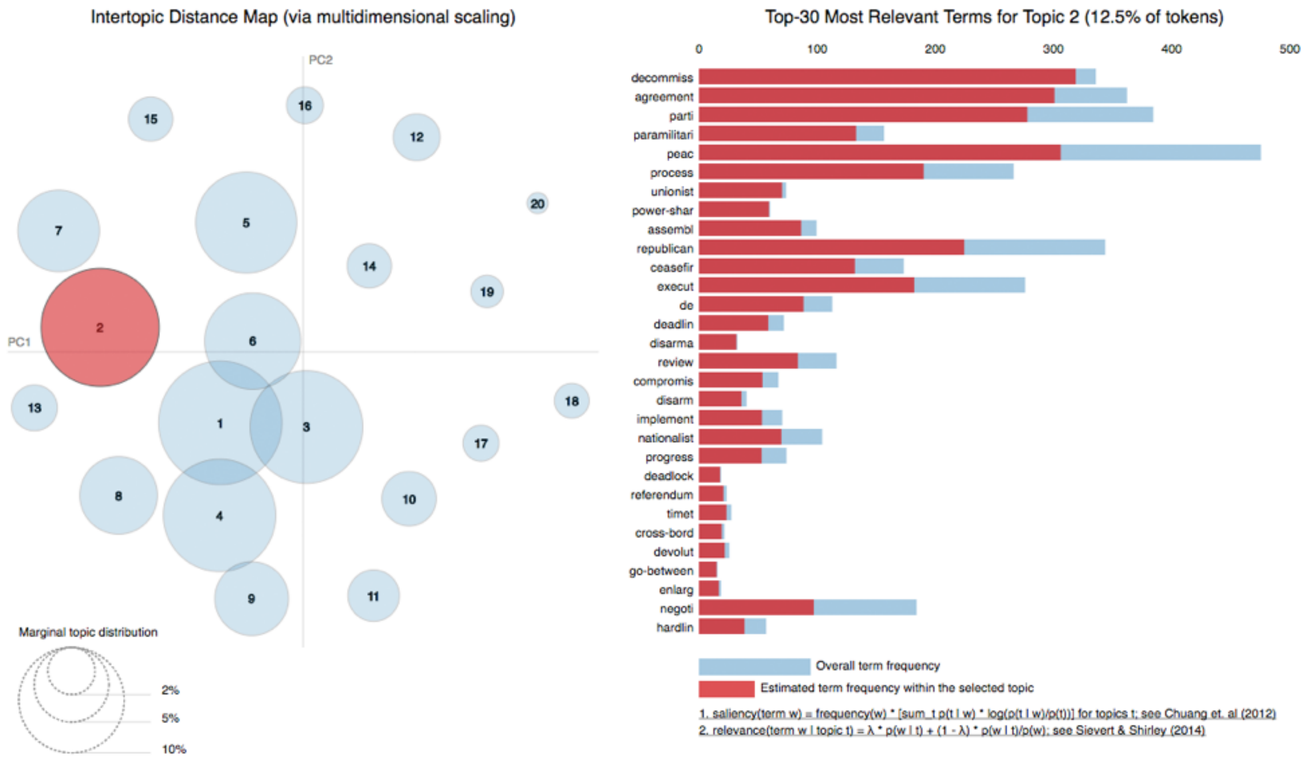
Generates random documents by choosing a mixture of 10 topics and generating individual words by choosing a random topic.

3

Topics Extraction

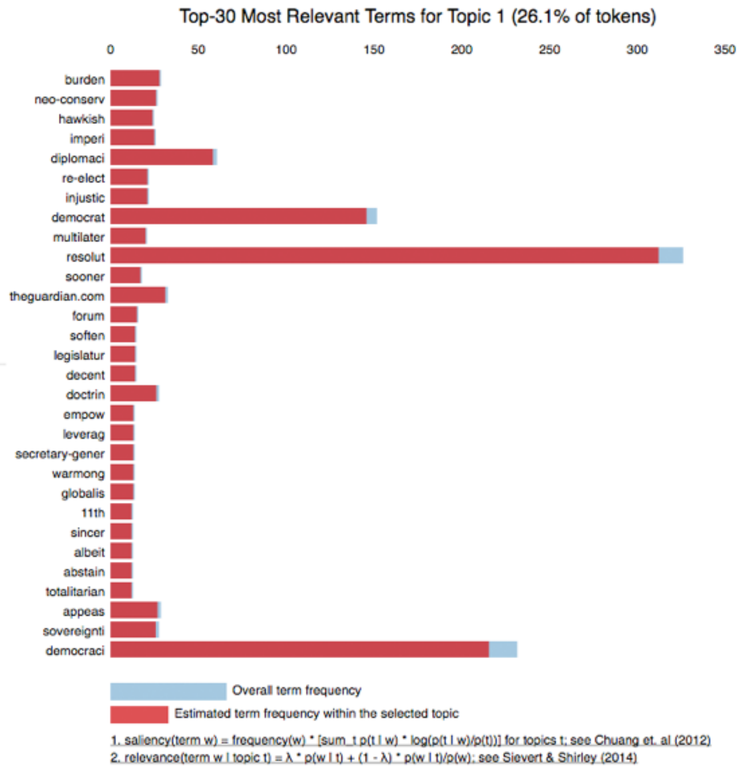
Comparing to the random documents generated in the previous step, estimate what combination of topics would generate documents similar to the articles.

Topics 1990 – 2000



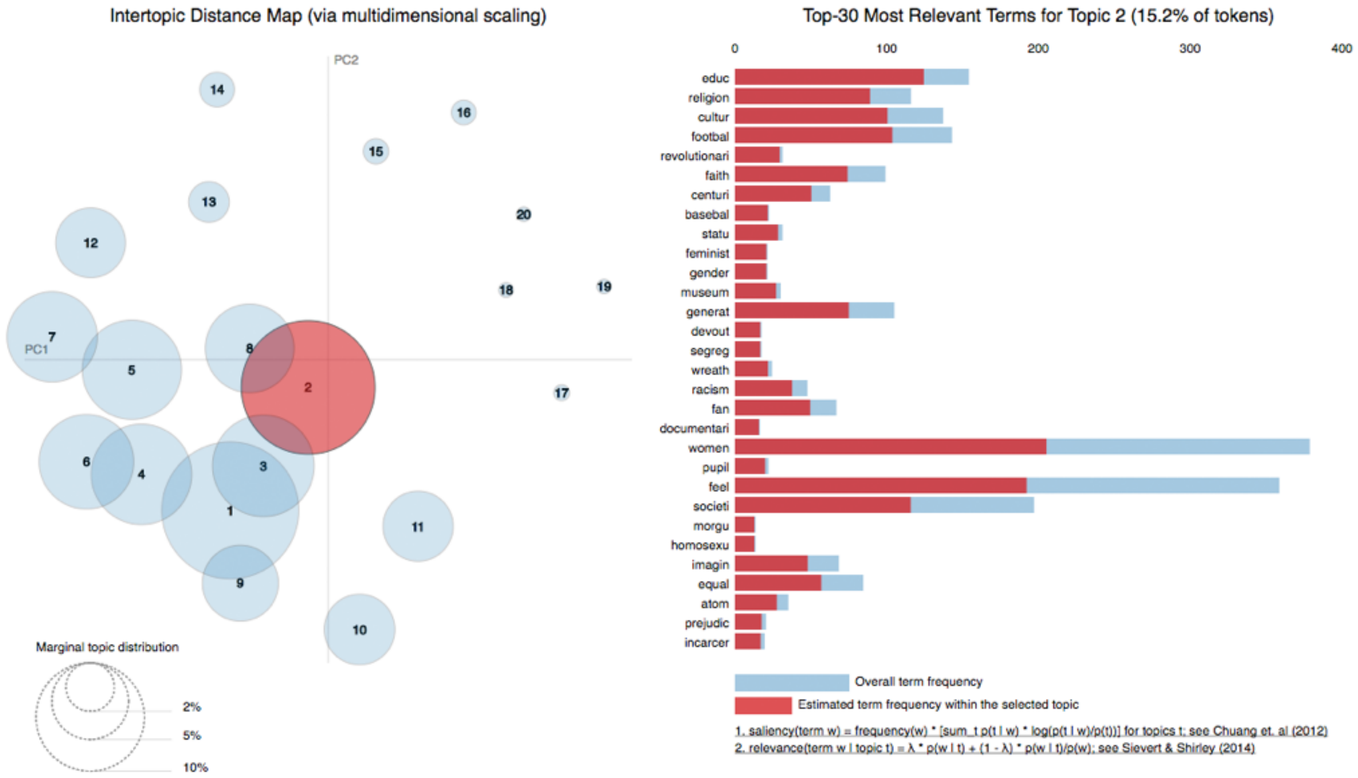
One of the most prominent topic in the 90s seems to be still ongoing conflict in Northern Ireland.

Topics 2002 – 2003



The most prominent of the early 2000s seems to be the neoliberal political movement in the US and related War on terror.

Topics 2016 – 2017



One of the most prominent topic of the mid 2010s seems to be Islamic terrorism and related conversations about religion and human rights.

Evaluating Results

01. Clusterization

Clustering analysis failed to uncover surprising insights. Seems like the articles are mostly clustered based on place of the attack.
Resulting clusters are not very well defined, with the exception of the Israel / Palestine cluster. This might be due to the general similarity of all the documents, since they all cover the same topic of terrorism.

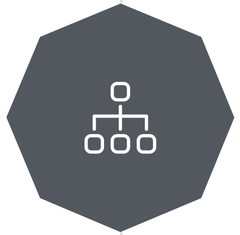
02. Topic Modelling

The analysis of the evolution of topics throughout years yields an interesting overview of the themes discussed.
While the raw result of the topic modelling analysis is hard to interpret, with the use data visualisation tools, the changing topics and different keywords associated with them are easily explored.

Conclusions

Clusterization

The document clusterization methods used don’t work very well for documents which are very similar.



Topical Modelling

Topical modelling proved useful in uncovering general themes across the documents.

Visualization Methods

Methods of data visualisation proved critical in evaluating results of the all the method of the information retrieval.