Case study:
(Almost) Real-Time
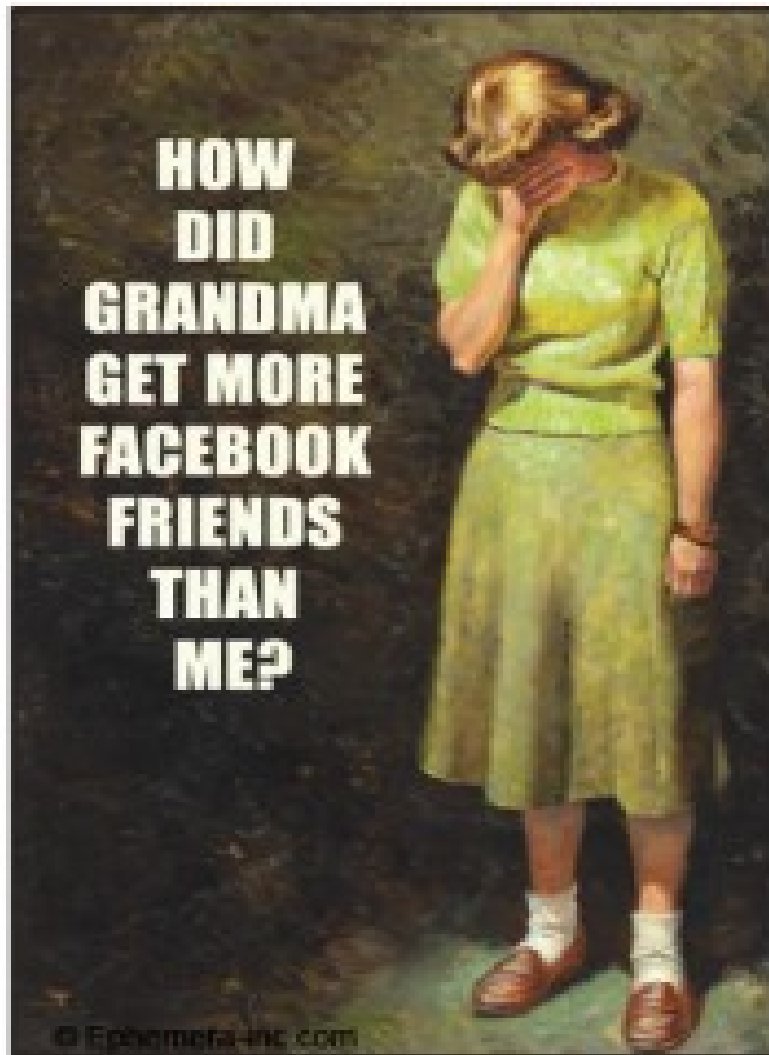# Social Media Analysis of Political Tweets

We are all connected to each other...

- Information, thoughts and opinions are shared prolifically on the social web these days

- 72% of online adults use social networking sites

Your grandmother is three times as likely to use a social networking site now as in 2009

There are hundreds of tools for social media analytics

- Most of them are commercial and not freely available

- The research tools tend to focus on specific topics and scenarios, and aren't easily adaptable

- The analysis they do often doesn't go much beyond number crunching, e.g.

  - look at number of tweets, retweets, favourites

  - filter by hashtag or keyword for topic categorisation

  - use off-the-shelf sentiment tools

  - use counts of word length, POS categories etc

  - very little semantics, don't deal with variation, ambiguity, slang, sarcasm etc.

Analysing Social Media is harder than it sounds

There are lots of things to think about!

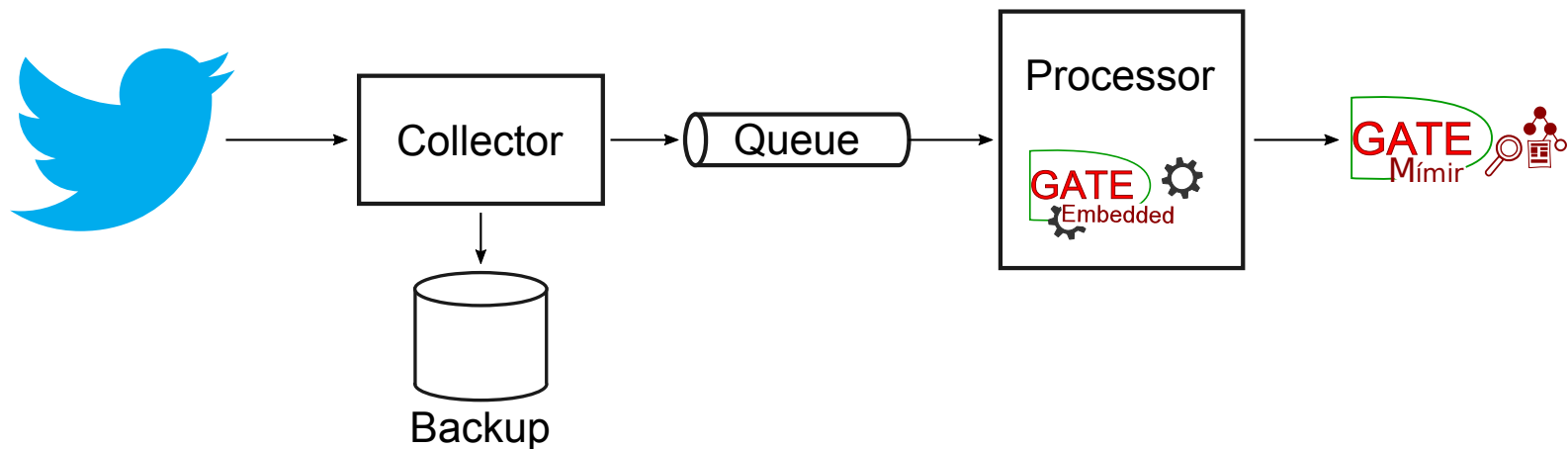# Let's search for keywords like "Arctic"

Oops!

Seems like we need something to help!

Framework

- Data collection (via Twitter streaming API)

- Documents stored as JSON and processed (annotated) via GCP (Gate Cloud Paralleliser)

- Documents indexed via MIMIR

- Search and visualisation via MIMIR/Prospector

Live streaming

- If we want to process the tweets in real time, we can use the Twitter streaming client to feed the incoming tweets to a message queue.

- A separate process then reads messages from the queue, annotates them and pushes them into Mimir.

- If the rate of incoming tweets exceeds the capacity of the processing side, we can simply launch more instances of the message consumer across different machines to scale the capacity.

- Query and visualisation can then be performed as before on whatever data we currently have available

PFT: the Political Futures Tracker Application

- Example of using the technology on a real scenario - analysing political tweets in the run-up to the UK elections

- Project funded by Nesta http://www.nesta.org.uk/

- Series of blog posts about the project, leading up to the election, see e.g.

  http://www.nesta.org.uk/news/political-futures-tracker

- Some more information from the technical side on the GATE website at

  https://gate.ac.uk/projects/pft/

The idea behind it

- Party campaigns for UK elections are now largely defined by their social media presence and the messages that they are pushing online

- This means we need to find innovative ways to examine these new and rich data sources.

- PFT is a tool providing near-real time analysis of political texts in the run up to the 2015 UK General election

- This includes both social media (Twitter) and party materials such as manifestos, political speeches and websites

- It aims to analyse key topics, future thinking and sentiment, as they unfold

First we had to collect the data

Twitter collection

- collected Tweets using Twitter's "statuses/filter" streaming API

- can follow up to 5000 user IDs and receive in real time

- collected all tweets and retweets posted by these users

- also retweets of, and replies to, any tweet posted by these users

- We also collected tweets specific to each of the TV debates, using some selected hashtags

Twitter collection (2)

- Initial list of 506 UK MPs' Twitter accounts, extracted from a CSV file made available by BBC News Labs and cleaned

- Also added list of UK election candidates collected and made available at https://yournextmp.com, and updated periodically

  – 2,991 by the time of the election (3630 candidates overall, but many without Twitter handles)

- Added 21 official party accounts

- Total number of accounts followed by the time of the election: 3,060

  – 461 MPs standing again included in both the MP and candidate lists

Tweets per hour collected

Douglas Carswell's accidental "Hello Kitty" tweet

Government U-turn on fracking
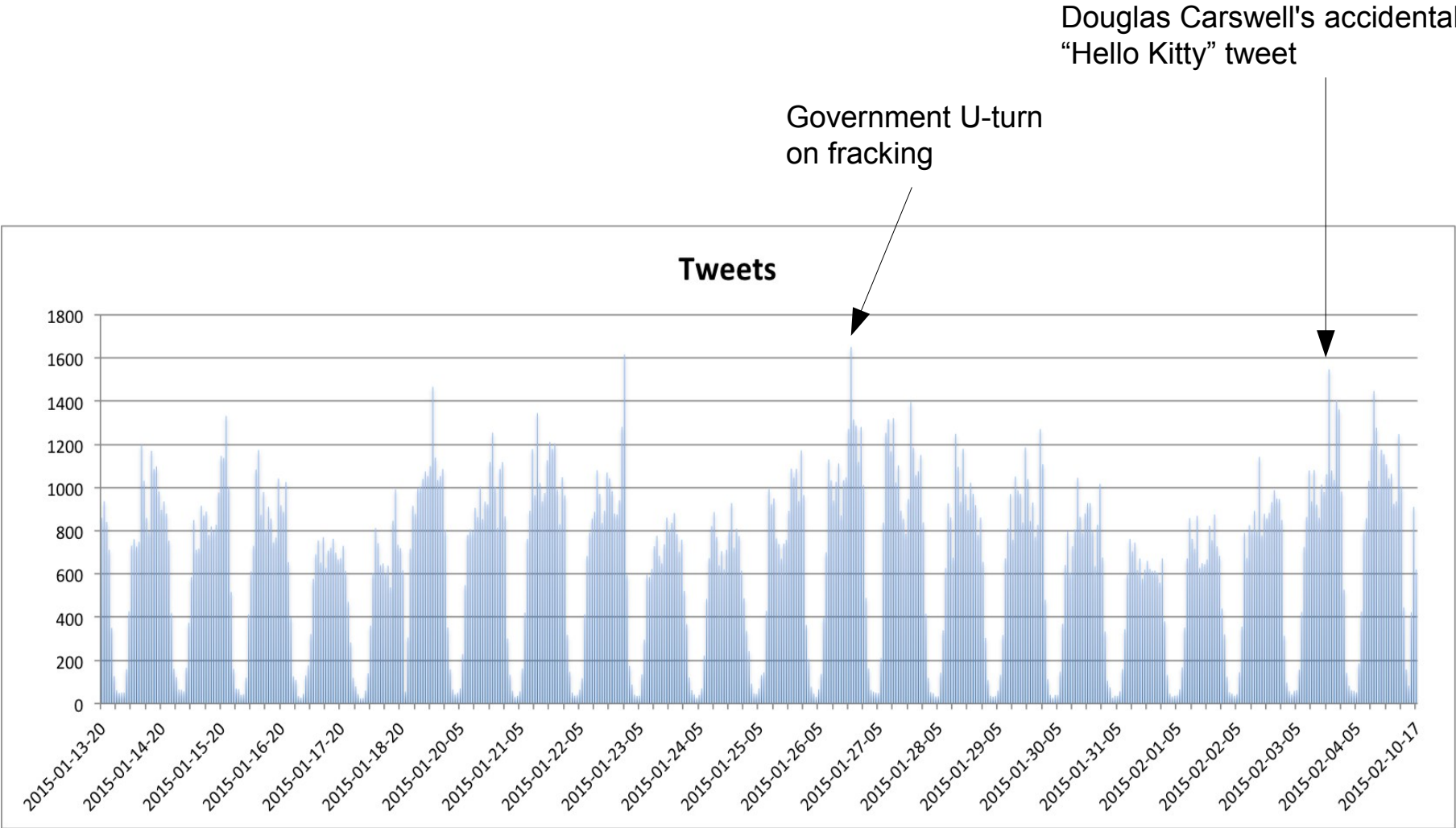
Longer web documents

- Also crawled websites of UK political parties (Con, Lab, LD, Green, UKIP, BNP, SNP, PC, plus the NI parties and various smaller parties)

- Initial crawl on 28th-29th October retrieved 375MB (compressed)

- Re-crawled regularly to pick up new pages

- Final crawl performed the day before the election (6th May) retrieved 535MB (compressed), including 27,500 usable HTML/PDF documents

Debate/hashtag-based Data Collection

- We also added functionality to the data collector to track Tweets by hashtags/keywords

- Used during TV debates to track the relevant debate hashtag (#leadersdebate, #bbcdebate) plus more general hashtags such as #GE2015

- Tweets processed live, with graphs of sentiment by theme produced on a 5 minute rolling cycle

  - George Windsor live-blogged some of these graphs during the debates

Politician and candidate annotation

- Acquired and corrected a list of UK MPs and election candidates and their affiliations, twitter accounts and DBpedia URIs

- Converted to gazetteers so that MPs in various forms (name or twitter handle) can be recognised in tweets and annotated with the relevant info (URI, full name, constituency etc.)

- Some MPs have now started changing their Twitter handles, removing MP from that. First example seen on Mar 23rd. Solution to this problem will be implemented by 30 Mar

    - https://twitter.com/BackBarwell/status/580053148435353600

    - "Due to election rules I've today changed my Twitter handle from @GavinBarwellMP to @BackBarwell. #Croydon"

Politician and candidate annotation

- Acquired and corrected a list of UK MPs and election candidates and their affiliations, twitter accounts and DBpedia URIs

- Converted to gazetteers so that MPs in various forms (name or twitter handle) can be recognised in tweets and annotated with the relevant info (URI, full name, constituency etc.)

- When Parliament was dissolved, many MPs changed their handles per parliamentary rules (there are no MPs while parliament is dissolved, so you must stop using the title)

  - https://twitter.com/BackBarwell/status/580053148435353600

  - "Due to election rules I've today changed my Twitter handle from @GavinBarwellMP to @BackBarwell. #Croydon"

- As long as the numeric account ID stays the same, we can track these changes, and automatically update gazetteers and KB

Recognition of MPs / Candidates

Topic Recognition

- A set of themes was taken from the categories used on http://www.gov.uk

- For each theme, a gazetteer list was developed containing typical keywords and phrases representative of that theme

- e.g. "asylum seeker" indicates the topic "borders and immigration"

- Each list was expanded via:

  - automatic term recognition (based on tf.idf) over a corpus of manifestos and other political documents

  - manual additions

- Each list also contains potential head terms and modifiers which can be expanded into longer terms on the fly from the text during analysis stage

- e.g. "terrorist" can modify many other words (terrorist attack, terrorst threat, ...)

Topic recognition

@Ed_Miliband "3m British jobs and thousands of British businesses rely on the EU. I won't be the Prime Minister that puts that at risk."

Topic			▼	
C root	▼	british job	▼	✕
C rule	▼	ModifierLookupTopic	▼	✕
C string	▼	british jobs	▼	✕
C subtheme	▼	job	▼	✕
C theme	▼	employment	▼	✕
C	▼		▼	✕

▶ Open Search & Annotate tool

This term is found by first recognising the head word "job" from a list under the theme "employment" and matching against its root form in the text, i.e. "job".

It is then extended to include the adjectival modifier "British", which is not present in a list anywhere.

Sentiment annotation

- Annotations are created over the whole sentence and contain the following features:

  - **sentiment_kind**: optimism / pessimism

  - **holder**: the person holding the opinion (MP's name)

  - **holder_UR**I: the URI fo the holder

  - **target**: the target of the opinion, e.g. MP or topic

  - **target_URI**: if appropriate, the URI of the target

  - **score**: a positive/negative value reflecting the strength of opinion

  - **sarcasm**: yes / no (whether sarcasm is present)

  - **sentiment_string**: the main word(s) that contain sentiment

- These annotations and features will be used as input to MIMIR to facilitate analysis/aggregation

Positive opinion about science and innovation

Annotation of Multiple Topics/Sentiments

- For every tweet, once annotated with basic sentiment and topic information, we collect the following information: number of sentiments, topics and the position of the sentiment in the tweet.

- Then we apply a context algorithm:

  - If the tweet contains one or more Topics and one or more Sentiments with the same polarity (pos or neg), then a SentenceSentiment annotation is created. If there are multiple sentiments, we use the one with the highest score, or failing that, the nearest one to the topic.

  - If the tweet contains one Topic and and more than one Sentiment with different polarity, then we take into account the Sentiment from the same sentence as the Topic.

  - If the tweet contains more then one Topic and and more than one Sentiment with different polarity in the same sentence, then we build a "Topic context".

Building the Topic Contexts

- We build topic contexts for each topic as follows: for the first topic encountered, the context is from the beginning of the sentence until the second topic. The rest of the Sentence will be the context for the second(or third etc) topic.

- We also use some phrase breaker words like "but", "because" etc. in order to delimit a phrase which should end the context, i.e. a context cannot span two phrases joined by such words.

- Some topics like "eu" and "eu budget" have different themes associated with them. In this case, they are treated as one entity in terms of context. Two SentimentSentences will be created (one for each Topic theme).

Expansion of core sentiment lexicons

- Expanded the core sentiment lexicon with new terms

- Over a corpus of political tweets, we checked every adjective/adverb/noun/verb for sentiment and for each one, we extracted synonyms from WordNet.

- The synonyms were checked for sentiment in our lists, and any new ones added to the lexicon

- If no synonyms were found, first order hyponyms were investigated in the same way.

GATE Mímir:
Answering Questions Google Can't

GATE Mimir

- can be used to index and search over text, annotations, semantic metadata (concepts and instances)

- allows queries that arbitrarily mix full-text, structural, linguistic and semantic annotations

- is open source

Show me:

- all documents mentioning a temperature between 30 and 90 degrees F (expressed in any unit)

- all abstracts written in French on Patent Documents from the last 6 months which mention any form of the word "transistor" in the English abstract

- the names of the patent inventors of those abstracts

- all documents mentioning steel industries in the UK, along with their location

Search news articles for politicians born in Sheffield

MIMIR: Searching Text Mining Results

- Searching and managing text annotations, semantic information, and full text documents in one search engine

- Queries over <u>annotation graphs</u>

- Regular expressions, Kleene operators

- Designed to be integrated as a web service in custom end-user systems with bespoke interfaces

- Demos at http://services.gate.ac.uk/mimir/

Hands-on with Semantic Search

Try these queries on the BBC News demo:
http://services.gate.ac.uk/mimir/gpd/search/index

- Gordon Brown

- Gordon Brown said

- Gordon Brown [0..3] root:say

- {Person} [0..3] root:say

- {Person.gender=female}[0..3] root:say

- Make sure you type the queries EXACTLY or they probably won't work!

- Try making up some of your own queries.

Try your hand with some SPARQL
(for the more adventurous!)

- {Person inst ="http://dbpedia.org/resource/Gordon_Brown"} [0..3] root:say

- {Person sparql="SELECT ?inst WHERE { ?inst a :Politician }"} [0..3] root:say

- {Person sparql = "SELECT ?inst WHERE { ?inst :party <http://dbpedia.org/resource/Labour_Party_%28UK%29> }" } [0..3] root:say

- {Person sparql = "SELECT ?inst WHERE {
    ?inst :party <http://dbpedia.org/resource/Labour_Party_%28UK%29> .
    ?inst :almaMater <http://dbpedia.org/resource/University_of_Edinburgh> }"
  } [0..3] root:say

Can you work out what these do?

What diseases are in these documents?

What Pathogens?

Disease vs Disease Co-ocurrences
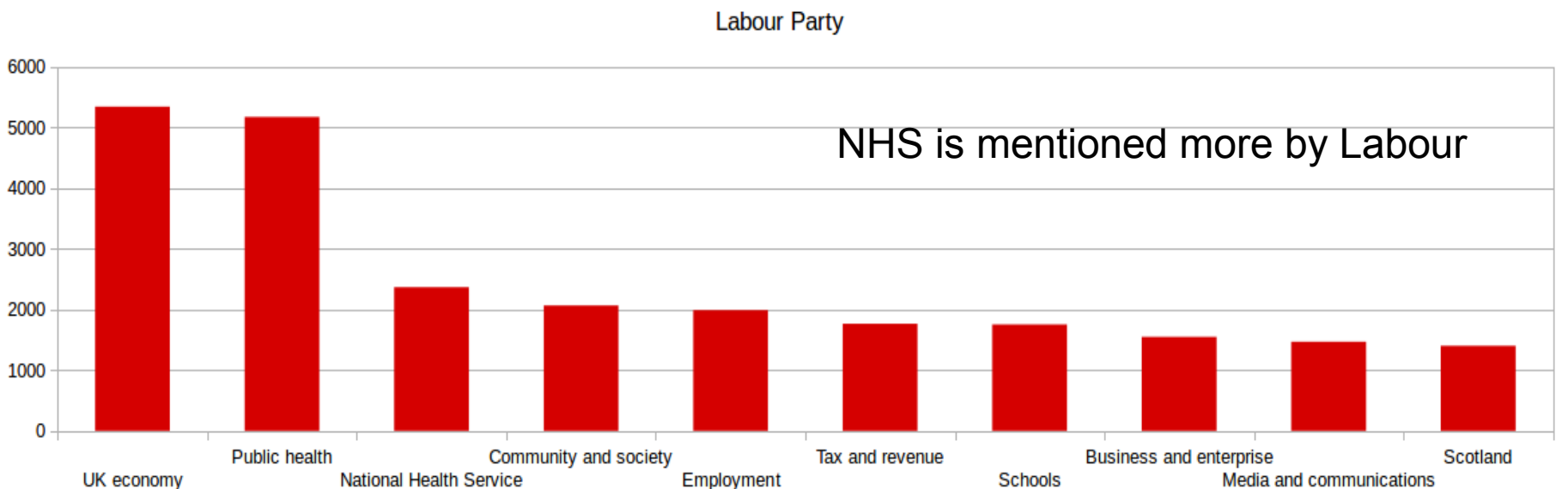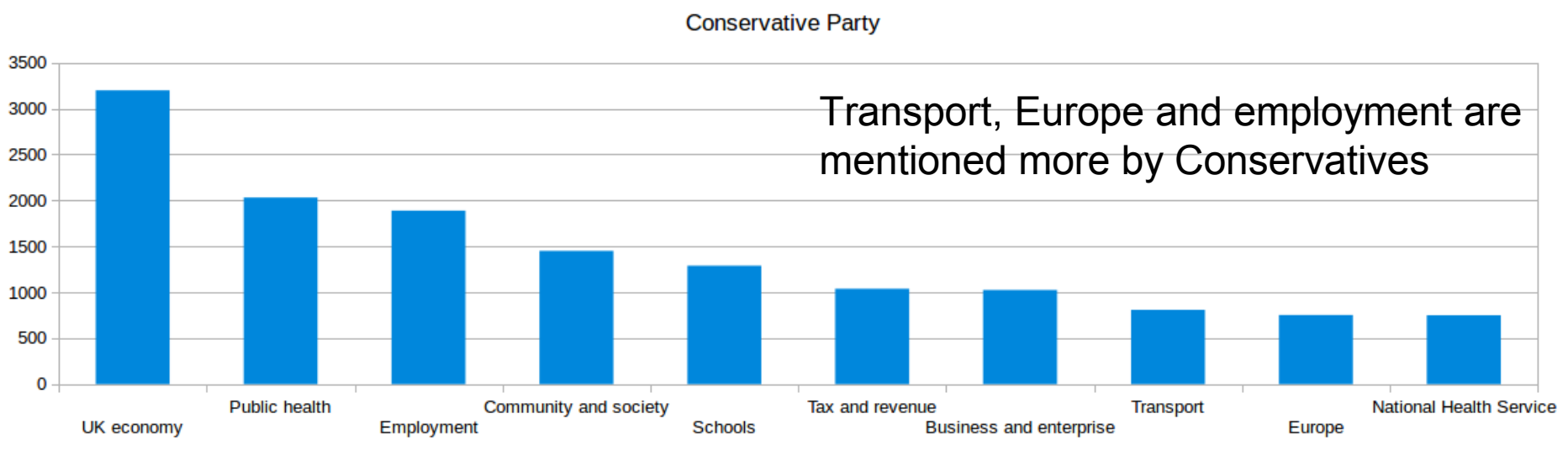
Diseases vs Pathogens

Document Indexing with MIMIR

- MIMIR allows for indexing and querying text, annotations and semantic knowledge

    - this gives a rich source of data for analysis

- Currently we have used MIMIR to index

    - the raw collected text

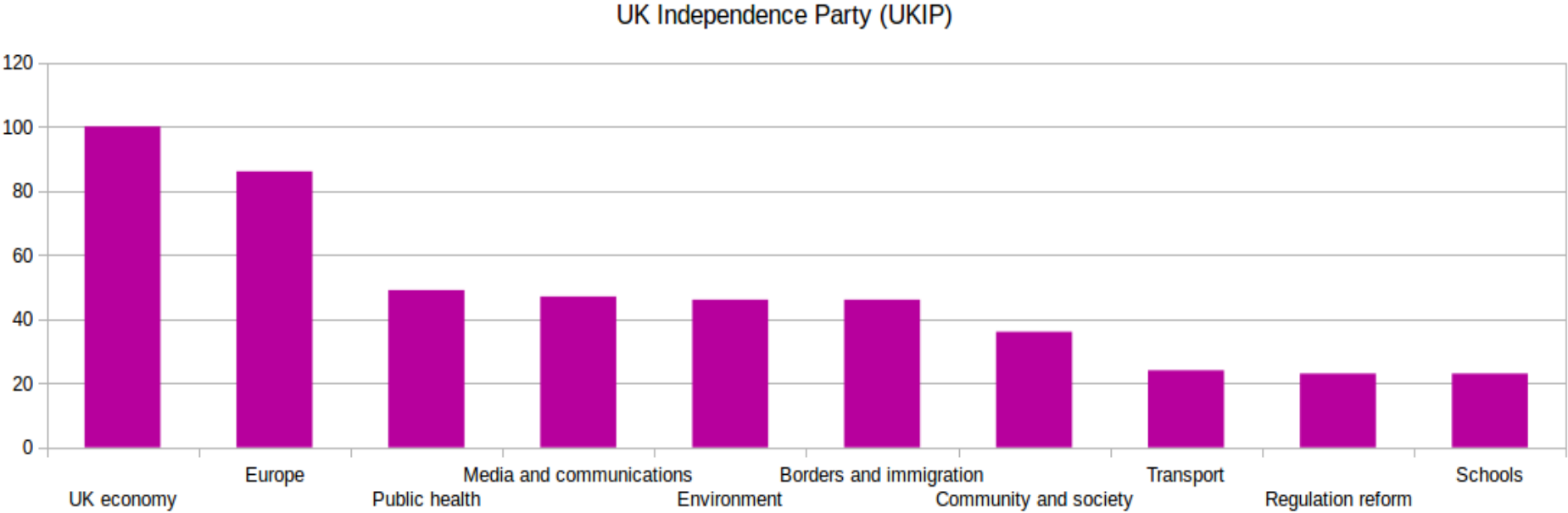   - annotations provided by Twitter and by the applications

Examples of Mimir queries on our corpus

- Get all documents which talk about the borders/immigration topic

  {Topic theme = "borders_and_immigration"}

- Get all documents where the author of the document is a candidate

  {DocumentAuthor sparql = "?c nesta:candidate ?author_uri"}

- Get all documents where the author is an MP standing for re-election for the same seat

  {DocumentAuthor sparql="?c nesta:candidate ?author_uri . ?c dbp-prop:mp ?author_uri"}

- Get all documents where the author is a candidate for the Sheffield Hallam constituency

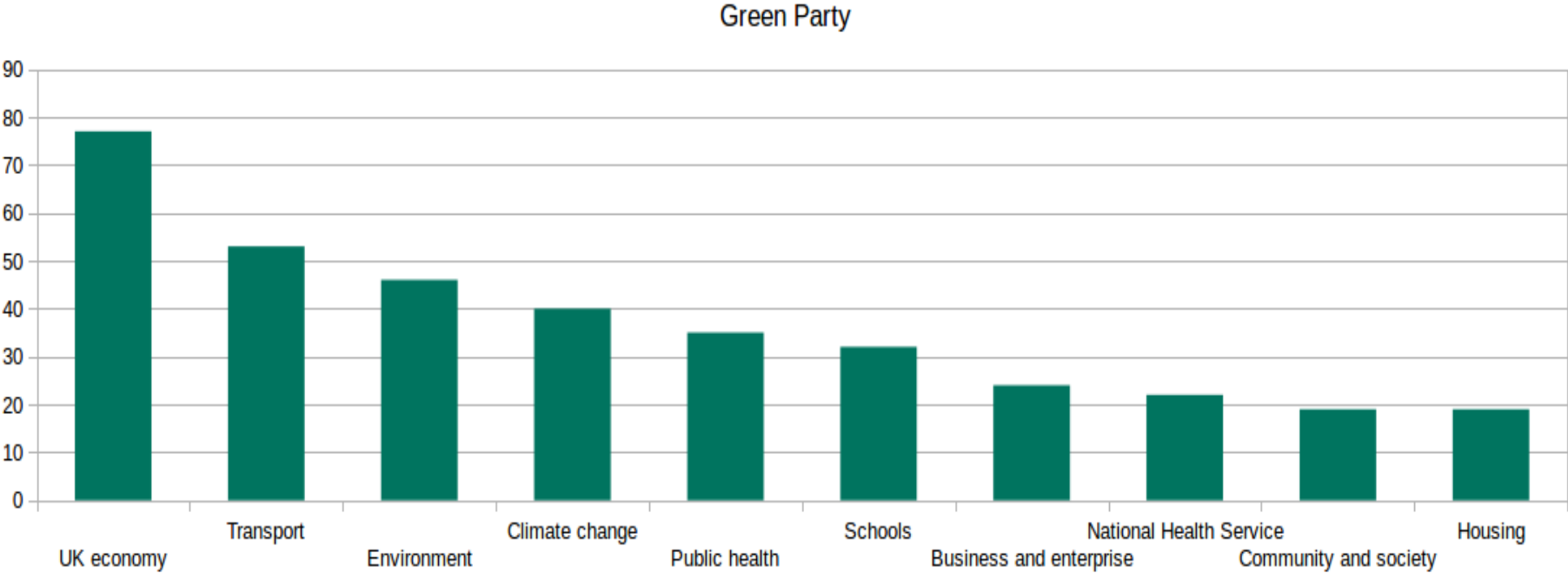 {DocumentAuthor sparql="<http://dbpedia.org/resource/Sheffield_Hallam_(UK_Parliament_constituency)> nesta:candidate ?author_uri"}

# What do the different parties talk about? Conservative vs Labour

Conservative Party

Transport, Europe and employment are mentioned more by Conservatives

Labour Party

NHS is mentioned more by Labour

Topics mentioned by UKIP

UK Independence Party (UKIP)
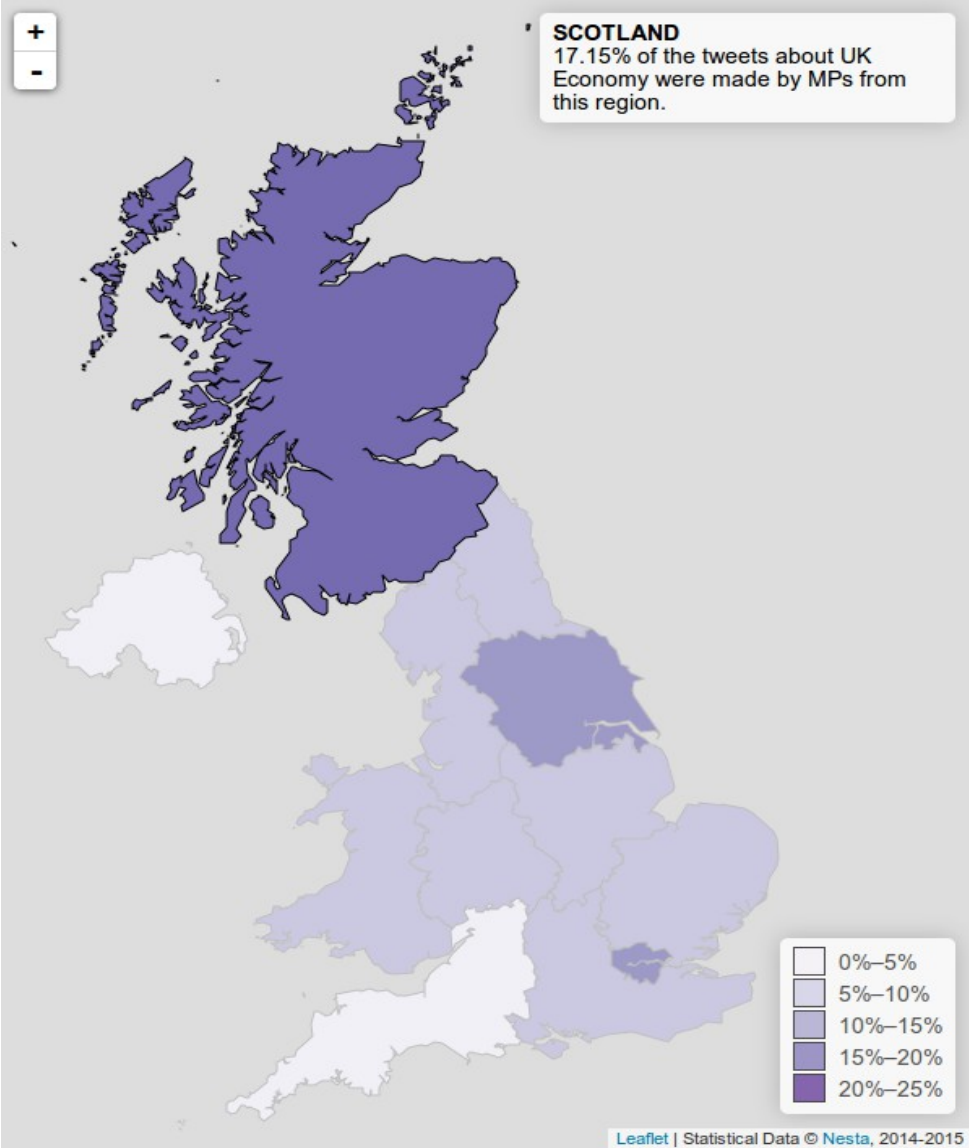
Topics mentioned by the Green Party

Expected topics are high on the list

# How old are MPs?

Average MP Age by Party
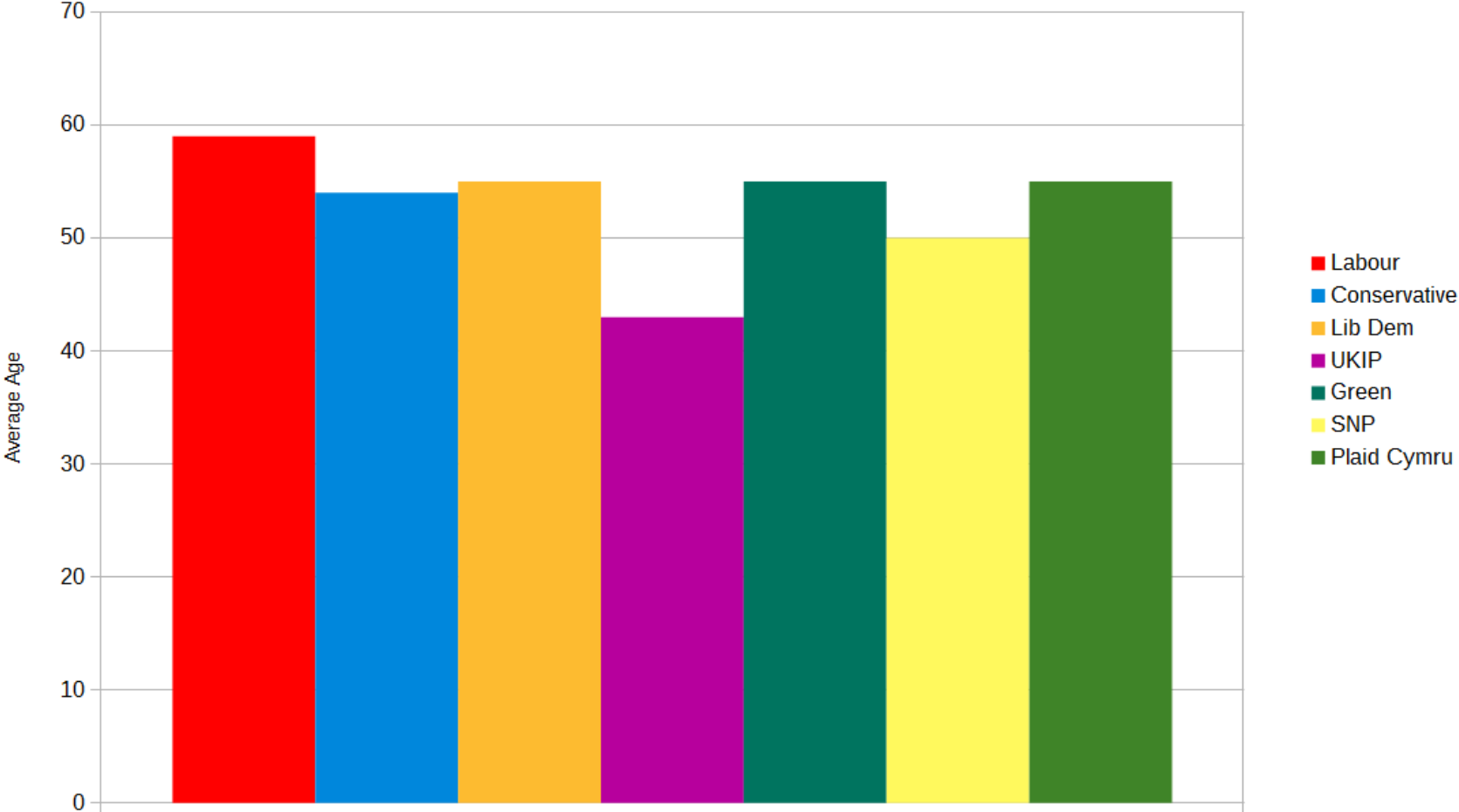
# Where did people tweet about the economy?

Aggregating Sentiment

- We can aggregate sentiment about tweets on each topic to get interesting perspectives.

- During the television debates, we analysed sentiment of the public (via their tweets) about the major topics discussed by each political leader

- We found that sentiment was broadly positive for Ed Milliband, but that when he talked about the NHS and the economy, the public reaction was more negative about these topics.
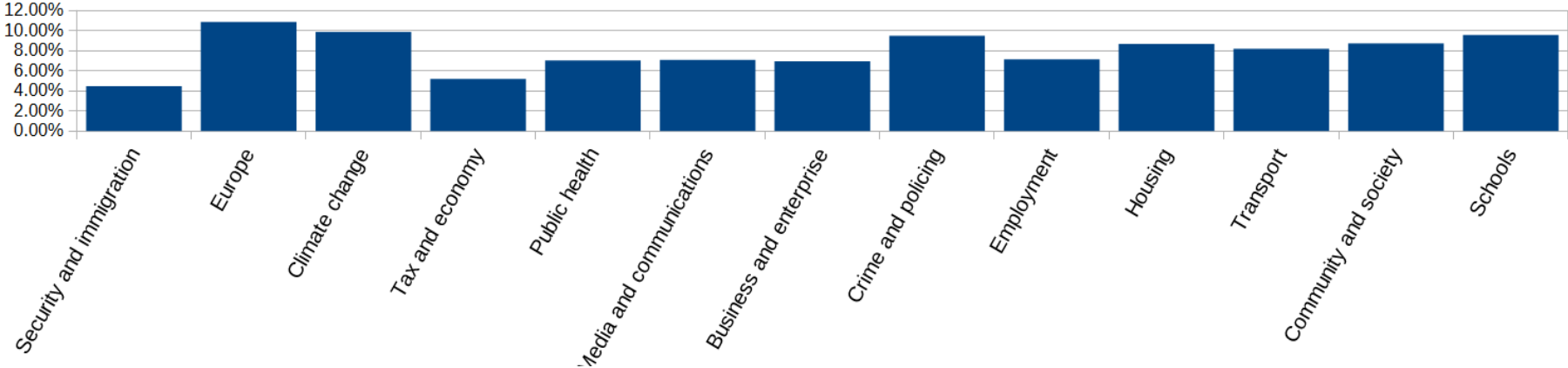
Issues that Milliband talked about
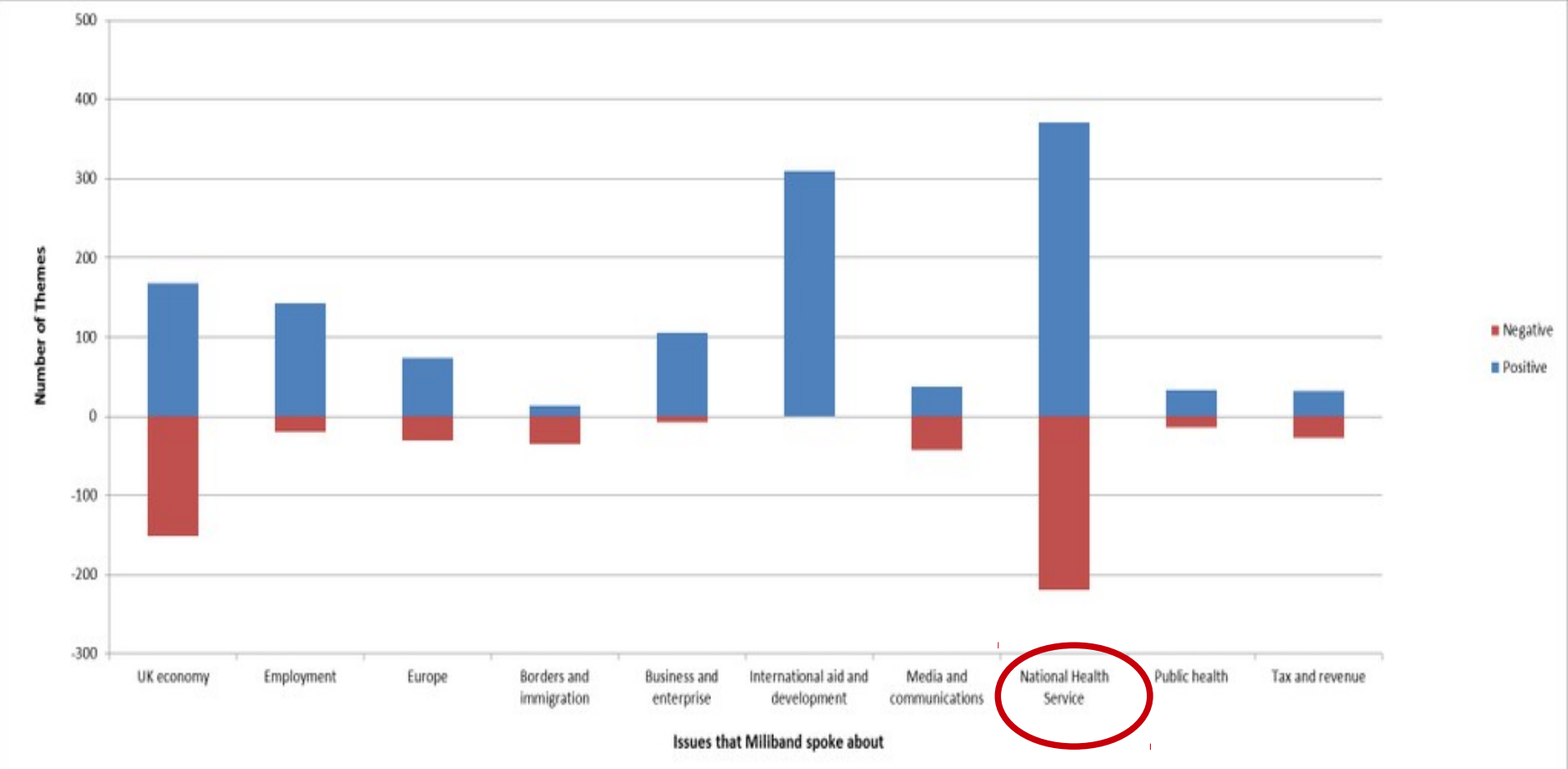
Using Prospector: terms that co-occur with immigration topic
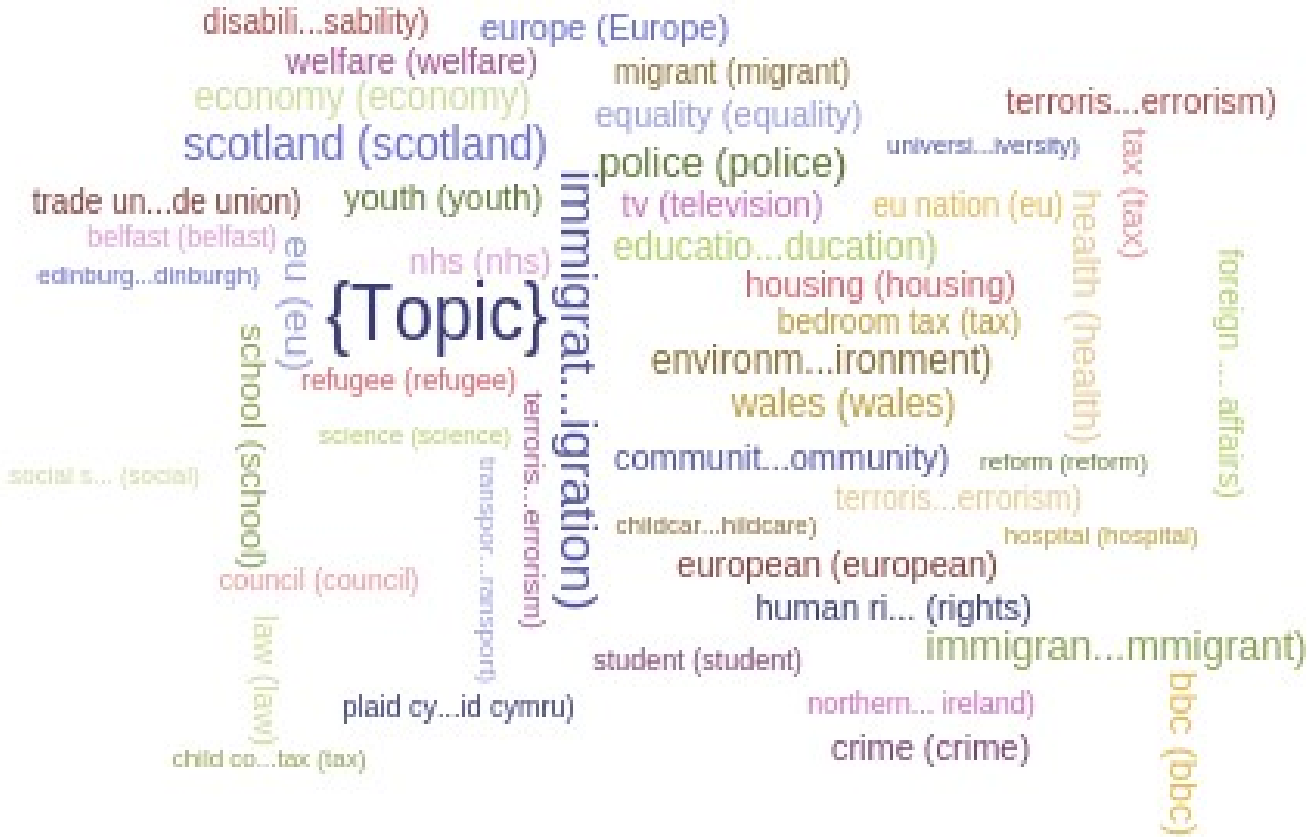
Which tweets express sentiment?

Engagement with Climate Change

- We did some experiments to try to prove a theory that people engage more with climate change than other political topics

- We used the same dataset, and looked at typical engagement metrics for each topic to see how they compared

- High levels of sentiment, positive sentiment, inclusion of URLs, number of retweets and replies all indicate engagement

- We found that while climate change and the environment weren't mentioned frequently by politicians (other than by the Green Party), they showed high levels of engagement

Summary

- Once you have the indexed data, you can carry on doing all kinds of interesting comparisons and analysis.

- Simple analysis tools can give you pretty pictures, but you can do much more interesting things when you delve a bit deeper and make use of information not explicit in the text

- For this you need both NLP and Linked Open Data

- We also used the same underlying technology to investigate the public response to tweets about climate change during the run-up to the election

- Here we showed that people engaged more with climate change than other topics, possibly because they felt it was something they could personally influence

Acknowledgements and more information

- Some publications about this:

  - A. Dietzel and D. Maynard. Climate Change: A Chance for Political Re-Engagement? In Proc. of the Political Studies Association 65th Annual International Conference, April 2015, Sheffield, UK.

  - D. Maynard, M. A. Greenwood, I. Roberts, G. Windsor, K. Bontcheva. Real-time Social Media Analytics through Semantic Annotation and Linked Open Data. Proceedings of WebSci 2015, Oxford, UK, June 2015 (forthcoming)

  - 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 (forthcoming).