

# terminology mining from new text - 2011

Jussi Karlgren

5 maj 2011, dsv

Jussi Karlgren

ph d in (computational) linguistics from stockholm

senior researcher in information access at sics, stockholm

docent in language technology at univ of helsinki

founding partner, gavagai ab, stockholm



- Independent non-profit research institute
- About 100-200 researchers
- ... networks, distributed systems, programming tools, collaborative environments, information access, design, digital art...



- recent startup company
- about 7-8 employees
- extracts actionable intelligence from very large text streams

Gavagai was here!



„terminology mining in dynamic and changing text streams”

an example task: *identification and analysis of attitude*

attitude / opinion / sentiment analysis

# applications

- commercial
- pr
- security
- journalism
- linguistics

# applications

- commercial
- pr
- security
- journalism
- linguistics

# applications

- commercial
- pr
- security
- journalism
- linguistics

# applications

- commercial
- pr
- security
- journalism
- linguistics

# applications

- commercial
- pr
- security
- journalism
- linguistics

# applications

- commercial
- pr
- security
- journalism
- linguistics

## a prototypical attitudinal expression

Expression	WHO	FEELS WHAT	ABOUT WHAT
I like sauerkraut	I	like	sauerkraut
Kissing is nice	?	nice	kiss
	<i>someone</i>	<i>sentiment term</i>	<i>topic</i>

is this picture true?

## a prototypical attitudinal expression

Expression	WHO	FEELS WHAT	ABOUT WHAT
I like sauerkraut	I	like	sauerkraut
Kissing is nice	?	nice	kiss
	<i>someone</i>	<i>sentiment term</i>	<i>topic</i>

is this picture true?

“It is this, I think, that commentators mean when they say glibly that the ‘world changed’ after Sept 11.”

“President Hafez Al-Assad has said that peace was a pressing need for the region and the world at large and Syria, considering peace a strategic option would take steps towards peace.”

“Mr Cohen, beginning an eight-day European tour including a Nato defence ministers’ meeting in Brussels today and tomorrow, said he expected further international action soon, though not necessarily military intervention.”

“It is this, I think, that commentators mean when they say glibly that the ‘world changed’ after Sept 11.”

“President Hafez Al-Assad has said that peace was a pressing need for the region and the world at large and Syria, considering peace a strategic option would take steps towards peace.”

“Mr Cohen, beginning an eight-day European tour including a Nato defence ministers’ meeting in Brussels today and tomorrow, said he expected further international action soon, though not necessarily military intervention.”

“It is this, I think, that commentators mean when they say glibly that the ‘world changed’ after Sept 11.”

“President Hafez Al-Assad has said that peace was a pressing need for the region and the world at large and Syria, considering peace a strategic option would take steps towards peace.”

“Mr Cohen, beginning an eight-day European tour including a Nato defence ministers’ meeting in Brussels today and tomorrow, said he expected further international action soon, though not necessarily military intervention.”

## the task in more detail

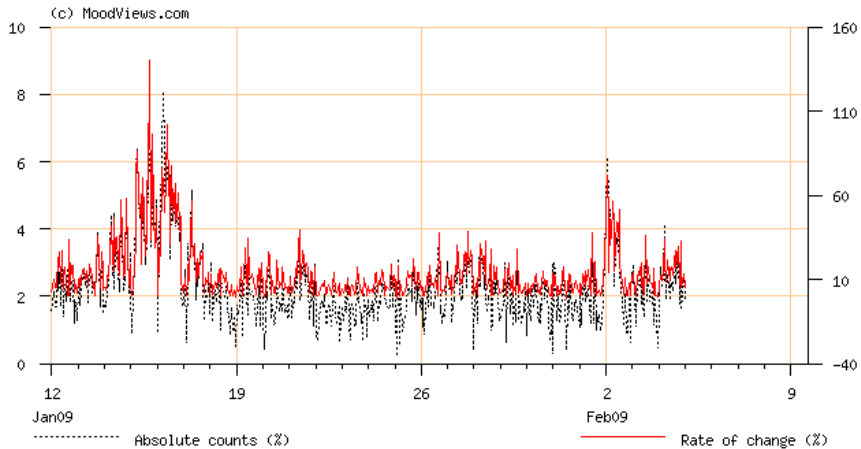
- identify attitude?
- identify change in attitude?
- with respect to some topic?
- identify polarity of attitude?
- identify intensity of attitude?
- understand the human condition better?

## examples

even simple applications can be useful and interesting!

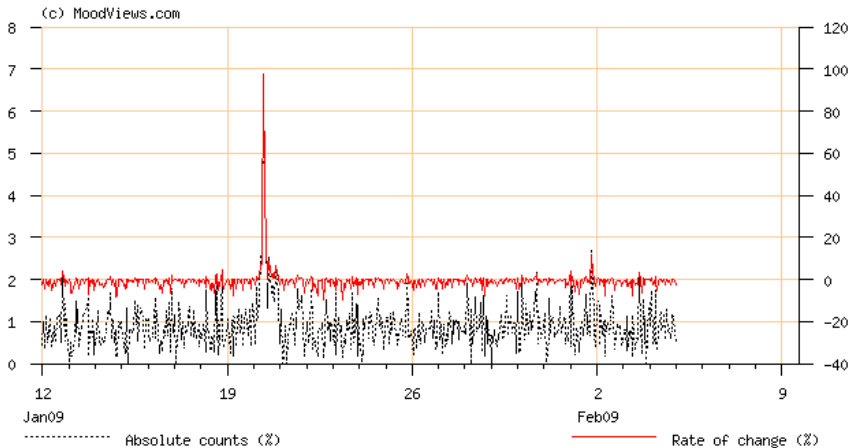
cold

### Changes in "cold" over the last 24 days



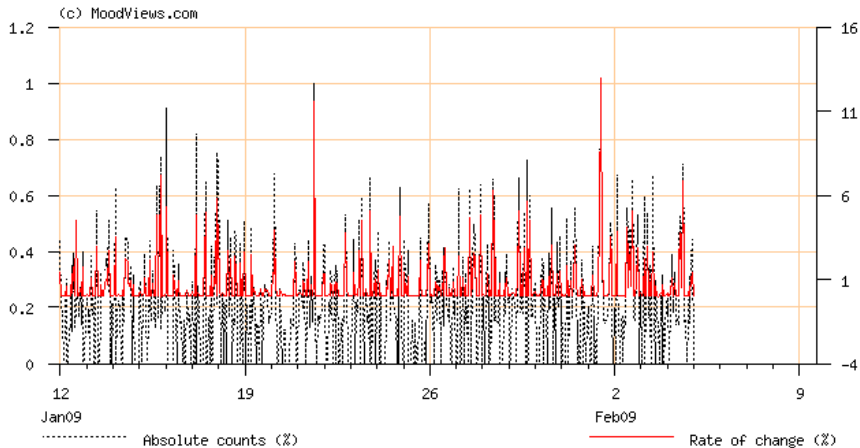
# ecstatic

Changes in "ecstatic" over the last 24 days



# embarrassed

Changes in "embarrassed" over the last 24 days



## example products

**buzz change alert** our brand name has been mentioned in a new context. **Alert!**

**competitor relative advantage** how does our brand compare to others with respect to some central theme (quality, reliability, sexiness etc).

**who are our competitors** what other brands are mentioned with ours? (should we try to change our profile?)

**new terminology** consumers talk about the product line using non-standard terminology: how can the consultant keep up with rapidly changing slang?

**e-fluencer ID** which contributors are most central and best informed? can we contribute to their understanding of our product?

## example products

**buzz change alert** our brand name has been mentioned in a new context. **Alert!**

**competitor relative advantage** how does our brand compare to others with respect to some central theme (quality, reliability, sexiness etc).

**who are our competitors** what other brands are mentioned with ours? (should we try to change our profile?)

**new terminology** consumers talk about the product line using non-standard terminology: how can the consultant keep up with rapidly changing slang?

**e-fluencer ID** which contributors are most central and best informed? can we contribute to their understanding of our product?

## example products

**buzz change alert** our brand name has been mentioned in a new context. **Alert!**

**competitor relative advantage** how does our brand compare to others with respect to some central theme (quality, reliability, sexiness etc).

**who are our competitors** what other brands are mentioned with ours? (should we try to change our profile?)

**new terminology** consumers talk about the product line using non-standard terminology: how can the consultant keep up with rapidly changing slang?

**e-fluencer ID** which contributors are most central and best informed? can we contribute to their understanding of our product?

## example products

**buzz change alert** our brand name has been mentioned in a new context. **Alert!**

**competitor relative advantage** how does our brand compare to others with respect to some central theme (quality, reliability, sexiness etc).

**who are our competitors** what other brands are mentioned with ours? (should we try to change our profile?)

**new terminology** consumers talk about the product line using non-standard terminology: how can the consultant keep up with rapidly changing slang?

**e-fluencer ID** which contributors are most central and best informed? can we contribute to their understanding of our product?

## example products

**buzz change alert** our brand name has been mentioned in a new context. **Alert!**

**competitor relative advantage** how does our brand compare to others with respect to some central theme (quality, reliability, sexiness etc).

**who are our competitors** what other brands are mentioned with ours? (should we try to change our profile?)

**new terminology** consumers talk about the product line using non-standard terminology: how can the consultant keep up with rapidly changing slang?

**e-fluencer ID** which contributors are most central and best informed? can we contribute to their understanding of our product?

example

what brands are like **my brand**?

# example

## preem

statoil 0.479 okq8 0.454 hemköp 0.351 apoteket 0.337 konsum 0.336 shell 0.313 statiol 0.310 macken 0.309 affären  
0.303 affärn 0.296 jysk 0.296 djuraffären 0.283 pressbyrån 0.281 bageriet 0.275 mobilia 0.269 granngården 0.265  
biltema 0.261 brändåsen 0.259 nöjesbutiken 0.258 ikea 0.258 willys 0.256 elgiganten 0.256 kiosken 0.256  
östenssons 0.254 stan 0.247 kupolen 0.246 djurmagasinet 0.244 7eleven 0.239 allum 0.237 citygross 0.236

# example

## preem

statoil 0.479 okq8 0.454 hemköp 0.351 apoteket 0.337 konsum 0.336 shell 0.313 statiol 0.310 macken 0.309 affären  
0.303 affärn 0.296 jysk 0.296 djuraffären 0.283 pressbyrån 0.281 bageriet 0.275 mobilia 0.269 granngården 0.265  
biltema 0.261 brändåsen 0.259 nöjesbutiken 0.258 ikea 0.258 willys 0.256 elgiganten 0.256 kiosken 0.256  
östenssons 0.254 stan 0.247 kupolen 0.246 djurmagasinet 0.244 7eleven 0.239 allum 0.237 citygross 0.236

attitude analysis can be done on any text source

blogs: unfettered discourse, wom, low publication threshold, no editorial control

but it's *new text* — new processing practice necessary

attitude analysis can be done on any text source

blogs: unfettered discourse, wom, low publication threshold, no editorial control

but it's *new text* — new processing practice necessary

reminiscent of (but not identical to) spoken language:

- rapid topical shifts
- slang, ad-hoc usage and coinage of terms
- inconsistent and incomplete
- dynamic, reactive and multilingual

# new challenges

we need to handle:

- scalability
- change
- noise

# lexical resources

in practice, all existing projects use term lists.

you try!

term lists are always incomplete and always out of date.

## lexical resources

in practice, all existing projects use term lists.

you try!

term lists are always incomplete and always out of date.

# lexical resources

in practice, all existing projects use term lists.

you try!

term lists are always incomplete and always out of date.

which means terminology mining is crucial for the task!

# internet terminology

“recommend”:

"recomend"	0.972	← spelling variation
"reccomend"	0.968	← spelling variation
"reccommend"	0.941	← spelling variation
"loathe"	0.880	
"remember"	0.879	
"regret"	0.877	
"hate"	0.876	
"despise"	0.873	
"hope"	0.873	
"bet"	0.872	
"promise"	0.871	
"presume"	0.870	
"looove"	0.870	← spelling variation for "love"
"loooove"	0.863	← spelling variation for "love"
"lurve"	0.850	← spelling variation for "love"
"love"	0.846	← correct spelling for "love"
"loooooove"	0.836	← spelling variation for "love"

# distribution $\approx$ meaning

"You shall know a word by the company it keeps!" (Firth)

"For a large class of cases ... though not for all ... in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language" (Wittgenstein).

"The value of the chess pieces depends upon their position upon the chess board, just as in the language each term has its value through its contrasts with all the other terms" (Saussure)

"...if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution" (Harris)

## word space models: semantic spaces

- collect distributional data for linguistic items
- compute vector representation of those data
- consult that vector representation using standard methods

## collect data

*context<sub>j</sub>* “- How can you have any pudding if you don't eat your meat?”

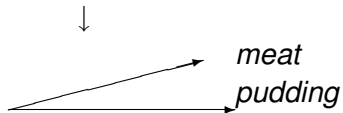
*context<sub>j</sub>* “The Yorkshire pudding is a staple of the British Sunday dinner and in some cases is eaten as a separate course prior to the main meat dish.”



$$\vec{v}_{meat} = [v_1, \dots, 1, \dots, 1, \dots, v_n]$$

# compute convenient representation

$$\vec{v} = [v_1, \dots, v_n]$$



↓  
consult

## Cosine representation using convenient methodology

$$d_{\cos}(x, y) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

**theory:** vector spaces are mathematically well defined and understood.

**implementation:** manageable implementational framework.

**aesthetics:** geometric interpretation is intuitively plausible.

**theory:** vector spaces are mathematically well defined and understood.

**implementation:** manageable implementational framework.

**aesthetics:** geometric interpretation is intuitively plausible.

**theory:** vector spaces are mathematically well defined and understood.

**implementation:** manageable implementational framework.

**aesthetics:** geometric interpretation is intuitively plausible.

associative relations - immediate relations to adjacent words:

*eat* → *food*

synonymy relations - second order relations to words that share contexts:

*eat*



*drink*

but

- building a word-by-document matrix is hard work
- needs to be repeated for every new item and every new document
- most cells are zero
- the dimensionality of the matrix is huge!

# dimensionality reduction

$$\vec{v} = [v_1, \dots, v_j, \dots, v_n]$$

- $v_j$ :  $f$ (frequency of  $v$  in context  $d_j$ )
- cooccurrence matrix is of order  $w \times n$ .
- typically enormous, very sparse and too specific.
- matrix algebraic operations are expensive and can only be done post hoc!

## research question

what is the appropriate (“intrinsic”, “latent”) dimensionality of linguistic data? are there large angles in semantic space?

$d_{\cos}(\text{" cardamom" }, \text{" tensor algebra" }) = ?$

# research question

how does a semantic space evolve?

## the Spinn3r data

≈ 44 million blog posts made between 01-08-2008 and 01-10-2008

Two subsets:

- ≈ 1G words
- ≈ 600M words

## nearest neighbors

non-standard synonyms, misspellings, polysemy

good		bad	
great	0.91	weird	0.86
prefect	0.83	sucky	0.86
perfect	0.83	scary	0.86
pristine	0.81	cool	0.85
stable	0.80	nasty	0.84
grat	0.80	dumb	0.84
fantastic	0.80	sad	0.84
flawless	0.79	lame	0.84
mint	0.79	creepy	0.84
immaculate	0.79	stupid	0.84

# evaluation

evaluation of mined terminology is notoriously difficult

properties of the semantic representation:

- stability of the semantic neighborhoods
- evolution of the semantic neighborhoods
- diversity of language use

# semantic evolution

how much do the semantic neighborhoods evolve when we add more data?

the number of new terms among the ten nearest neighbors

## semantic evolution

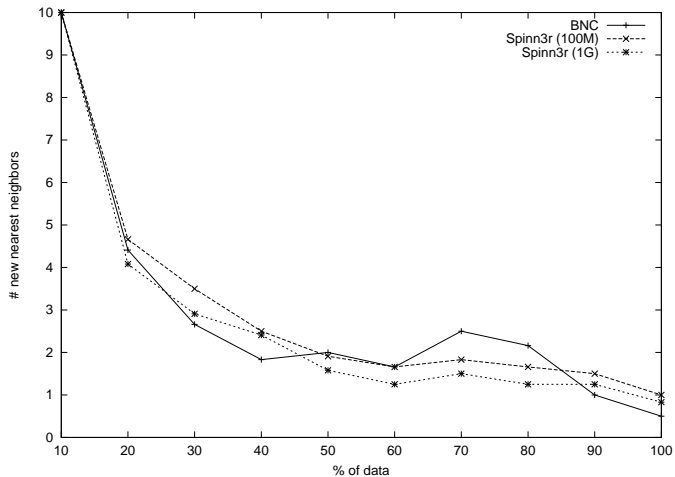
nearest neighbors to “love” for 90% of the Spinn3r data:

hate hope loathe loooove adore loooooove looove **loved** detest miss

and when having seen 100% of the data:

hate hope loathe loooove adore **loooooooooove** looove loooooove miss  
detest

# semantic evolution



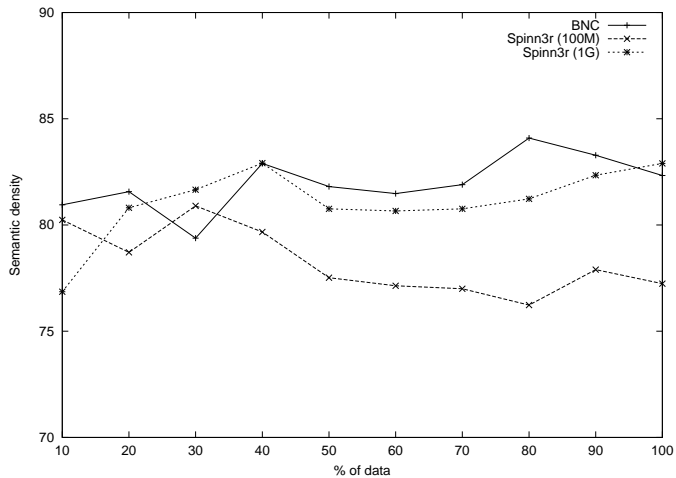
## semantic density

how homogeneous is the data?

- 1 for each target word, extract the  $n$  nearest neighbors
- 2 for each neighbor, extract the  $n$  nearest neighbors
- 3 count the number of unique words thus extracted

a high number indicates variable language use, while a low number indicates that the data is homogeneous

# semantic density



## a word of caution

build your knowledge representation for the task.

don't optimise your processing for a suboptimal knowledge representation.

... but instead, select your representation so that it learns from task space!

## a word of caution

build your knowledge representation for the task.

don't optimise your processing for a suboptimal knowledge representation.

... but instead, select your representation so that it learns from task space!

## a word of caution

build your knowledge representation for the task.

don't optimise your processing for a suboptimal knowledge representation.

... but instead, select your representation so that it learns from task space!

## conclusions



- **attitude is not only words!**  
(but may be approximated by words reasonably well)
- **words change!** (but this can be modelled)
- **data is more difficult if you leave the lab** (you knew that already!)
- **the task target is unclear** (allows leeway to define it)
- **don't optimise an experiment!** (this goes for whatever you'll be doing)