Text Mining with R

Rob Zinkov

October 19th, 2010

Outline

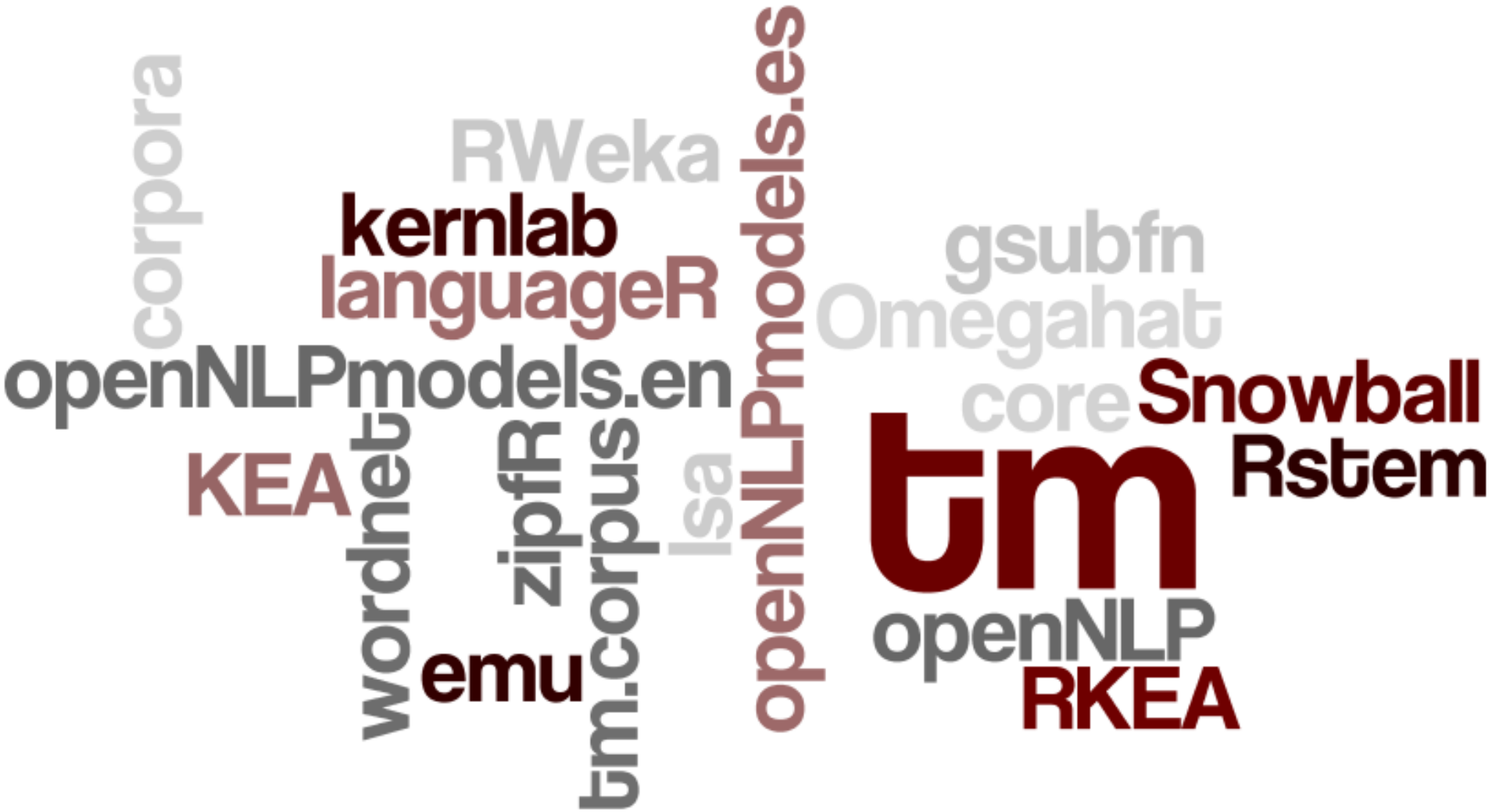
What is Text Mining?

Text mining is any process or program that:

Raw human written text

Structured information

R is very good for this



Themes

- R is a great glue language
- CRAN already has a lots of packages that work well together

Caveats

- I use outside libraries more than necessary
- Many of these algorithms could be written completely in R
- There are nicer ways to integrate these libraries
- Text mining is a vast field that can't be covered in 40 minutes

Readability

Preprint typeset in JHEP style. - HYPER VERSION

SITP-10/xx

Cascades with Adjoint Matter: Adjoint Transitions

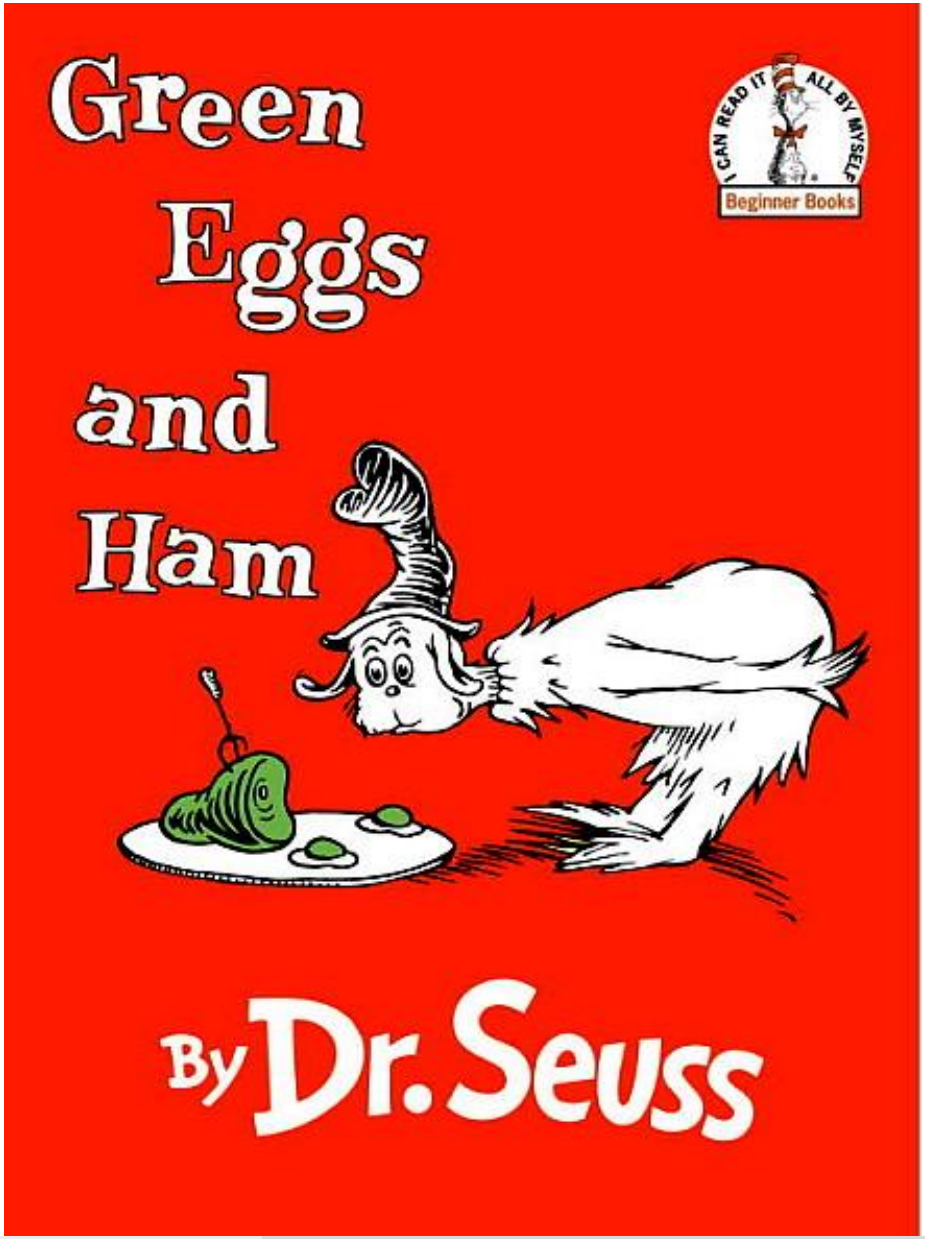
Dušan Simić[1,2]

[1]*Department of Physics, Stanford University*
Stanford, CA 94305 USA

[2] *Theory Group, SLAC National Accelerator Laboratory*
Menlo Park, CA 94025 USA

simic@stanford.edu

Abstract: A large class of duality cascades based on quivers arising from non-isolated singularities enjoy adjoint transitions - a phenomenon which occurs when the gauge coupling of a node possessing adjoint matter is driven to strong coupling in a manner resulting in a reduction of rank in the non-Abelian part of the gauge group and a subsequent flow to weaker coupling. We describe adjoint transitions in a simple family of cascades based on a $\mathbb{Z}_2$-orbifold of the conifold using field theory. We show that they are dual to Higgsing and produce varying numbers of $U(1)$ factors, moduli, and monopoles in a manner which we calculate. This realizes a large family of cascades which proceed through Seiberg duality and Higgsing. We briefly describe the supergravity limit of our analysis, as well as a prescription for treating more general theories. A special role is played by $\mathcal{N} = 2$ SQCD. Our results suggest that additional light fields are typically generated when UV completing certain constructions of spontaneous supersymmetry breaking into cascades - potentially leading to instabilities.



- Readability gives us an idea of the difficulty of the document
- It also gives a rough measure of the quality

Flesch-Kincaid readability test

Readability can be roughly measured with

$$206.876 - 1.015\left(\frac{w}{s}\right) - 84.6\left(\frac{y}{w}\right)$$

where

w = total words

y = total syllables

s = total sentences

Score	Notes
90.0-100.0	easily understandable by an average 11-year-old student
60.0-70.0	easily understandable by 13- to 15-year-old students
0.0-30.0	best understood by university graduates

Flesch-Kincaid Reading Age

$$(0.39 * ASL) + (11.8 * ASW) - 15.59$$

where ASL = average sentence length
where ASW = average syllables per word

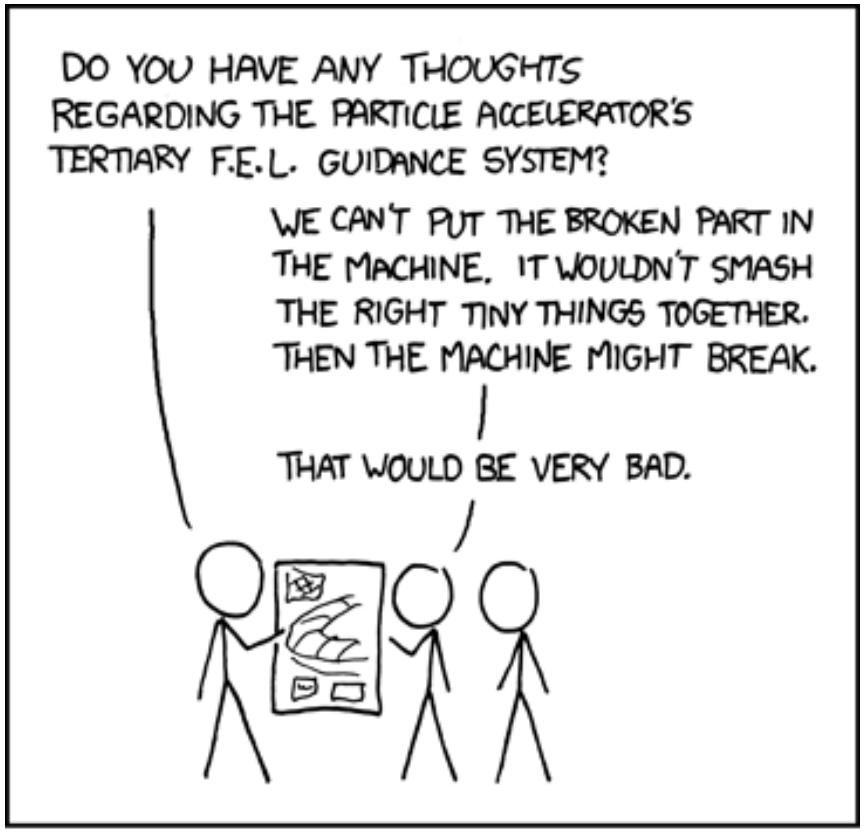
```
system(paste("java -jar CmdFlesh.jar","test_review.txt"))
```

Demo!

Notes

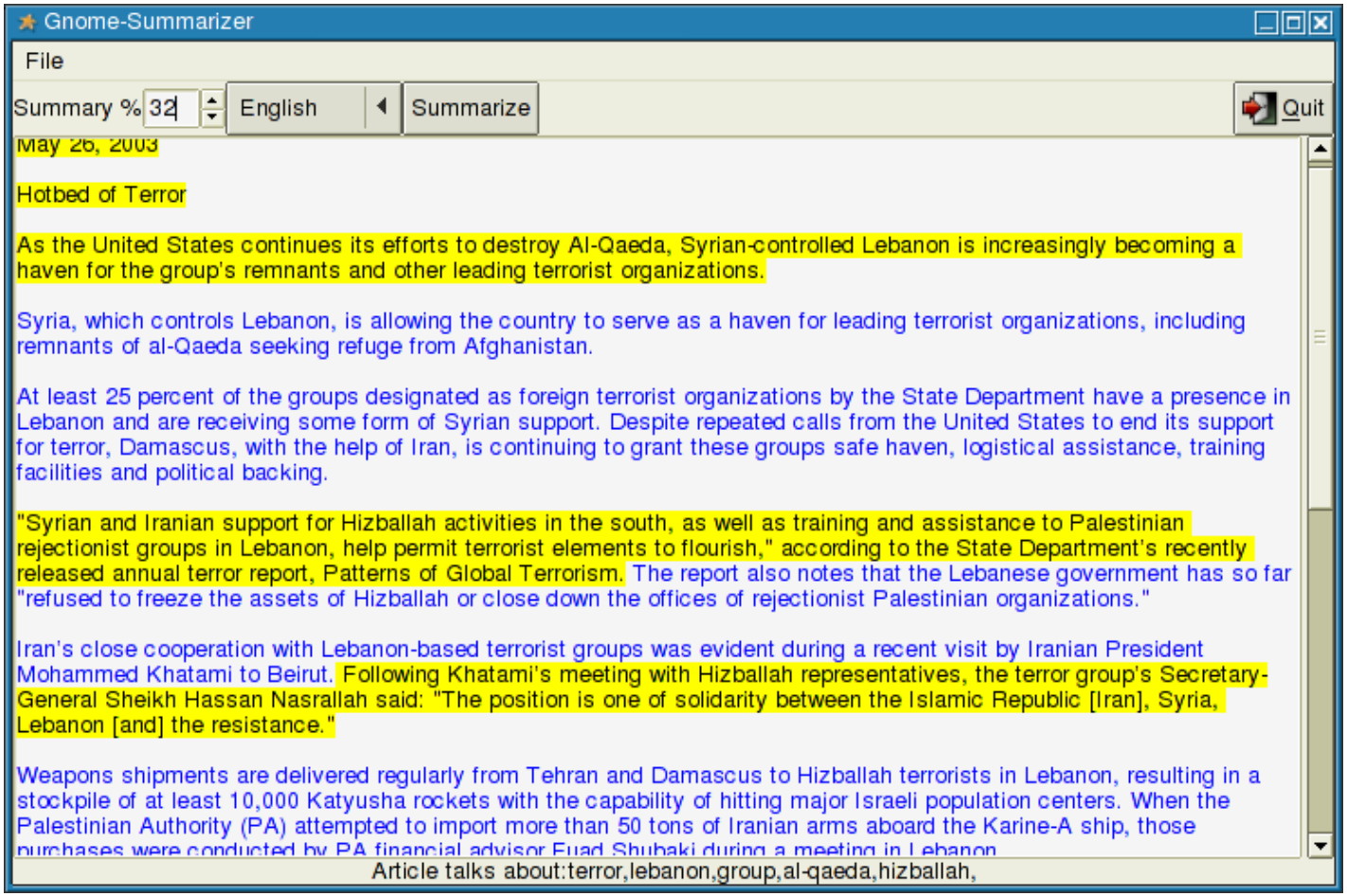
This algorithm isn't hard to implement.
Quick trick. Count vowel clusters in words to estimate syllables

Summarization is about a succient distinct distillation of the relevant content in a document.


DO YOU HAVE ANY THOUGHTS REGARDING THE PARTICLE ACCELERATOR'S TERTIARY F.E.L. GUIDANCE SYSTEM?
WE CAN'T PUT THE BROKEN PART IN THE MACHINE. IT WOULDN'T SMASH THE RIGHT TINY THINGS TOGETHER. THEN THE MACHINE MIGHT BREAK.
THAT WOULD BE VERY BAD.
I SPENT ALL NIGHT READING SIMPLE.WIKIPEDIA.ORG, AND NOW I CAN'T STOP TALKING LIKE THIS.

Most approaches involve selecting out the most relevant sentences
The simplest technique is to just look for sentences with popular terms

I use libots, as an example but there is much better work



Demo!

Topic Modeling

Topic Modeling is a way to group and categorize documents
Usually unsupervised approach

217	274	126	63	200
INSECT	SPECIES	GENE	STRUCTURE	FOLDING
MYB	PHYLOGENETIC	VECTOR	ANGSTROM	NATIVE
PHEROMONE	EVOLUTION	VECTORS	CRYSTAL	PROTEIN
LENS	EVOLUTIONARY	EXPRESSION	RESIDUES	STATE
LARVAE	SEQUENCES	TRANSFER	STRUCTURES	ENERGY
42	**2**	**280**	**15**	**64**
NEURAL	SPECIES	SPECIES	CHROMOSOME	CELLS
DEVELOPMENT	GLOBAL	SELECTION	REGION	CELL
DORSAL	CLIMATE	EVOLUTION	CHROMOSOMES	ANTIGEN
EMBRYOS	CO2	GENETIC	KB	LYMPHOCYTES
VENTRAL	WATER	POPULATIONS	MAP	CD4
112	**210**	**201**	**165**	**142**
HOST	SYNAPTIC	RESISTANCE	CHANNEL	PLANTS
BACTERIAL	NEURONS	RESISTANT	CHANNELS	PLANT
BACTERIA	POSTSYNAPTIC	DRUG	VOLTAGE	ARABIDOPSIS
STRAINS	HIPPOCAMPAL	DRUGS	CURRENT	TOBACCO
SALMONELLA	SYNAPSES	SENSITIVE	CURRENTS	LEAVES
39	**105**	**221**	**270**	**55**
THEORY	HAIR	LARGE	TIME	FORCE
TIME	MECHANICAL	SCALE	SPECTROSCOPY	SURFACE
SPACE	MB	DENSITY	NMR	MOLECULES
GIVEN	SENSORY	OBSERVED	SPECTRA	SOLUTION
PROBLEM	EAR	OBSERVATIONS	TRANSFER	SURFACES

Topic Modeling - continued

CRAN includes a package for topicmodeling
This package using LDA and CTM

LDA - Latent Dirchilet Allocation

$$\theta_{k|j} \sim D[\alpha]$$

$$\phi_{w|k} \sim D[\beta]$$

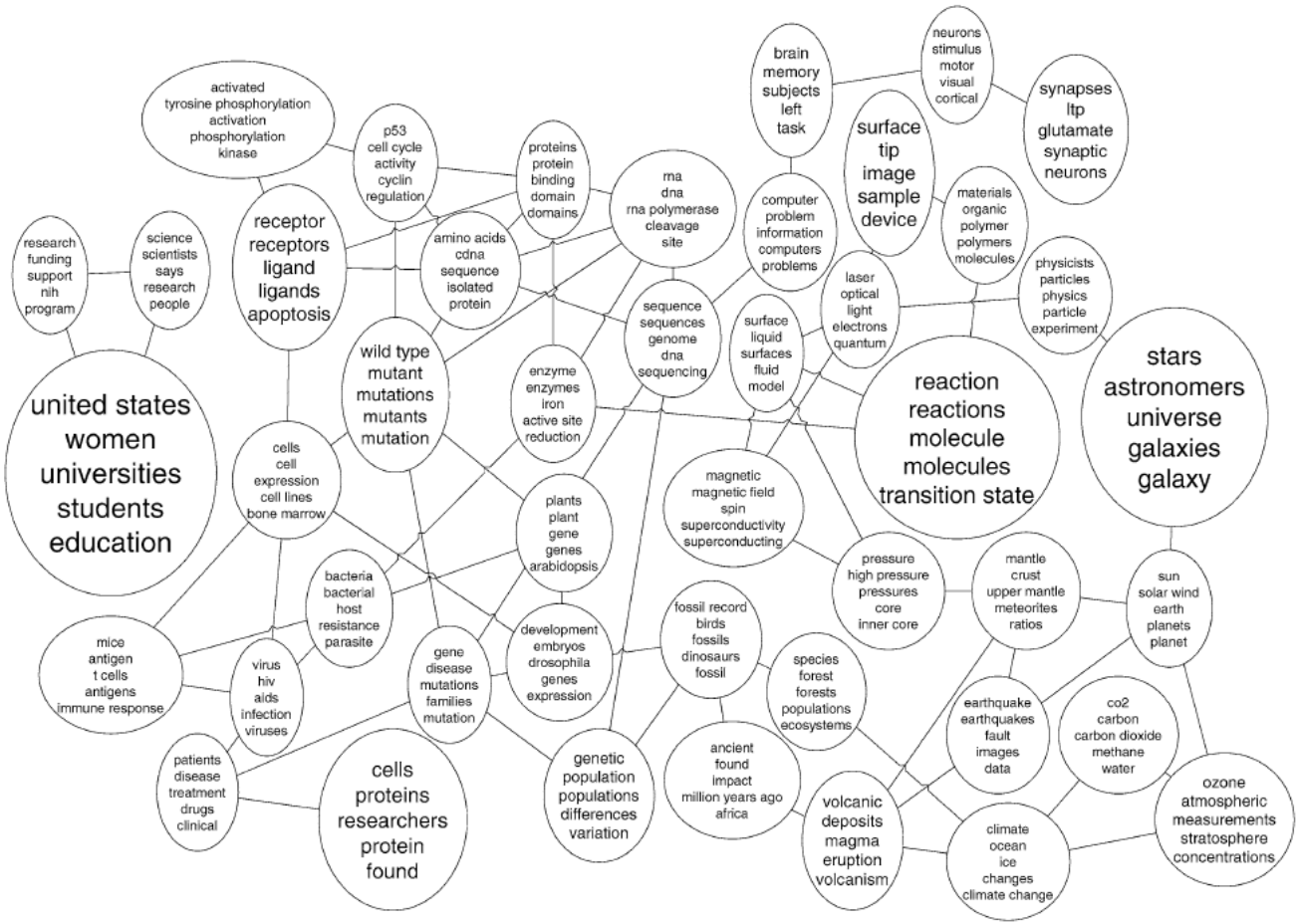
$$z_{ij} \sim \theta_{k|j}$$

$$x_{ij} \sim \phi_{w|z_{ij}}$$

$$p(z_{ij} = k|\mathbf{z}^{\neg ij}, \mathbf{x}, \alpha, \beta) \propto (\alpha + n_{k|j}^{\neg ij})(\beta + n_{x_{ij}|k}^{\neg ij})(W\beta + n_{k}^{\neg ij})^{-1}$$

$$n_{jkw} = \#\{i : x_{ij} = w, z_{ij} = k\}$$

CTM - Coorelated Topic Models



CTM - Coorelated Topic Models

$$\theta_{k|j} \sim log(N(\mu, \Sigma))$$

$$\phi_{w|k} \sim D[\beta]$$

$$z_{ij} \sim \theta_{k|j}$$

$$x_{ij} \sim \phi_{w|z_{ij}}$$

Demo!

Sentiment analysis is about gauging mood based on the text.

7 of 7 people found the following review helpful:

★★★★★ **This Milk Changed My Life**, August 8, 2010

By **Robert D. Queen "itcbob"** (Springfield, VA) - See all my reviews

REAL NAME™

This review is from: Tuscan Whole Milk, 1 Gallon, 128 fl oz (Misc.)

The Tuscan whole milk is the most amazing drink I have ever had. I used to be an alcoholic, but after one drink of this amazing milk, alcohol has never touched my lips again. Why drink bourbon when this amazing milk from the hills of Tuscany is now available to us all. Nothing short of the Second Coming compares to the sight of Tuscan Milk. Less than $100 per gallon is a steal. Don't miss out on the amazing opportunity to experience Tuscany as its finest.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comment

14 of 19 people found the following review helpful:

★☆☆☆☆ **No Protection at All**, August 8, 2007

By **J. McArthur** - See all my reviews

REAL NAME™

This review is from: JL421 Badonkadonk Land Cruiser/Tank

My wife and kids were playing in my JL421, and I thought I would give them a bit of a scare as a joke, so a shot a few rounds at the side with a rather large gun that I have and the bullets penetrated right through and killed them all! I am so disappointed with the quality of this land cruiser. I called the manufacturer and they said it wouldn't be covered under warranty because I did it intentionally. I'm never buying from this company again.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comment

Opinion corpus available at:

Wiebe's corpora
http://www.cs.pitt.edu/mpqa/
Sentiwordnet:
http://sentiwordnet.isti.cnr.it/

For more sophistication

- Best solved using a Conditional Random Field
- This area is still new
- No R libraries
- Entity Extraction needed for more fine-grained sentiment

Demo!

Named Entity Recognition

The purpose of NER is to extract out and label phrases in a sentence

Bill Clinton arrived at the United Nations Building in Manhattan.

Challenges

- People and locations may be referred to in ambiguous ways.
- Entity may never have been seen before
- Entity may be referred to with pronouns
- Wikipedia and Capitalization heuristics aren't good enough

I use the Illinois Named Entity Extractor
http://cogcomp.cs.illinois.edu/page/software_view/4

Demo!

Conclusions

There are lots of interesting things you can do with text mining
R is very good at integrating all of them.

Questions?