An open source GATE toolkit for social media analysis

Diana Maynard
University of Sheffield, UK

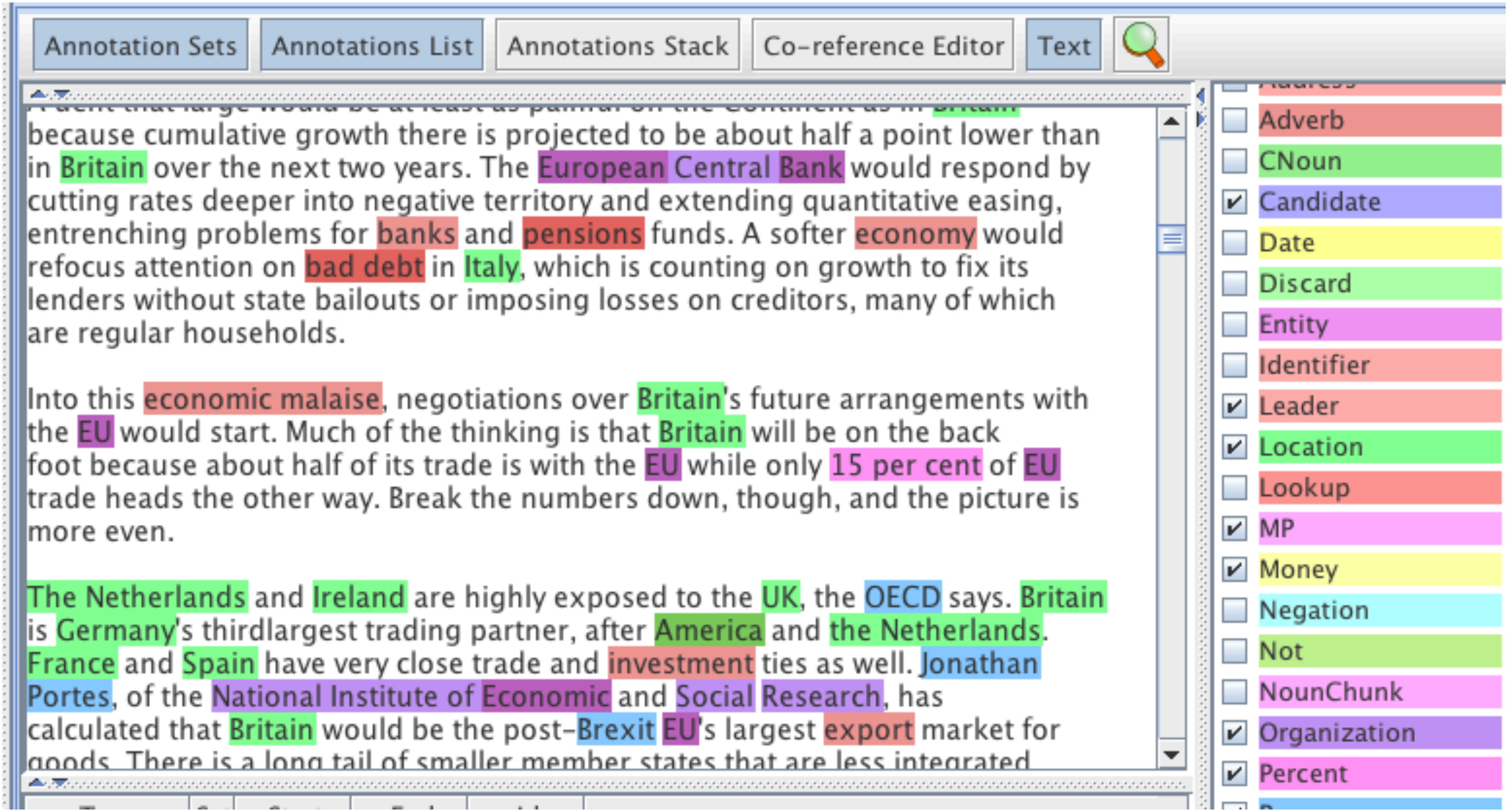
GATE for text engineering

- Tool for developing and deployment of Text Mining technology
- http://gate.ac.uk
- 20 years old (but not past it yet!)
- established large developer community, incl. industrial committers (Ontotext, Intellius, Text Mining Solutions)
- Used worldwide by many organisations to build bespoke solutions, e.g. TNA, Press Association, BBC, NHS
- A free open source framework (LGPL) and GUI
- Includes Information Extraction tools in many languages
- Research team of 15 people
- Lots of tools also for social media analysis

GATE open-source social media tools

Name/Description	Description / Domain	License
GATE	Numerous tools for text analytics, including social media	LGPL + plugin specific
TwitIE	NER for tweets	LGPL
GATECloud	Various cloud-based GATE services, including social media analytics	SaaS model
Sentiment analysis	Generic twitter-based sentiment analysis	LGPL
Tweet Collector	Social media analytics	SaaS model
YODIE (in development)	Cloud-based named entity disambiguation and linking; version for social media	SaaS model

GATE Annotation



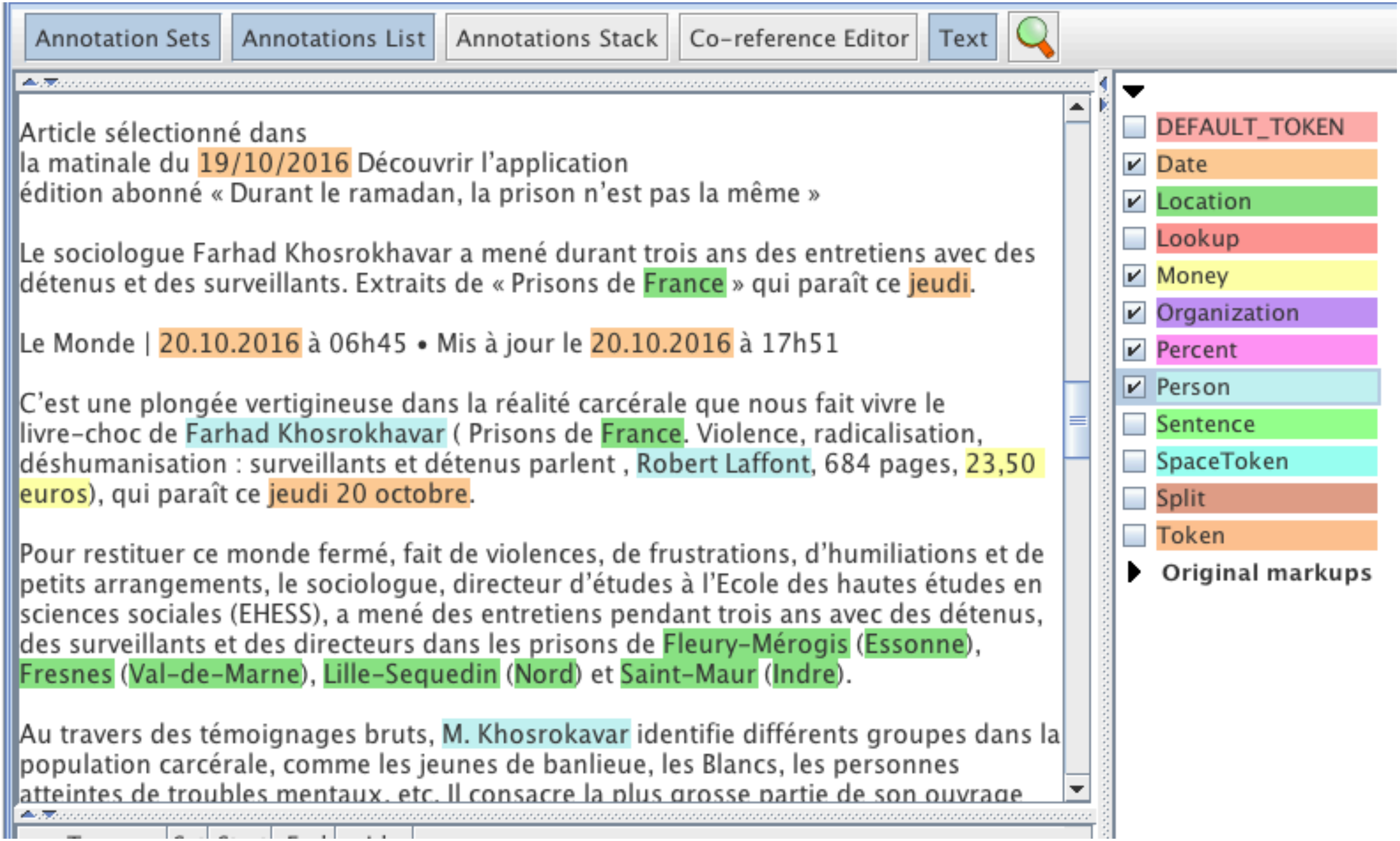
because cumulative growth there is projected to be about half a point lower than in Britain over the next two years. The European Central Bank would respond by cutting rates deeper into negative territory and extending quantitative easing, entrenching problems for banks and pensions funds. A softer economy would refocus attention on bad debt in Italy, which is counting on growth to fix its lenders without state bailouts or imposing losses on creditors, many of which are regular households.

Into this economic malaise, negotiations over Britain's future arrangements with the EU would start. Much of the thinking is that Britain will be on the back foot because about half of its trade is with the EU while only 15 per cent of EU trade heads the other way. Break the numbers down, though, and the picture is more even.

The Netherlands and Ireland are highly exposed to the UK, the OECD says. Britain is Germany's thirdlargest trading partner, after America and the Netherlands. France and Spain have very close trade and investment ties as well. Jonathan Portes, of the National Institute of Economic and Social Research, has calculated that Britain would be the post-Brexit EU's largest export market for goods. There is a long tail of smaller member states that are less integrated

- ☐ Adverb
- ☐ CNoun
- ☑ Candidate
- ☐ Date
- ☐ Discard
- ☐ Entity
- ☐ Identifier
- ☑ Leader
- ☑ Location
- ☐ Lookup
- ☑ MP
- ☑ Money
- ☐ Negation
- ☐ Not
- ☐ NounChunk
- ☑ Organization
- ☑ Percent

NER in French



Article sélectionné dans
la matinale du 19/10/2016 Découvrir l'application
édition abonné « Durant le ramadan, la prison n'est pas la même »

Le sociologue Farhad Khosrokhavar a mené durant trois ans des entretiens avec des détenus et des surveillants. Extraits de « Prisons de France » qui paraît ce jeudi.

Le Monde | 20.10.2016 à 06h45 • Mis à jour le 20.10.2016 à 17h51

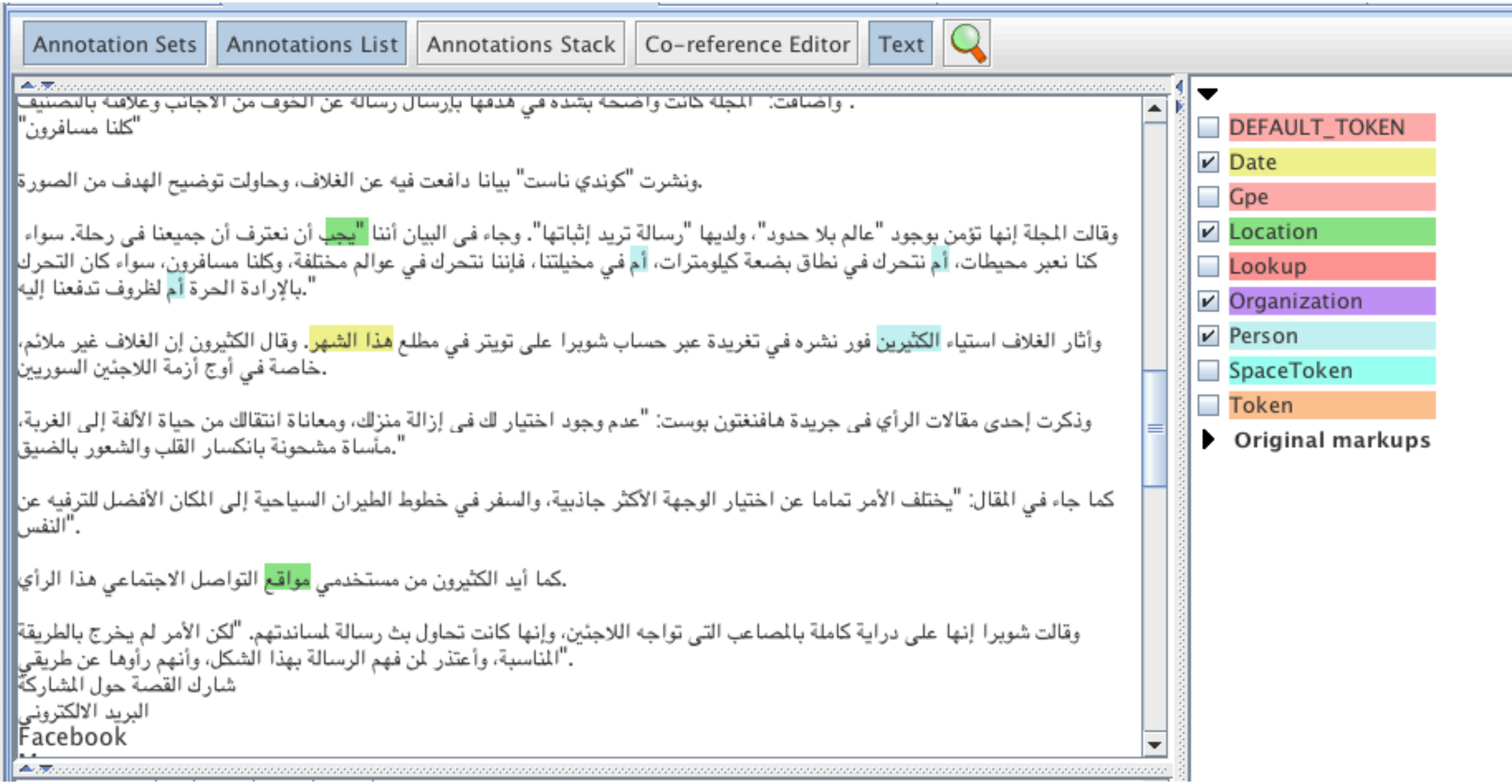
C'est une plongée vertigineuse dans la réalité carcérale que nous fait vivre le livre-choc de Farhad Khosrokhavar (Prisons de France. Violence, radicalisation, déshumanisation : surveillants et détenus parlent , Robert Laffont, 684 pages, 23,50 euros), qui paraît ce jeudi 20 octobre.

Pour restituer ce monde fermé, fait de violences, de frustrations, d'humiliations et de petits arrangements, le sociologue, directeur d'études à l'Ecole des hautes études en sciences sociales (EHESS), a mené des entretiens pendant trois ans avec des détenus, des surveillants et des directeurs dans les prisons de Fleury-Mérogis (Essonne), Fresnes (Val-de-Marne), Lille-Sequedin (Nord) et Saint-Maur (Indre).

Au travers des témoignages bruts, M. Khosrokavar identifie différents groupes dans la population carcérale, comme les jeunes de banlieue, les Blancs, les personnes atteintes de troubles mentaux, etc. Il consacre la plus grosse partie de son ouvrage

□ DEFAULT_TOKEN
☑ Date
☑ Location
□ Lookup
☑ Money
☑ Organization
☑ Percent
☑ Person
□ Sentence
□ SpaceToken
□ Split
□ Token
▶ Original markups

NER in Arabic



وأضافت: "المجلة كانت واضحة بشدة في هدفها بإرسال رسالة عن الخوف من الأجانب وعلاقته بالتصنيف .
"كلنا مسافرون"

ونشرت "كوندي ناست" بيانا دافعت فيه عن الغلاف، وحاولت توضيح الهدف من الصورة.

وقالت المجلة إنها تؤمن بوجود "عالم بلا حدود"، ولديها "رسالة تريد إثباتها". وجاء فى البيان أننا "يجب أن نعترف أن جميعنا فى رحلة. سواء
كنا نعبر محيطات، أم نتحرك في نطاق بضعة كيلومترات، أم في مخيلتنا، فإننا نتحرك في عوالم مختلفة، وكلنا مسافرون، سواء كان التحرك
بالإرادة الحرة أم لظروف تدفعنا إليه."

وأثار الغلاف استياء الكثيرين فور نشره في تغريدة عبر حساب شويرا على تويتر في مطلع هذا الشهر. وقال الكثيرون إن الغلاف غير ملائم،
خاصة في أوج أزمة اللاجئين السوريين.

وذكرت إحدى مقالات الرأي فى جريدة هافنغتون بوست: "عدم وجود اختيار لك فى إزالة منزلك، ومعاناة انتقالك من حياة الألفة إلى الغربة،
مأساة مشحونة بانكسار القلب والشعور بالضيق."

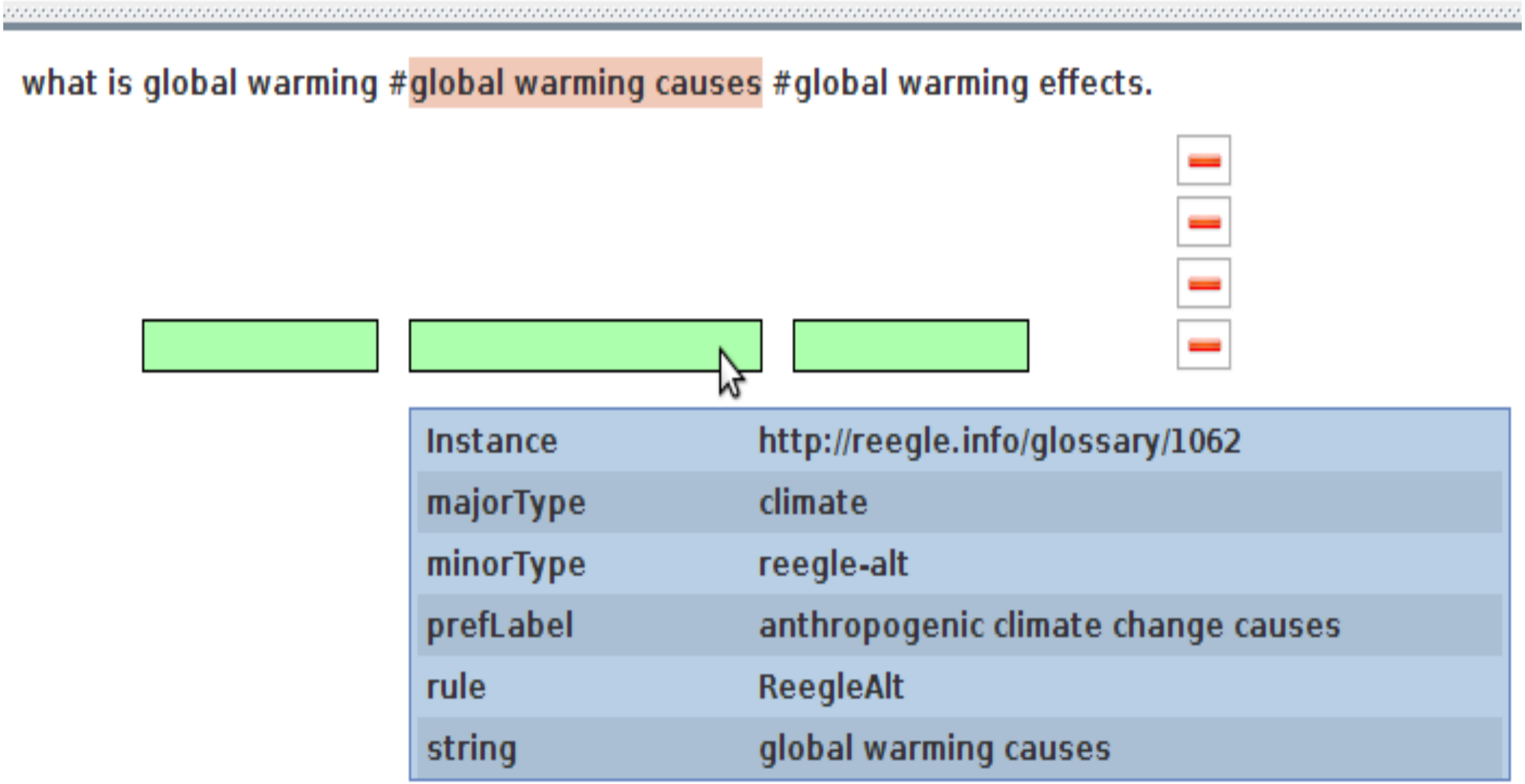
كما جاء في المقال: "يختلف الأمر تماما عن اختيار الوجهة الأكثر جاذبية، والسفر في خطوط الطيران السياحية إلى المكان الأفضل للترفيه عن
النفس".

كما أيد الكثيرون من مستخدمي مواقع التواصل الاجتماعي هذا الرأي.

وقالت شويرا إنها على دراية كاملة بالمصاعب التى تواجه اللاجئين، وإنها كانت تحاول بث رسالة لمساندتهم. "لكن الأمر لم يخرج بالطريقة
المناسبة، وأعتذر لمن فهم الرسالة بهذا الشكل، وأنهم رأوها عن طريقي".
شارك القصة حول المشاركة
البريد الالكتروني
Facebook

☐ DEFAULT_TOKEN
☑ Date
☐ Gpe
☑ Location
☐ Lookup
☑ Organization
☑ Person
☐ SpaceToken
☐ Token
▶ Original markups

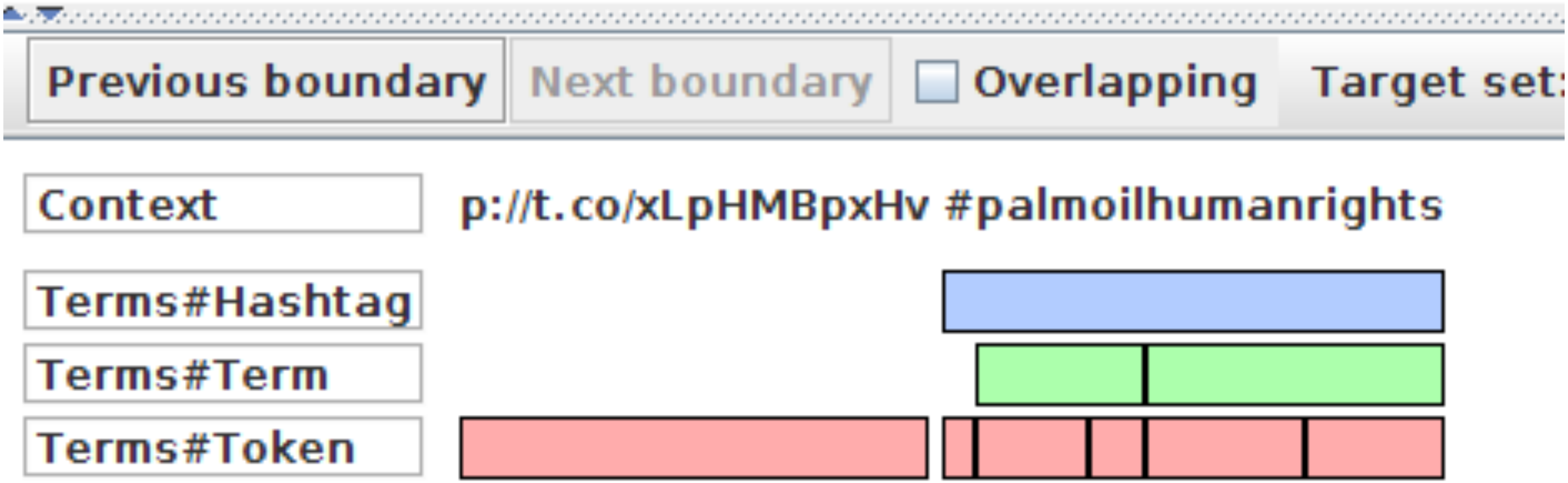
Environmental term recognition



what is global warming #global warming causes #global warming effects.

Instance	http://reegle.info/glossary/1062
majorType	climate
minorType	reegle-alt
prefLabel	anthropogenic climate change causes
rule	ReegleAlt
string	global warming causes

Hashtag analysis



- We need to chop up the hashtags into individual words, and find any terms within them
  - #palmoilhumanrights --> palm oil + human rights

GATE Cloud Tools

- You can try various demos of our tools on the GATE Cloud
- http://cloud.gate.ac.uk
- You can process small amounts of text for free
- For larger volumes, you can set up an account and pay for the time you use

Collect your own Tweets!

- Twitter provides APIs to download tweets in real time based on various filtering conditions
- They can be challenging to set up if you're not a programmer
- We have tools on the GATE Cloud to let you set up a query to collect tweets by particular people, matching certain keywords, geo-tagged with particular locations, and so on.
- You don’t need to install any software!
- More details available on GATE Cloud pages

Exploring the collected data

- You can also get 6 reports on your data collected so far, updated in real time.

1. View the top hashtags as a bar chart or tag cloud, and add hashtags to the tracking list in the collector.
2. View a line graph of tweet frequency per hour.
3. View a bar chart or cloud of the top topics according to lists of terms. The lists can be edited from a link on the topics graph page. Topics can be added as terms to the tracker.
4. View a bar chart or cloud of the most mentioned users. These can also be added for the collector to track.
5. View a bar chart or cloud of the frequency of the terms you are tracking.
6. View a bar chart or cloud of the top words of one grammatical type (adjective, verb, noun etc.) (only for English)

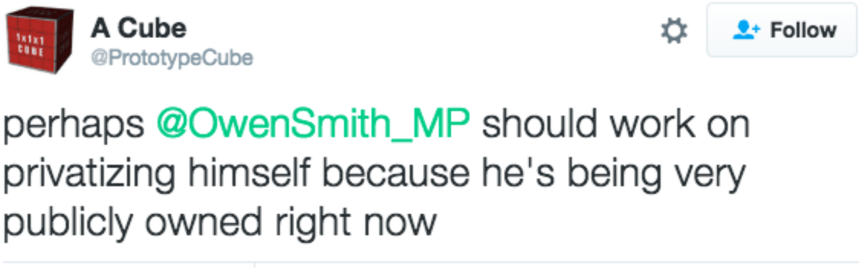
Demo

Analysing Polarised Societal Debates

- Public, polarised debates affect policy making, the course of a country, voting intents
- Historically, the public debates took place on TV and involved only politicians or media figures

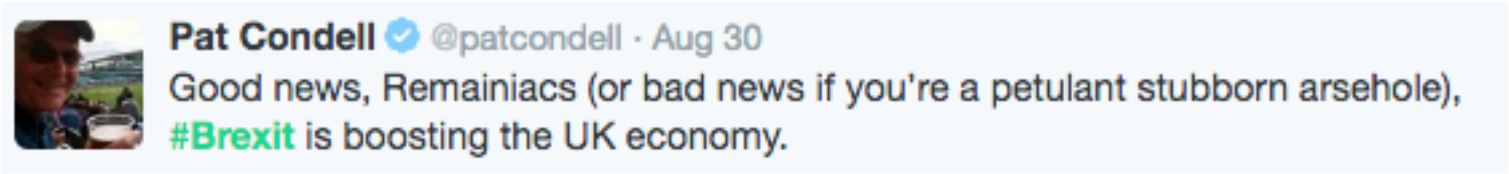
Motivation

- Today, polarised debates also take place publicly on the internet, and anyone can participate
- Analysing these public statements tells us:
  - What participants are thinking
  - What arguments they are making
  - Who is thinking what
  - Gain a “big picture” view of the response to events or policies.



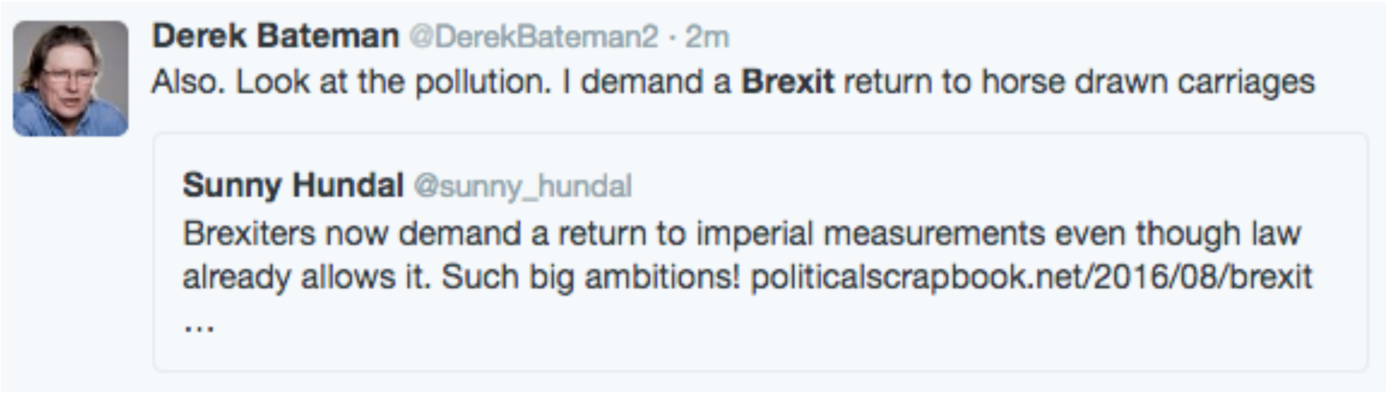
A Cube @PrototypeCube

perhaps @OwenSmith_MP should work on privatizing himself because he's being very publicly owned right now



Pat Condell @patcondell · Aug 30

Good news, Remainiacs (or bad news if you're a petulant stubborn arsehole), #Brexit is boosting the UK economy.



Derek Bateman @DerekBateman2 · 2m

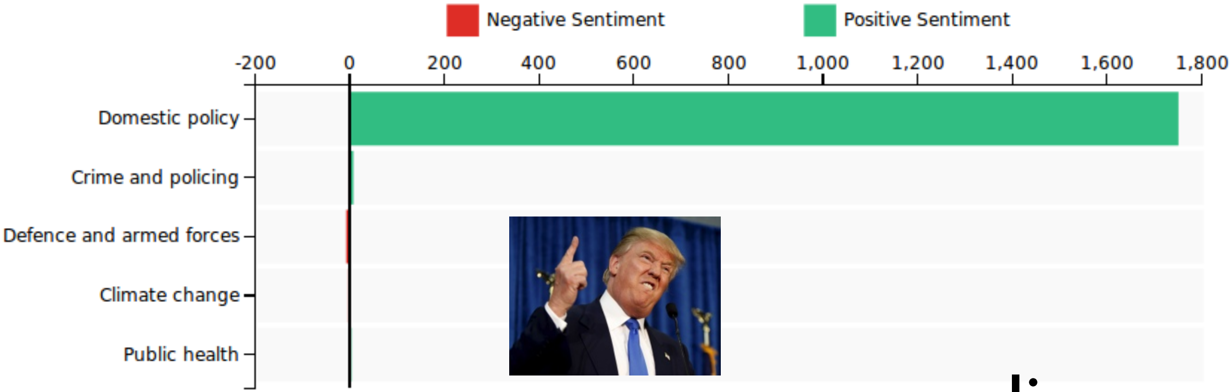
Also. Look at the pollution. I demand a **Brexit** return to horse drawn carriages

> **Sunny Hundal** @sunny_hundal
>
> Brexiters now demand a return to imperial measurements even though law already allows it. Such big ambitions! politicalscrapbook.net/2016/08/brexit ...

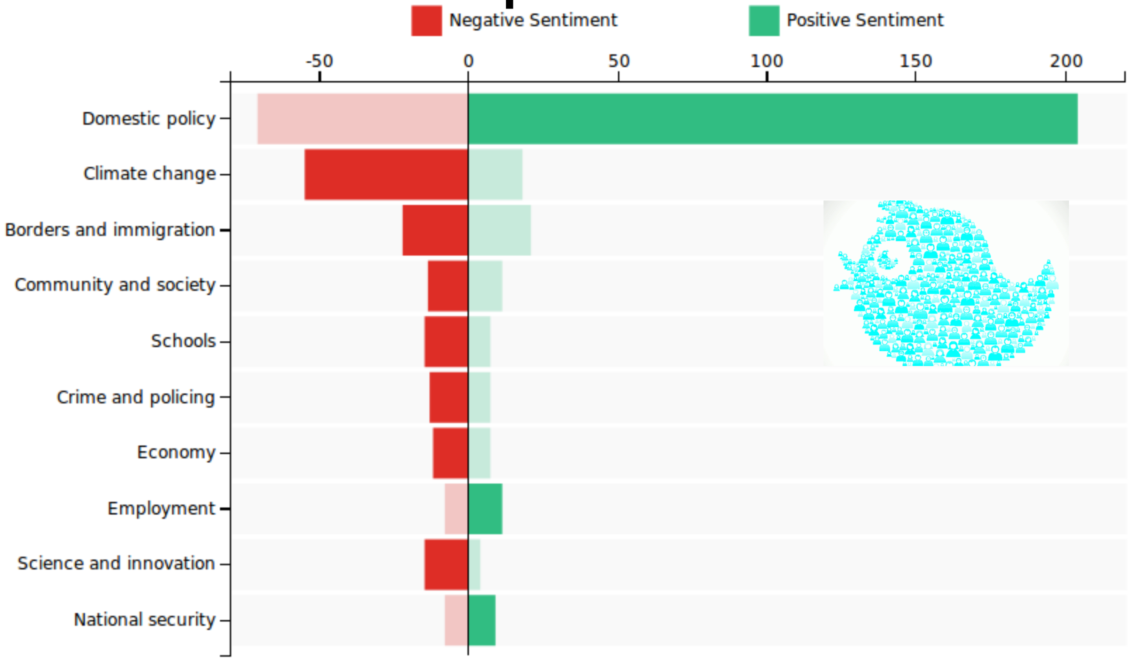
The Story

- Analyse social media surrounding polarised societal debates
- Index the results to understand, visualise and explore:
  - What topics are being discussed?
  - What sentiments are being expressed?
  - Who is participating in the debate?
  - How do debates evolve over time?

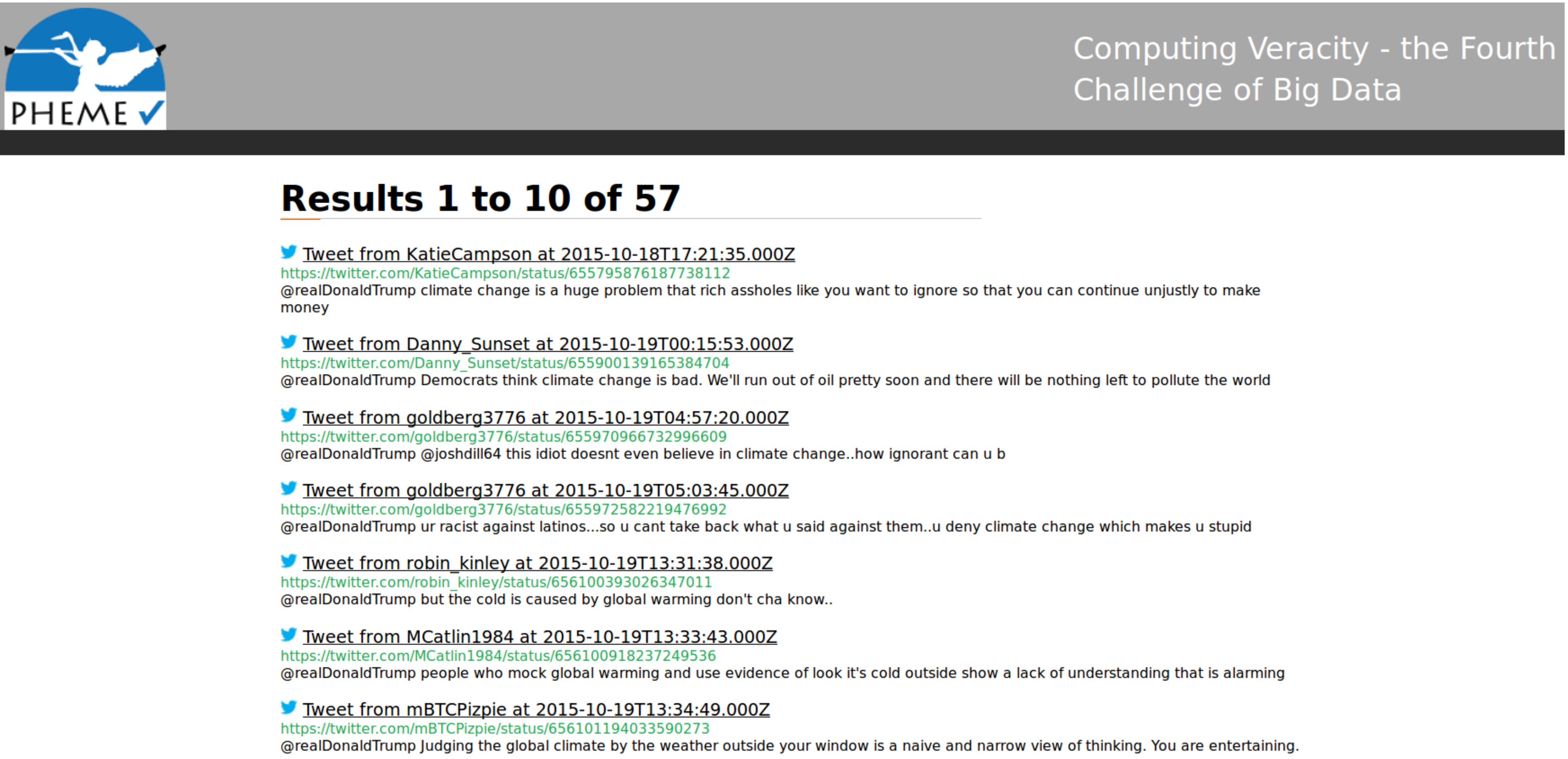
Real-time Opinion Monitoring



vs replies



Replies to Trump re: climate change



Computing Veracity - the Fourth Challenge of Big Data

Results 1 to 10 of 57

Tweet from KatieCampson at 2015-10-18T17:21:35.000Z
https://twitter.com/KatieCampson/status/655795876187738112
@realDonaldTrump climate change is a huge problem that rich assholes like you want to ignore so that you can continue unjustly to make money

Tweet from Danny_Sunset at 2015-10-19T00:15:53.000Z
https://twitter.com/Danny_Sunset/status/655900139165384704
@realDonaldTrump Democrats think climate change is bad. We'll run out of oil pretty soon and there will be nothing left to pollute the world

Tweet from goldberg3776 at 2015-10-19T04:57:20.000Z
https://twitter.com/goldberg3776/status/655970966732996609
@realDonaldTrump @joshdill64 this idiot doesnt even believe in climate change..how ignorant can u b

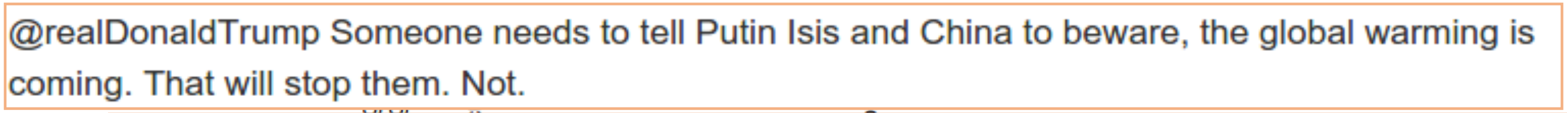
Tweet from goldberg3776 at 2015-10-19T05:03:45.000Z
https://twitter.com/goldberg3776/status/655972582219476992
@realDonaldTrump ur racist against latinos...so u cant take back what u said against them..u deny climate change which makes u stupid

Tweet from robin_kinley at 2015-10-19T13:31:38.000Z
https://twitter.com/robin_kinley/status/656100393026347011
@realDonaldTrump but the cold is caused by global warming don't cha know..

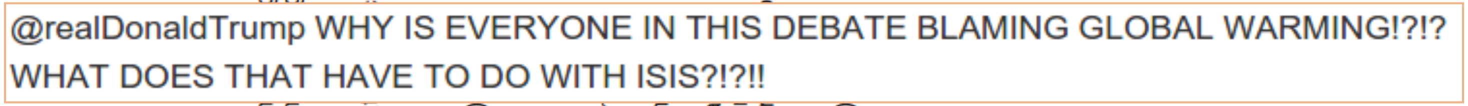
Tweet from MCatlin1984 at 2015-10-19T13:33:43.000Z
https://twitter.com/MCatlin1984/status/656100918237249536
@realDonaldTrump people who mock global warming and use evidence of look it's cold outside show a lack of understanding that is alarming

Tweet from mBTCPizpie at 2015-10-19T13:34:49.000Z
https://twitter.com/mBTCPizpie/status/656101194033590273
@realDonaldTrump Judging the global climate by the weather outside your window is a naive and narrow view of thinking. You are entertaining.

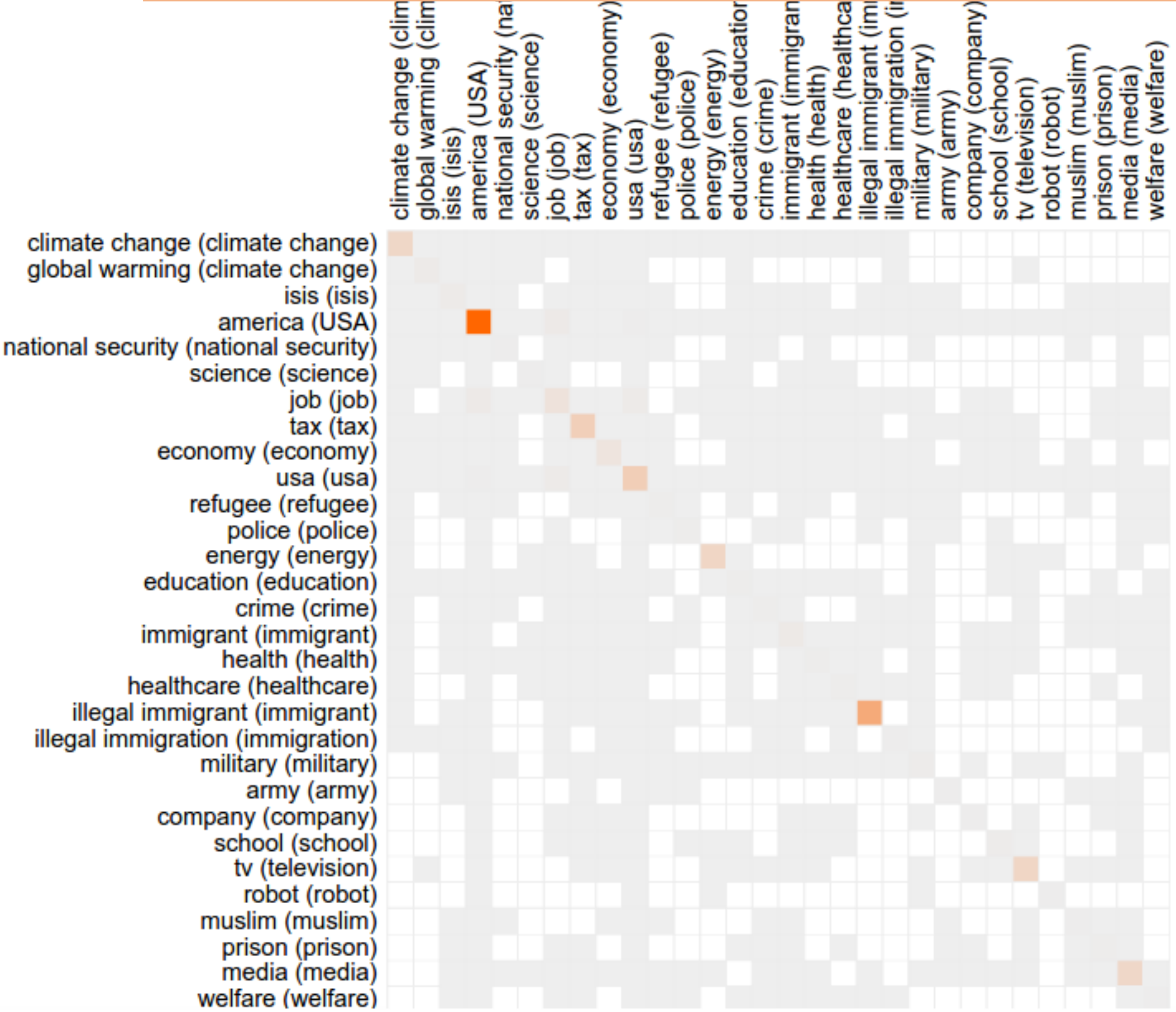
Climate change, ISIS and Trump



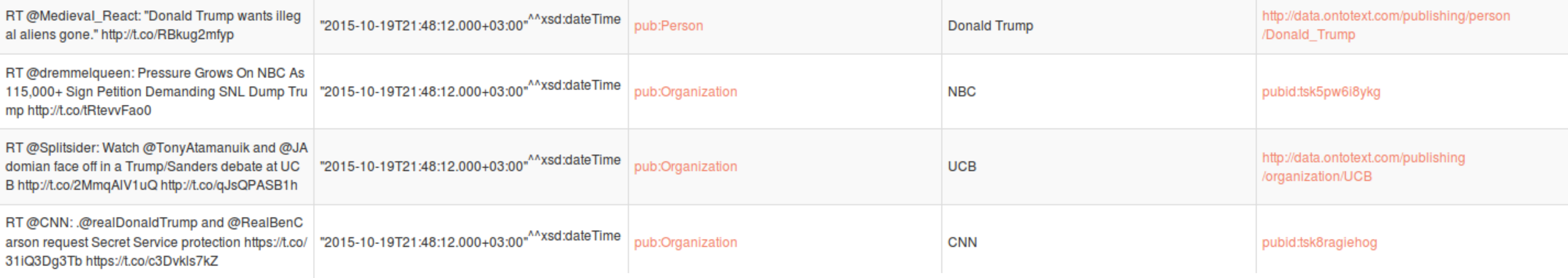
@realDonaldTrump Someone needs to tell Putin Isis and China to beware, the global warming is coming. That will stop them. Not.



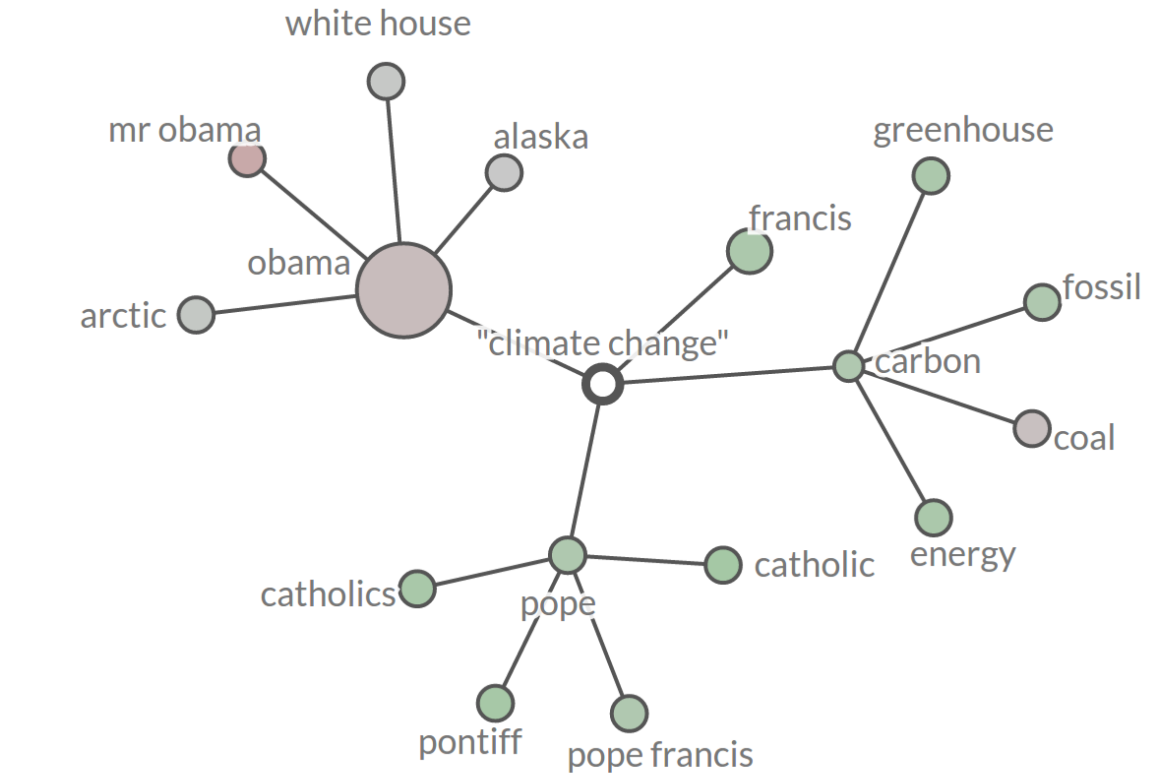
@realDonaldTrump WHY IS EVERYONE IN THIS DEBATE BLAMING GLOBAL WARMING!?!? WHAT DOES THAT HAVE TO DO WITH ISIS?!?!!



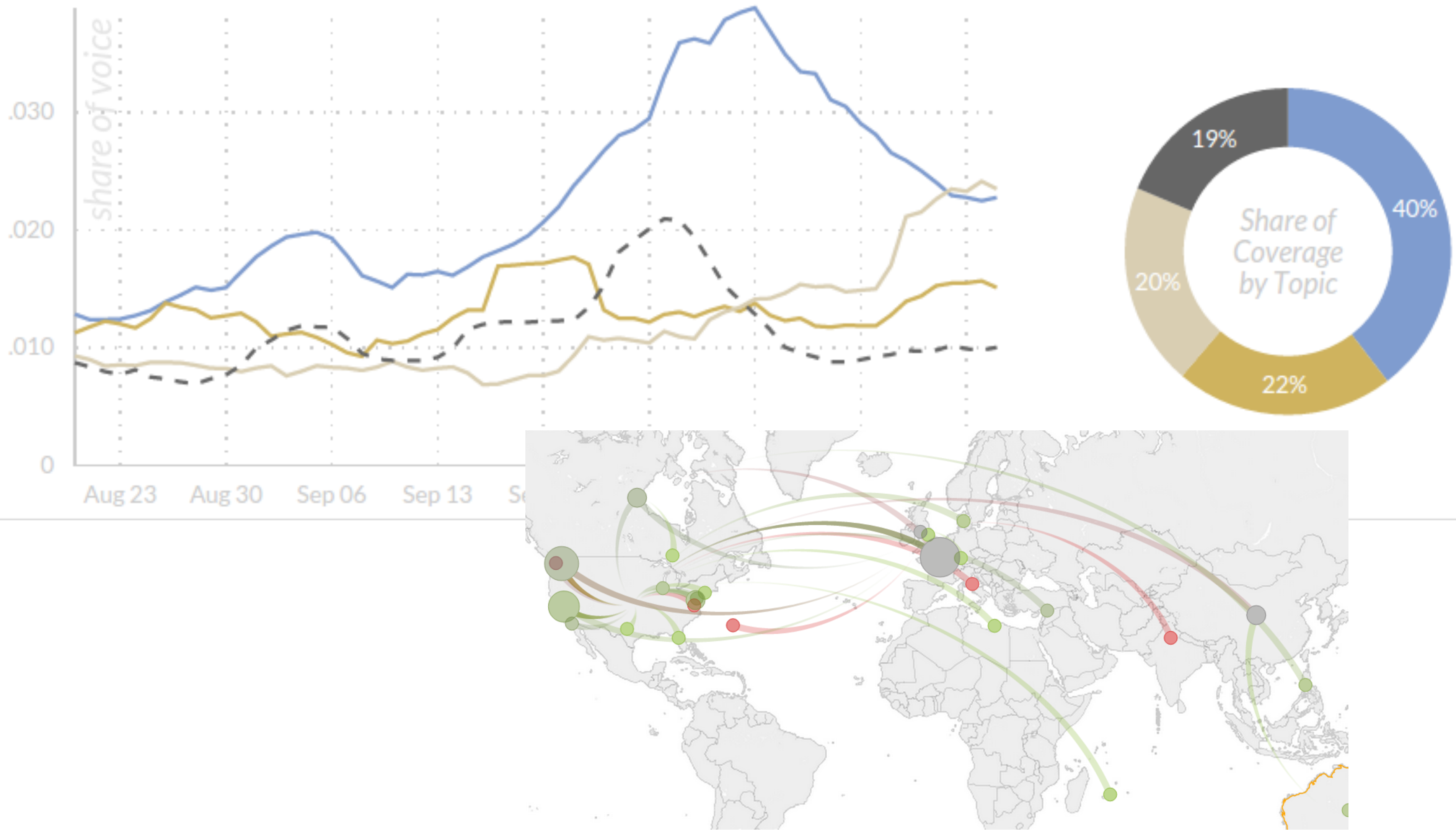
Topic relations



RT @Medieval_React: "Donald Trump wants illegal aliens gone." http://t.co/RBkug2mfyp	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Person	Donald Trump	http://data.ontotext.com/publishing/person/Donald_Trump
RT @dremmelqueen: Pressure Grows On NBC As 115,000+ Sign Petition Demanding SNL Dump Trump http://t.co/tRtevvFao0	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	NBC	pubid:tsk5pw6i8ykg
RT @Splitsider: Watch @TonyAtamanuik and @JAdomian face off in a Trump/Sanders debate at UCB http://t.co/2MmqAlV1uQ http://t.co/qJsQPASB1h	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	UCB	http://data.ontotext.com/publishing/organization/UCB
RT @CNN: .@realDonaldTrump and @RealBenCarson request Secret Service protection https://t.co/31iQ3Dg3Tb https://t.co/c3Dvkls7kZ	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	CNN	pubid:tsk8ragiehog
RT @CNN: .@realDonaldTrump and @RealBenCarson request Secret Service protection https://t.co/31iQ3Dg3Tb https://t.co/c3Dvkls7kZ	"2015-10-19T21:48:1			/publishing/organization



Dynamics Over Time/Location



SoBigData

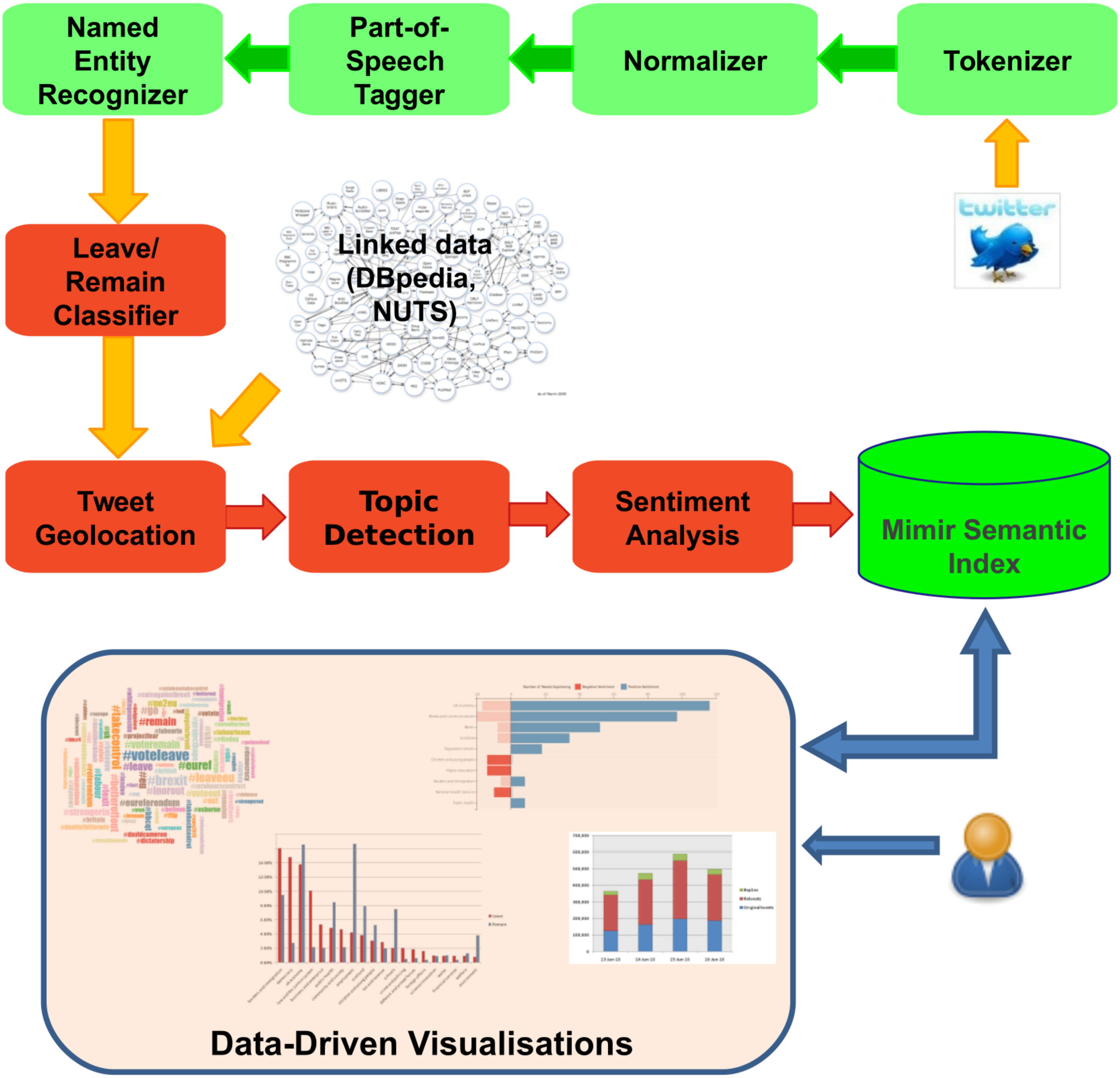
www.sobigdata.eu
@sobigdata

- European Research Infrastructure for Big Data and Social Mining
- 7 European countries, more than 100 researchers
- Collections of datasets, creation of tools
- Trans-national access programme (come and visit us!)
- Research organised in vertical themes (exploratories) on top of SBD infrastructure
- Perform cross-disciplinary cutting-edge social media mining research
  - City of Citizens
  - Well-being & Economy
  - Societal Debates
  - Migration Studies

brexit: a case study

Photo credit: https://www.flickr.com/photos/armydre2008

Brexit analyser



Named Entity Recognizer

Part-of-Speech Tagger

Normalizer

Tokenizer

twitter

Leave/ Remain Classifier

Linked data (DBpedia, NUTS)

Tweet Geolocation

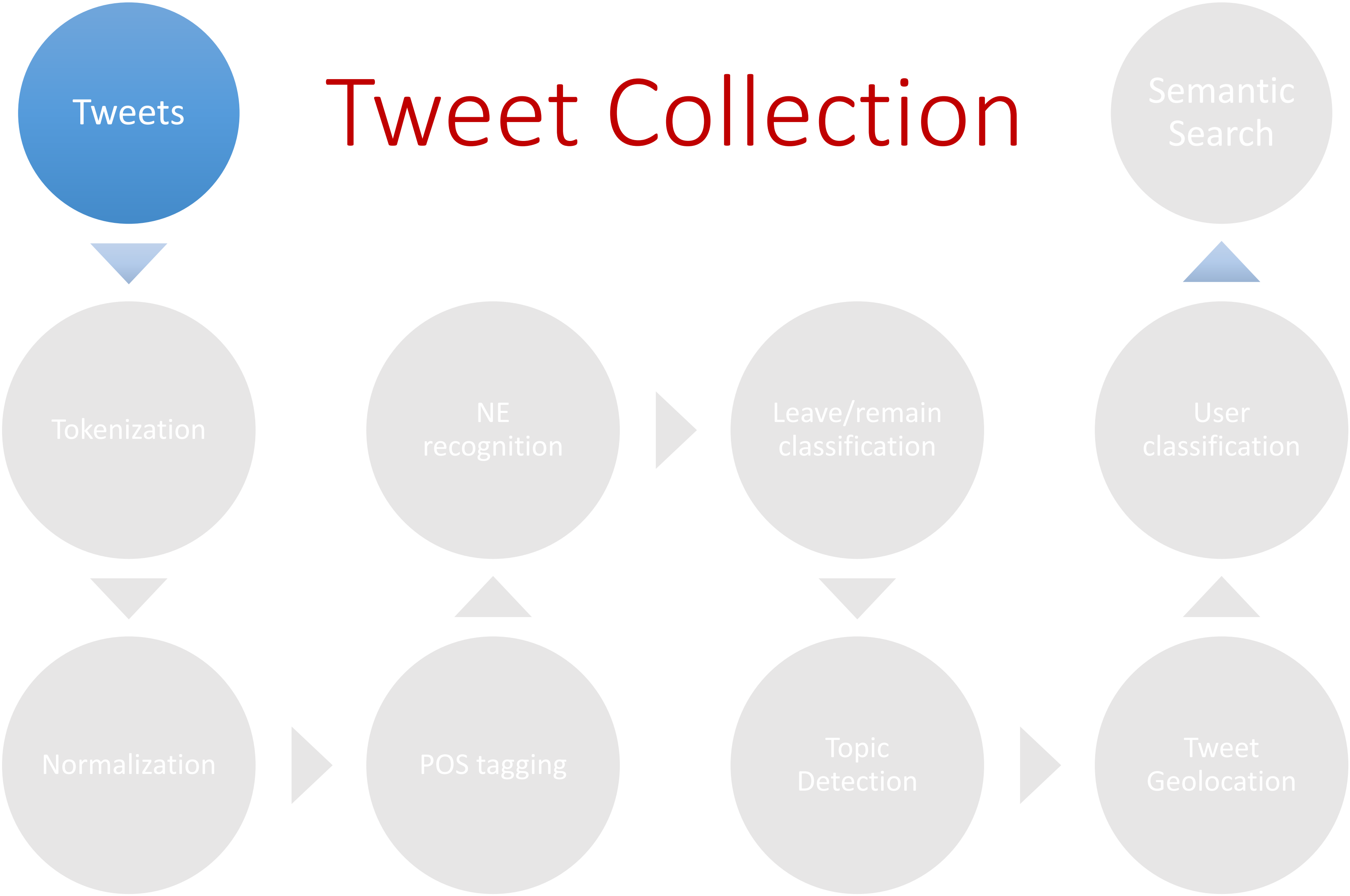
Topic Detection

Sentiment Analysis

Mimir Semantic Index

Data-Driven Visualisations

Tweet Collection



Tweets

Tokenization

Normalization

POS tagging

NE recognition

Leave/remain classification

Topic Detection

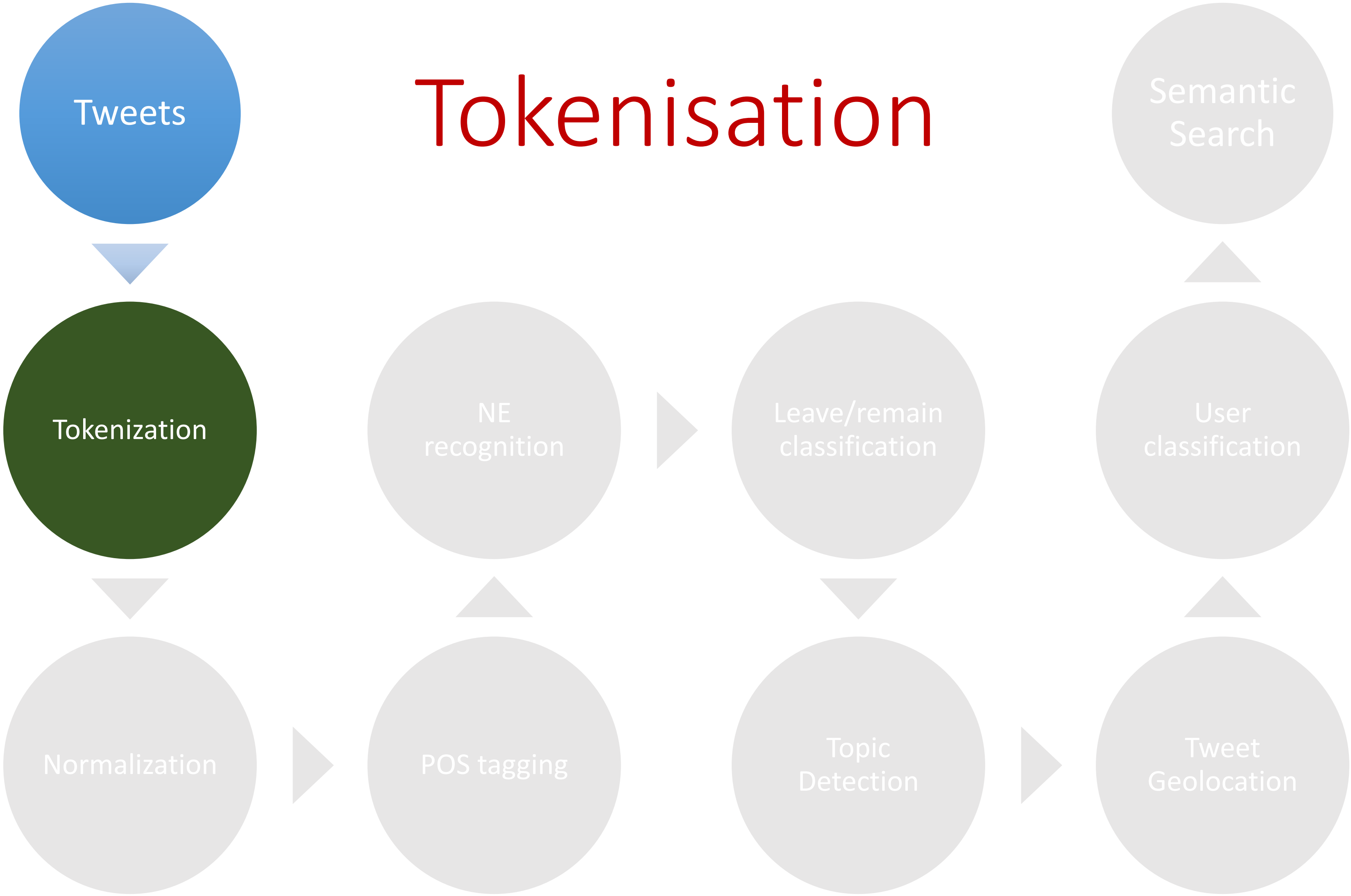
Tweet Geolocation

User classification

Semantic Search

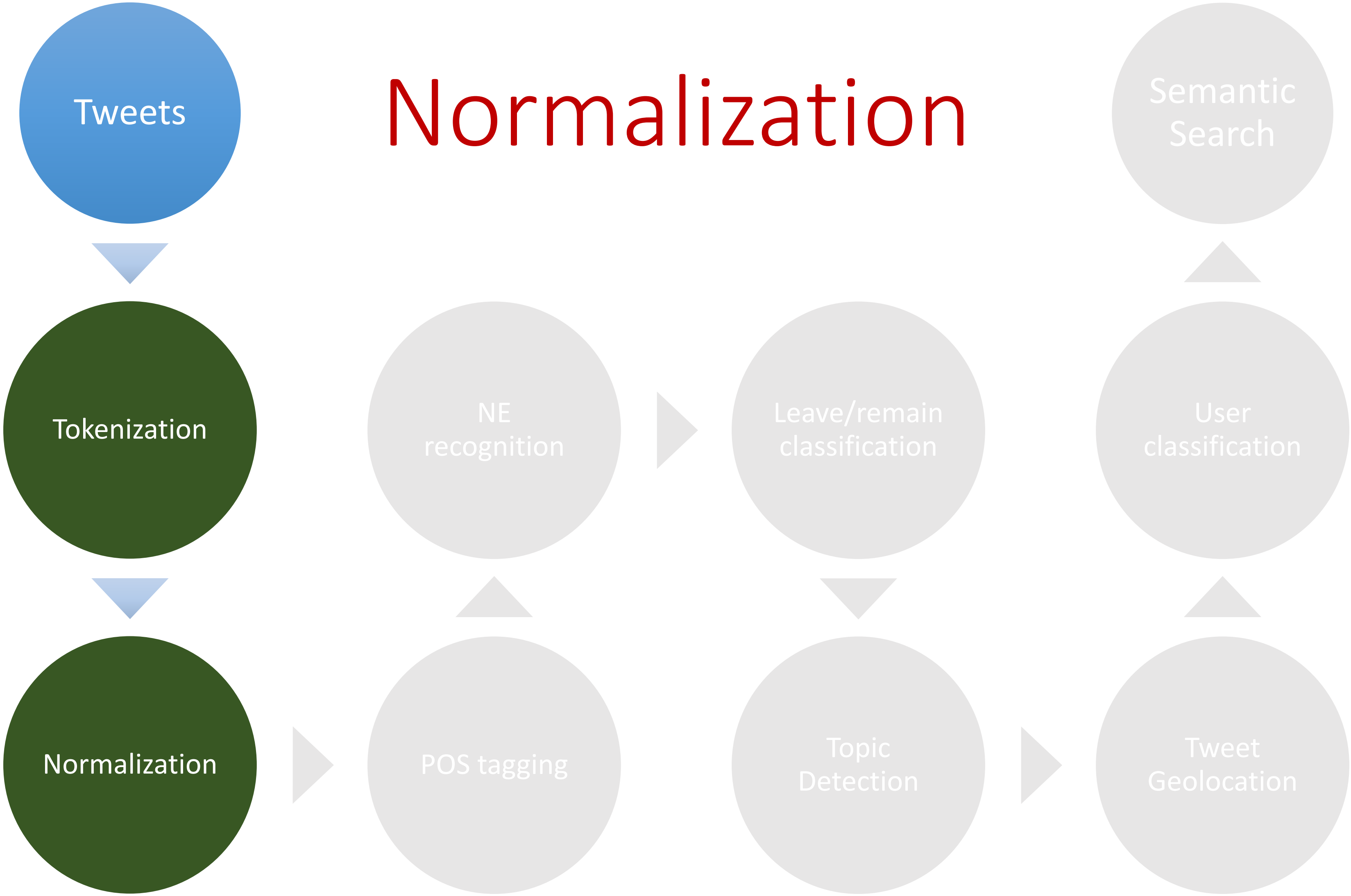
Tweets are tracked in real time using the streaming API

Tokenisation



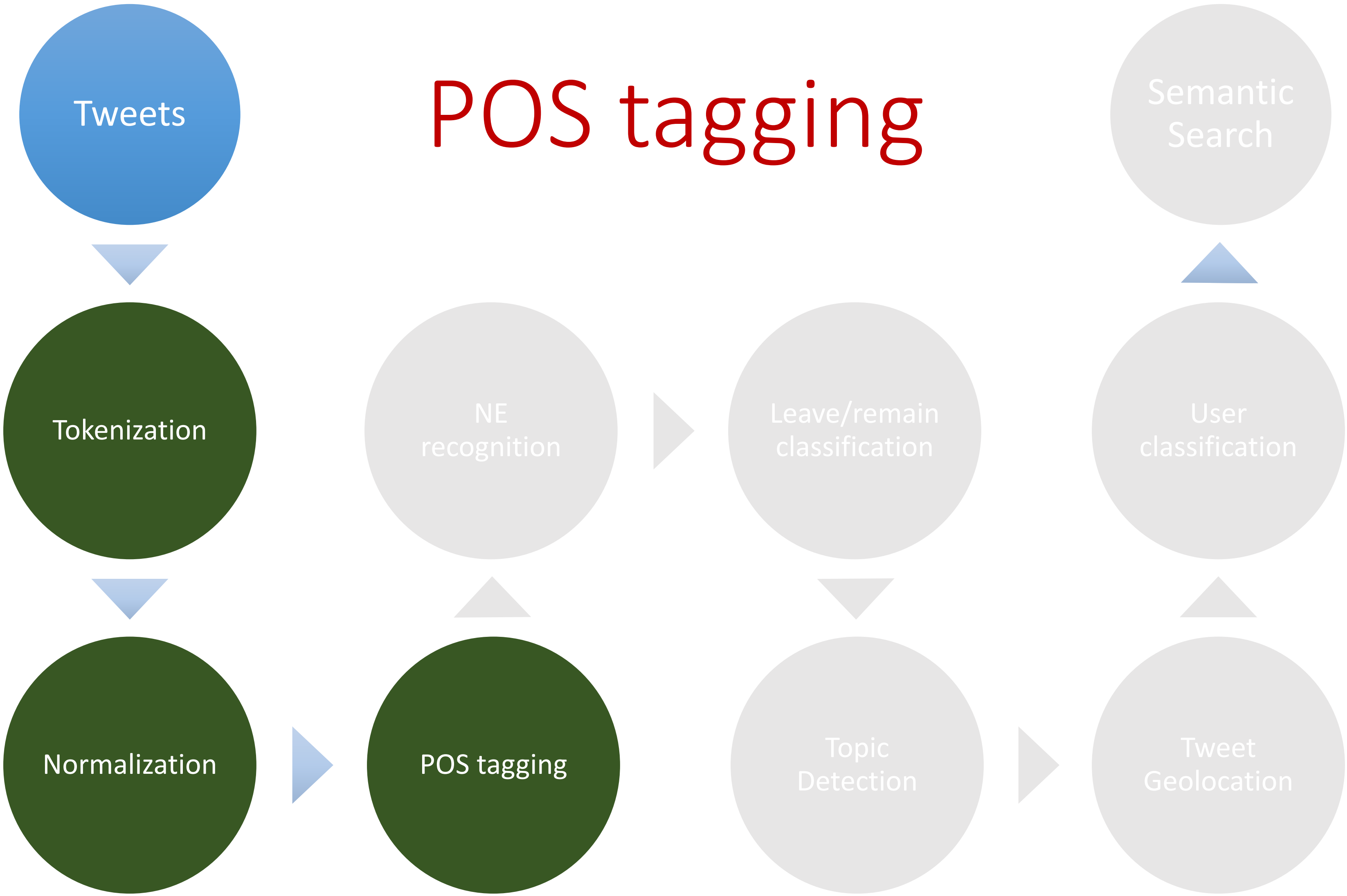
Individual tokens (analogous to terms) are extracted

Normalization



Spelling and abbreviations are normalised to help linguistic processing tools

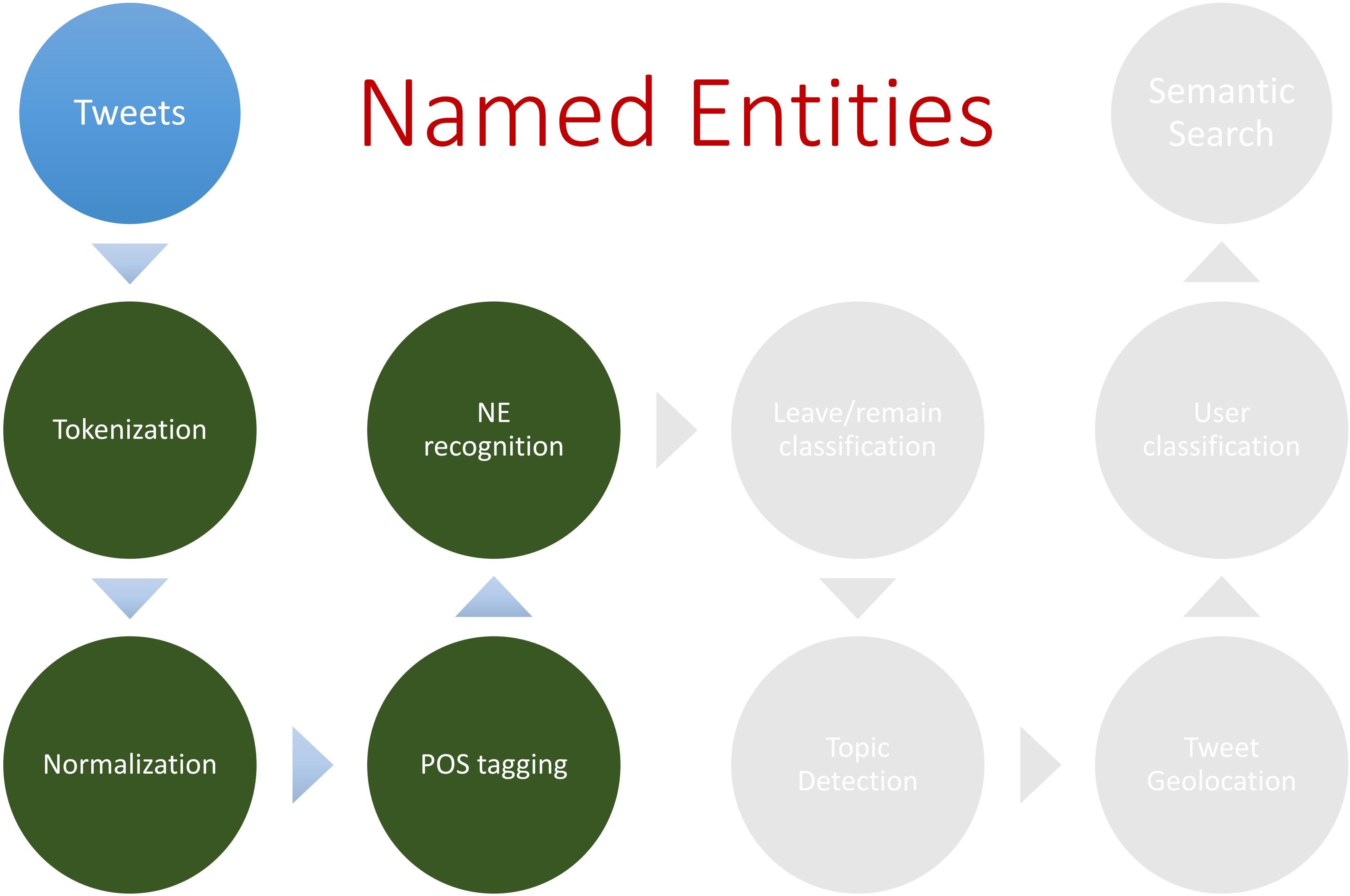
POS tagging



Parts-of-speech are identified.

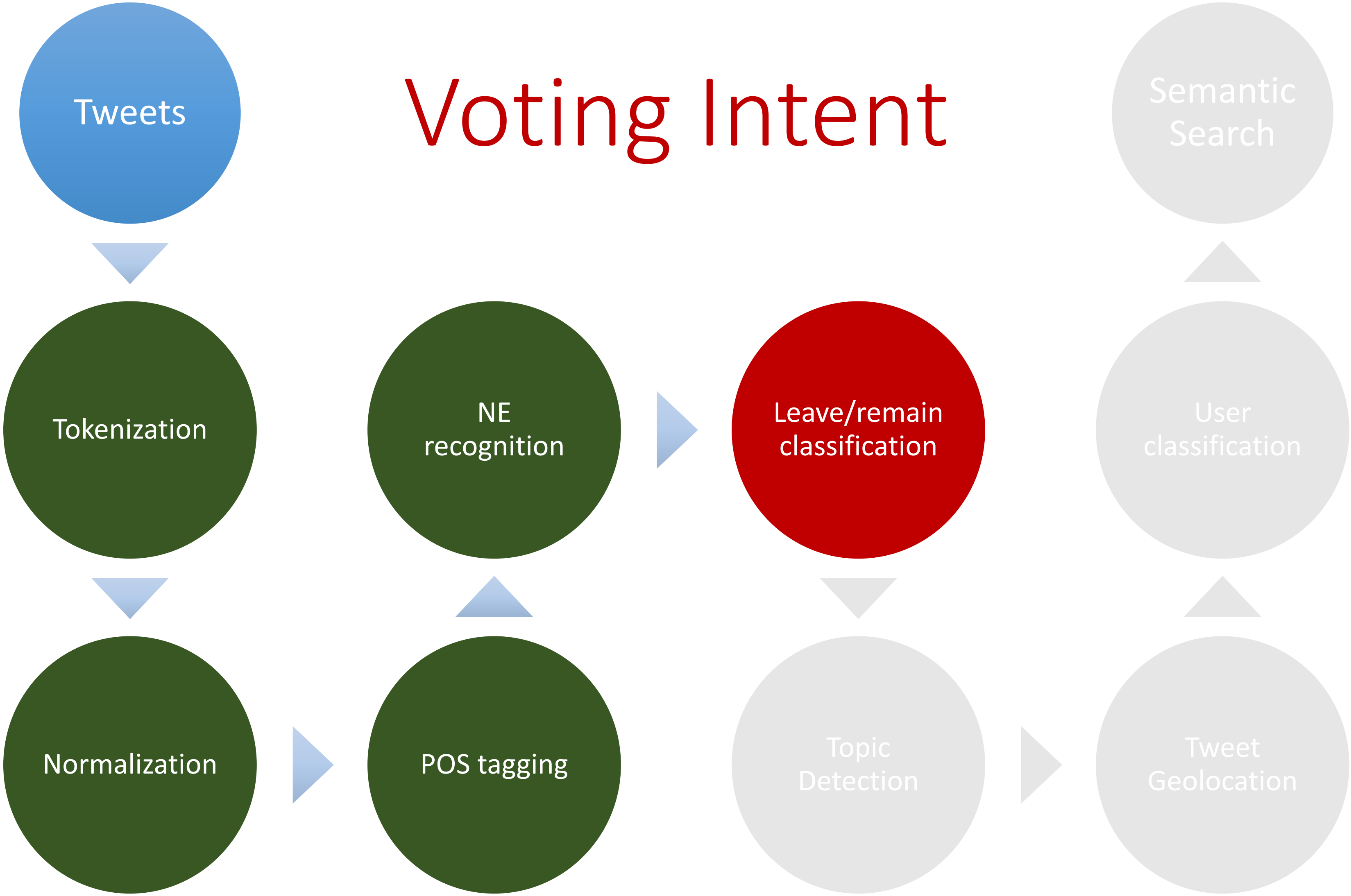
This is necessary to support later processing

Named Entities



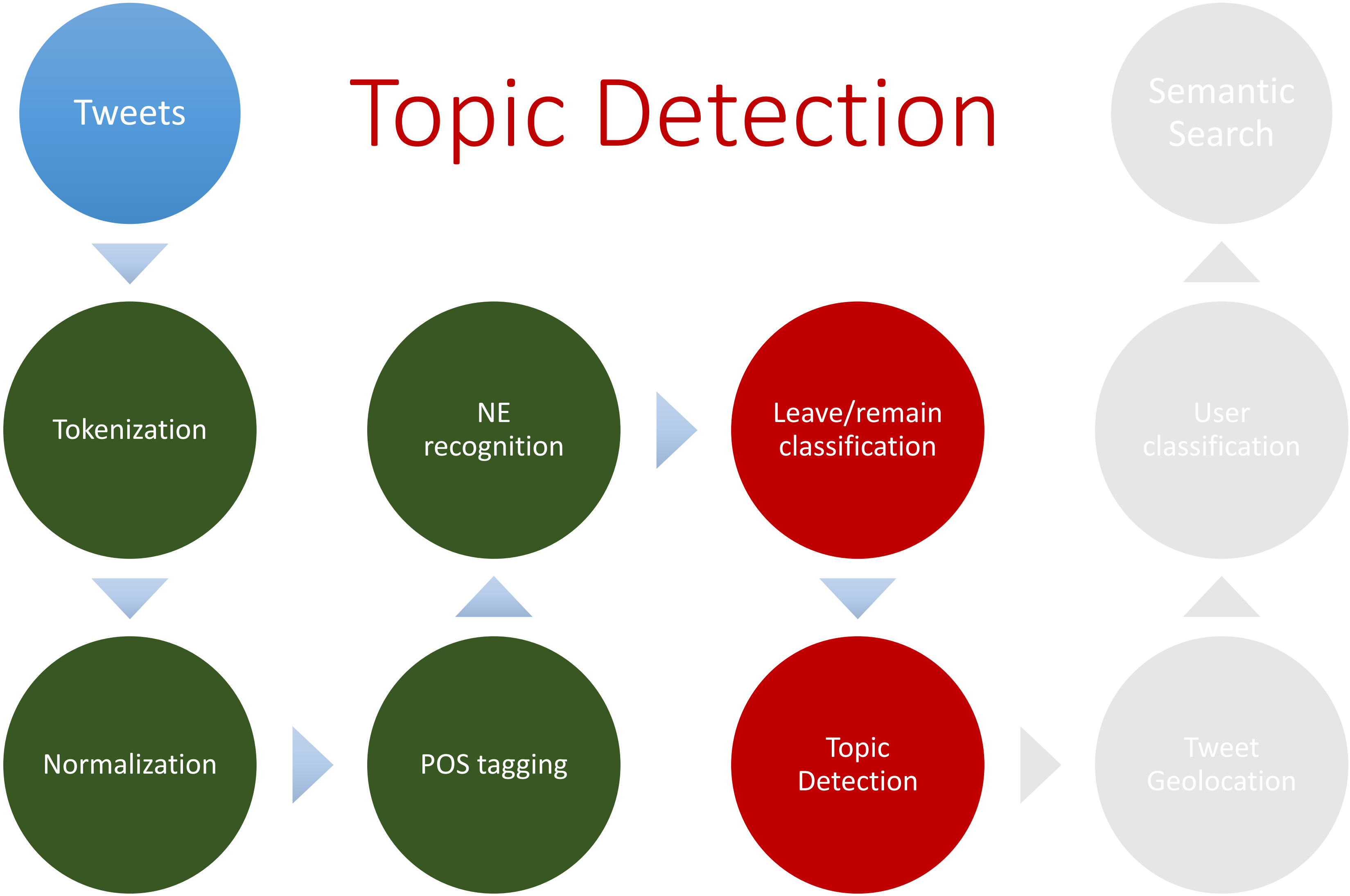
We discover mentions of entities such as people, locations, organisations and products

Voting Intent



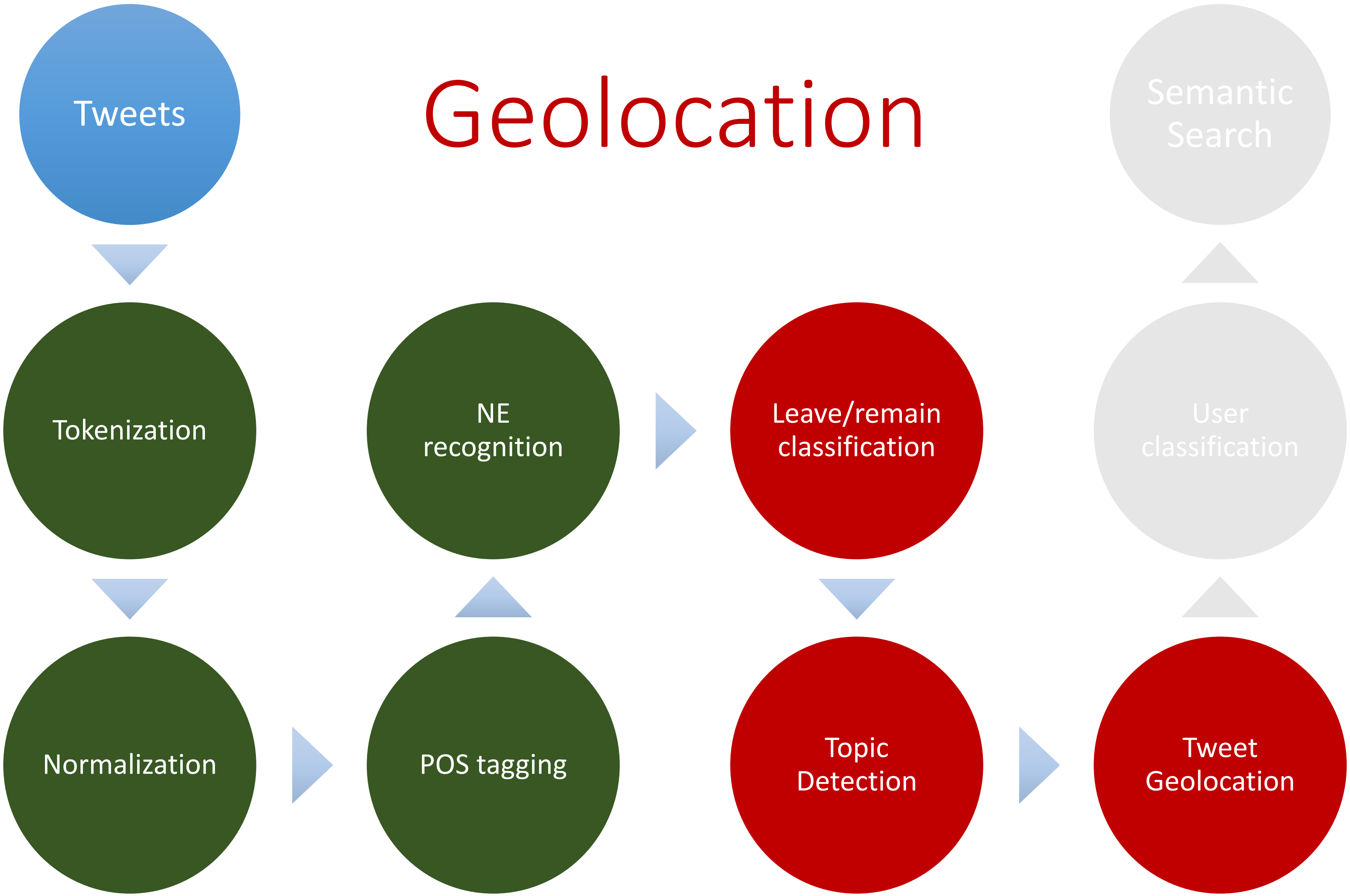
Intent to vote leave or remain classified by #hashtag

Topic Detection



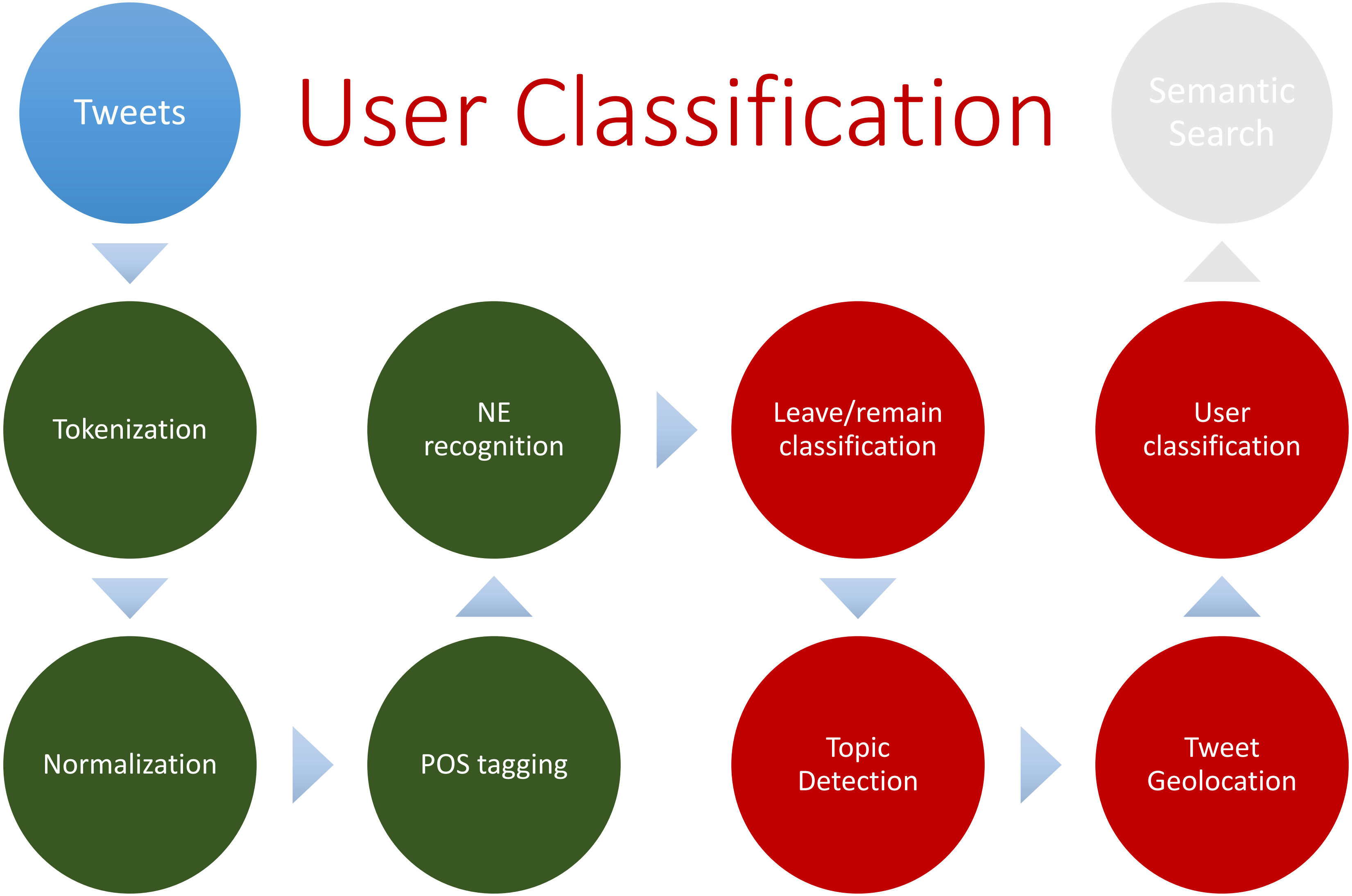
Tweets are matched against a detailed list of topics

Geolocation



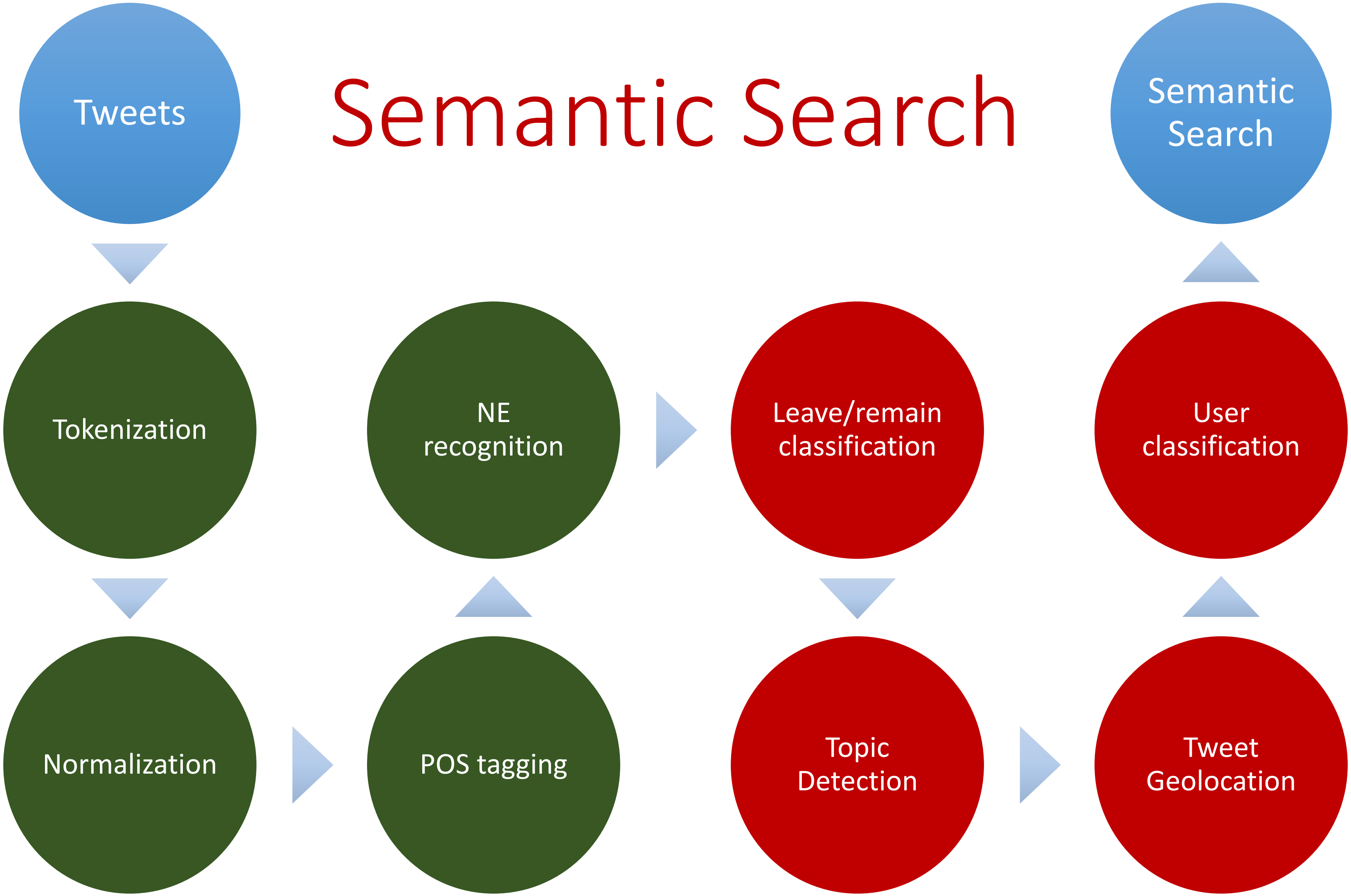
Tweets are linked to NUTS regions based on place tags and user home locations

User Classification



We classify users into eg journalist, charity, member of public

Semantic Search



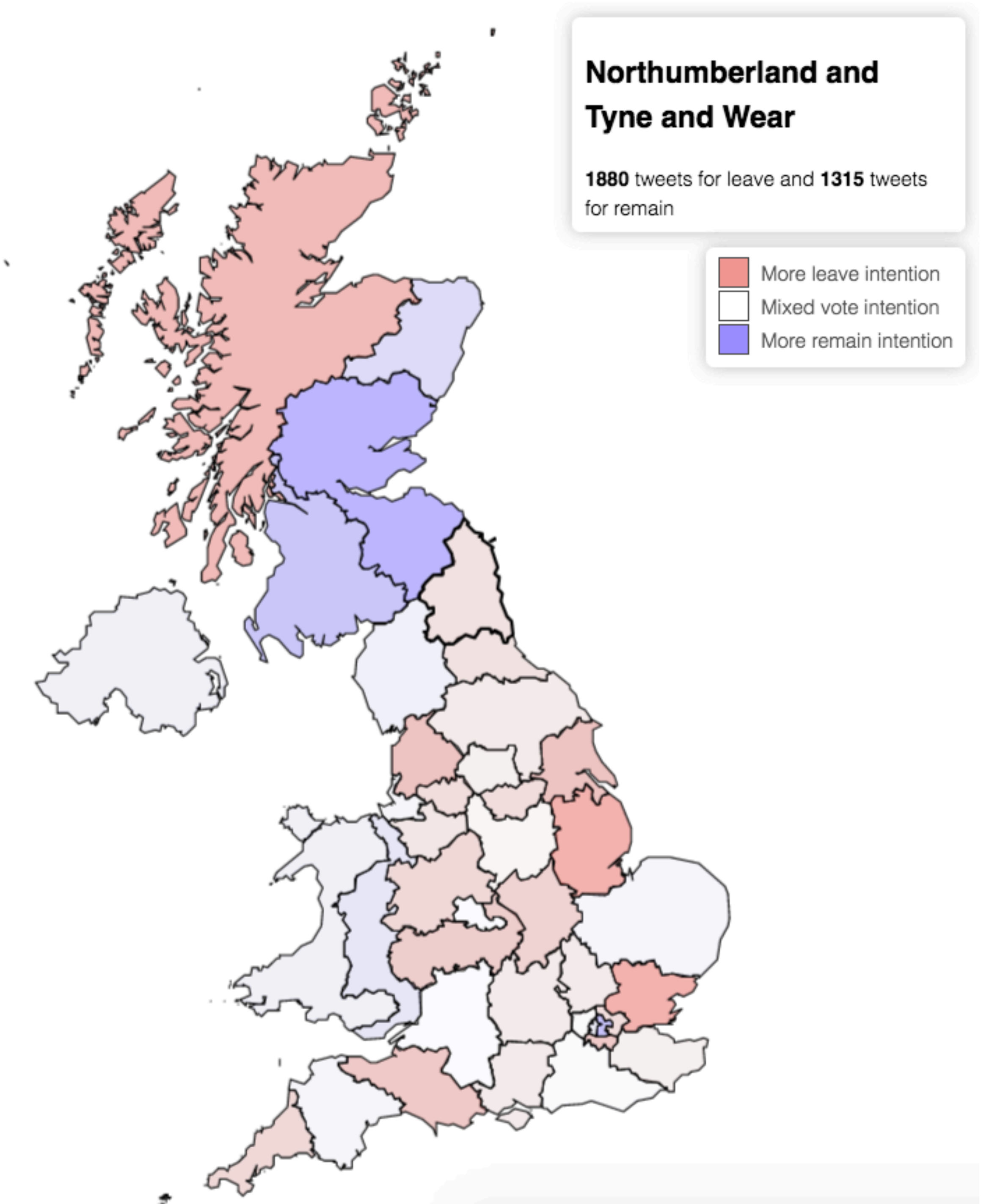
Tweet text and annotations are indexed in semantic search engine Mímir for search and visualisation

Semantic search with Mímir

- Mímir: Multiparadigm Indexing and Retrieval
- Complex queries – can search over annotations like

```
{DocumentTimestamp hour_timestamp >= 2016062223 hour_timestamp <
2016062323} OVER ( {DocumentKind tweet_kind = original} AND {NUTS2
NUTS2 = "UKE3"})
```

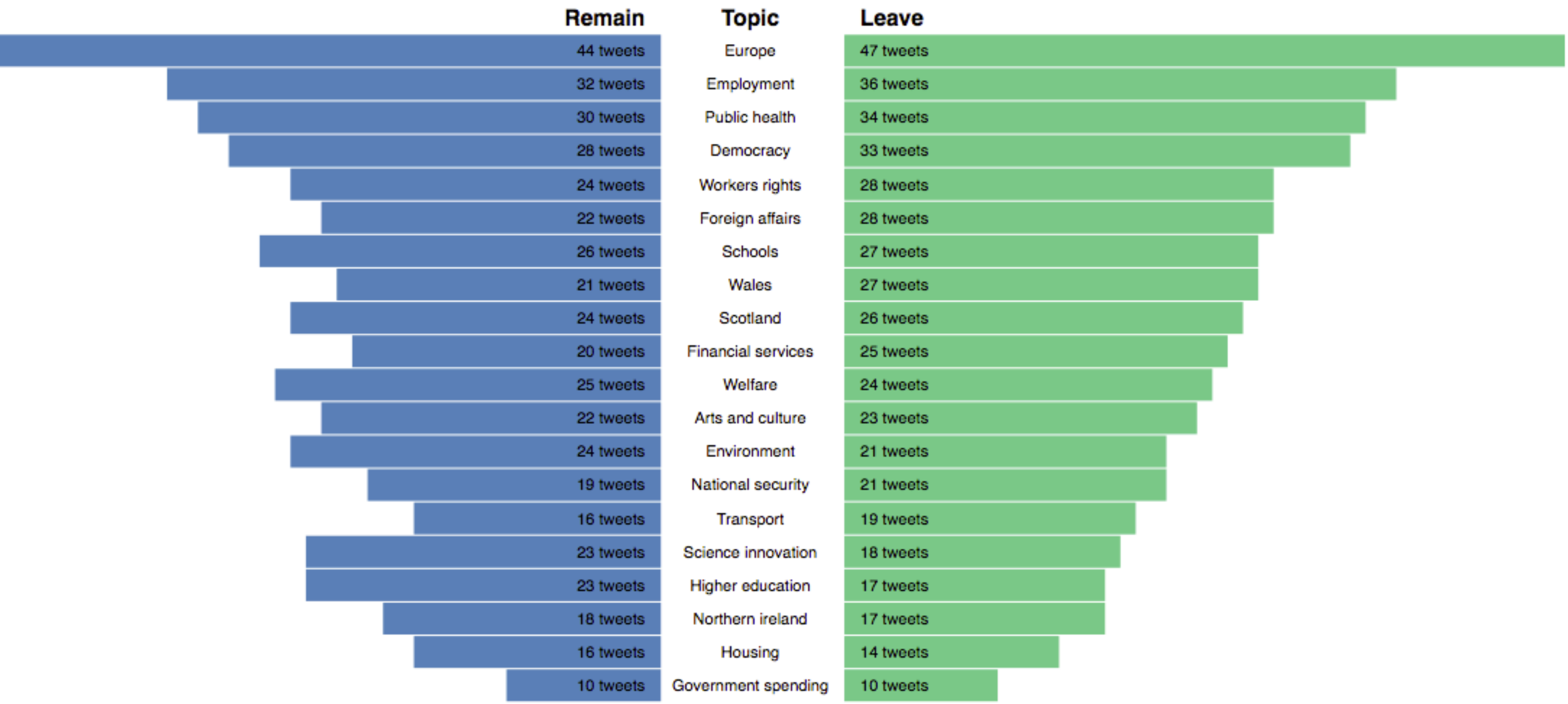
- Can also mix in full text and semantic queries. Very powerful!
- Which politicians from the North of England over the age of 40 talked most positively about climate change?
- Allows us to drill down and easily see tweets with certain properties



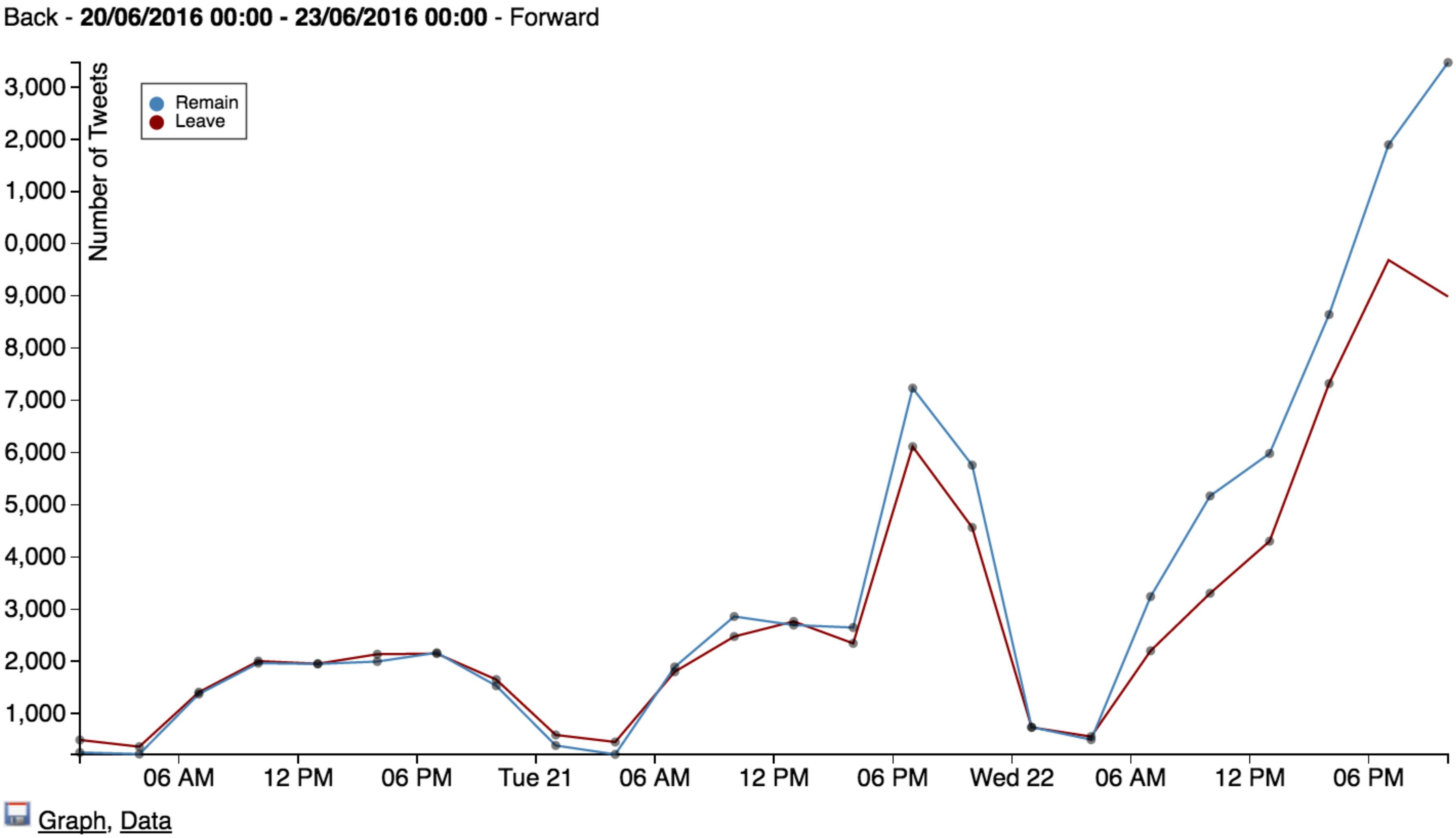
Northumberland and Tyne and Wear

1880 tweets for leave and **1315** tweets for remain

Voting intention per topic



Voting Intention over time



Back - **20/06/2016 00:00 - 23/06/2016 00:00** - Forward

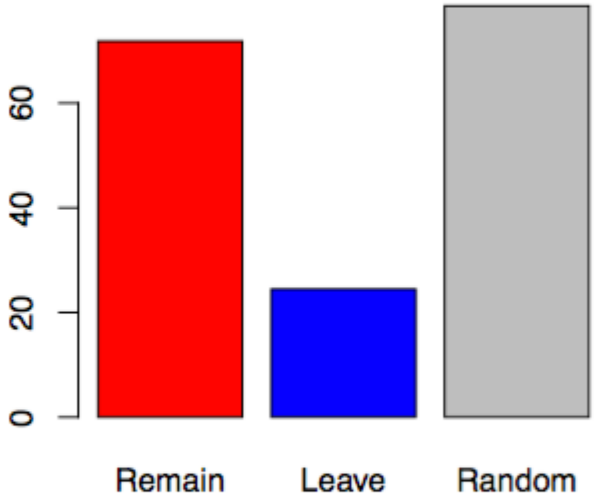
Graph, Data

Ego-network analysis

- Analysis using previous work on selected Brexit data
    - Valerio Arnaboldi
    - Institute of Informatics and Telematics (IIT)
    - Italian National Research Council (CNR)
- Understanding the quantity and quality of relationships of debate participants
- Remain, leave and neutral users were selected by Sheffield using their pipeline and processed by CNR-IIT
- Nice example of bringing several systems together

Remainers are more social

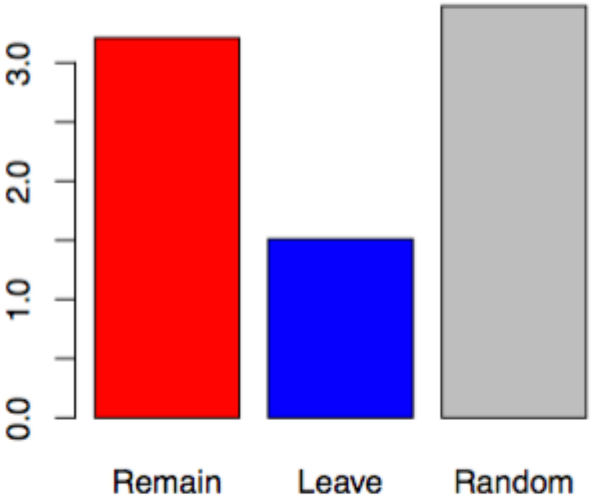
- They maintain larger ego networks
- Larger active network size
  - number of people actively contacted by the egos
  - with a frequency of at least one direct tweet per year
  - considering mentions, replies, and retweets
- Effect remains even after filtering out antisocial accounts



Active network size in users

Leavers had fewer social circles

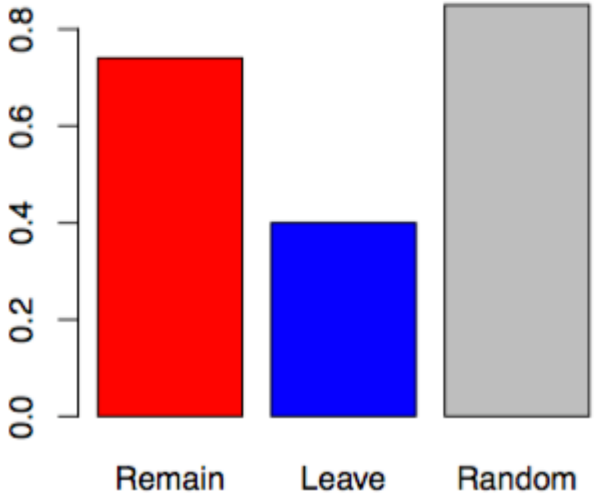
- Social circle:
  - Group of users in an ego-network
  - With similar levels of interaction to one another
  - Most frequent interactions with a smaller social circle
  - More occasional interactions with a larger circle



Average number of social circles

Many “leave” accounts didn’t socialise at all

- A lot of accounts were not used socially
- This is reflected in the number of accounts with social circles
- The leave twitterers were well below the random sample.

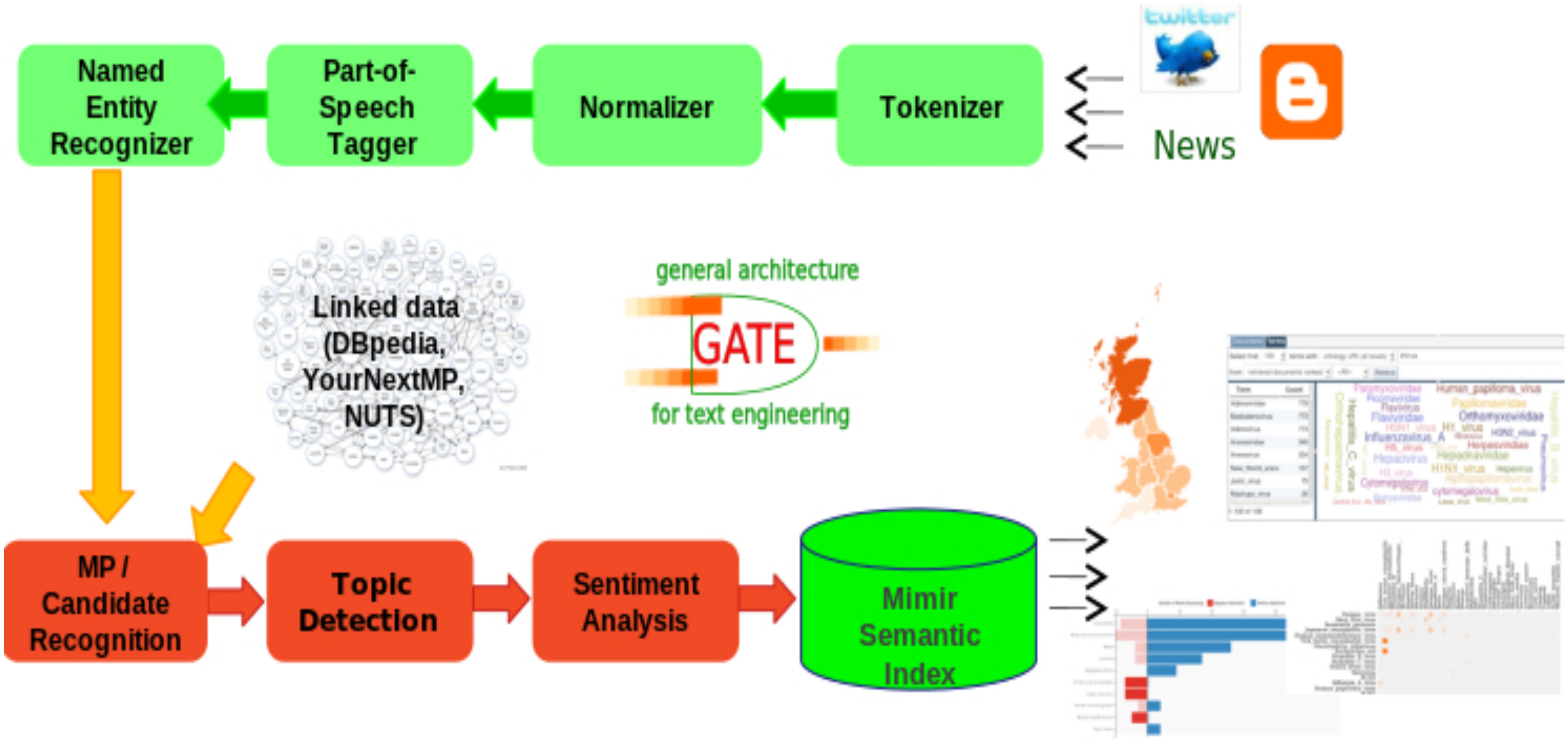


% of accounts with at least 1 social circle

Brexit Demo

- http://demos.gate.ac.uk/sobigdata/brexit/
- More cool visualisations (work in progress)

GATE social media analysis toolkit: analysing the UK elections



News

Tokenizer

Normalizer

Part-of-Speech Tagger

Named Entity Recognizer

Linked data (DBpedia, YourNextMP, NUTS)

general architecture GATE for text engineering

MP / Candidate Recognition

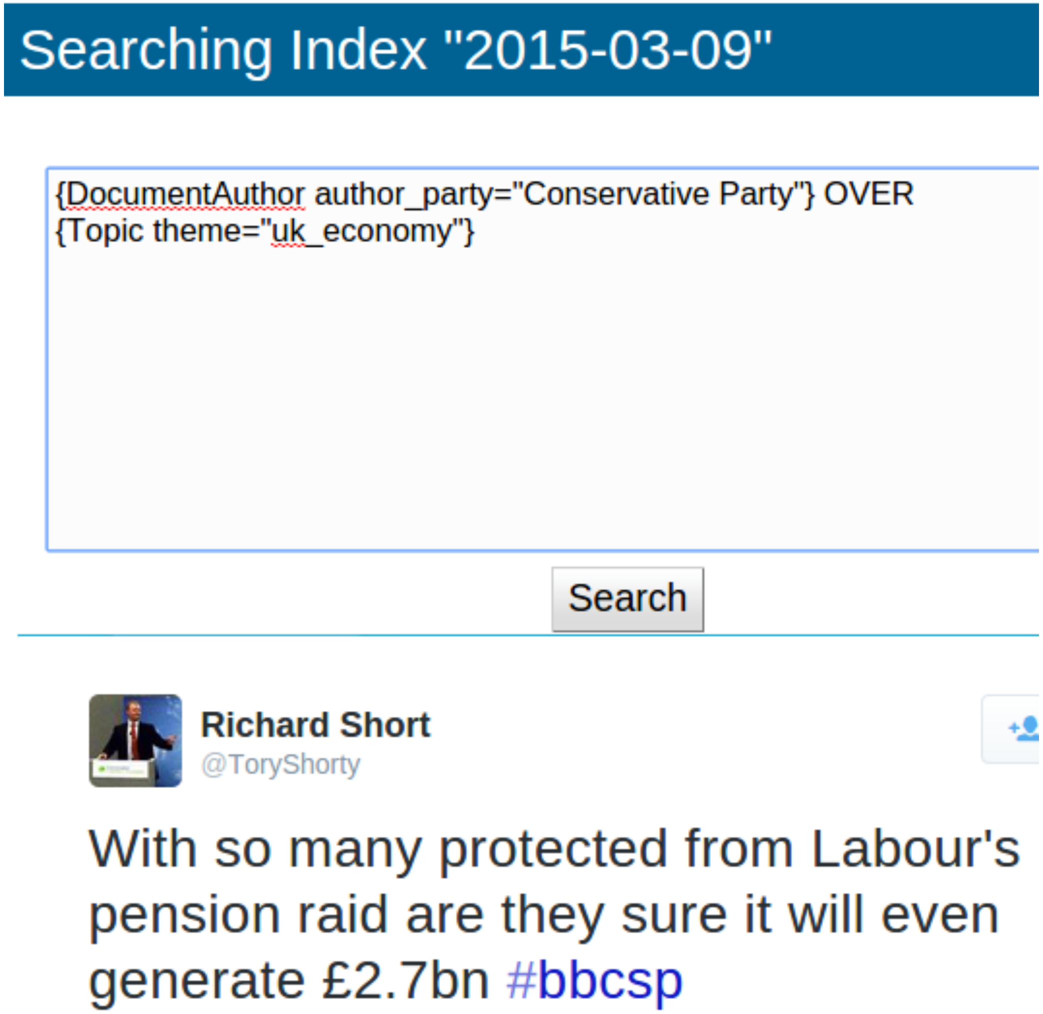
Topic Detection

Sentiment Analysis

Mimir Semantic Index

MIMIR query

- Dataset: every tweet by MP / Candidate / Party, plus all replies/retweets
- Find all tweets where a Conservative MP talked about the economy



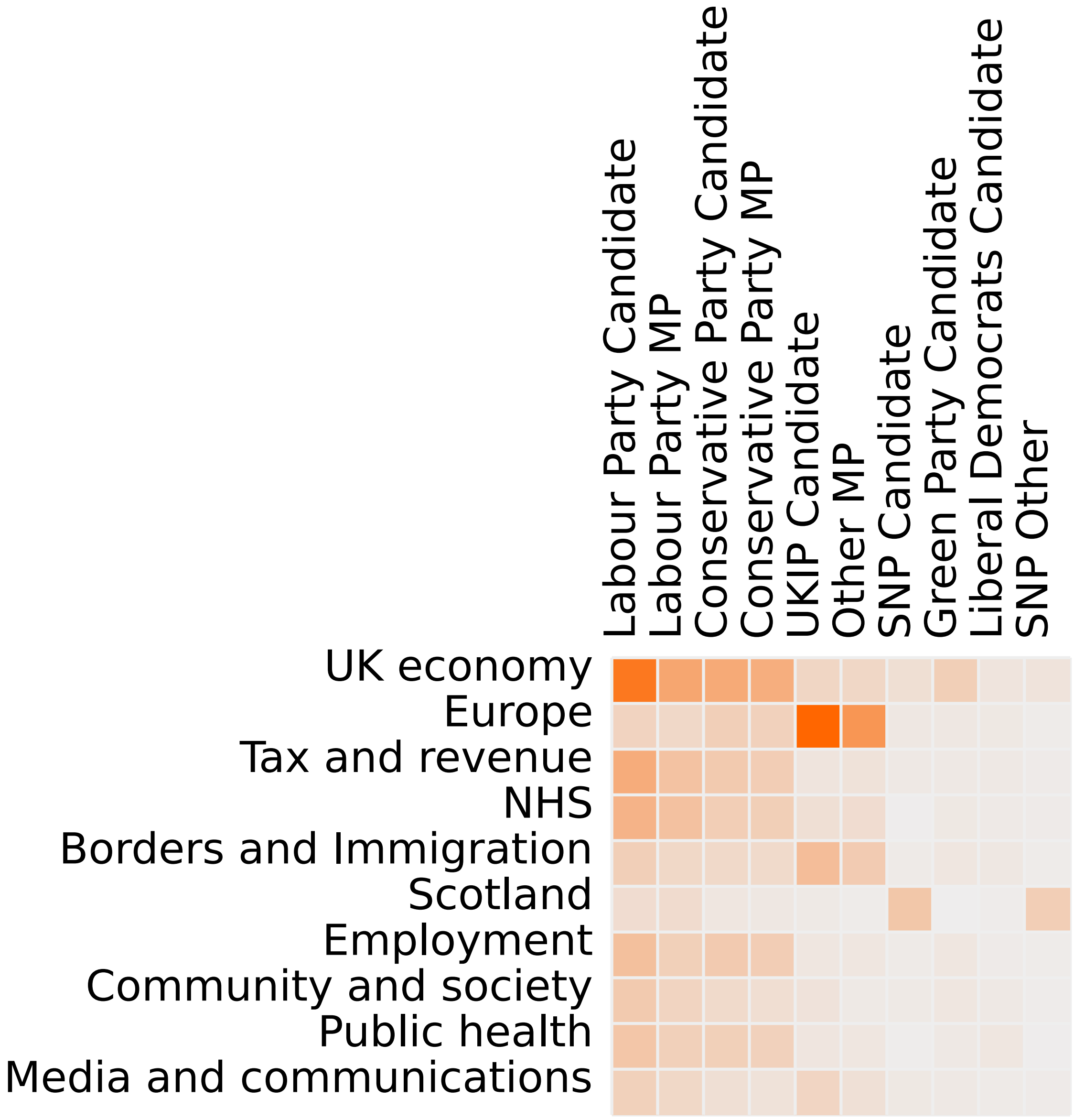
Searching Index "2015-03-09"

```
{DocumentAuthor author_party="Conservative Party"} OVER
{Topic theme="uk_economy"}
```

Richard Short
@ToryShorty

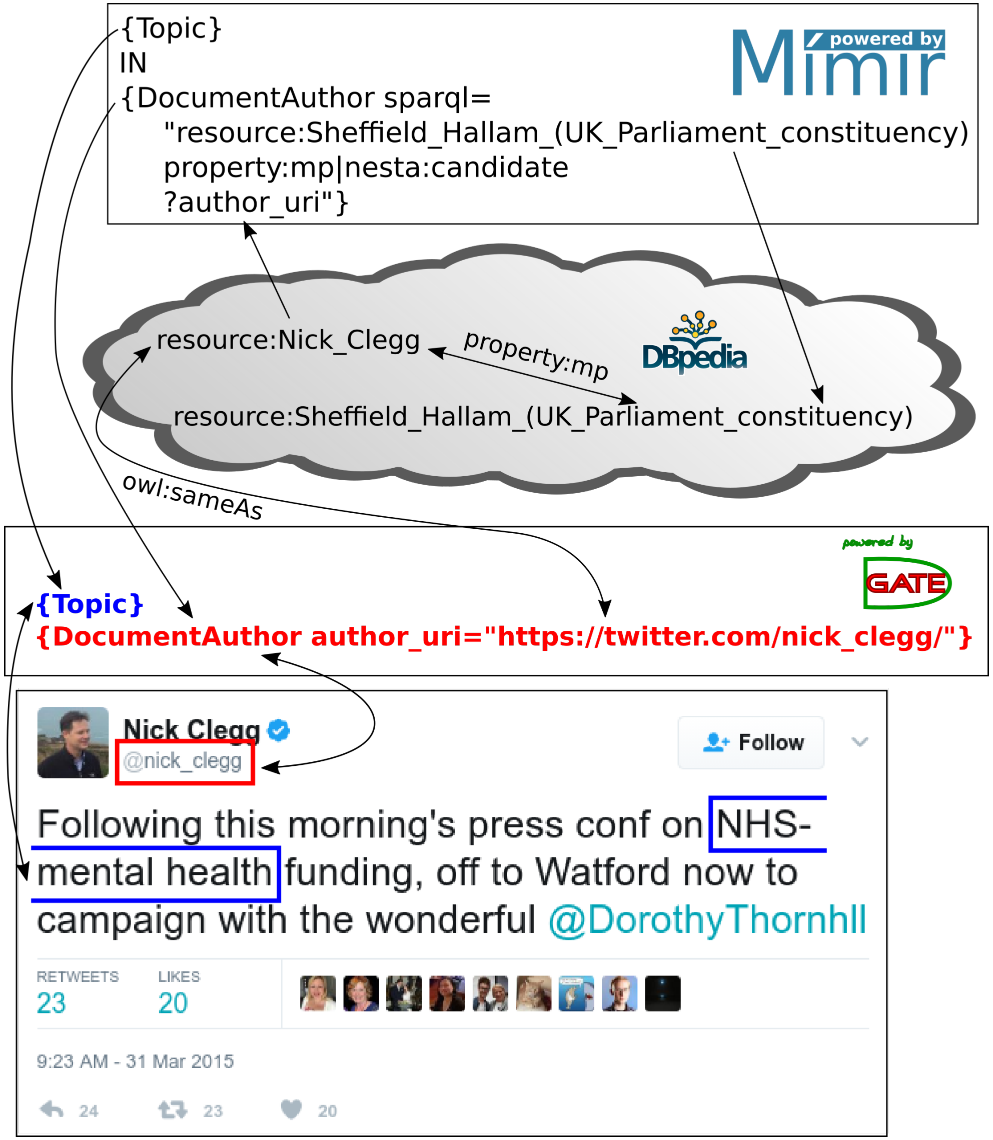
With so many protected from Labour's pension raid are they sure it will even generate £2.7bn #bbcsp

Parties / themes co-occurrence



Semantic querying via DBpedia

- Querying for information not explicit in the text
- “Find all topics in tweets by the MP from Sheffield Hallam”
- We don’t need to know who this MP is
- If the Sheffield Hallam MP changed at some point, this would be captured



```
{Topic}
IN
{DocumentAuthor sparql=
    "resource:Sheffield_Hallam_(UK_Parliament_constituency)
    property:mp|nesta:candidate
    ?author_uri"}
```

{Topic}

{DocumentAuthor author_uri="https://twitter.com/nick_clegg/"}

Nick Clegg
@nick_clegg

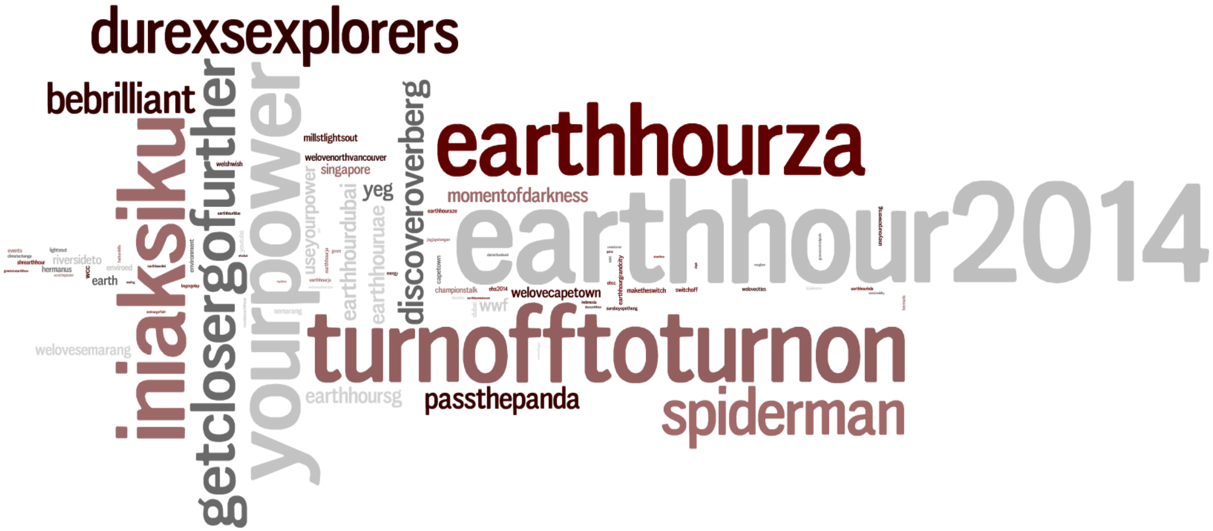
Following this morning's press conf on NHS-mental health funding, off to Watford now to campaign with the wonderful @DorothyThornhll

RETWEETS 23
LIKES 20

9:23 AM - 31 Mar 2015

Analysis of the EarthHour campaign

- Analysis of hashtags and topics mentioned
- The main activities and themes of the campaign drove most of the social media conversations
- Users engaged in the campaign but did not necessarily engage with climate change and sustainability issues.
- Lack of correlation between Durex campaign and climate change engagement



Behaviour Analysis

- Based on the assumption that users in different behavioural stages communicate differently (different emotions, directives, etc.)

Pajarito @lindopajarito . 2h

Our building needs 40% of all energy consumed in Switzerland! ☹

Desirability: Negative sentiment (expressing personal frustration-anger/sadness)

DJPajarito @DJPajaritoGenial . 12h

I'm so proud when I remember to save energy and I know however small it's helping.

Buzz: Positive sentiment (happiness/joy). I/we + present tense

You are an organisation, not an individual!

HotelPajarito @HotelPajarito . 18h

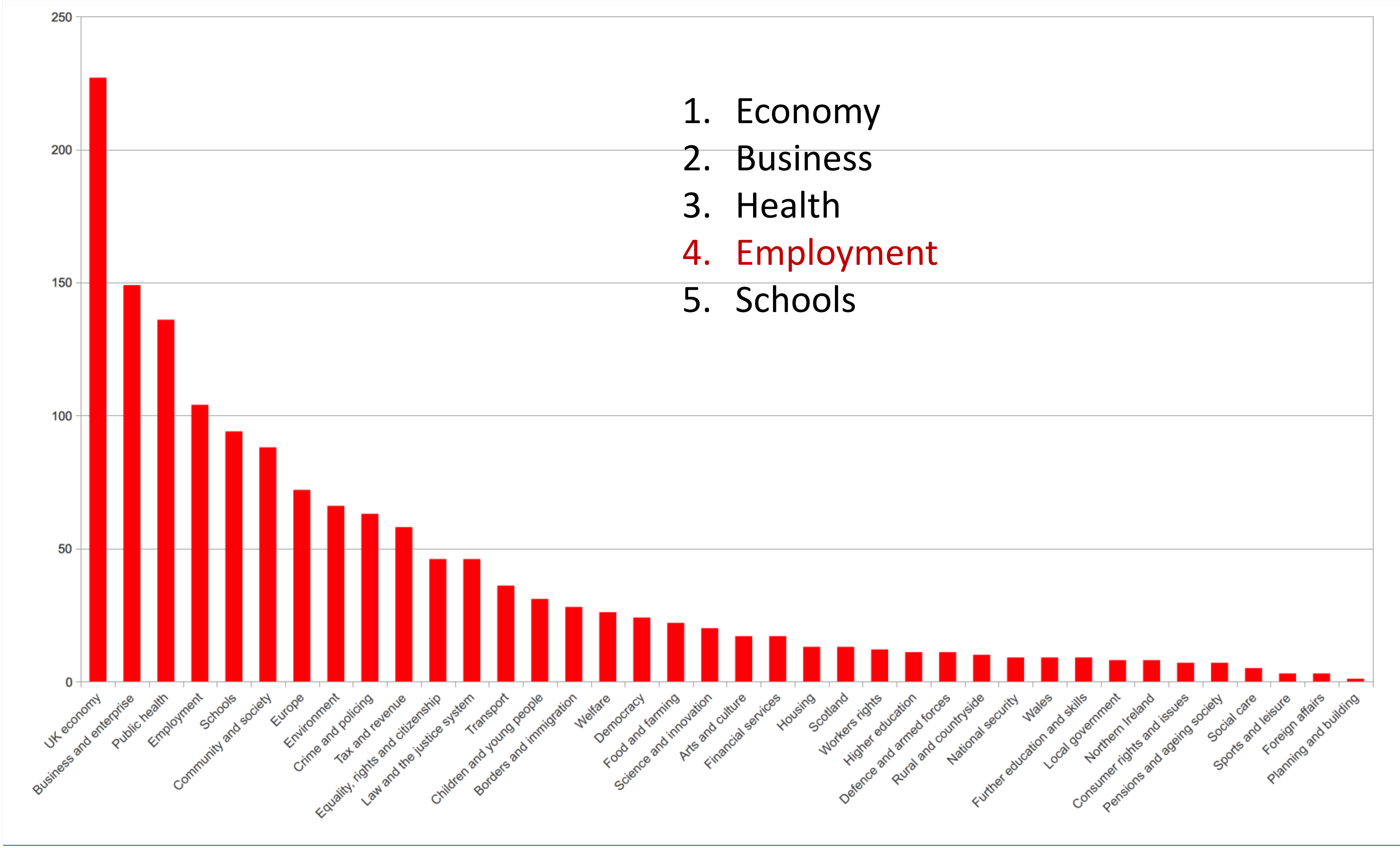
Join us today today to switch of a light for EH! ☺

Invitation: Positive sentiment (happy) + use of vocatives

What’s next?

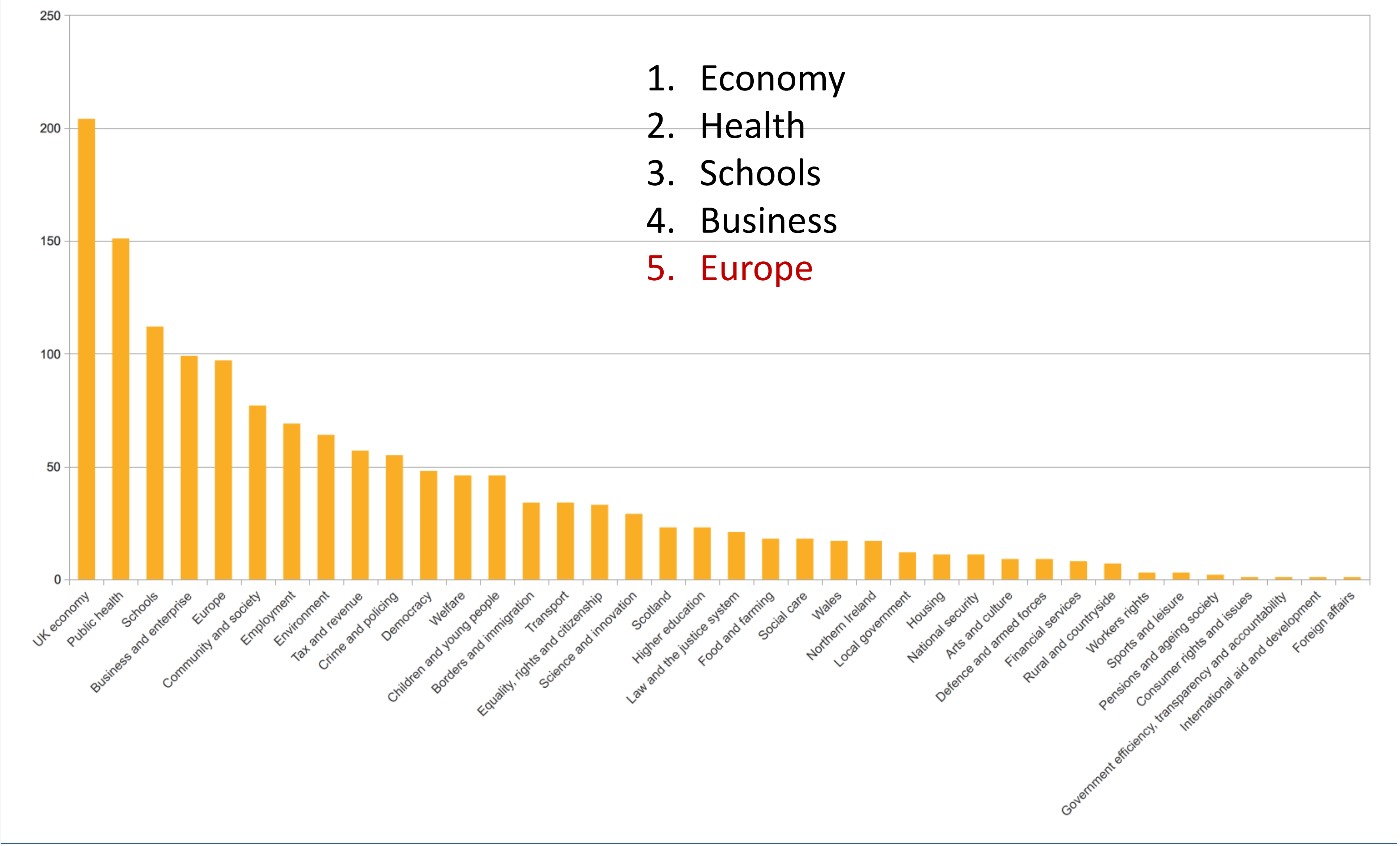
- A number of tools have already been brought together around Brexit and other debates
- More tools still being developed
- Tools available through the SoBigData VRE and GATECloud
- Focus on analysis rather than predicting results
- 2017 UK elections: imminent results coming this week on party manifestos!
- Other current work analysing hate speech around Brexit, and correlating social media and news media discussions

What are Labour talking about?



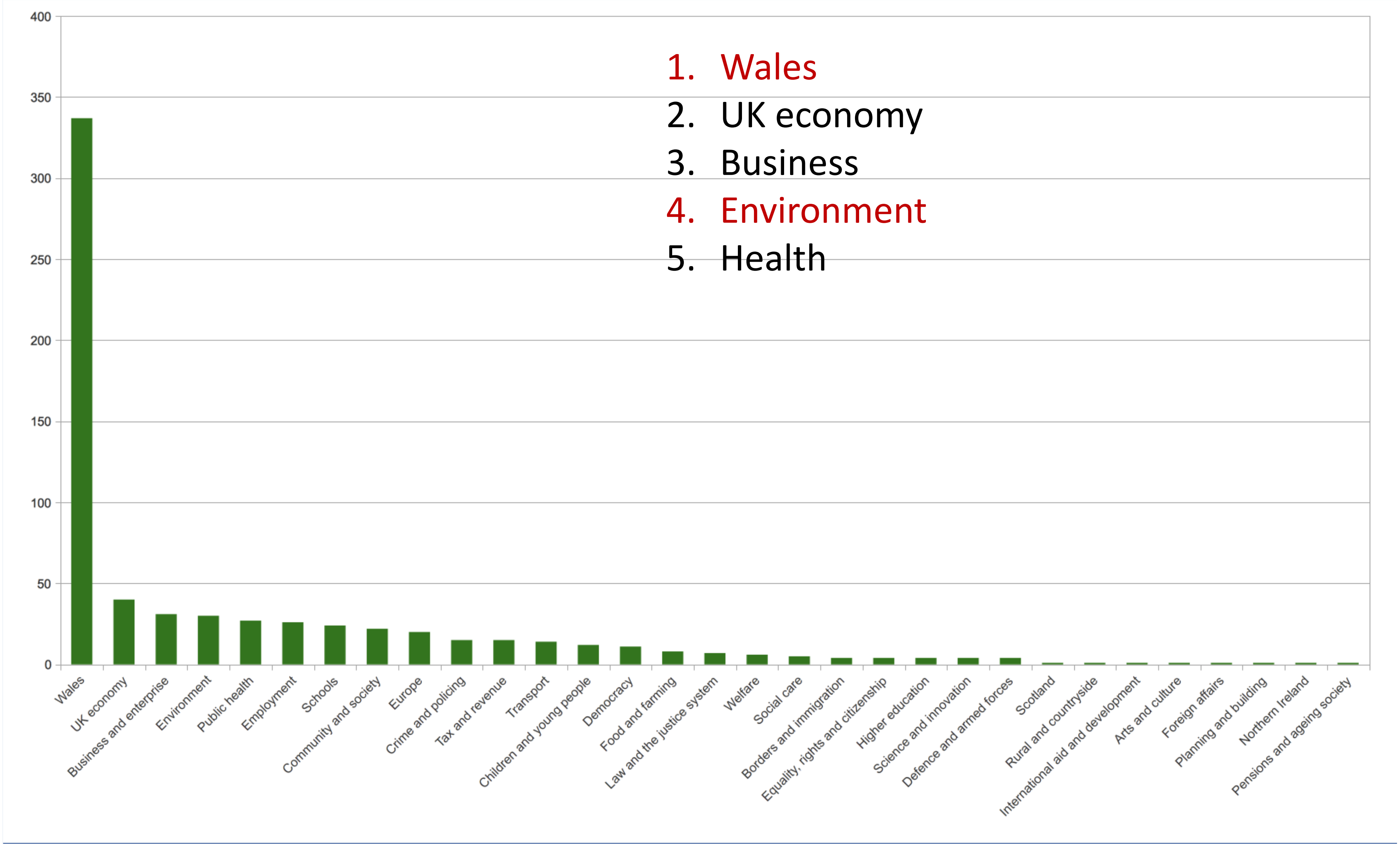
1. Economy
2. Business
3. Health
4. Employment
5. Schools

What are Lib Dems talking about?



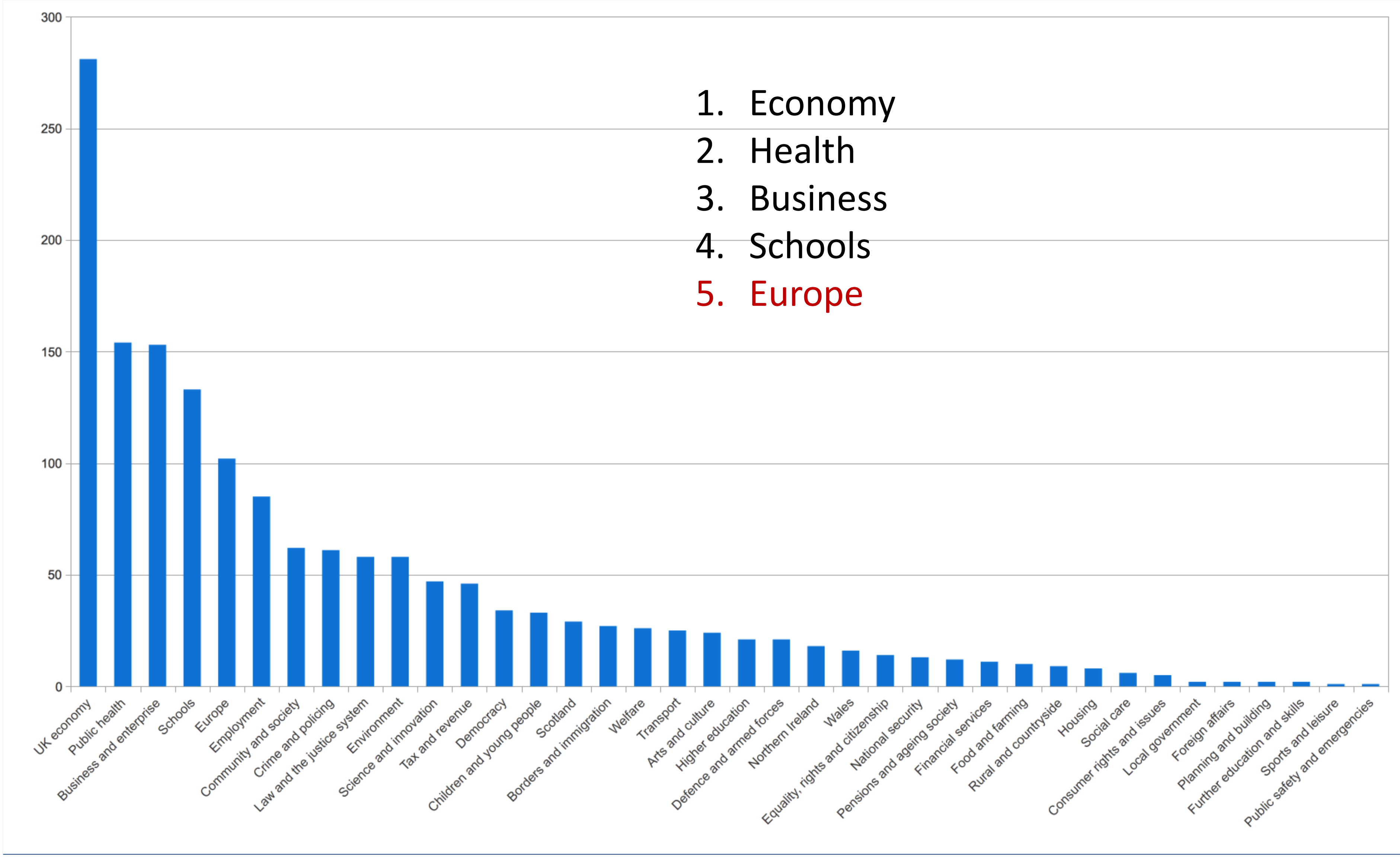
1. Economy
2. Health
3. Schools
4. Business
5. Europe

What are Plaid Cymru talking about?



1. Wales
2. UK economy
3. Business
4. Environment
5. Health

\what are Conservatives talking about?



1. Economy
2. Health
3. Business
4. Schools
5. Europe

Links for more info

- GATE at http://gate.ac.uk
- GateCloud at https://cloud.gate.ac.uk
- Come on a GATE training course in June!
- Brexit Visualisations at http://demos.gate.ac.uk/sobigdata/brexit/
- Brexit study blog post from NESTA at http://www.nesta.org.uk/blog/network-analysis-top-eu-referendum-tweeters
- Brexit study blog posts from Sheffield at http://gate4ugc.blogspot.co.uk/search/label/Brexit
- UK elections monitor at http://gate.ac.uk/projects/pft

Publications

- Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout and Kalina Bontcheva. A Framework for Real-time Semantic Social Media Analysis. Web Semantics: Science, Services and Agents on the World Wide Web, 2017.
- K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013).
- D. Maynard and K. Bontcheva. Understanding climate change tweets: an open source toolkit for social media analysis. In Proc. of EnviroInfo 2015, Copenhagen, Sep. 2015
- Diana Maynard, Kalina Bontcheva, Isabelle Augenstein. Natural Language Processing for the Semantic Web. Morgan and Claypool, December 2016. ISBN: 9781627059091 (contains a chapter on social media analysis)
- Available (and more) at https://gate.ac.uk/gate/doc/papers.html

Acknowledgements

This work supported by:

- the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework and H2020 Programmes for R&D
  - KNOWMAK (726992) https://gate.ac.uk/projects/knowmak/
  - DecarboNet (610829) http://www.decarbonet.eu
  - Pheme (611233) http://www.pheme.eu
  - SoBigData (654024) http://www.sobigdata.eu
  - COMRADES (687847) http://www.comrades-project.eu
  - OpenMinted (654021) http://openminted.eu
  - Kconnect (644753) http://www.kconnect.eu/
- Nesta http://nesta.org.uk