



The
University
Of
Sheffield.



Social media analysis with GATE

Dr. Diana Maynard
University of Sheffield, UK



Outline

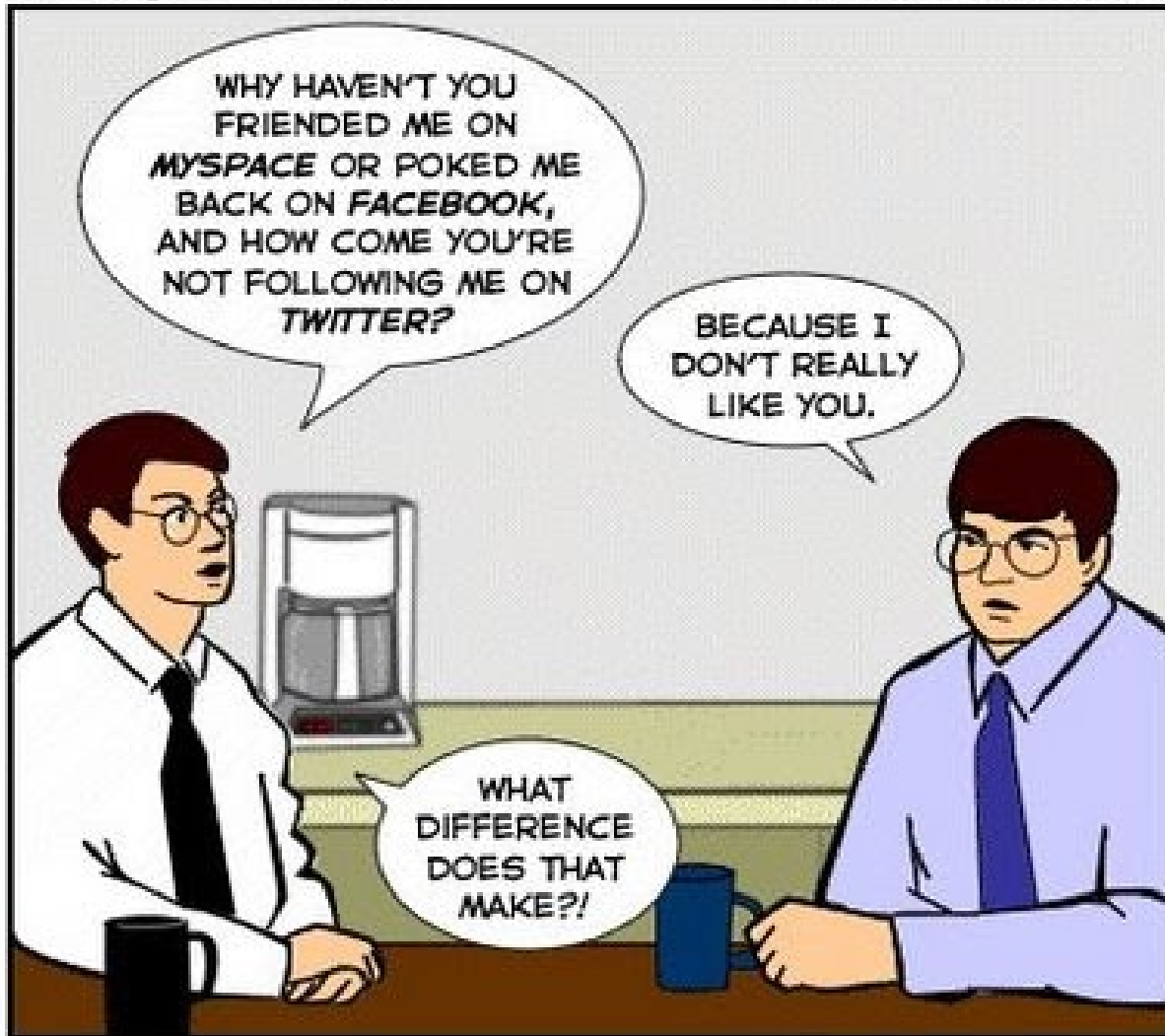
- Introduction to social media analysis
- NLP and Information Extraction
- GATE and other tools
- Analysing social media
- Sentiment analysis for social media
- Application: analysing climate change In DecarboNet

1. Introduction to Social Media Analysis

We are all connected to each other...

The Joy of Tech™

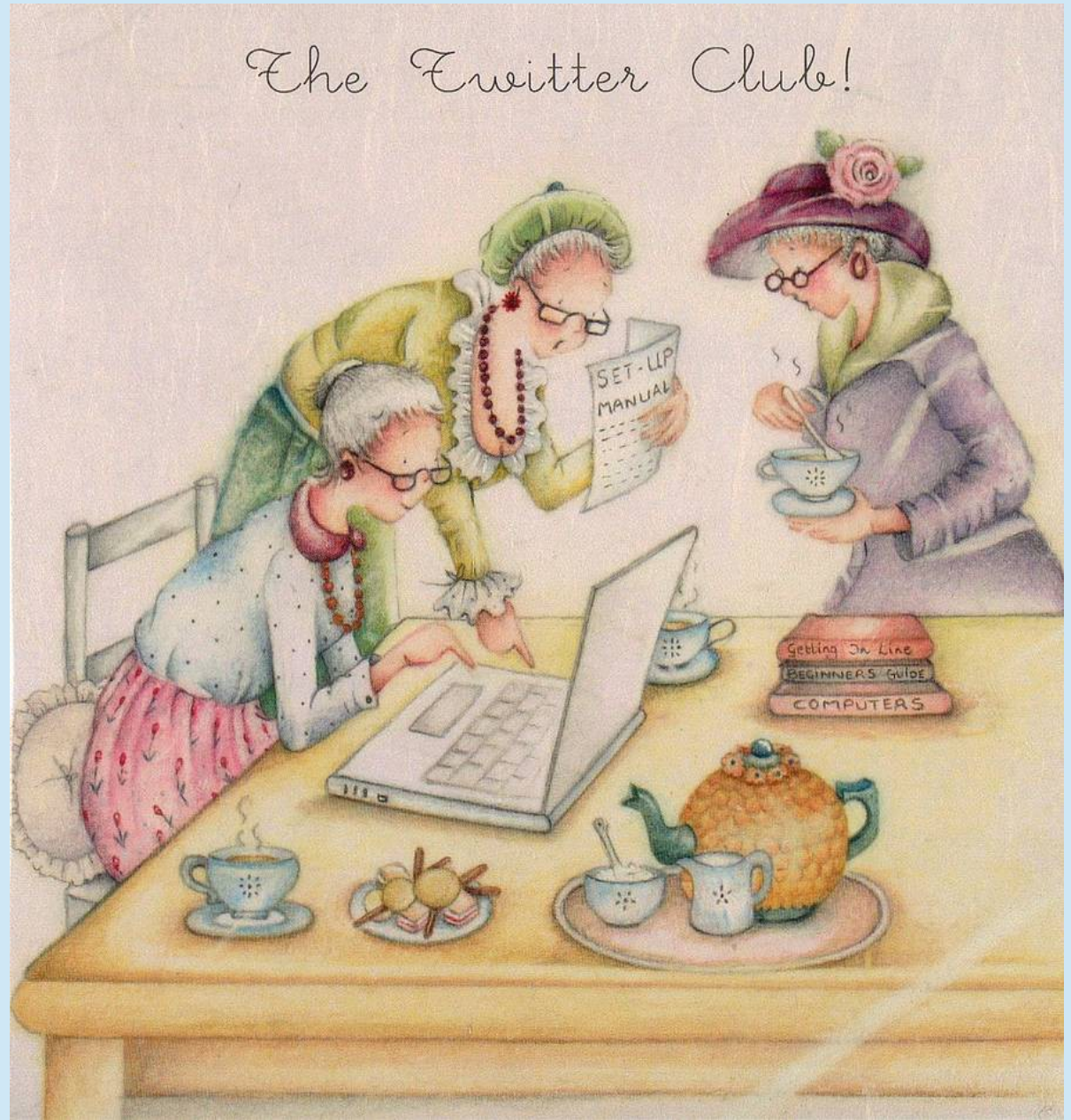
by Nitrozac & Snaggy



joyoftech.com

- Information, thoughts and opinions are shared prolifically on the social web these days
- 72% of online adults use social networking sites

Social media
isn't just for
young people
either!



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



Time Spend by Average Social networking user per month



Pinterest

405 mins



twitter

89 mins



LinkedIn

21 mins










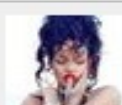


Google+

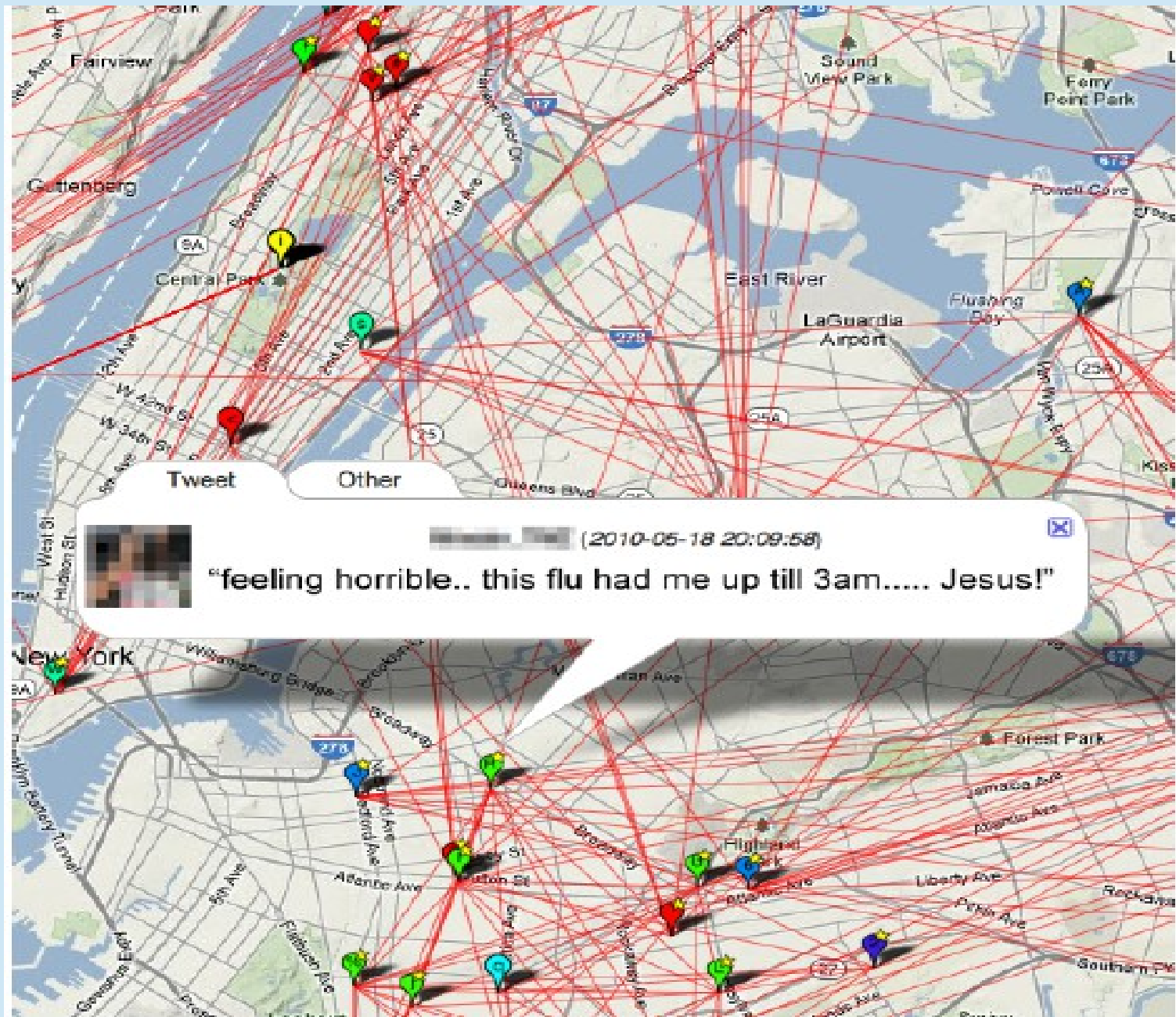
3 mins

- In Britain and the US, approx 1 hour a day on social media
- 90% of marketers use social media channels for business


What are people reading about?

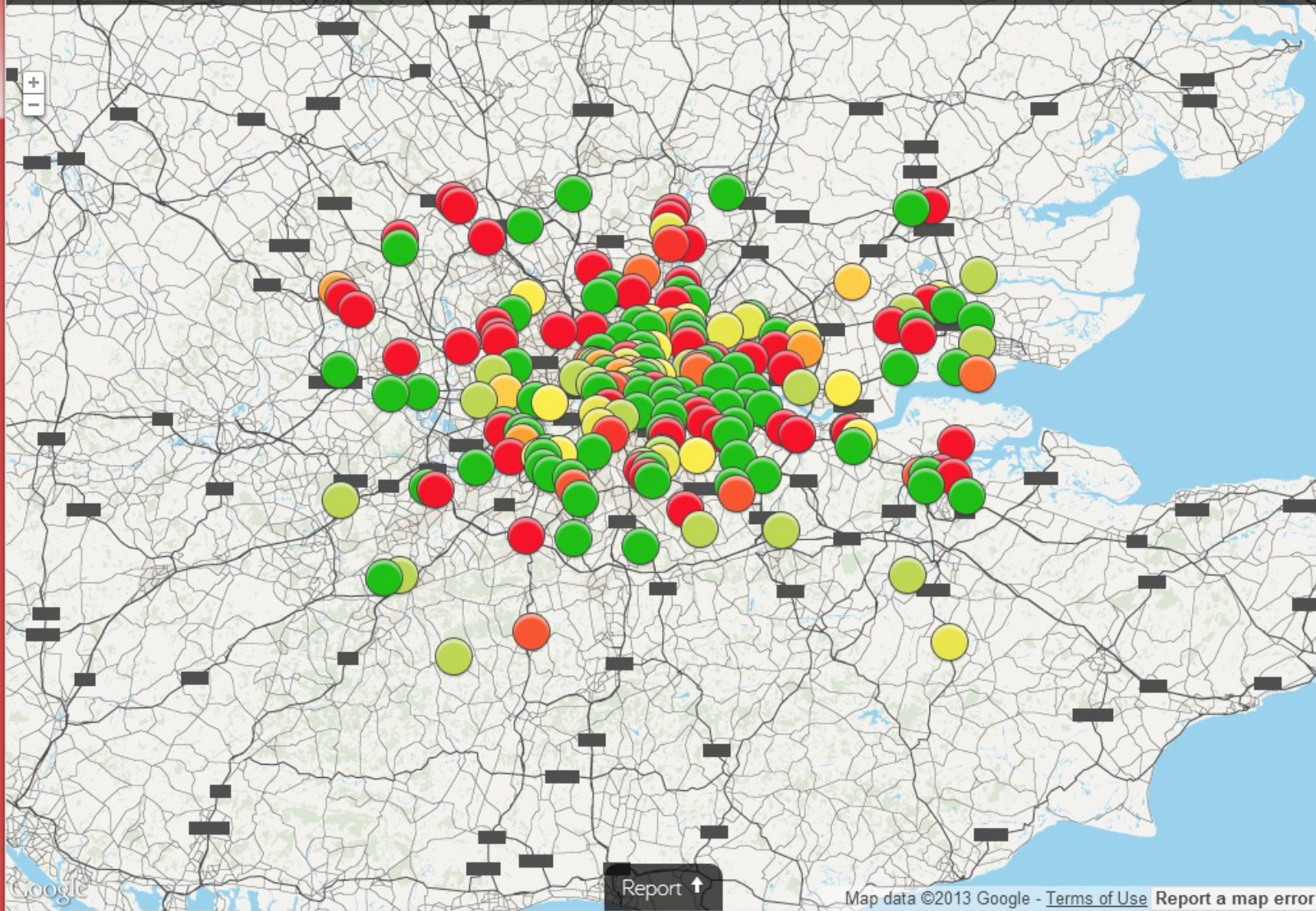
- Of the top 10 Twitter accounts with the highest number of followers:
 - 7 pop stars
 - 2 social media sites
 - and Barack Obama
- As you can imagine, there's a lot of mindless drivel on social media sites

#	Name (Screen Name)
1.	 Katy Perry @katyperry
2.	 Justin Bieber @justinbieber
3.	 Lady Gaga @ladygaga
4.	 Barack Obama @BarackObama
5.	 Taylor Swift @taylorswift13
6.	 YouTube @YouTube
7.	 Britney Spears @britneyspears
8.	 Rihanna @rihanna
9.	 Instagram @instagram
10.	 Justin Timberlake @jtimberlake



Health Risk: **Low**

 Tweet





- One in six people have used social media to get information about an emergency
- 75% of people would use Facebook to post eyewitness information on an emergency or newsworthy event; 22% would use blogs, 21% would use Twitter
- During an emergency, one in two people would use social media to let loved ones know they are safe
- We need good real-time analysis tools to help process this data

Eoghan Rice @rice_e

This man survived by clinging to the roof of a house. Waves dragged him up to roof level. #yolanda #haiyan
pic.twitter.com/GsYu9JTq7S

Expand

Reply

Retweet

Favorite

More



Why analyse social media?

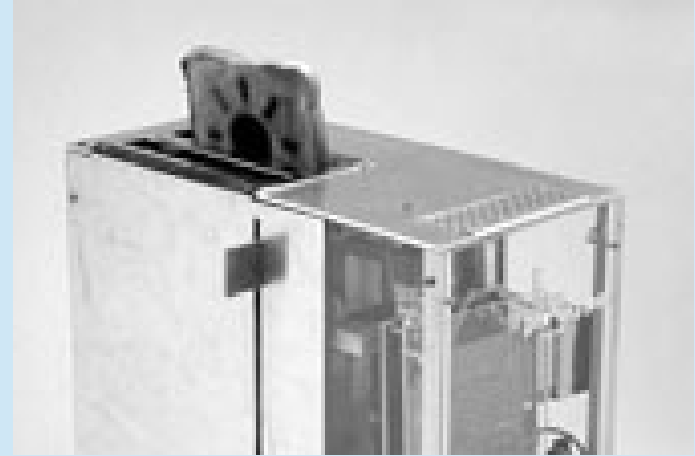
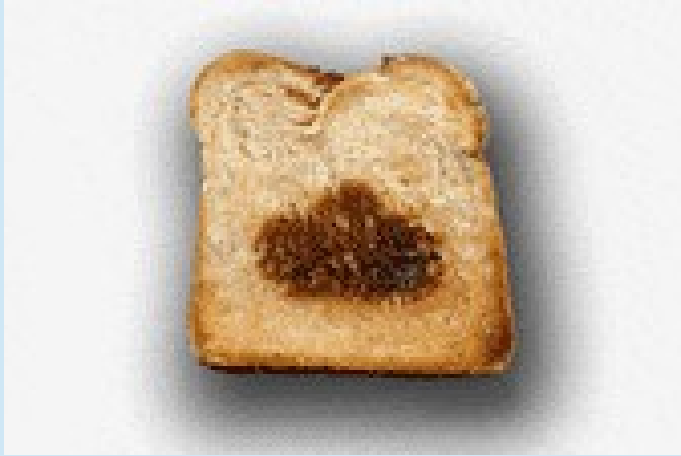
- Understanding customer reviews and so on is a huge business
- But also:
 - Tracking political opinions: what events make people change their minds?
 - How does public mood influence the stock market, consumer choices etc?
 - How are opinions distributed in relation to demographics?
 - What makes users engage with topics, e.g. climate change campaigns?
- NLP tools are crucial in order to make sense of all the information

2. NLP and Information Extraction

Oddly enough, people have successfully combined information and toast...



The weather-forecasting toaster



- This weather-forecasting toaster, connected to a phone point, was designed in 2001 by a PhD student
- It accessed the MetOffice website via a modem inside the toaster and translated the information into a 1, 2 or 3 for rain, cloud or sun
- The relevant symbol was then branded into the toast in the last few seconds of toasting

With tools such as these, why do we need text mining?



- It turns out that toast isn't actually a very good medium for finding information

It is difficult to access unstructured information efficiently

Information extraction tools can help you:

- Save time and money on **management of text** and data from multiple sources
- Find **hidden links** scattered across huge volumes of diverse information
- Integrate **structured data** from variety of sources
- **Interlink** text and data
- **Collect information** and extract new facts



Why are Entities and Events Useful?

- They can help answer the “Big 5” journalism questions (who, what, when, where, why)
- They can be used to categorise the texts in different ways
 - look at all texts about David Cameron (or Justin Bieber).
- They can be used as targets for opinion mining
 - find out what people think about David Cameron (or Justin Bieber)
- When linked to an ontology and/or combined with other information, they can be used for reasoning about things not explicit in the text
 - seeing how opinions about different prime ministers (or pop stars) have changed over the years

Finding things not in the text

GUS - GATE Unified Sear... x +

dem... demos.gate.ac.uk/mimir/gpd/search/gus#page=1

 **Search**  powered by Mimir

Searching index: News Demo

```
{Person sparql = "SELECT ?inst WHERE { ?inst :party  
<http://dbpedia.org/resource/Labour_Party_%28UK%29>"}"  
root:say) IN ({{Document date > 20110000} OVER {DocumentClassification sparql =  
"SELECT ?inst WHERE { ?inst a bbc:Classification . FILTER (?inst = bbc:Scotland)}}")
```

Search

Results 1 - 5 of 5

Trident nuclear fleet cuts ruled out by Liam Fox (cached)
bn. Former Prime Minister **Gordon Brown said** in 2009 that he would

Trident nuclear fleet cuts ruled out by Liam Fox (cached)
bn. Former Prime Minister **Gordon Brown said** in 2009 that he would

Councils set out budget savings (cached)
'priorities' Council leader **Gordon Matheson said**: "I've always

Scottish election: Green policy and business on agenda (cached)
businesses. And former chancellor **Alistair Darling said** Labour would put families first

Libya: 'Sick' Moussa Koussa has 'no Lockerbie secrets' (cached)
Labour's shadow home secretary **Yvette Cooper said** the questions for Mr Koussa

And with a snazzier interface

The screenshot shows a web browser window with the title 'People in the News'. The address bar contains the URL: `demos.gate.ac.uk/pin/?name=&bornIn=Sheffield&famousAs=Politician|OfficeHolder&after=01%2F04%2F2011&before=30%2F`. The page has a red header with the text 'PEOPLE IN THE NEWS' and a world map background. Below the header are two search panels. The left panel, 'Looking For...', has fields for 'Name:', 'Fuzzy Name Matching' (checkbox), 'Born In:' (text box with 'Sheffield'), and 'Famous As:' (dropdown menu with 'Politician'). The right panel, 'In Articles...', has fields for 'Published Between' (date range '01/04/2011' to '30/04/2011'), 'Classified As:' (dropdown menu with 'Scotland'), and 'Ignore Boilerplate Text' (checkbox). A 'Search' button is at the bottom right of the search panels. Below the search panels is a red bar with 'Results 1 to 2 of 2' on the left and 'Show Underlying Mimir Query' with a dropdown arrow on the right. The search results show a link to 'Scottish election: Respect Coalition Against Cuts profile' with a URL and a 'Cached' status. Below the link are two lines of text: '... Bow - whose sitting MP Oona King had voted for the war ...' and '... success came when Galloway overturned Oona King's 10,000- ...'. At the bottom of the page, it says 'Powered by GATE Mimir' and '© The University of Sheffield, 2011'.

People in the News

demos.gate.ac.uk/pin/?name=&bornIn=Sheffield&famousAs=Politician|OfficeHolder&after=01%2F04%2F2011&before=30%2F

PEOPLE IN THE NEWS

Looking For...

Name:

Fuzzy Name Matching

Born In:

Famous As:

In Articles...

Published Between and

Classified As:

Ignore Boilerplate Text

Search

Results 1 to 2 of 2 [Show Underlying Mimir Query](#)

[Scottish election: Respect Coalition Against Cuts profile](http://www.bbc.co.uk/news/uk-scotland-13048761) - [Cached](#)

... Bow - whose sitting MP Oona King had voted for the war ...

... success came when Galloway overturned Oona King's 10,000- ...

Powered by [GATE Mimir](#)
© The University of Sheffield, 2011

3. GATE and other tools

What is GATE?

GATE is an NLP toolkit developed at the University of Sheffield over the last 20 years

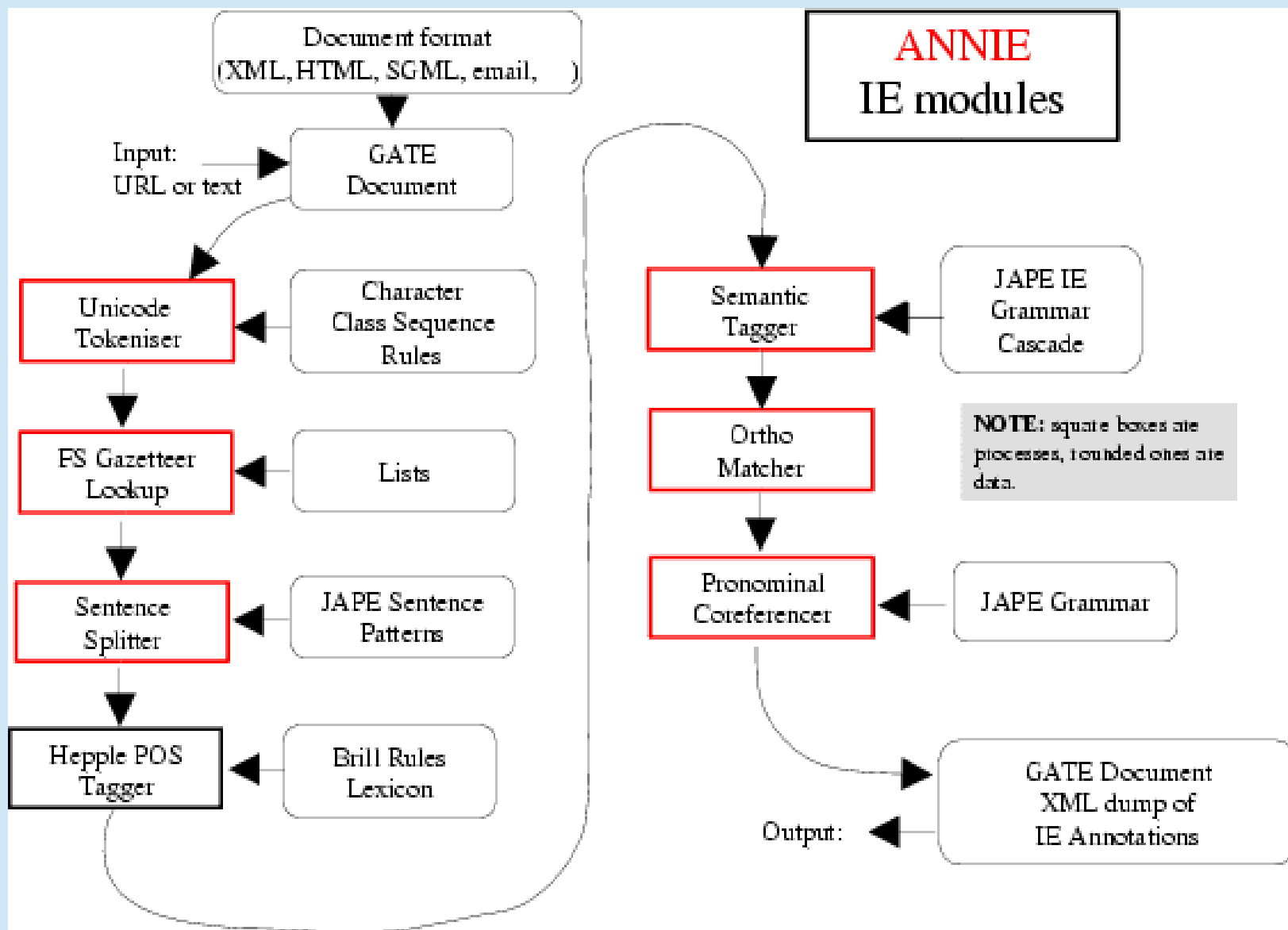
It includes:

- **components** for language processing, e.g. parsers, machine learning tools, stemmers, IR tools, IE components for various languages...
- tools for **visualising** and **manipulating** text, annotations, ontologies, parse trees, etc.
- **various information extraction** tools
- **evaluation** and **benchmarking** tools

ANNIE

- **ANNIE** is GATE's rule-based IE system
- It uses the language engineering approach (though we also have tools in GATE for ML)
- Distributed as part of GATE
- Uses a finite-state pattern-action rule language, JAPE
- ANNIE contains a reusable and easily extendable set of components:
 - generic preprocessing components for tokenisation, sentence splitting etc
 - components for performing NE on general open domain text

ANNIE Modules



Document with Tokens

The screenshot shows a software interface for linguistic analysis. At the top, there are several tabs: "Annotation Sets", "Annotations List", "Annotations Stack", "Class", "Co-reference Editor", "Instance", and "Text". The "Text" tab is active, displaying a document with several lines of text. The first line, "Union Appeals For Talks To End BA Strike", and the first few words of the second line, "Skip to navigation, Skip to content", are highlighted in green. Below the text, a table lists the features for the first five tokens. The "Token" feature is checked in the right-hand panel.

Union Appeals For Talks To End BA Strike

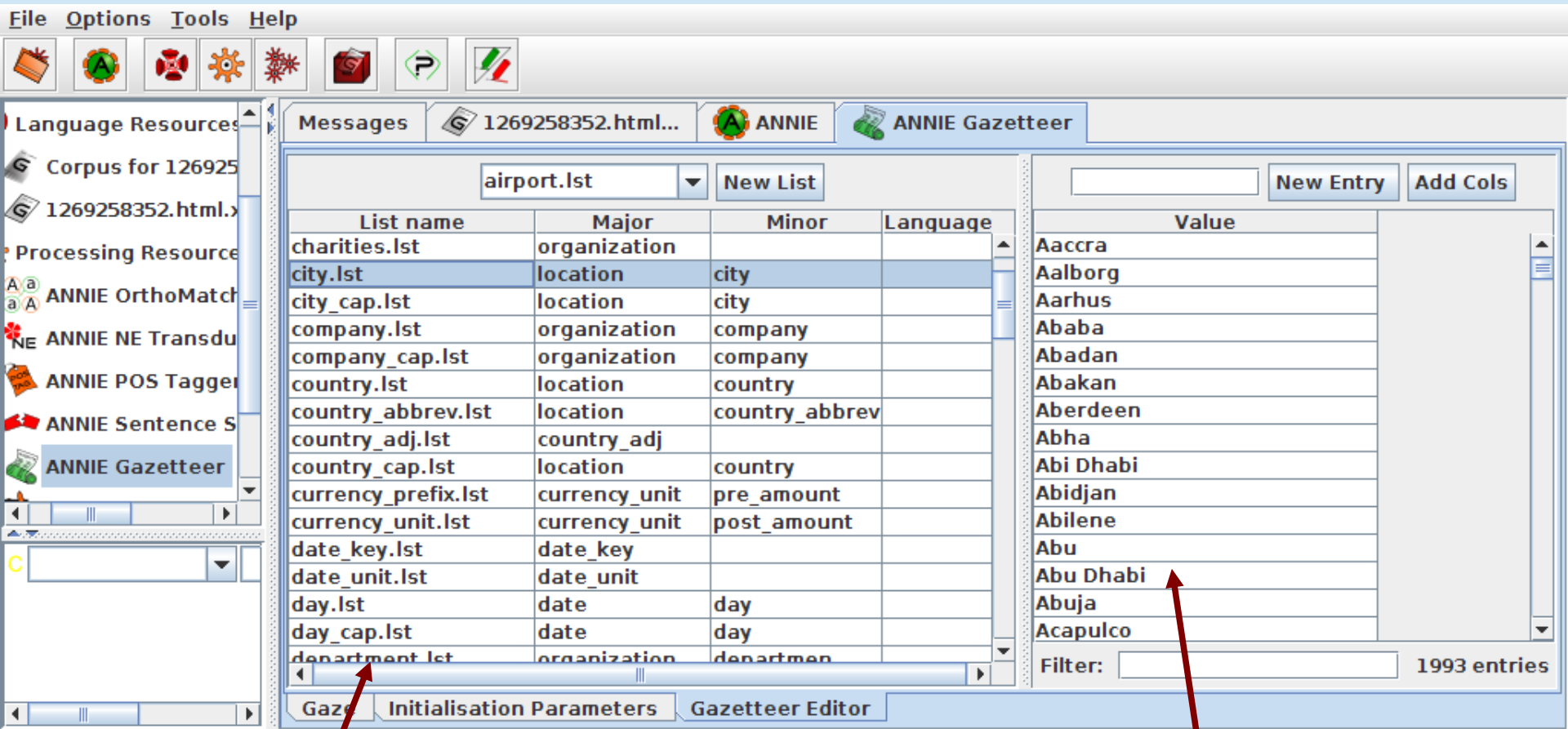
Skip to navigation, Skip to content,
Home | Contact Us | News Search;
HubPage
Airwise News
Airport Guide
Airwise Travel
Search
Union Appeals For Talks To End BA Strike
March 22, 2010

Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers.

Type	Features
Token	{ category=NNP, kind=word, length=5, orth=upperInitial, string=Union}
Token	{ category=NNPS, kind=word, length=7, orth=upperInitial, string=Appeals}
Token	{ category=IN, kind=word, length=3, orth=upperInitial, string=For}
Token	{ category=NNS, kind=word, length=5, orth=upperInitial, string=Talks}
Token	{ category=TO, kind=word, length=2, orth=upperInitial, string=To}

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- Original markups

Gazetteer editor



definition file entries

entries for selected list

Document with NEs

The screenshot displays a document editor window with a toolbar at the top containing buttons for 'Annotation Sets', 'Annotations List', 'Annotations Stack', 'Co-reference Editor', 'Text', and a search icon. The main text area contains several paragraphs with various words highlighted in different colors, indicating named entities. A floating configuration panel is overlaid on the text, showing a 'Date' field, a 'kind' dropdown set to 'year', and a 'C' icon. Below the panel is a button labeled 'Open Search & Annotate tool'. On the right side, a vertical list of entity types is shown, with checkboxes next to each. The 'Entities' section is expanded, and several types are checked: Date, Location, Organization, and Person. Other types like EmbeddedHead1, EmbeddedHead2, EnglishContent, FirstPerson, GermanContent, Head, JobTitle, MOD, MultiWord, Sentence, SingleWord, and SpaceToken are unchecked. At the bottom of the window, there are tabs for 'Document Editor' and 'Initialisation Parameters', and a 'New' button.

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

In so far as a political party in the **United States** can "decide" anything, the party decided not to have the fight it needed to have between reality-based Republicans and the other kind. And so **1960** but in disguised form. The **Republican** coalition season,

Given the state of our political **kind** understood with a theory like this. There is **year** some of the key phrases.

1.) When I say "reality-based Republicans" I mean those who recognize the danger in trying to make descriptions of the world conform to their wishes. By the "other kind" I mean those who don't. Or: members of the **Republican** coalition who exhibit certain behaviors **F.A. Hayek** wrote about in **1960**. This quotation was dug up by Chis Mooney , author of The **Republican** War on Science . It is from Hayek's essay, "Why I am Not a Conservative."

Personally, I find that the most objectionable feature of the conservative attitude is its

English Entities
 Date
 EmbeddedHead1
 EmbeddedHead2
 EnglishContent
 FirstPerson
 GermanContent
 Head
 JobTitle
 Location
 Lookup
 MOD
 MultiWord
 Organization
 Person
 Sentence
 SingleWord
 SpaceToken

Open Search & Annotate tool

Document Editor Initialisation Parameters New

Coreference

Annotation Sets Annotations List Annotations Stack Class **Co-reference Editor** Instance T

Completion of the **National Air Traffic Services** deal comes at a critical time for the government as it tries to push through the PPP for the **London** Underground.

The sale to a strategic investor of a 46 per cent stake in **Nats** is the first time in Europe that management control of en route air traffic services has passed into private hands.

It has been carried out despite a pledge by Labour before the 1997 general election that **UK** air was "not for sale."

Under the terms of the deal, which was approved by the European competition authorities in May, the government has retained a 49 per cent stake and a golden share, while a 5 per cent stake is to be allocated to **Nats**' 5,700 staff.

Sets : **Default**

Types : **Organization** Show

Co-reference Data
♀ Default

- National Air Traffic Services**
- Airline Group**
- UK**
- London**
- March**

Document Editor Initialisation Parameters

4. Analysing Social Media

Analysing language in social media is hard

- Grundman:politics makes #climatechange scientific issue,people don't like knowitall rational voice tellin em wat 2do
- @adambation Try reading this article , it looks like it would be really helpful and not obvious at all. <http://t.co/mo3vODoX>
- Want to solve the problem of #ClimateChange? Just #vote for a #politician! Poof! Problem gone! #sarcasm #TVP #99%
- Human Caused #ClimateChange is a Monumental Scam!
<http://www.youtube.com/watch?v=LiX792kNQeE> ... F**k yes!!
Lying to us like MOFO's Tax The Air We Breath! F**k Them!

Challenges for NLP

- Noisy language: unusual punctuation, capitalisation, spelling, use of slang, sarcasm etc.
- Terse nature of microposts such as tweets
- Use of hashtags, @mentions etc causes problems for tokenisation #thisistricky
- Lack of context gives rise to ambiguities
- NER performs poorly on microposts, mainly because of linguistic pre-processing failure
 - Running standard IE tools (ANNIE) on 300 news articles – 87% F-measure
 - Running ANNIE on some tweets - < 40% F-measure
 - Other tools (e.g. Stanford NER) can reach even lower scores

Lack of context causes ambiguity

Branching out from Lincoln park after dark ... Hello Russian Navy, it's like the same thing but with glitter!



??

Getting the NEs right is crucial

Branching out from Lincoln park after dark ... Hello Russian Navy, it's like the same thing but with glitter!



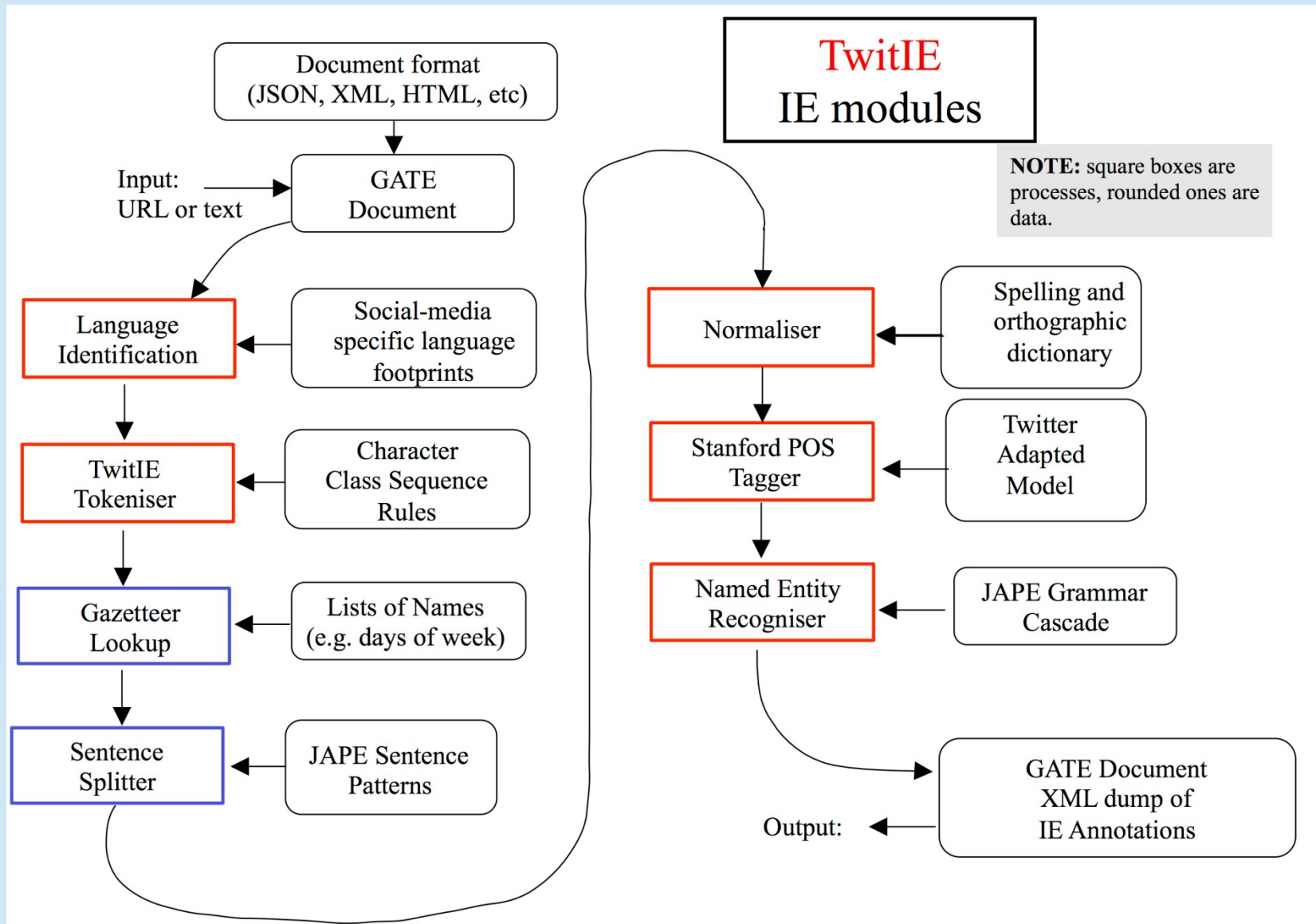
How do we deal with this kind of text?

- Typical NLP pipeline means that degraded performance has a knock-on effect along the chain
- Short sentences confuse language identification tools
- Linguistic processing tools have to be adapted to the domain



- Retraining individual components on large volumes of data
- Adaptation of techniques from e.g. SMS analysis
- Development of new Twitter-specific tools (e.g. GATE's TwitIE)
- But....lack of standards, easily accessible data, common evaluation etc. are holding back development

TwitIE to the rescue



5. Sentiment Analysis for Social Media

Opinion Mining

- Along with NER, opinion mining is a key component in social web analysis
- NER: names of people, organisations, locations
- Opinion mining: what sentiments are being expressed?



Why do we want to find opinions?

- Opinion mining allows us to answer questions such as:
 - What are the opinions on crucial social events and the key people involved?
 - How are these opinions distributed in relation to demographic user data?
 - How have these opinions evolved?
 - Who are the opinion leaders?
 - What is their impact and influence?

Finding Opinions

A positive opinion about Romney



Romney was the perfect candidate, and he was the President this country needs.

A negative opinion about the Republican volunteers



Such apathy among the Republican volunteers is disgusting.

We analyse the texts and classify opinionated statements with:

- a polarity (positive or negative)
- a score (strength of opinion)
- a target (which entity or event the opinion is about)

Finding Opinions is not trivial

- We can use sentiment dictionaries to look up words like “disgusting” and “perfect” and match them to a sentiment
- But this isn't enough on its own.
- We have to make sure to match the sentiment to the correct target (entity)
- We have to deal with negative words and their scope
 - “Happy” and “not happy” have opposite sentiment
 - But “not great” does not imply negative sentiment
- We have to deal with things like sarcasm, especially in tweets.

“Aahh how sweet it is to wake up to ignorance and stupidity :-)”

But there are lots of tools that “analyse” social media already....

- Streamcrab <http://www.streamcrab.com/>
- Semantria <http://semantria.com>
- Social Mention <http://socialmention.com/>
- Sentiment140: <http://www.sentiment140.com/>
- TipTop: <http://feeltiptop.com/>

Why are these sites unsuccessful?

- They don't work well at more than a very basic level
- They mainly use dictionary lookup for positive and negative words
- Or they use ML, which only works for text that's similar in style to the training data
- Things like sarcasm which occur less frequently may not get picked up
- They classify the tweets as positive or negative, but not with respect to the keyword you're searching for
 - keyword search just retrieves any tweet mentioning it, but not necessarily about it as a topic
 - no correlation between the keyword and the sentiment

“Positive” tweets about fracking

- *Help me stop fracking. Sign the petition to David Cameron for a #frack-free UK now!*
- *I'll take it as a sign that the gods applaud my new anti-fracking country love song.*
- *#Cameron wants to change the law to allow #fracking under homes without permission. Tell him NO!!!!*

Death confuses opinion mining tools

- Opinion mining tools are good for a general overview, but not for some situations



Whitney Houston wasn't very popular...

Twitter Sentiment

 Tweet < 273

 Like < 319

 +1 < 20

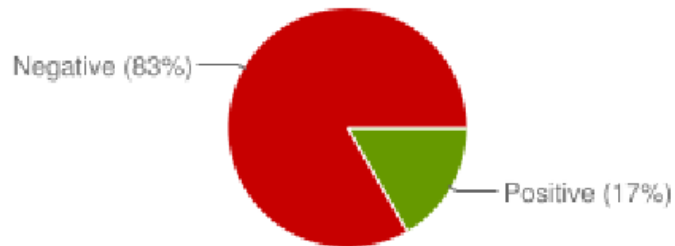
"Whitney Houston"

Search

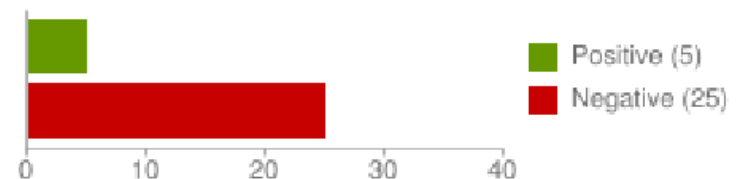
[Save this search](#)

Sentiment analysis for "Whitney Houston"

Sentiment by Percent



Sentiment by Count



Or was she?

Tweets about: "Whitney Houston"

bazyboy25: **Whitney houston**...too soon? #CelebritiesThatLookLikeTheyStank

Posted 5 minutes ago

TeghanSimone: Radio playing **Whitney Houston**.. I swear I'm about to cry... So sad

Posted 5 minutes ago

JB3LL: hoes about to get **whitney houston**'d tonight! #TheWalkingDead

Posted 5 minutes ago

derickaadamss: "@indreamville_ : Twitter I'm curious who do you think had more problems Michael Jackson or **Whitney Houston**???"
<<<< **Whitney Houston!**

Posted 5 minutes ago

charlottesteer4: Listening to **Whitney Houston** loveeeee songsss <3 she's amazing <3

Posted 5 minutes ago

DionneHeraty40: @Sbarry25 The reason why **Whitney Houston** died at only 41 <http://t.co/JJKRDjbj>

Posted 5 minutes ago

ShortySoooFine: #musicwasbestwhen legends like James brown, Michael Jackson, **Whitney Houston** still lived.

Posted 5 minutes ago

CarlmannJohnson: Pray for Bobby Brown!!! He lost his ex-wife **Whitney Houston** and his dad Herbert Brown... Prayers up for you!!

Posted 5 minutes ago

LonelySpaceman: Is it bad that I thought **Whitney Houston** was already dead?

Posted 5 minutes ago

eatmy_CHOCLATE: My aunt in there playing **Whitney Houston** making me sad

Posted 5 minutes ago

Case study:
Rule-based Opinion Mining on Tweets

Why Rule-based?

- Although ML applications are typically used for Opinion Mining, there are advantages to using a rule-based approach when training data isn't easily available
- For example, working with multiple languages and/or domains
- Rule-based system is more easily adaptable
- Novel use of language and grammar makes ML hard
- ML struggles to deal with the nitty-gritty

GATE Components

- TwitIE
 - structural and linguistic pre-processing, specific to Twitter
 - includes language detection, hashtag retokenisation, POS tagging, NER
- (Optional) term recognition using TermRaider
- Sentiment gazetteer lookup
- JAPE opinion detection grammars
- (Optional) aggregation of opinions
 - includes opinion interestingness component

Basic approach for opinion finding

- Find sentiment-containing words in a linguistic relation with terms/entities (opinion-target matching)
 - life flourishing in Antarctica
- Dictionaries give a starting score for sentiment words
- Use a number of linguistic sub-components to deal with issues such as negatives, adverbial modification, swear words, conditionals, sarcasm etc.

A positive sentiment list


- awesome category=adjective score=0.5
- beaming category=adjective score=0.5
- belonging category=noun score=0.5
- benefic category=adjective score=0.5
- benevolently category=adverb score=0.5
- caring category=noun score=0.5
- charitable category=adjective score=0.5
- charm category=verb score=0.5

A negative sentiment list

Examples of phrases following the word “go”:

- down the pan
- down the drain
- to the dogs
- downhill
- pear-shaped

A positive tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Life flourishing in Antarctic fjords despite climate change <http://t.co/7TWq00>

◀ ▶ ✖ ✎ ▶ ▶ 🔔 ✕


SentenceSentiment ▼

<input type="radio"/>	comment	▼	▼	✖
<input checked="" type="radio"/>	entity_string	▼	Antarctic ▼	✖
<input checked="" type="radio"/>	polarity	▼	positive ▼	✖
<input checked="" type="radio"/>	rule	▼	SentenceEntitySentiment ▼	✖
<input type="radio"/>	sarcasm	▼	no ▼	✖
<input checked="" type="radio"/>	score	▼	0.5 ▼	✖
<input checked="" type="radio"/>	sentiment_string	▼	flourishing ▼	✖
<input checked="" type="radio"/>		▼	▼	✖




Type
SentenceSen ▶ Open Search & Annotate tool

=positive, rule=Sentence

A negative tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Parody video exposes P&G's use of dirty palm oil <http://t.co/Pnb3eLmTHE> #Climate #Change #Solar


◀ ▶  ◀ ▶  

SentenceSentiment ▼



<input type="radio"/> comment	▼	▼	✗	
<input checked="" type="radio"/> polarity	▼	negative	▼	✗
<input checked="" type="radio"/> rule	▼	SentenceSentiment	▼	✗
<input type="radio"/> sarcasm	▼	no	▼	✗
<input checked="" type="radio"/> score	▼	-0.5	▼	✗
<input checked="" type="radio"/> sentiment_string	▼	dirty	▼	✗
<input checked="" type="radio"/>	▼	▼	▼	✗

▶ Open Search & Annotate tool

A Sarcastic Tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Nice to know the people in charge, who have so much power, are making smart decisions... #sarcasm #climatechange
[http://www.theguardian.com/environment/planet-oz/2014/feb/24/climate-change-dick-warburton-sceptic-australia-renewable-energy-target-review?CMP=tw_tfd ...](http://www.theguardian.com/environment/planet-oz/2014/feb/24/climate-change-dick-warburton-sceptic-australia-renewable-energy-target-review?CMP=tw_tfd)

SentenceSentiment

<input type="radio"/>	comment			<input type="checkbox"/>
<input checked="" type="radio"/>	polarity	negative		<input type="checkbox"/>
<input checked="" type="radio"/>	rule	SentenceEntitySentiment		<input type="checkbox"/>
<input type="radio"/>	sarcasm	yes		<input type="checkbox"/>
<input checked="" type="radio"/>	score	-0.5		<input type="checkbox"/>
<input checked="" type="radio"/>	sentiment_string	Nice		<input type="checkbox"/>
<input checked="" type="radio"/>				<input type="checkbox"/>

Features

rule-SentenceEntitySentiment sarcasm=yes score

Analysing Hashtags



What's in a hashtag?

- Hashtags often contain smushed words
 - #SteveJobs
 - #CombineAFoodAndABand
 - #southamerica
- For NER we want the individual tokens so we can link them to the right entity
- For opinion mining, individual words in the hashtags often indicate sentiment, sarcasm etc.
 - #greatidea
 - #worstdayever



How to analyse hashtags?

- Camelcasing makes it relatively easy to separate the words, using an adapted tokeniser, but many people don't bother
- We use a simple approach based on dictionary matching the longest consecutive strings, working L to R
 - #lifeisgreat -> #-life-is-great
 - #lovinglife -> #-loving-life
- It's not foolproof, however
 - #greatstart -> #-greats-tart
- To improve it, we could use contextual information, or we could restrict matches to certain POS combinations (ADJ+N is more likely than ADJ+V)

My friend Barry likes Apple products



Barry Norton

cannot wait to see what Apple's new products are

Or does he?



Barry Norton

cannot wait to see what Apple's new products are #sarcasm



Like · Comment · Share · 3 hours ago · 👤

👍 3 people like this.



Barry Norton (The tag is for Diana - sometimes she can't tell)

3 hours ago · Unlike · 👍 5

What does sarcasm do to polarity?

- Sarcasm often indicated by hashtags in tweets such as #sarcasm, #irony, #whoknew etc.
- It's very hard to identify sarcasm outside these parameters
- In general, when someone is being sarcastic, they're saying the opposite of what they mean
- So as long as you know which bit of the utterance is the sarcastic bit, you can simply reverse the polarity/ But it's not that easy.

Eating breakfast food for lunch. Living the dream.

#toast #rebel #sarcasm

- If there is no polarity on the original statement, the sarcastic version is probably negative

It's not like I wanted to eat breakfast anyway #sarcasm

- Sarcasm can be positive. “Having a terrible time here in Italy.”

Using Machine Learning for Opinion Mining

- If we can collect enough manually annotated training data, we can also use an ML approach for this task
- Product reviews: use star-based rating (but these have flaws)
- Other domains, e.g. politics: classify sentences or tweets (the ML *instances*), many of which do not contain opinions.
- So the ML *classes* will be *positive, neutral and negative*
- (Some people classify *neutral* and *no opinion* as distinct classes, but we find the distinction too difficult to make reliably)

Training on tweets

- You can use hashtags as a source of classes
 - Example: collect a set of tweets with the **#angry** tag, and a set without it, and delete from the second set any tweets that look angry
 - Remove the **#angry** tag from the text in the first set (so you're not just training the ML to spot the tag)
 - You now have a corpus of manually annotated angry/non-angry data
- This approach can work well, but if you have huge datasets, you may not be able to do the manual deletions
- You can also train on things like **#sarcasm** and **#irony**

6. Applications

The Decarbonet Project

- Scientists predict adverse consequences to our climate unless stronger actions are taken
- Collective awareness about many climate change issues is still problematic
- We are exposed to vast amounts of conflicting information
- Hard to know what is accurate and relevant
- DecarboNet: “A Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioural Change”
- 3-year EU project, started October 2013



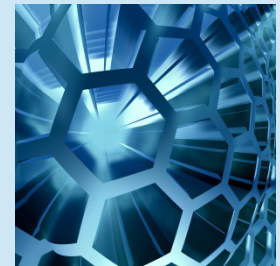
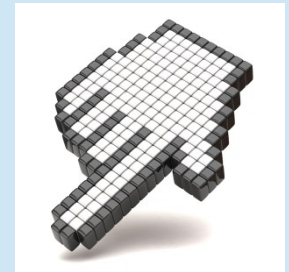
waag society



This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 610829.

DecarboNet Objectives

- Raise Individual and Collective Awareness
- Trigger Behavioural Change and Foster Social Innovation
- Analyse Behavioural Patterns and Information Diffusion



Social media analysis for climate change

- Do we really know what people mean when they tweet?
- NLP tools for the automatic discovery of new insights, by automatically extracting information from social media.
- Extracted information can be linked together to form new facts or to allow new hypotheses to be explored further.
 - What arguments for and against man-made causes of climate change develop in social media?
 - What impact does this information have?
 - How do people's opinions change over time?
 - What kinds of topics are most engaging for social media users?

Media Watch for Climate Change

www.ecoresearch.net/climate/

Climate Change Media Watch

Search: "climate change" 0 - 50 out of 6147 25 Jul 2014 - 25 Sep 2014

News Social Media Eco-NGOs Fortune 1000 Cluster Map Geo Map Tag Cloud

General

- Adaptation 1158
- Climate Change 6147
- Climate Policy 140
- Climate Science 156
- Mitigation 518

Renewable Energy

- Biomass 494
- Geothermal 345
- Hydro 369
- Solar 1937
- Wind 1138

World Summits

- Durban (COP17) 406
- Qatar (COP18) 1372
- UNCSD (Rio+20) 8
- Warsaw (COP19) 6

Events and Campaigns

- Earth Day 76
- Earth Hour 10
- IPCC Report 250

Frequency Sentiment Disagreement

40% 55%

■ Negative ■ Neutral ■ Positive

Geo Map

Associations Search History

- "climate change" 6147
- summit 5094
- ban 7264
- secretary-general ban [...] 806
- dicaprio 354
- islamic 8412
- iraq 9596
- secretary-general ban 866
- climate summit 677
- secretary-general 1805
- un general 168
- brown 11852
- treaty 1519
- obama 14937
- syria 6945
- un climate 271

Documents Quotes Word Tree Entities Entity Map Sources Source Map

09/24 0.0 **11 ways climate change affects us**
 « 11 ways climate change affects the world. << Previous Next >>. The consequences of climate change, Topic A this week at the U.N. Climate Summit in New York, go far beyond warming temperatures, which scientists say are melting the polar ice caps and raising sea levels. Here's a look at 10 other key effects of climate change, some of which may surprise you. »
 edition.cnn.com

09/23 -0.8 **Companies propose Wyoming wind farm plan to power Los Angeles**
 map - which countries are responsible for climate change? 3. 3. UN climate change summit in N...
 theguardian.com

09/23 -0.6 **Google to cut ties with rightwing lobby group over climate change 'lies'**
 Street protesters 3. 3. Flood Wall Street climate change protest holds ground in lower Manhatt...
 theguardian.com

09/22 0.0 **Police face off with Flood Wall Street protesters in climate change march**
 climate change 2. 2. Flood Wall Street climate change protest holds ground in lower Manhattan ...
 theguardian.com

09/24 0.0 **Climate Change Is Hurting Food Security: Why Our Crops Are Threatened**
 business-news.thestreet.com/philly/story/climate-change-hurting-food-security-why-our-cro...
 philly.com

09/22 0.0 **Prince Charles urges action against climate change 'before it is too late'**
 York as they meet for major international climate change talks.
 telegraph.co.uk

09/23 +0.8 **Video: Watch Leonardo DiCaprio's UN speech on climate change**
 Watch Leonardo DiCaprio's UN speech on climate change. Hollywood superstar and prominen...
 theglobeandmail.com

09/22 -1.0 **Emma Thompson: climate change deniers are 'bonkers' video**
 Emma Thompson says climate change deniers are 'bonkers'. She speaks as she joins thousands ...
 theguardian.com

Tag Cloud

abbott actor alec algae audubon august blasio book border brown california celsius church clinton coal copenhagen coral crist drought extremism extremist film flood govern humanitarian ice immigration india iraq islam klein labour lake layer marriage naomi oba paris party peace poverty reef ret rockefell scotland scott scottish summit syria tories treaty wall water

Keywords Edges: 5 ▾ +

Summary

- Opinion mining and social media analysis are **hard** and therefore **error-prone** (despite what vendors will tell you about how great their product is)
- It's very unlikely that an off-the-shelf tool will do exactly what you want, and even if it does, performance may be low
- Opinion mining tools need to be **customised** to the task and domain
- For tasks that mainly look at aggregated data, basic tools work quite well, but it's hard to know what they're really doing
- Lots of exciting tools being developed, but still plenty of research to be done
- Which is handy as it keeps us all in a job :-)

Acknowledgements and further Information

- Research partially supported by the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D (FP7) DecarboNet (610829)
- DecarboNet project <http://www.decarbonet.eu>
- GATE website <http://gate.ac.uk>
- Slides from GATE course on Opinion Mining
- Opinion mining demo: <http://demos.gate.ac.uk/arcomem/opinions/>

