

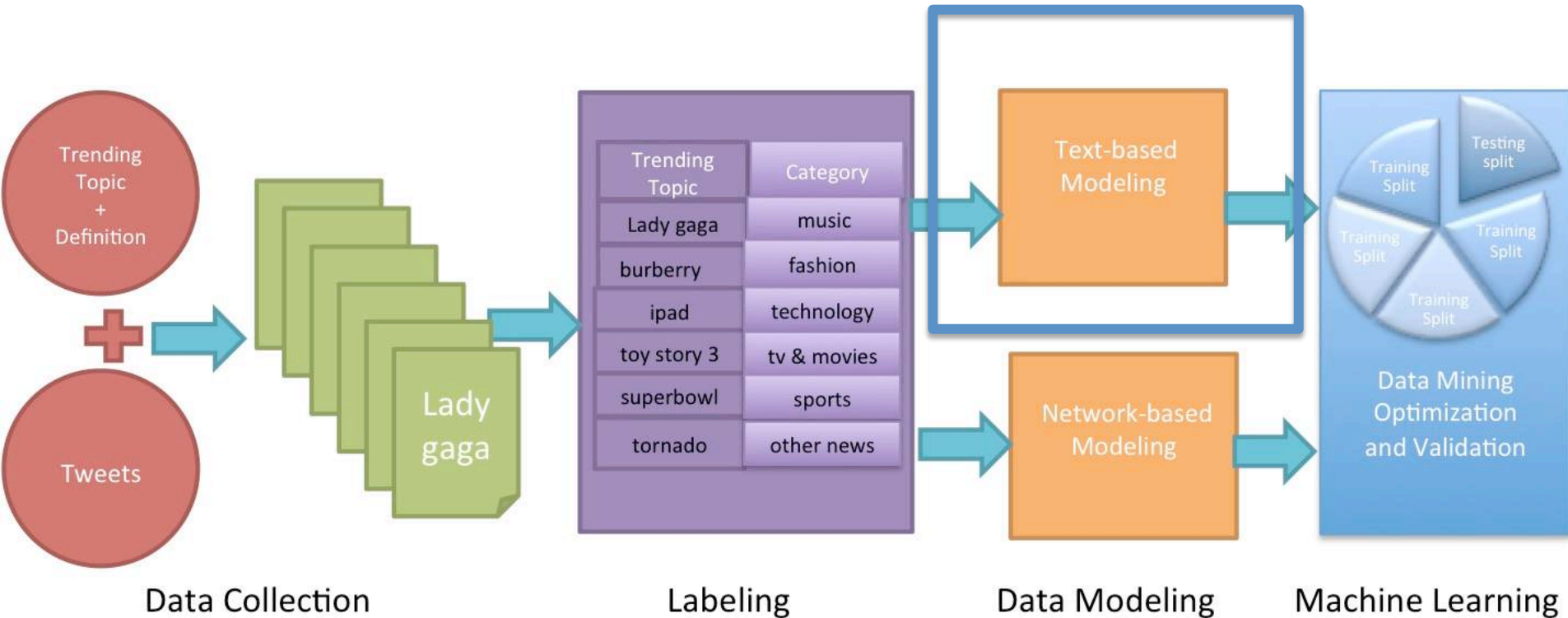
Social Analytics for BI



PART C

Text mining

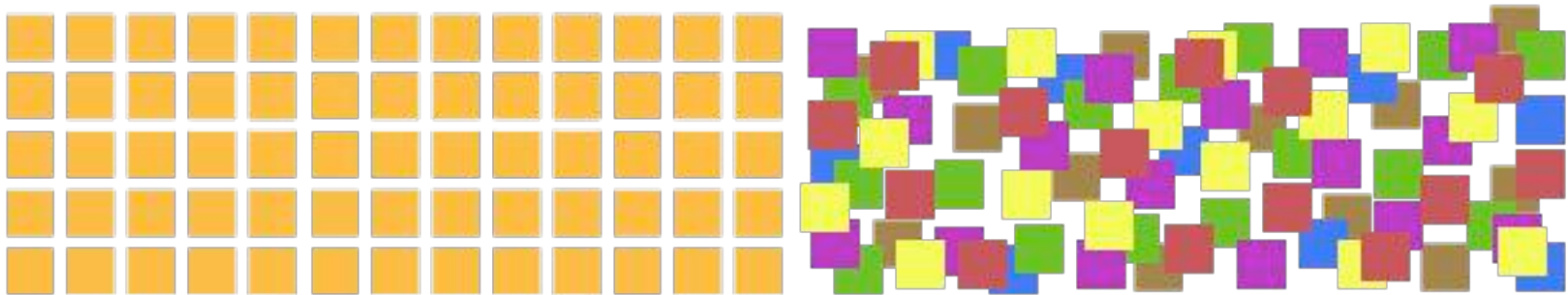
Extracting Social Network data for BI: workflow



Two types of data: network-based and **text-based**

The power of unstructured data

“80% of business-relevant information originates in unstructured form, primarily text.”

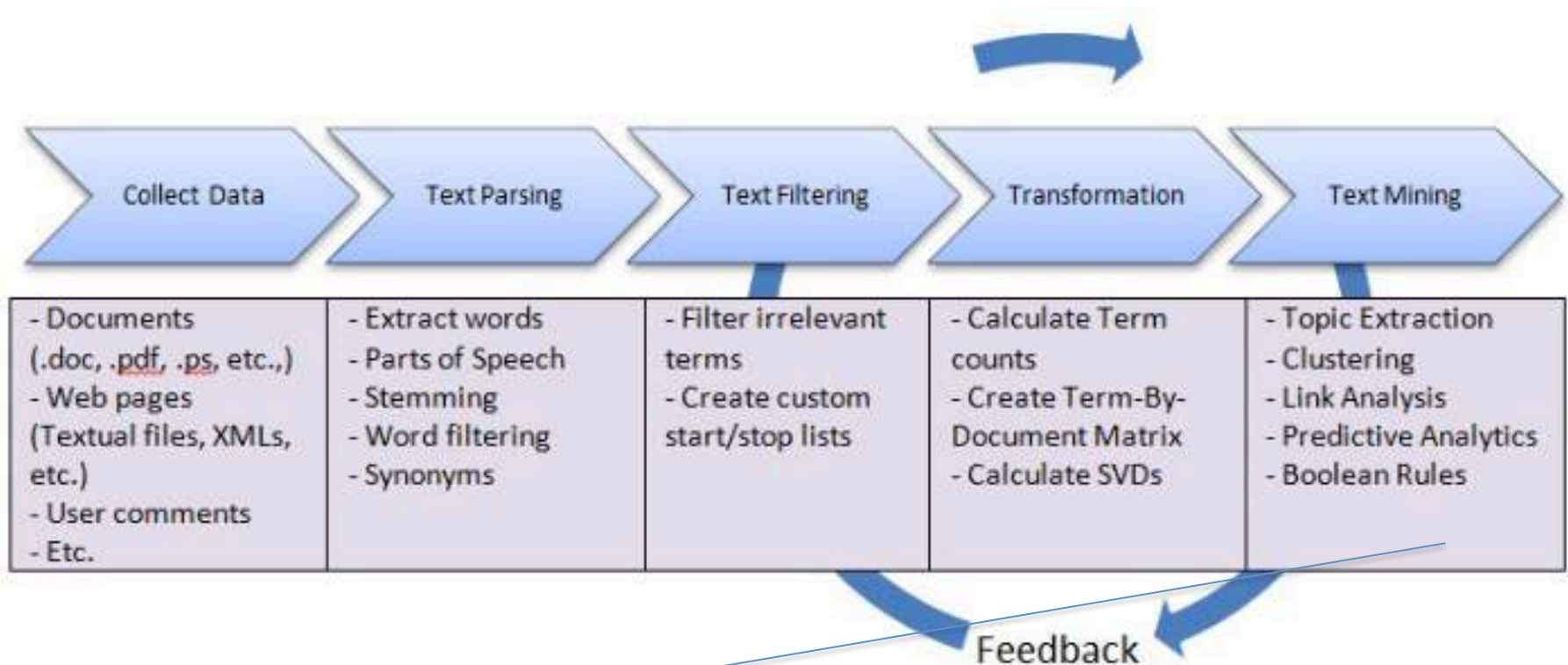


Structured Data vs. **Unstructured Data**

Text Mining

- Text mining is an emergent technology attempting to extract useful information (and knowledge) from unstructured data
- Text mining is an extension of data mining to textual data
- Social networks contain a lot of information in textual form, such as posts, links, blogs, news articles, emails..

Text processing workflow (more in detail)



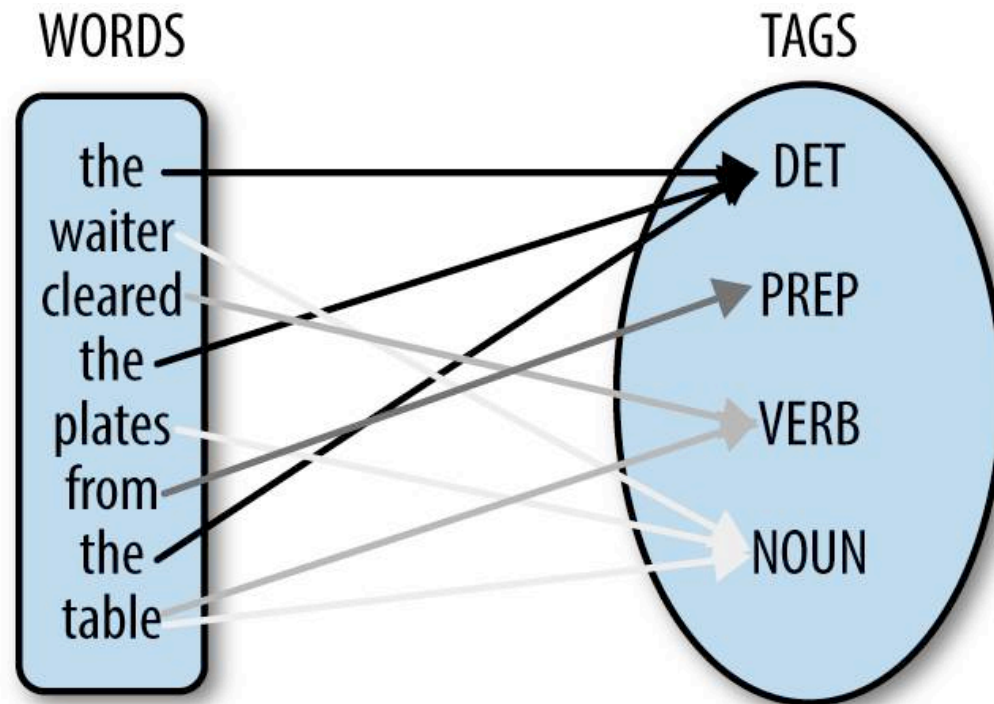
TEXT MINING
APPLICATIONS

Basics of text mining

- Steps of text analytics
- Text representation

Basic processing steps: POS

- Part of speech tagging: labelling words with part of speech



Basic processing steps: lemmas

- Reduce inflections or variant forms to base form
 - *am, are, is* → *be*
 - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*

Why is useful?

LEMMA QUERY RESULT

QUERY

l:happy

happy, happier, happiest

l:go

go, goes, going, gon(na), went, gone

l:man

man, men

l:un*

understand, unit, under, ...

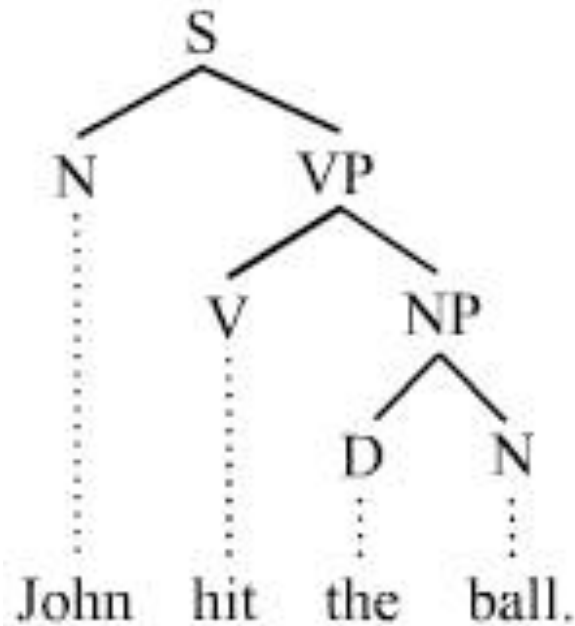
l:*ion

organization, organizations, mention, mentioned, ...

l:s.ng

sing, sings, sang, sung, singing, song, songs

Basic processing steps: syntactic analysis



Constituency-based parse tree

This is a rather complex task. Most often, systems recognize simpler structures, e.g., subject-verb-object, adjective-noun,

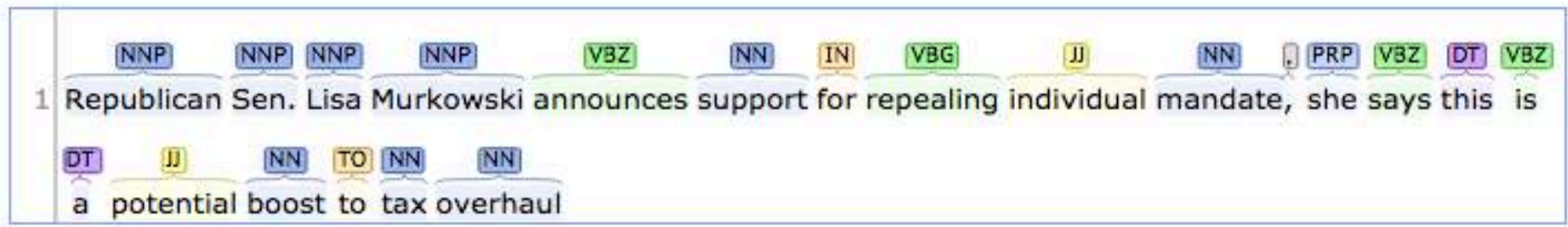
Stanford CoreNLP

Output format:

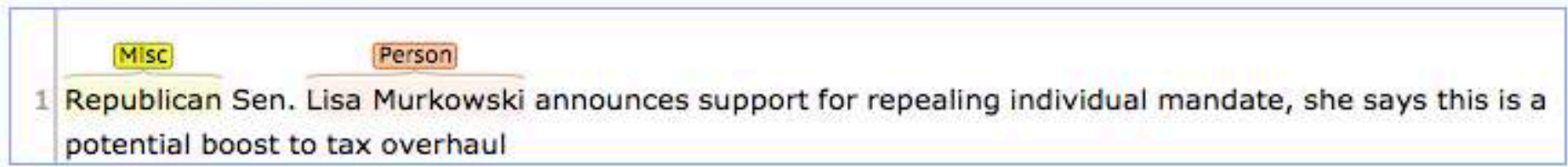
Please enter your text here:

Republican Sen. Lisa Murkowski announces support for repealing individual mandate, she says this is a potential boost to tax overhaul

Part-of-Speech:



Named Entity Recognition:

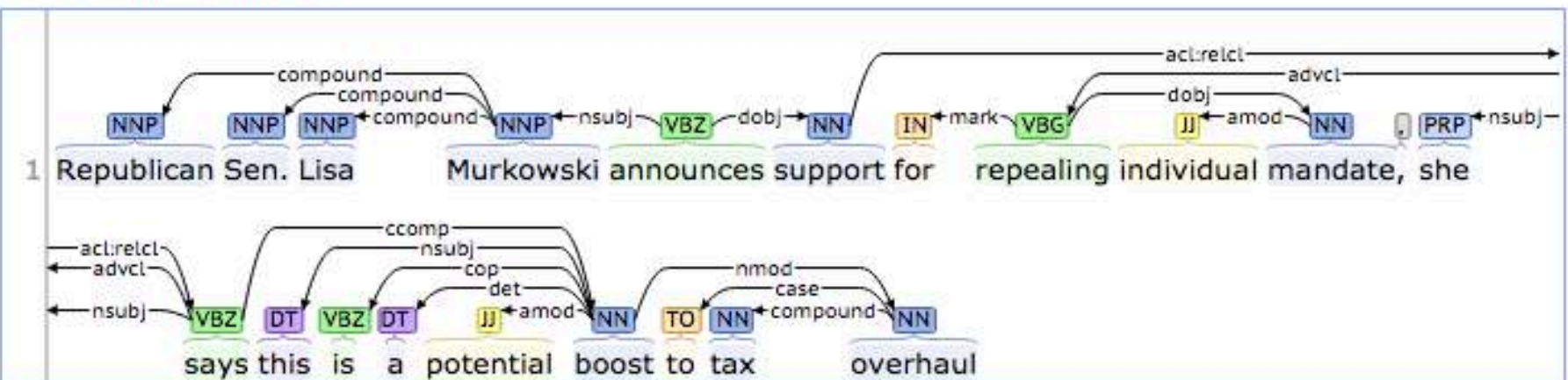


Putting all together (2)

Coreference:



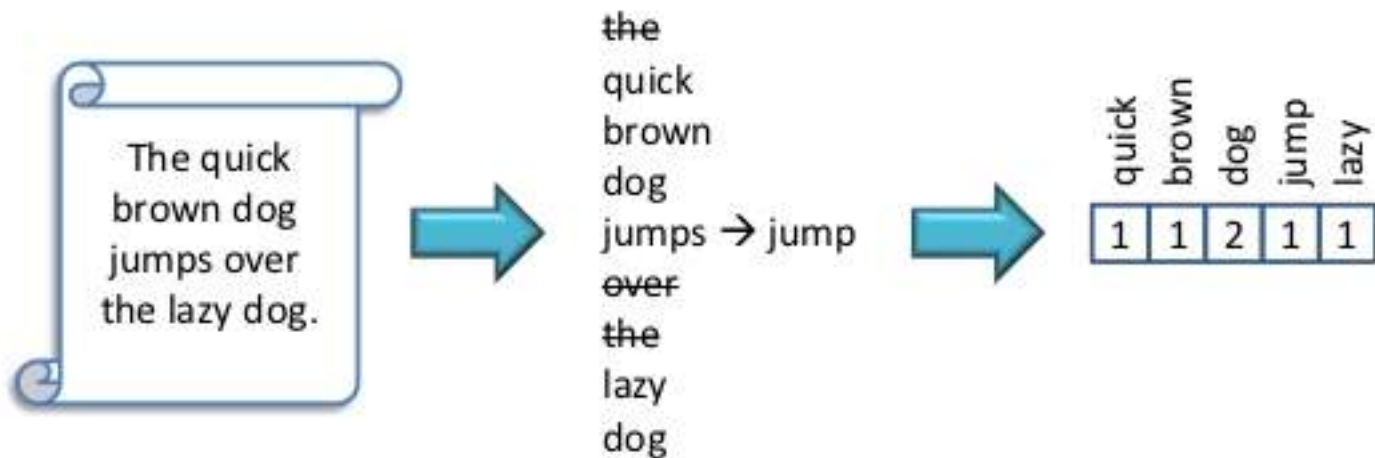
Basic Dependencies:



How do we represent texts? Bag of words

Bag of words

- Tokenize
- Remove stop words
- Lemmatize
- Compute weights



Seems simple.. however..

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach our eyes. For a long time, the retinal image was considered as a movie screen. Hubel and Wiesel discovered that the visual centers in the brain analyze the image in a more complex way following the discovery of the various cells of the cortex. Hubel and Wiesel have demonstrated that the message about the image falling on the retina undergoes a cell-by-cell analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.



**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be created by a predicted 30% increase in exports to \$750bn, compared with \$575bn in 2004. China's trade surplus of \$660bn. The government has also announced that it will allow the yuan to rise against the dollar, but it will take its time and tread carefully before allowing the yuan to rise further in value.



**China, trade,
surplus, commerce,
exports, imports, US,
yuan, bank, domestic,
foreign, increase,
trade, value**

Why bag of words?

– Basic idea:

- Keywords are extracted from texts.
- These keywords describe the (usually) topical content of Web pages and other text contributions.
- Each unique word in a corpus of documents (web pages, social messages..) = one attribute
- Each document is a record with non-zero weight for each word in that document, zero weight for other words

➔ Words become “attributes”, whose values can be binary (the word *is* or **is not** in a text), or real numbers (e.g. the relative frequency of a word in the text)

Example

the dog is on the table

0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the

The example considers a vocabulary of 8 words – in the reality the vocabulary has millions of words – document records have millions of attributes

Which words should we care about? A complex problem

- E.g.: Companies assume that people refer to them by name
- Big mistake!!
- There are multiple dimensions for reference: hashtags, names of people (e.g., managers), products, and for each: abbreviations, initials, nicknames
- Additional problems: ambiguity, synonyms..

Example: search “Watson” on Twitter Search

The screenshot displays a Twitter search interface for the term "Watson". At the top, two profile cards are shown side-by-side. The left card is for Emma Watson, featuring her profile picture, a "Segui" button, her name with a verified badge, her handle @EmmaWatson, and a bio that identifies her as an actor and UN Women Global Goodwill Ambassador. The right card is for Deshaun Watson, also with a "Segui" button, his name with a verified badge, handle @deshaunwatson, and a bio that includes his NFL team and social media handles. Below these profiles, a tweet from CommentWise (@oncommentwise) is visible, mentioning "Watson" and linking to a video. The bottom portion of the image shows a tweet from Digital Marketing (@DollyRayDigital) featuring a video thumbnail with the text "IBM Watson Education Personalizing the teaching ...".

Emma Watson ✓
@EmmaWatson
Actor & @UN_Women Global Goodwill Ambassador. Facebook: EmmaWatson Instagram: EmmaWatson Goodreads: OurSharedShelf

Deshaun Watson ✓
@deshaunwatson
God 1st! •815™ •GodSpeed •Memo™ •six. #NEGU For Football Inquiries Contact: @DavidMulugheta For Marketing & Business Inquiries...

CommentWise @oncommentwise · 4 min
Evidence, my dear **Watson!**: Kajal Iyer As a country we have grown up on fictional detectives who solve cases by... dlvr.it/PvJqQD

Digital Marketing 🙌 @DollyRayDigital · 2 h
#IBM #Watson Education Personalizing the teaching & learning experience youtu.be/ZvGhbJ8V8eA #Cognitive #ArtificialIntelligence #AI #IoT

IBM Watson Education Personalizing the teaching ...
youtube.com

So.. rather than “bag of words”, bag of concepts

The screenshot shows a software interface for text annotation. At the top, there are tabs for "Annotation Sets", "Annotations", "Co-reference Editor", and "Text", along with a magnifying glass icon. The main text area contains several paragraphs with various words and phrases highlighted in different colors. A vertical list on the right side of the interface shows a set of concepts, each with a colored background and a checkbox. The concepts include Address, DEFAULT_TOKEN, Date, FirstPerson, JobTitle, Location, Lookup, Money, Organization, Percent, Person, Sentence, SpaceToken, Split, Temp, TempLocation, Title, Token, and Unknown. The "Date" and "Person" checkboxes are checked. Below the text area, there is a table with columns for Type, Set, Start, End, and Features. The table lists several annotations with their corresponding features. At the bottom left, it says "86 Annotations (1selected)" and at the bottom right, there is a "New" button.

Background to and purpose of visit

The enquiry had come via **Jean Coldham** of The British **Midlands** and originally from the UKTI in the British Consulate in **Chicago**. UKTI had been alerted of the possible expansion project via **Sterigenics HQ** in **Chicago**, who had supplied **Ron Peacock**'s contact details. **Ron** had previously helped emda with a number of research projects, agreeing to answer questions and give background information etc.. This was the first indication, however, of a possible expansion project for the Company in **Somercotes**, part of the EM SFI grant assisted area. Local contact would be followed by a meeting with the Group FCO in the **US**, involving UKTI staff and **Jean**.

NB **Ron Peacock** originally set the Company up in **1992**, is nominally based in **European HQ** in **Belgium**, although he spends the bulk of his time in **UK**. He is now responsible for special projects and capital investments.

Company Background

The Company was set up in **1992** as a spin off of **Griffith Laboratories** (next door) as **Griffith Microscience**. In **1998** it was floated on **Nasdaq** and eventually acquired by **IBA S & I**. IBA then decided that the Company was 'non-core', and sold it in turn to **Sterigenics**, a **US** owned sterilisation company. Relatively recently the **Sterigenics Group** was acquired by the **PPM Capital**, the venture capital group.

World HQ is in **Chicago**, **European HQ** in **Leuven, Belgium** and the Group has recently built new facilities in **Germany** and **China**.

Type	Set	Start	End	Features
Location		1684	1695	{locType=[null], matches=[2982, 2955, 2970], rule1=LocationPost, rule2=LocFinal}
Location		1700	1707	{locType=[null], matches=[2956, 2971], rule1=InLoc1, rule2=LocFinal}
Location		1752	1754	{locType=[null], matches=[2977, 2979, 2957, 2974], rule1=InLoc1, rule2=LocFinal}
Organization		1836	1861	{matches=[2958, 2980, 2975, 3097, 3099], orgType=[null], rule1=OrgXEnding, rule2=OrgFi
Date		1876	1880	{kind=date, matches=[2959, 2954], rule1=YearContext1, rule2=DateOnlyFinal}
Organization		1898	1919	{orgType=[null], rule1=OrgXBase, rule2=OrgFinal}

86 Annotations (1selected)

New

Classification example: mobile apps users



In classification, categories are assigned. Based on messages, classify user in one of these cats

Text Mining: popular applications

- Named entities recognition: Who? What? Where?
- Topics: what is the “buzz”?
- Categories/clusters: what is this about? What is in common?
- Sentiment: how do they feel about?
- Information extraction: what are the relevant facts?

Message Clustering: example

related themes

over the last 50 minutes

#wikipedia	wolfram alpha
wolframalpha.com/	computational
fructose in onion	computational
with your input	knowledge
google killer	engine
wolfram	via @time
wolfram alpha	been alive
@time	bit.ly/sa6be
days old	got the number
meaning of life	wrong answer
i have tried	queries failed
on the day	eg
was born	i was born
engine	search engine
#fail	input

tweets by theme

“queries failed”

: not finding #WolframAlpha all I had hoped, data-driven queries failed every time ("ave # job applicants interviewed to fill position") *peterjwolfgang*

So far #WolframAlpha has failed for the two queries that I have tried that weren't a play test #fail *mercad* [show 1 similar tweet »](#)

(drilldown 1)

“google killer”

#WolframAlpha is going to be huge, but it's not a google killer, and most people will never use it *briggs1*

@charlesf11 Yeah, it's bugging me that people are calling it a Google killer. They're not the same animal. #wolframalpha #wikipedia #google

mattbramanti

#WolframAlpha is no Google killer, but it's the smartest calculator ever.

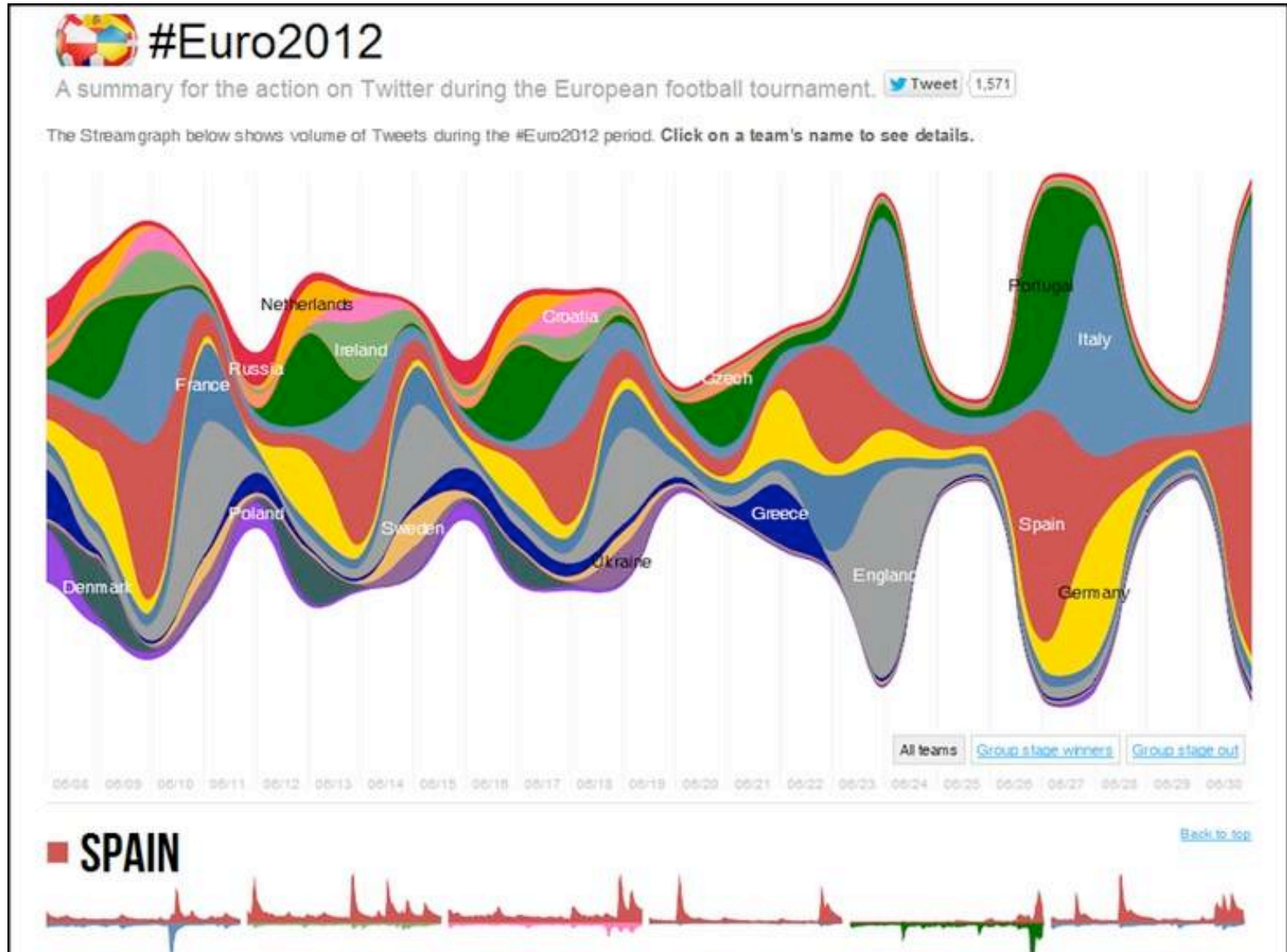
PrestonStahley

Remember: in clustering we have no known categories. Messages are clustered based on similarity of attributes (bag of words)

Topic extraction

- In topic extraction, the purpose is to detect group of words that are representative of a “trending” discussion topic
- Based on the analysis of terms that co-occur in many messages
- Example: *Watson, cognitive, IBM, AI*
- Example 2: *#WomenBoycottTwitter, women, sexual, solidarity*

Tracing topics (stream graphs)



Tracing Topics (TamTamy-Reply)



Home



Social Network



Social Buzz

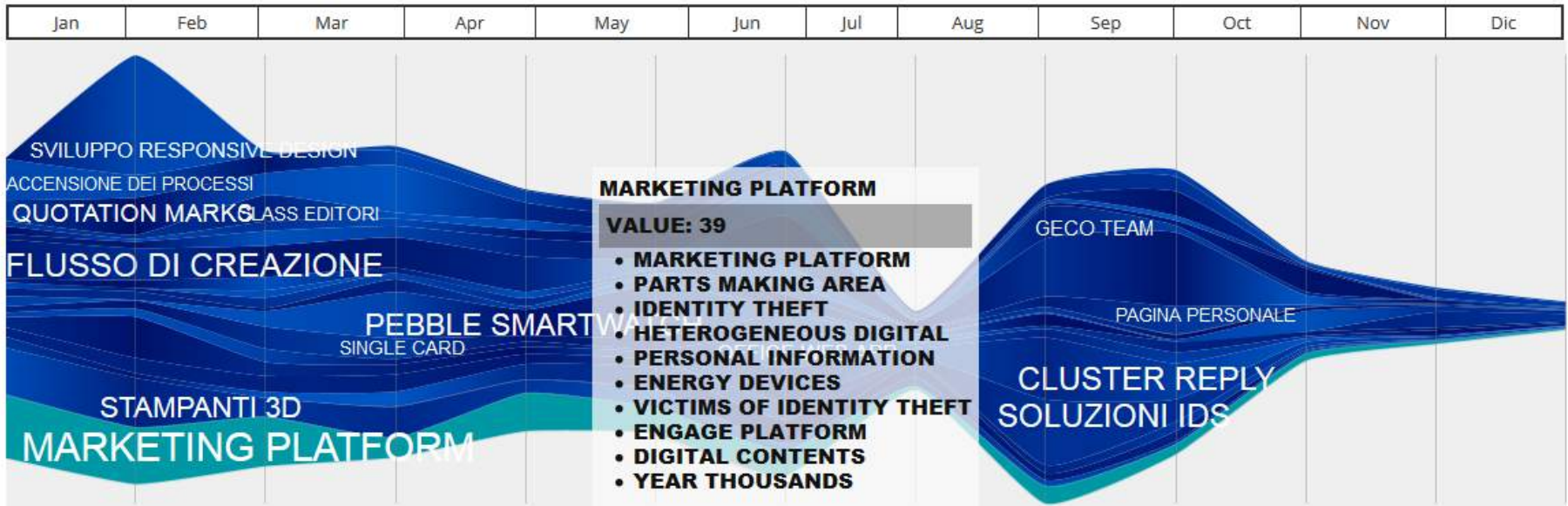


Statistics



People

Personalize Streamgraph Parameters



Opinion mining



Introduction – facts and opinions

- Two main types of textual information on the Web.
 - Facts and Opinions
- Current search engines search for facts (assume they are true)
 - Facts can be expressed with topic keywords.
- Search engines do not search for opinions
 - Opinions are hard to express with a few keywords
 - What do people think of Motorola Cell phones?
 - Current search ranking strategy is not appropriate for opinion retrieval/search.

Opinions are user-generated content

- **Word-of-mouth on the Web**
 - One can express personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs ... (called the user generated content.)
 - They contain valuable information
 - Web/global scale!!
 - **Our interest:** to mine opinions expressed in the user-generated content
 - A very challenging problem.
 - Practically very useful.

Applications

- **Businesses and organizations:** product and service benchmarking. Market intelligence.
 - Business spends a huge amount of money to find consumer sentiments and opinions.
 - Consultants, surveys and focused groups, etc
- **Individuals:** interested in other' s opinions when
 - Purchasing a product or using a service,
 - Finding opinions on political topics,
- **Ads placements:** Placing ads in the user-generated content
 - Place an ad when one praises a product.
 - Place an ad from a competitor if one criticizes a product.
- **Opinion retrieval/search:** providing general search for opinions
 - Predicting behaviours and trends in finance, medicine, politics

Blah blah blah blah blah blah
blah. **Talking about cell
phones. I bought the new
[redacted] today. It's so
beautiful!** blah blah blah blah
blah blah blah blah blah
blah blah blah blah I **also like**
[redacted]. **However I cannot
recommend the [redacted]
due to its weak battery. It's
a shame.** blah blah blah
blah



Talking about cell phones. I
bought the new [redacted]
today. It's so **beautiful!**

behavior



I also like [redacted] However I **cannot
recommend** the [redacted] due
to its **weak battery.** It's a shame.

brand



feature/attribute



Impact

- 81% of Internet users have done online research on a product 20% do so on a typical day
- Among readers of online reviews between **73% and 87%** report that reviews had a significant influence on their purchase
- Consumers report being willing to pay from **20% to 99%** more for a 5-star-rated item than a 4-star-rated item (the variance stems from what type of item or service is considered);
- 32% have provided a rating on a product, service, or person via an online ratings system, and 30% have posted an online comment or review regarding a product or service.

A formalization of the opinion mining task

- Basic components of an **opinion**:
 - **Opinion holder**: The person or organization that holds a specific opinion on a particular object.
 - **Object**: on which an opinion is expressed (it can be described by features, e.g. for an hotel room: dimension, clean, silent, cost,..)
 - **Opinion**: a view, attitude, or appraisal on an object (or object feature) from an opinion holder.



Opinion mining “grain”

- At the **document (or review) level**:
 - Task: sentiment classification of reviews
 - Classes: positive, negative, and neutral
 - Assumption: each document (or review) focuses **on a single object** (not true in many discussion posts) and contains opinion from a single opinion holder.
 - Example: Movie reviews
- At the **sentence level**:
 - Task 1: identifying subjective/opinionated sentences
 - Classes: objective and subjective (opinionated)
 - Task 2: sentiment classification of sentences
 - Classes: positive, negative and neutral.
 - Assumption: **a sentence contains only one opinion**; not true in many cases.
 - Then we can also consider clauses or phrases.
 - Example: hotel reviews

Opinion Mining Tasks (cont.)

- At the **feature level** (Example: product reviews, usually you want to know opinions on various features of the product to improve or to compare)
 - *Task 1: Identify and extract object features that have been commented on by an opinion holder*
 - *Task 2: Determine whether the **opinions on the features** are positive, negative or neutral.*
 - *Task 3: Group feature synonyms.*
- **Opinion holders**: identify holders is also useful, e.g., in news articles, etc, but they are usually known in the user generated content, i.e., **authors of the posts.**

Feature-Based Opinion Summary

“I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

Feature-Based Summary:

Feature1: Touch screen

Positive: 212

- The touch screen was really cool.
- The touch screen was so easy to use and can do amazing things.

...

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

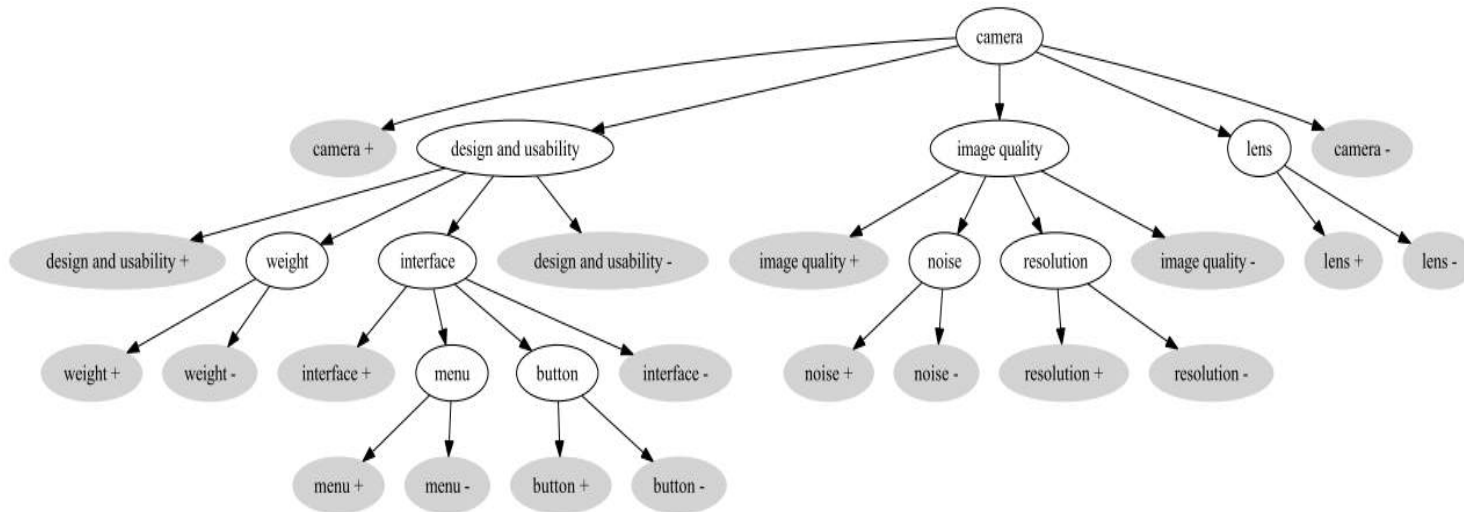
...

Feature2: battery life

...

....

Needs “knowledge” to represent object features



Opinion Analysis: Methods

- Two approaches to the problem:
 1. Machine-Learning (ML) solutions
 2. Lexicon-based solutions
 3. Hybrid solutions
- Each has advantages and disadvantages...

Data Analytics solutions

- Classification
 - Provide an algorithm with lots of **examples**
 - Documents that have been *manually/semi-automatically* annotated with a category
 - *Supervised* learning
 - In our case: e.g., positive/negative reviews (e.g. Tripadvisor)
 - Algorithm extracts “characteristic patterns” for each category and builds a *predictive model*
 - Apply *model* to new text -> get prediction

ML for document classification

- **Bag-of-words** document representation:
document \rightarrow vector (“opinion” words can be considered, or, any word)
 - Example:
 - d_1 = “good.... average... excellent.. good..”
 - d_2 = “okay ..good.. average.. fine..”
 - d_3 = “good... okay.. okay...”
 - Then **Vocabulary** = {“good”, “average”, “excellent”, “fine”, “okay”} and d_1 will be represented as:
 - $d_1 = \{2, 1, 1, 0, 0\}$ if features are frequency-based or
 - $d_1 = \{1, 1, 1, 0, 0\}$ if boolean-based

Example: Happiness

The screenshot shows a Mozilla browser window with the title "happy birthday jenny!!! - Mozilla". The address bar contains "http://community.livejournal.com/birthdaybuckle/". The browser's menu bar includes File, Edit, View, Go, Bookmarks, Tools, Window, and Help. The toolbar shows navigation buttons (Back, Forward, Reload, Stop) and a search button. The browser's status bar at the bottom displays several bookmarks: Home, Bookmarks, Google, GMX - Login zum Mitgli..., Google Scholar, SIGIR 2006—Seattle, mozilla.org, mozillaZine, and mozdev.org.

The main content area displays a birthday message thread. The top message is from an anonymous user, stating "I'm sorry I can't be there to see you open your gift, but I'll see you soon! Much, much love to you. ::kisses::". The "Current Mood" is "chipper". Below the message is a "Link" button and a "Leave a comment" link.

The second message is dated "Fri, Nov. 19th, 2004, 03:36 pm" and is from "phineasjones" with the mood "smurfy!". The message text reads: "whreeeeee!!! it's your birthday!!! ::dances:: this makes three whole birthdays of yours that i've known you for. can you believe it? wacky. i truly hope this one is the very best of all of them. and that this year brings you still more happiness and more love and more fun - all the good things. you deserve them all, dear friend. i love you! ♥". A small profile picture of a man with glasses is shown to the right of the text. The "Current Mood" is "happy". Below the message is a "Link" button and a "Leave a comment" link.

The third message is dated "Fri, Nov. 19th, 2004, 10:15 am" and is from "wilderreentide" with the mood "happy happy day!".

On the left side of the page, there is a mood selection menu with a scrollable list of moods: envious, exanimate, excited, exhausted, flirty, frustrated, full, geeky, giddy, giggly, gloomy, good, grateful, groggy, grumpy, guilty, happy, high, hopeful, and horny. Below the list is an "Options:" section with "Security:" checked, "Auto-Format HTML:", "Current Location:", "Music:", "Mood:", and "Tags:".

Since messages are tagged with a mood by the message author, a machine learning classifier can learn to classify new untagged messages

Bag-of-sentiment-words

I **love** this movie. It's **sweet** but with **satirical** humor. The dialogue is **great** and the adventure scenes are **great fun**...It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times as I **love** it so much, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet.

great	2
love	2
recommend	1
laugh	1
happy	1
.....

Not only words

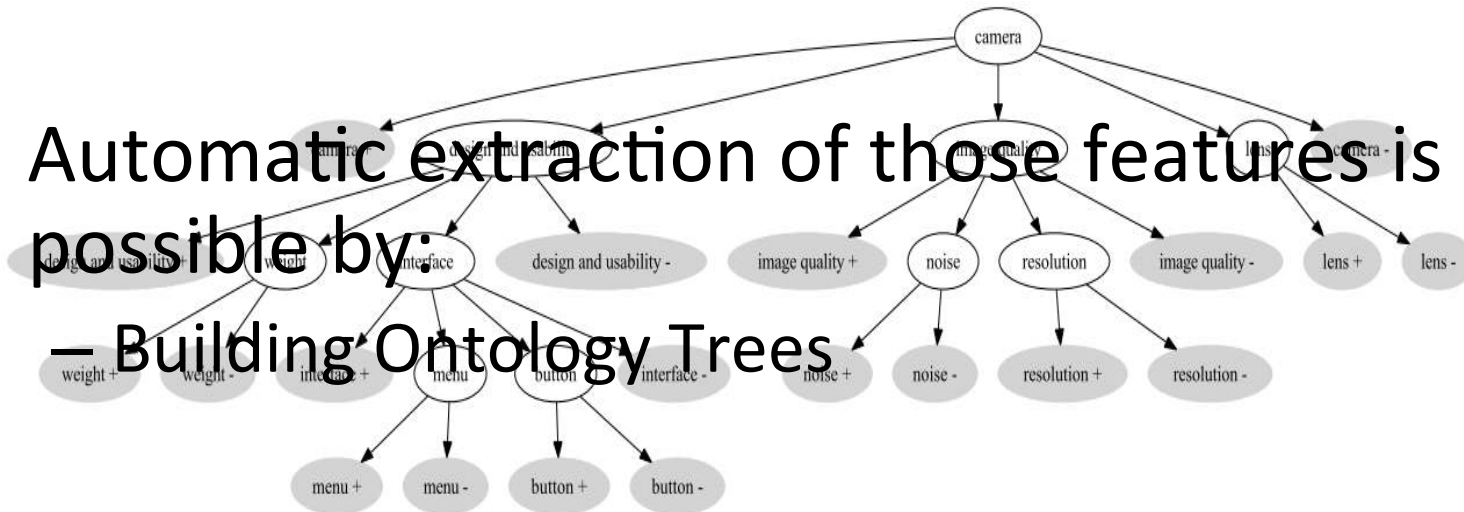
- Typical extensions, focus on *extending/enhancing* the document representation. Instead of/in addition to bag-of-words features, can use:
 - Extra features for *emphasised words* , *special symbols*
 - *wooooow*
 - *exclamations: !!!! ??*
 - *Emoticons* 😊

Feature-based Opinion Analysis

- As discussed, often the *Opinion Object* has different features
 - e.g., camera: lens, quality, weight.
- Often, such an aspect-based analysis is more valuable than a general +/-

- Automatic extraction of those features is possible by:

– Building Ontology Trees



Pros/Cons of the approach

- Advantages:
 - Tend to attain good predictive accuracy
 - Assuming you avoid the typical ML mishaps (e.g., over/underfitting)
- Disadvantages:
 - Need for **training corpus**
 - Solution: automated extraction (e.g., Amazon reviews, Rotten Tomatoes) or crowdsourcing the annotation process (e.g., Mechanical Turk)
 - **Domain sensitivity**
 - Trained models are well-fitted to particular product category (e.g., electronics) **but underperform if applied to other categories** (e.g., tourism)
 - Solution: train a lot of domain-specific models or apply *domain-adaptation* techniques
 - Particularly for Opinion Retrieval, you'll also need to identify the domain of the query!

Example:
"small" is
positive for a
camera,
negative for an
hotel room

Lexicon-based solutions

- Detect/extract the polarity of opinions, based on **affective** dictionaries
- Word-lists where each token is annotated with an ‘emotional’ value
 - e.g., positive/negative words or *words that express anger, fear, happiness, etc.*

Examples of affective dictionaries follow...

- Add **syntactic** and **prose** rules to estimate the overall polarity of text:
 - Negation detection: “the movie **wasn’t** good”
 - Exclamation detection: “great show**!!**”
 - Emoticon detection: “went to the movies 😊”
 - Emphasis detection: “You are go**oooo**d”
 - Intensifier, diminisher word detection: “**Very** good movie” vs. “good movie”

(Basic) lexicon-based approach

1. Detect emotion in two independent dimensions (numbers are weights of positive/negative opinionated words):
 1. Positive: $D_{\text{pos}}: \{1, 2, \dots, 5\}$
 2. Negative: $D_{\text{neg}}: \{-5, -4, \dots, -1\}$
2. (optional) Predict overall polarity by comparing them :
 - If $D_{\text{pos}} > |D_{\text{neg}}|$ then positive
 - Example: “He is brilliant but boring”
 - Emotion(‘brilliant’)=+3
 - Emotion(‘boring’)=-2

$D_{\text{pos}} = +3, D_{\text{neg}} = -2 \Rightarrow \text{positive}$
3. Negation detection: “He isn’t brilliant and he is boring”
 - *Emotion(NOT ‘brilliant’) = -2*
 - *Decreased by 1 and sign reversed*
4. Exclamation detection: “He is brilliant but boring!!”
 - *Increase weight of emphasized words*
 - *‘boring’=-3*

Pros/Cons of the approach

- Advantages:
 - Can be fairly accurate independent of environment
 - No need for training corpus
 - Can be easily extended to new domains with additional affective words
 - e.g., “amazeballs”
 - More often used in Opinion Retrieval
- Disadvantages:
 - Compared to a well-trained, in-domain ML model they typically underperform
 - Sensitive to affective dictionary coverage

Affective Lexicons

- They have been extensively used in the field either for lexicon-based approaches or in machine-learning solutions
 - Additional features
 - Bootstrapping: unsupervised solutions (see previous)
- Can be created manually, automatically or semi-automatically
- Can be domain-dependent or independent
- A lot of them are already available:
 - Manual
 - LIWC: Linguistic Inquiry and Word Count
 - ANEW: Affective norms for English words
 - Automatic:
 - WordNet-Affect
 - SentiWordNet ...

LIWC: Linguistic Inquiry and Word Count

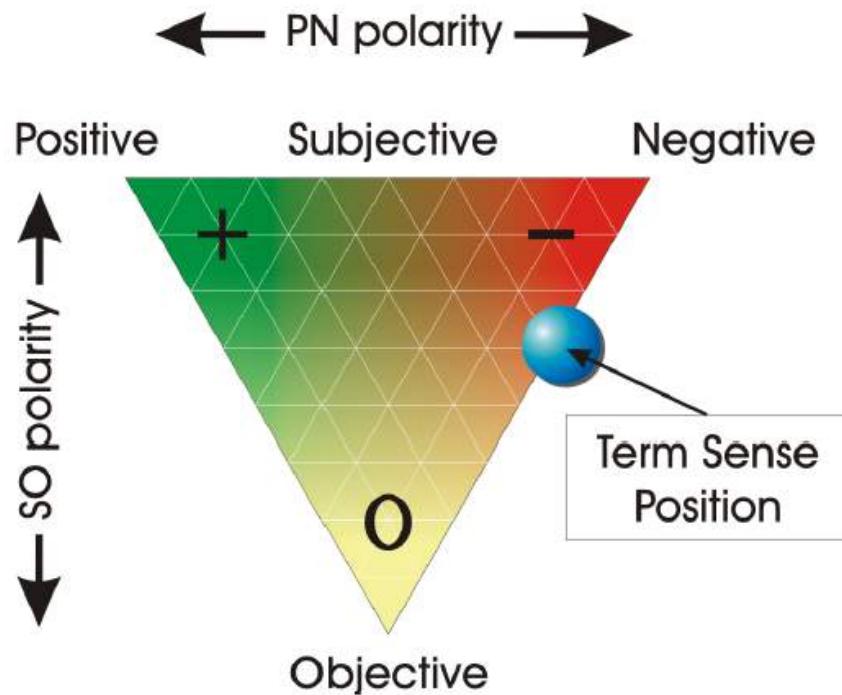
125 Affect						126 Posemo			127 Negemo			
abandon*	damn*	fume*	kind*	privleg*	supporting	accept	freed*	partie*	abandon*	enrag*	maddening	snob*
abuse*	danger*	fuming	kiss*	prize*	supportive*	accepta*	freeing	party*	abuse*	envie*	madder	sob
abusi*	daring	fun	laidback	problem*	supports	accepted	freely	passion*	abusi*	envious	maddest	sobbed
accept	darlin*	funn*	lame*	profit*	suprem*	accepting	freeness	peace*	ache*	envy*	maniac*	sobbing
accepta*	daze*	furious*	laugh*	promis*	sure*	accepts	freer	perfect*	aching	evil*	masochis*	sobs
accepted	dear*	fury	lazie*	protest	surpris*	active*	frees*	play	advers*	excruciat*	melanchol*	solemn*
accepting	decay*	geek*	lazy	protested	suspicio*	admir*	friend*	played	afraid	exhaust*	mess	sorrow*
accepts	defeat*	genero*	liabilit*	protesting	sweet	ador*	fun	playful*	aggravat*	fail*	messy	sorry
ache*	defect*	gentle	liar*	proud*	sweetheart*	advantag*	funn*	playing	aggress*	fake	miser*	spite*
aching	defenc*	gentler	libert*	puk*	sweetie*	adventur*	genero*	plays	agitat*	fatal*	miss	stammer*
active*	defens*	gentlest	lied	punish*	sweetly	affection*	gentle	pleasant*	agoniz*	fatigu*	missed	stank
admir*	definite	gently	lies	radian*	sweetness*	agree	gentler	please*	agony	fault*	misses	startl*
ador*	definitely	gigg*	like	rage*	sweets	agreeab*	gentlest	pleasing	alarm*	fear	missing	steal*
advantag*	degrad*	giver*	likeab*	raging	talent*	agreed	gently	pleasur*	alone	feared	mistak*	stench*
adventur*	delectabl*	giving	liked	rancid*	tantrum*	agreeing	gigg*	popular*	anger*	fearful*	mock	stink*
advers*	delicate*	glad	likes	rape*	tears	agreement*	giver*	positiv*	angr*	fearing	mocked	strain*
affection*	delicious*	gladly	liking	raping	teas*	agrees	giving	prais*	anguish*	fears	mockers*	strange
afraid	deligh*	glamor*	livel*	rapist*	tehe	alright*	glad	precious*	annoy*	feroc*	mocking	stress*
aggravat*	depress*	glamour*	LMAO	readiness	temper	amaz*	gladly	prettie*	antagoni*	feud*	mocks	struggl*
aggress*	depriv*	gloom*	LOL	ready	tempers	amor*	glamor*	pretty	anxi*	fiery	molest*	stubborn*
agitat*	despair*	glori*	lone*	reassur*	tender*	amus*	glamour*	pride	apath*	fight*	mooch*	stunk
agoniz*	desperat*	glory	longing*	rebel*	tense*	aok	glori*	privleg*	appall*	fired	moodi*	stunned
agony	despis*	goddam*	lose	reek*	tensing	appreciat*	glory	prize*	apprehens*	flunk*	moody	stuns
agree	destroy*	good	loser*	regret*	tension*	assur*	good	profit*	argh*	foe*	moron*	stupid*
agreeab*	destruct*	goodness	loses	reject*	terribl*	attachment*	goodness	promis*	argu*	fool*	mourn*	stutter*
agreed	determina*	gorgeous*	losing	relax*	terrific*	attract*	gorgeous*	proud*	arrogan*	forbid*	murder*	submissive*
agreeing	determined	gossip*	loss*	relief	terrified	award*	grace	radian*	asham*	fought	nag*	suck
agreement*	devastat*	grace	lost	reliev*	terrifies	awesome	graced	readiness	assault*	frantic*	nast*	sucked
agrees	devil*	graced	lous*	reluctan*	terrify	beaut*	graceful*	ready	asshole*	freak*	needy	sucker*
alarm*	devot*	graceful*	love	remorse*	terrifying	beloved	graces	reassur*	attack*	fright*	neglect*	sucks
alone	difficult*	graces	loved	repress*	terror*	benefic*	graci*	relax*	aversi*	frustrat*	nerd*	sucky
alright*	digni*	graci*	lovely	resent*	thank	benefit	grand	relief	avoid*	fuck	nervous*	suffer

ANEW: Affective norms for English words

Description	Word No.	Valence Mean(SD)	Arousal Mean(SD)	Dominance Mean (SD)	Word Frequency
abduction	621	2.76 (2.06)	5.53 (2.43)	3.49 (2.38)	1
abortion	622	3.50 (2.30)	5.39 (2.80)	4.59 (2.54)	6
absurd	623	4.26 (1.82)	4.36 (2.20)	4.73 (1.72)	17
abundance	624	6.59 (2.01)	5.51 (2.63)	5.80 (2.16)	13
abuse	1	1.80 (1.23)	6.83 (2.70)	3.69 (2.94)	18
acceptance	625	7.98 (1.42)	5.40 (2.70)	6.64 (1.91)	49
accident	2	2.05 (1.19)	6.26 (2.87)	3.76 (2.22)	33
ace	626	6.88 (1.93)	5.50 (2.66)	6.39 (2.31)	15
ache	627	2.46 (1.52)	5.00 (2.45)	3.54 (1.73)	4
achievement	3	7.89 (1.38)	5.53 (2.81)	6.56 (2.35)	65
activate	4	5.46 (0.98)	4.86 (2.56)	5.43 (1.84)	2
addict	581	2.48 (2.08)	5.66 (2.26)	3.72 (2.54)	1
addicted	628	2.51 (1.42)	4.81 (2.46)	3.46 (2.23)	3
admired	5	7.74 (1.84)	6.11 (2.36)	7.53 (1.94)	17
adorable	6	7.81 (1.24)	5.12 (2.71)	5.74 (2.48)	3
adult	546	6.49 (1.50)	4.76 (1.95)	5.75 (2.21)	25
advantage	629	6.95 (1.85)	4.76 (2.18)	6.36 (2.23)	73
adventure	630	7.60 (1.50)	6.98 (2.15)	6.46 (1.67)	14
affection	7	8.39 (0.86)	6.21 (2.75)	6.08 (2.22)	18
afraid	8	2.00 (1.28)	6.67 (2.54)	3.98 (2.63)	57

sentiwordnet.isti.cnr.it/

SentiWordNet



Opinion-Mining Tools



Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter

Less Happy  More Happy

<http://www.ccs.neu.edu/home/amislove/twittermood>

<http://www.ccs.neu.edu/home/amislove/twittermood/>

Twitter Sentiment Visualization

https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/



sentiment viz

Tweet Sentiment Visualization

Sentiment

Topics

Heatmap

Tag Cloud

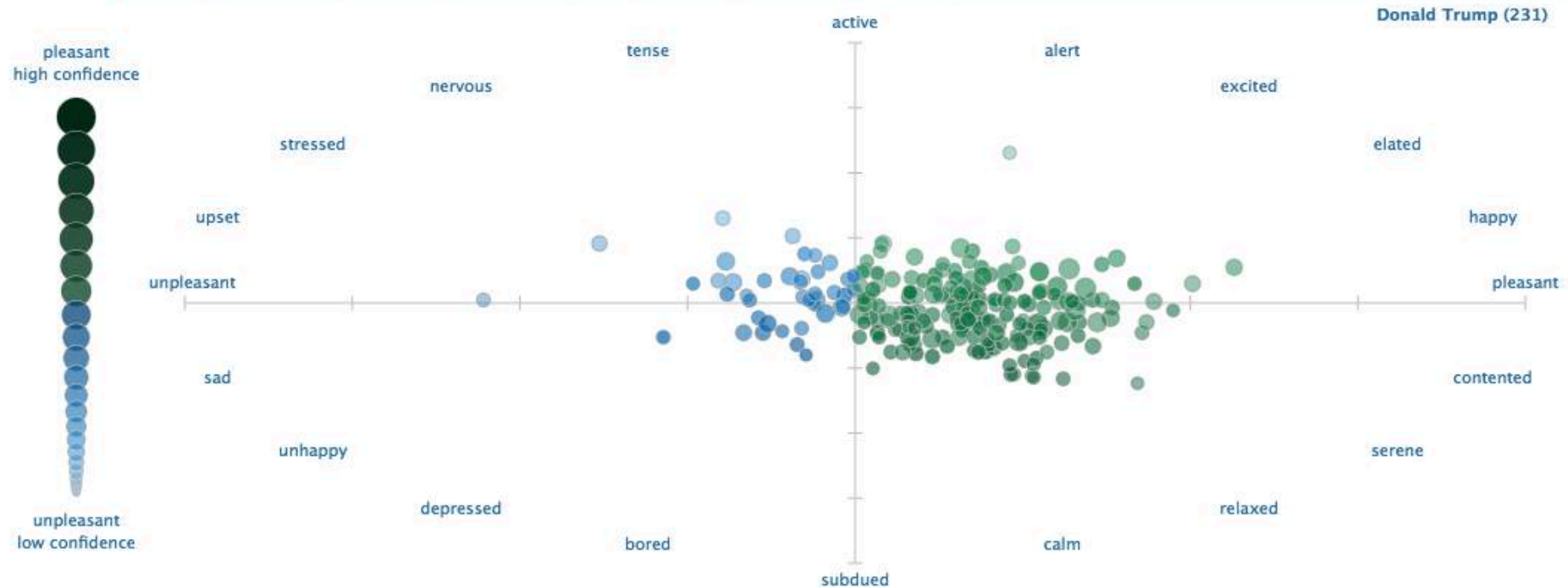
Timeline

Map

Affinity

Narrative

Tweets

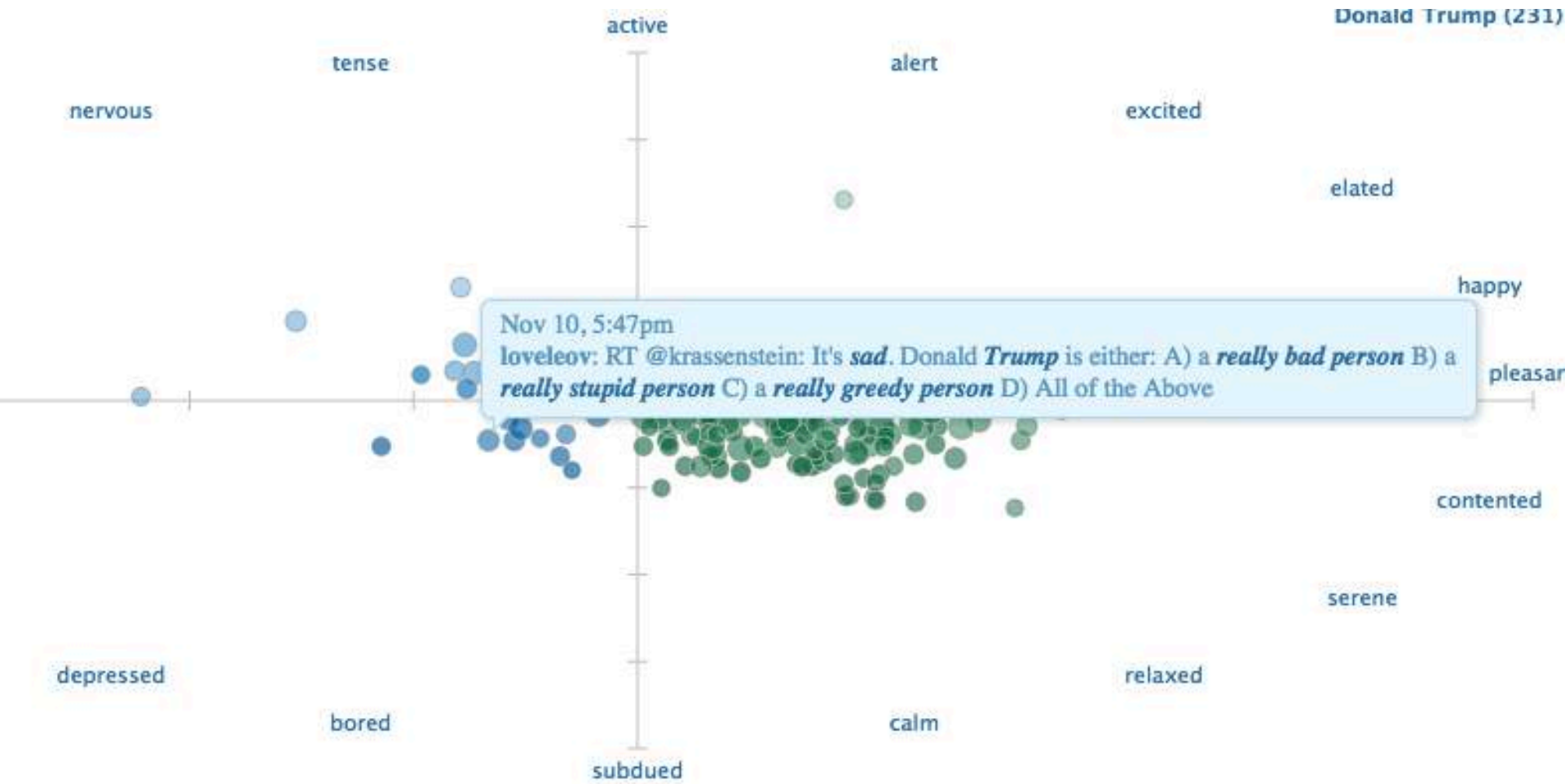


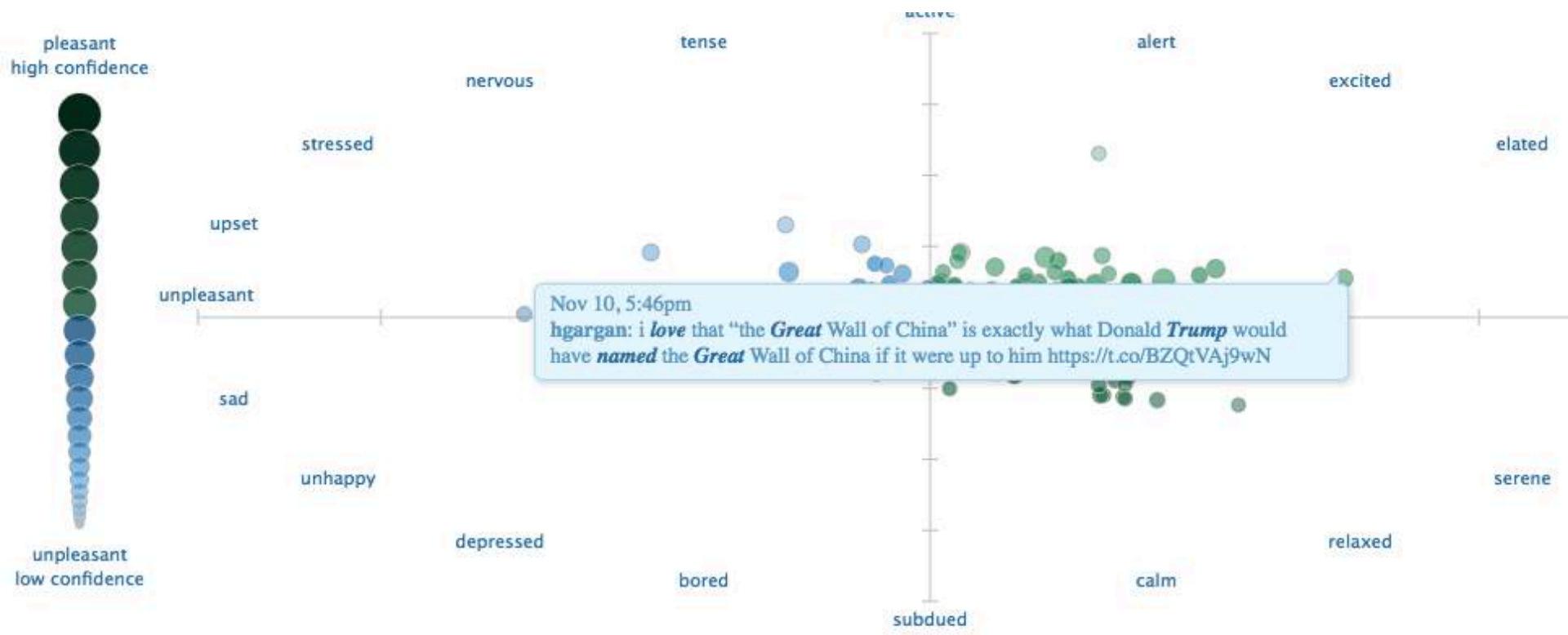
Keywords:

Query

What Do I Do?

Click to see tweets

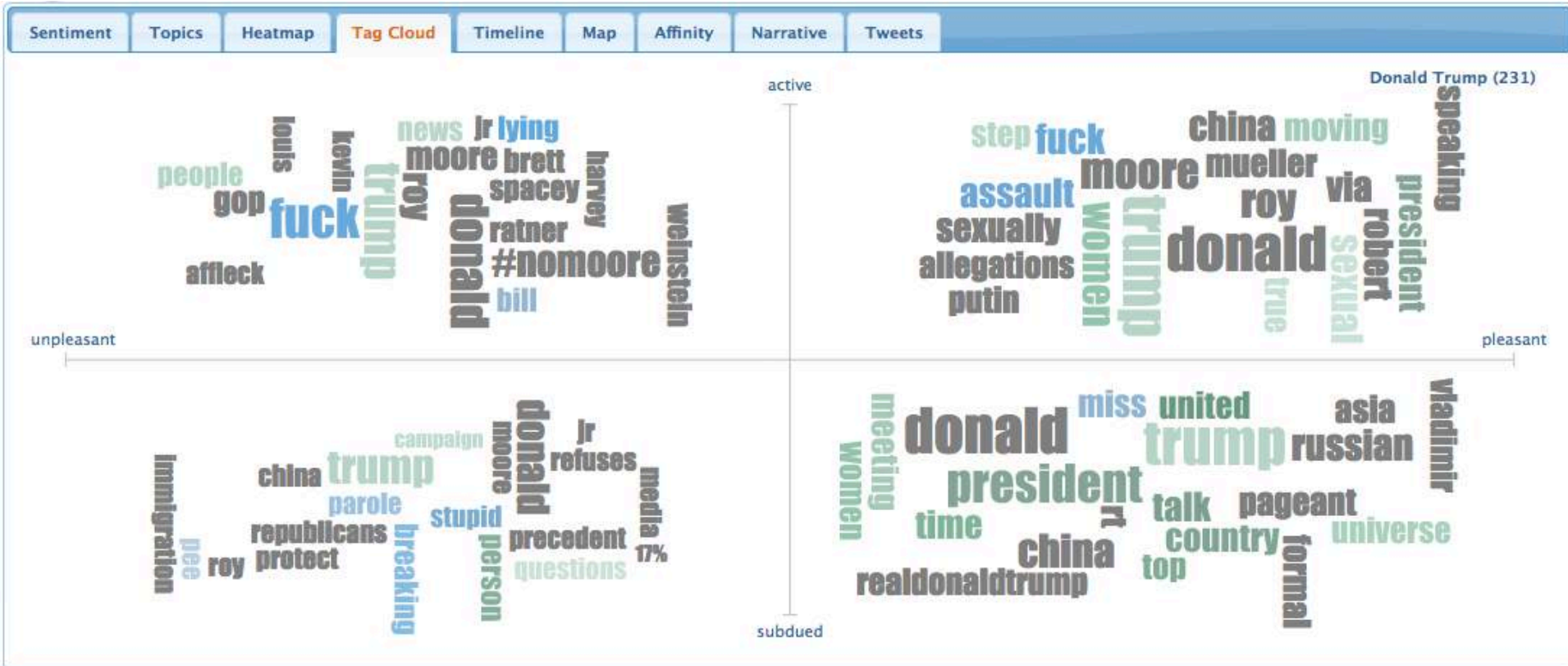




words: Donald Trump

Query

Tag cloud

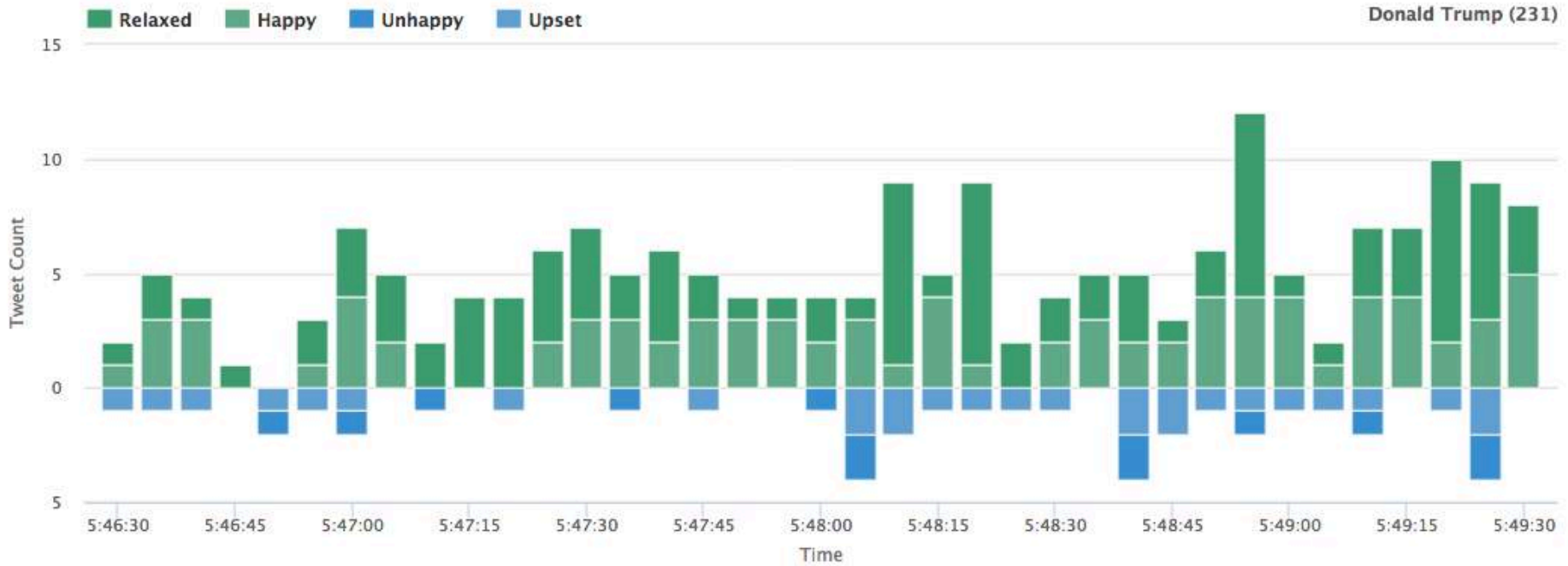


Timeline



sentiment viz
Tweet Sentiment Visualization

Sentiment Topics Heatmap Tag Cloud **Timeline** Map Affinity Narrative Tweets



Keywords:

Sentiment 140

www.sentiment140.com

ks | Fiordaliso | Provincia di Foggia... | Outlook Web App | Benvenuti al servizi... | Importati da IE | WebHome < ApprA... | WebHome < Estrinf... | WebHome < Bu

Sentiment140

Discover the Twitter sentiment for a product or brand.

 Sign in with Twitter

Twitter now requires all searches to be authenticated. Please login to authorize Sentiment140 to search Twitter. We promise to never spam your account.

 Tweet  Like 972  G+

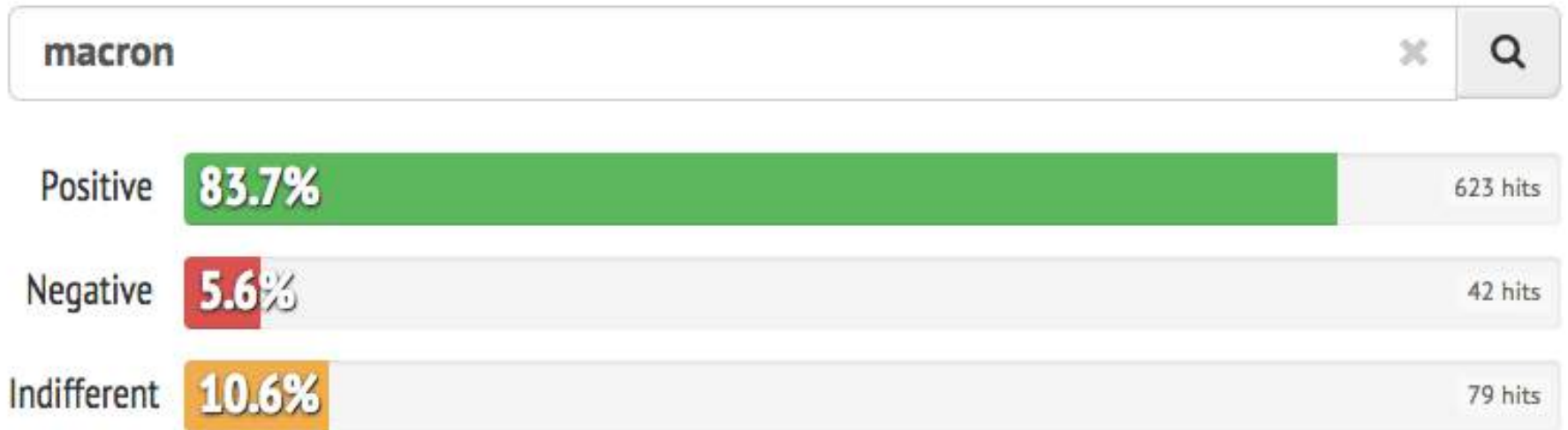
[About](#) | [API](#) | [Contact](#)

Copyright 2013

What does the internet think?

http://

www.whatdoestheinternetthink.net/



INFORMATION EXTRACTION FROM TEXTS

Information Extraction

- Text mining methods that can work together with BI and supplement business intelligence are mainly **information extraction** (IE) methods.
- Information extraction (IE) is the technology based on natural language processing in order to obtain some new information.
- The process takes text, or in some cases speech, as input and generates **unambiguous result in predefined, structured or semi-structured, format.**
- That result can be directly presented to users, or stored in the database for further usage and analyses.

Information Extraction: what is it?

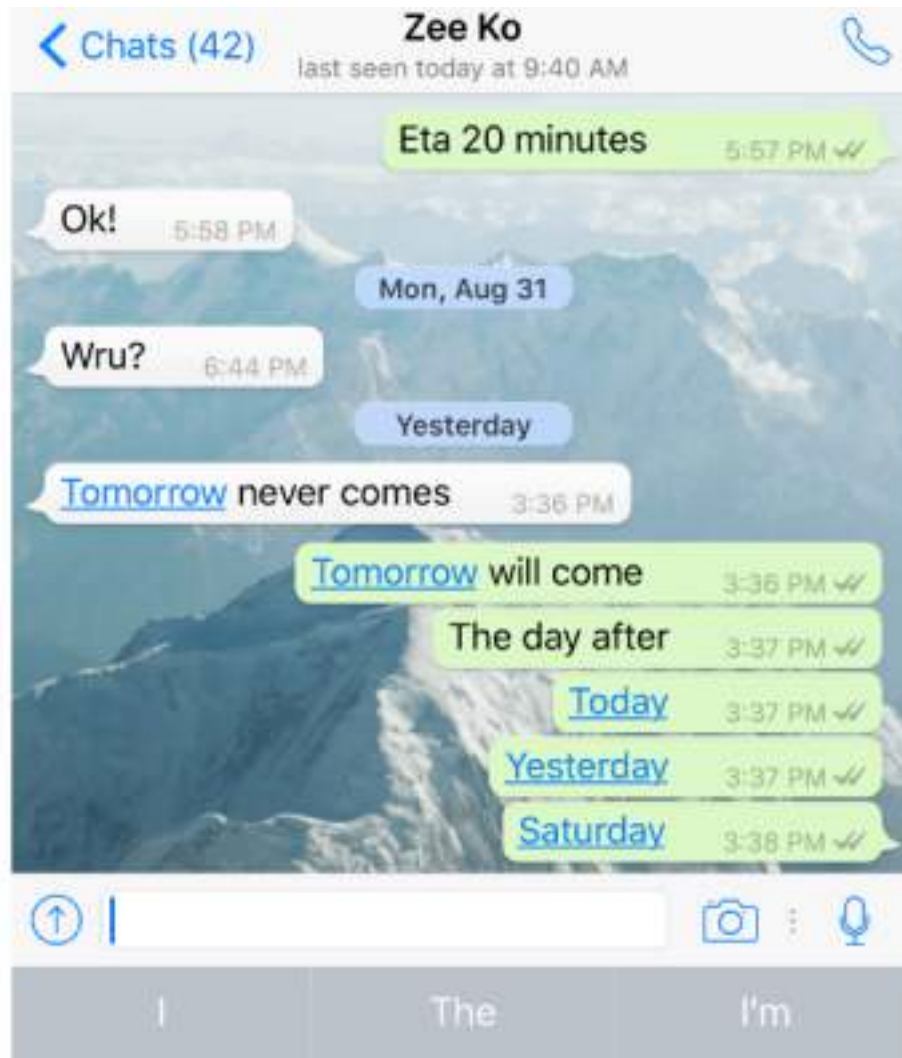
- Information extraction (IE) systems
 - Find and understand **limited relevant parts of texts**
 - Produce a **structured** representation of relevant information:
 - *relations* (in the database sense), a.k.a.,
 - *a knowledge base*
 - Goals:
 1. Organize information so that it is useful to people
 2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

- IE systems extract clear, **factual** information from texts
 - Roughly: *Who did what to whom when?*
- E.g.,
 - Gathering earnings, profits, board members, headquarters, etc. from company reports
 - Example:
 - “The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia. “
 - It extracts:
headquarters(“BHP Biliton Limited”, “Melbourne, Australia”)
(reads: the headquarter of BHP BL are in Melbourne, Australia)
 - *headquarter(x,y)* is a relation type between two entities, x and y

Low-level (simple) information extraction

- Is now an application on WhatsApp



ail,

The Full Task of Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

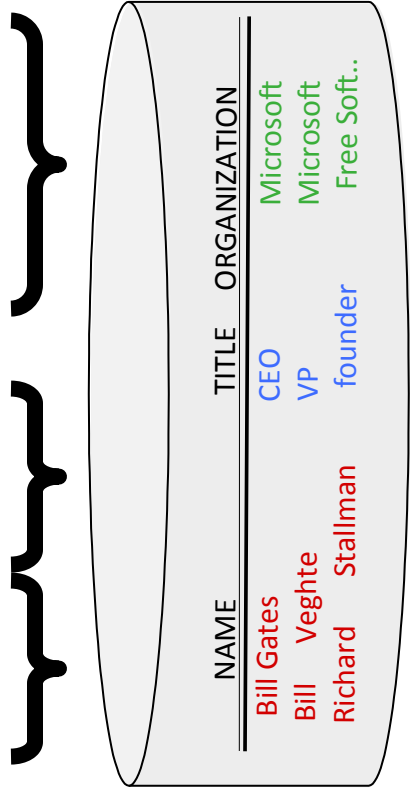
Now [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

In this example, the task is to extract names of company managers from web news wires

Microsoft Corporation CEO Bill Gates Gates
Microsoft Bill Veghte Microsoft
VP Richard Stallman founder Free Software Foundation



Slide by Andrew McCallum.

Example 2: Extracting Job Openings from the Web

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites

Back Forward Stop

Address <http://www.foodsci...>

Links AMEX Rewards T

OPUS: Job Listings - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address http://www.foodscience.com/jobs_midwest.html#top

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

Job Listings

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

Welcome

About OPUS

Executive Staff

Job Listings

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

**Test Kitchen-
Consumer Food Relations**

Major food manufacturer in Chicago area seeks a consumer food professional to write and test recipes. Will make presentations to marketing; will be a key player on a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field, plus a minimum three years applicable experience.
Contact Moira: e-mail
1-800-488-2611

Ice Cream Guru

If you dream of cold creamy chocolate or coochy coochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.
Contact Susan: e-mail
1-800-488-2611

Title: Ice Cream Guru

Description: If you dream of cold creamy...

Contact: susan@foodscience.com

Category: Travel/Hospitality

Function: Food Services



Products catalogues



Ballroom Dance Shoe

1 new from \$49.95

★☆☆☆☆ (5)

> Show only So Danca items

X-Strap Ballroom Dance Shoe

1 new from \$49.95

★★★★★ (5)

> Show only So Danca items



Dynex™ - 32" Class / 720p / 60Hz / LCD HDTV

Model: DX-32L150A11 | SKU: 9558089

★★★★☆ 3.8 of 5 (180 reviews)

[Check Shipping & Availability >](#)

Compare



Dynex™ - 24" Class / 1080p / 60Hz / LCD HDTV

Model: DX-24L150A11 | SKU: 9848048

★★★★☆ 4.3 of 5 (54 reviews)

[Check Shipping & Availability >](#)

Compare

Product	Type	Price
Dynex 32"	LCD TV	\$1000
...	...	

Another Example: Classify Advertisements (Real Estate)

<ADNUM>2067206v1</ADNUM>

<DATE>March 02, 2017</DATE>

<ADTITLE>MADDINGTON \$89,000</ADTITLE>

<ADTEXT>

OPEN 1.00 - 1.45

U 11 / 10 BERTRAM ST

NEW TO MARKET Beautiful

3 brm freestanding villa, close to shops & bus

Owner moved to Melbourne

ideally suit 1st home buyer,

investor & 55 and over.

Brian Hazelden 0418 958 996

R WHITE LEEMING 9332 3477

</ADTEXT>

date	address	price	bedrooms	type	...
02-03-2017	11 / 10 BERTRAM ST	89,000	3	villa	..

IE is DIFFERENT from Information retrieval (i.e., web search)

- Information retrieval detects documents of the specific interest and presents them to the users. **It is up to the user** to read the proposed document and extract information manually.
- On the other hand information extraction fetches only **specific information** from text which is relevant to the user.
- It is often a domain-dependent task

Why is IE difficult?

- There are many ways of expressing the same fact
E.g, detecting changes of management:
 - BNC Holdings Inc **named** Ms G Torretta **as its new chairman**.
 - Nicholas Andrews **was succeeded** by Gina Torretta as chairman of BNC Holdings Inc.
 - Ms. Gina Torretta **took the helm** at BNC Holdings Inc.
 - After a long **boardroom struggle**, **Mr Andrews** stepped down as chairman of BNC Holdings Inc. He was succeeded by Ms Torretta.

Why is it difficult?

What you search for in real estate advertisements application:

- Town/suburb. You might think it is easy, but:
 - **Real estate agents:** Coldwell Banker, Mosman
 - **Phrases:** Only 45 minutes from Parramatta
 - **Multiple property ads have different suburbs in one ad**
- Money: **you want a range not a textual match**
 - **Multiple amounts:** was \$155K, now \$145K
 - **Variations:** offers in the high 700s [*but not* rents for \$270]
- Bedrooms: **similar issues:** br, bdr, beds, B/R

Why is IE even harder on the web?

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

EMAIL THIS PAGE TO A FRIEND

- English Books
- German Books
- Spanish Books
- Sheet Music
- Musical Supplies
- US/World Maps
- Sports Memorabilia
- Videos/Posters

Established Phoenix 1994
NetStoreUSA.com

English Books > Antiques/Collectibles > Toys > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

Search

ADVANCED SEARCH >>

<< PREVIOUS TITLE | NEXT TITLE >> <<NEW RELEASES >>

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

Author: Meisenheimer, Lucky 1; Editor: T Brown &

CHECK THE AVAILABILITY

Home

To Order

Privacy

ADD TO CART

ABOUT US

Contact

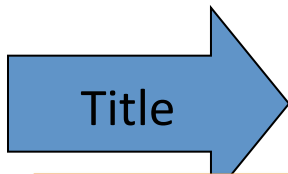
PRODUCT CODE: 0966761200

- USA/Canada: US\$ 43.40
- Australia/NZ: A\$ 124.50
- Other Countries: US\$ 80.90

[convert to your currency](#)

VIEW CART CHECKOUT

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it



Title

You have the additional problem of parsing the web page layout: where is the relevant content?



Need this price

BI applications of Information Extraction

- Collecting Company Information from multiple multilingual sources (English, German, Italian) to provide up-to-date information on competitors
- Identifying Chances of success in regions in a particular country
- Identify appropriate partners to do business with
- Creation of a Joint Ventures Database from multiple sources
- Tracking competitors: new products, management changes, joint ventures..

Basic steps

- Define a template (what type of information you wish to extract)
- Extract named entities
- Resolve co-reference
- Identify relations
- Fill the template

1. Define a template

- Unless we aim at open domain relation extraction (any interesting relation on the web), IE is domain dependent: management changes, real estates, job openings, new launched products..
- Since the objective is to extract text segments and fill a table, we need to define the structure of the table, e.g.:

Person	Position	Company

2. Named Entity Recognition (NER)

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

3. Co-reference resolution

VIENNA, January 20. /TASS/. **Russian gas producer Gazprom** expects the price of gas supplies to Europe at the level of \$184 per 1,000 cubic meters versus \$284 in the same period last year, Chairman of its board of directors Viktor Zubkov said on Wednesday.

"In the first quarter of 2015 the price was \$284 while according to [preliminary] estimates in the first quarter of this year it will be \$184, which means we lose \$100," he said.

According to Zubkov, **Gazprom** plans to supply 160 bcm of gas to Europe in 2016.

"Over the last 3 years Gazprom supplied 467.5 bln cubic meters [of gas] to Europe, which accounts for a third of consumed volumes. **We** expect 160 bln cubic meters this year," he said.

In 2015, the company increased its gas export to non-CIS countries by 8% compared with previous year - to 159.4 bln cubic meters. Average price for Russian gas was expected at \$237-242 per 1,000 cubic meters for supplies to Europe in 2015.

Particularly, gas supplies to Germany rose by 17.1% in 2015, to Italy - by 12.6%, to France - by 36.8%, to Britain - by 10.2%, to Austria - by 11.5%.

4. Relation extraction

Entities:

PER: Jose Mourinho
POSITION: trainer
ORG: Chelsea

Relation

Jose Mourinho
↓
Trainer
↓
Chelsea



Category type of trainer

Category type of trainee

Relation name

trainer(**person**, **team**)

5.Template filling

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

Activity:

Type: PRODUCTION

Company:

Product: *golf clubs*

Start-date:

The joint venture, Bridgestone Sports Taiwan Cp., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Activity:

Type: PRODUCTION

Company: *Bridgestone Sports Taiwan Co*

Product: *iron and "metal wood" clubs*

Start-date: DURING *1990*

Watson IE features

laser1.watson.ibm.com/sire/ie2.php

IBM Watson Developer Cloud - Relationship Extraction Service Demo

Services Docs App Gallery Community

Select Domain English News (KLUE3)

Submit Clear

download output as DOCXML

Please mouse over **colored** text to see mentions and entities, press CTRL while mouse-over **italicized** text to see relations. Click **P** to toggle parse tree display.

This on-site investigation focused on the performance of the Certified Advanced 208-Compliant air bag system in a 2005 Ford Escape XLT 4x4 sport utility vehicle. This two-vehicle crash occurred in July 2014 at 1539 hours in the state of Colorado. The crash occurred on a curved portion of a three-lane interstate roadway. The Ford Escape lost control on an interstate highway and struck a concrete barrier on the right side of the roadway. The impact resulted in sufficient longitudinal deceleration of the Escape to command the deployment of the frontal air bag system and actuation of the driver's seat belt pretensioner. The vehicle rotated out from the initial wall impact and was subsequently struck by a 2013 BYD Qin pulling a single trailer. The restrained 48-year-old male driver of the Escape appears to have sustained a minor facial injury. The driver of Qin did not report any injuries.

EVENT LEGAL DATE FACILITY ORGANIZATION EVENT_DISASTER PERSON
VEHICLE AGE EVENT_COMMUNICATION MEASURE TIME GPE

- P** This on-site **investigation** focused on the performance of the Certified Advanced 208-Compliant air bag system in a **2005 Ford Escape XLT 4x4** sport utility **vehicle**.
- P** This two-vehicle crash occurred in **July 2014** at **1539** **ORGANIZATION** of **Colorado**.
- P** The **crash** occurred on a curved portion of a three-lane interstate **roadway**.
- P** The **Ford Escape** lost control on an interstate **highway** and **struck** a concrete barrier on the right side of the **roadway**.
- P** The impact resulted in sufficient longitudinal deceleration of the **Escape** to command the deployment of the frontal air bag system and actuation of the **driver**'s seat belt pretensioner.
- P** The **vehicle** rotated out from the initial wall impact and was subsequently **struck** by a **2013 BYD Qin** pulling a single **trailer**.
- P** The restrained **48-year-old male driver** of the **Escape** appears to have sustained a minor facial injury.
- P** The **driver** of **Qin** did not **report** any injuries.

Watson NE recognition and classification

IBM Watson Knowledge Studio

View Details Attribute View View Guidelines Completed Close Alpha... 14pt 1

2004-49-168A.txt

1 V1, a 1999 Toyota Camry, was traveling southbound in the second lane of a four-lane divided (seven lanes overall, divided by raised median), concrete roadway, approaching an intersection.

2 V2, a 2004 Mercedes S430, was northbound in the fourth lane of a four-lane, divided (seven lanes overall, divided by raised median), concrete roadway, about to turn left into westbound traffic at the same intersection.

3 As both vehicles entered the intersection, the front of V1 impacted the front of V2.

4 V1 rotated clockwise as V2 rotated counter-clockwise, and the left side of V1 impacted the right side of V2 in a sideslap configuration.

5 Both vehicles moved southwest to final rest.

6 Both vehicles were towed due to damage.

7 The unrestrained driver of V1 was hospitalized with foot and rib fractures as well as a liver laceration.

8 The restrained driver of V2 was treated and released with minor abrasion and contusion as well as a finger fracture.

9 The restrained front right passenger in V2 was pronounced brain dead two days later from multiple brain injuries.

10 V1 was equipped with redesigned dual frontal airbags, which deployed

Entity		Mention
Type	Subtype	Role
a	ACCIDENT_CAUSE	
o	ACCIDENT_OUTCOME	
-	CONDITION	
i	IMPACT	
f	MANUFACTURER	
m	MODEL	
y	MODEL_YEAR	
r	PART_OF_CAR	
p	PERSON	
s	STRUCTURE	
H	VEHICLE	

Other available systems

- Stanford CORE NLP

<https://stanfordnlp.github.io/CoreNLP/repl.html>

- Sheffield GATE IE framework:

<https://gate.ac.uk/ie/> and their business implementation:

<https://gate.ac.uk/business.html>