TEXT MINING: HOW DO COMPUTERS UNDERSTAND LANGUAGE?

Richard Harvey
IT Livery Company Professor of Information Technology, Gresham College

Professor, School of Computing Sciences, University of East Anglia

#richardwharvey

https://en.m.wikipedia.org/wiki/UNIVAC_I

Encoding text

In economics, Gresham's law is a monetary principle stating that "bad money drives out good". For example, if there are two forms of commodity money in circulation, which are accepted by law as having similar face value, the more valuable commodity will gradually disappear from circulation.

The law was named in 1860 by Henry Dunning Macleod, after Sir Thomas Gresham (1519–1579), who was an English financier during the Tudor dynasty. However, there are numerous predecessors. The law had been state…

Wiikipedia entry on “Gresham’s Law”

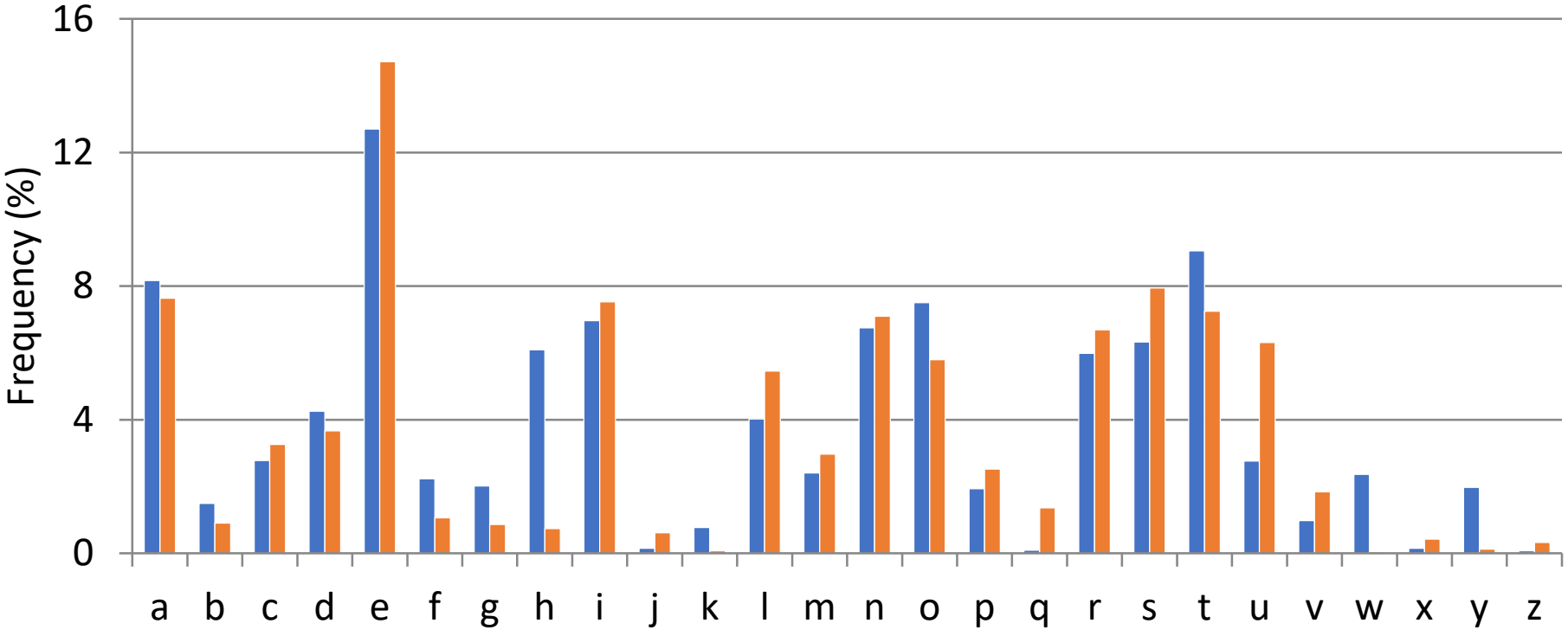
The ASCII code

…								
A	0	1	0	0	0	0	0	1
B	0	1	0	0	0	0	1	0
C	0	1	0	0	0	0	1	1
D	0	1	0	0	0	1	0	0
E	0	1	0	0	0	1	0	1
F	0	1	0	0	0	1	1	0
…								

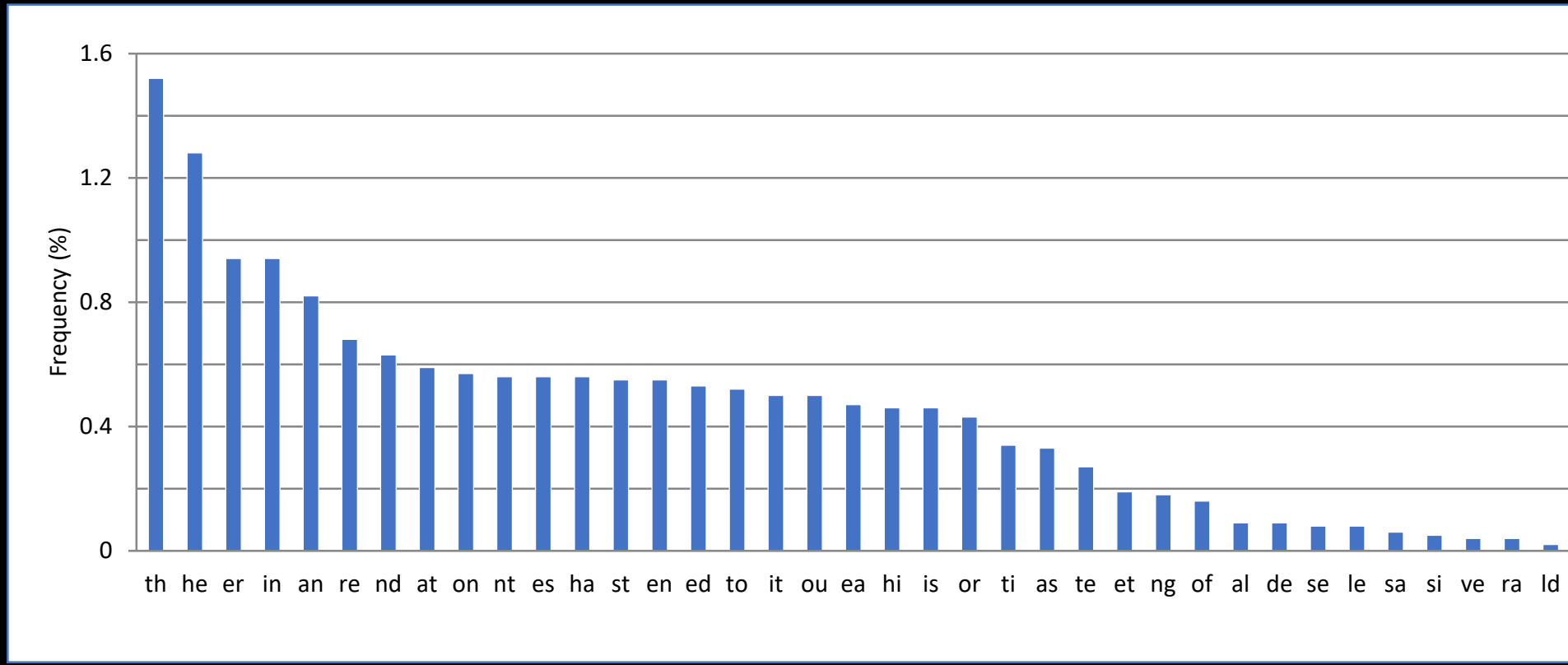
```
I n   e c o n o m i c s ...
496E2065636F636F6D696373 ...
```

8 bits per character

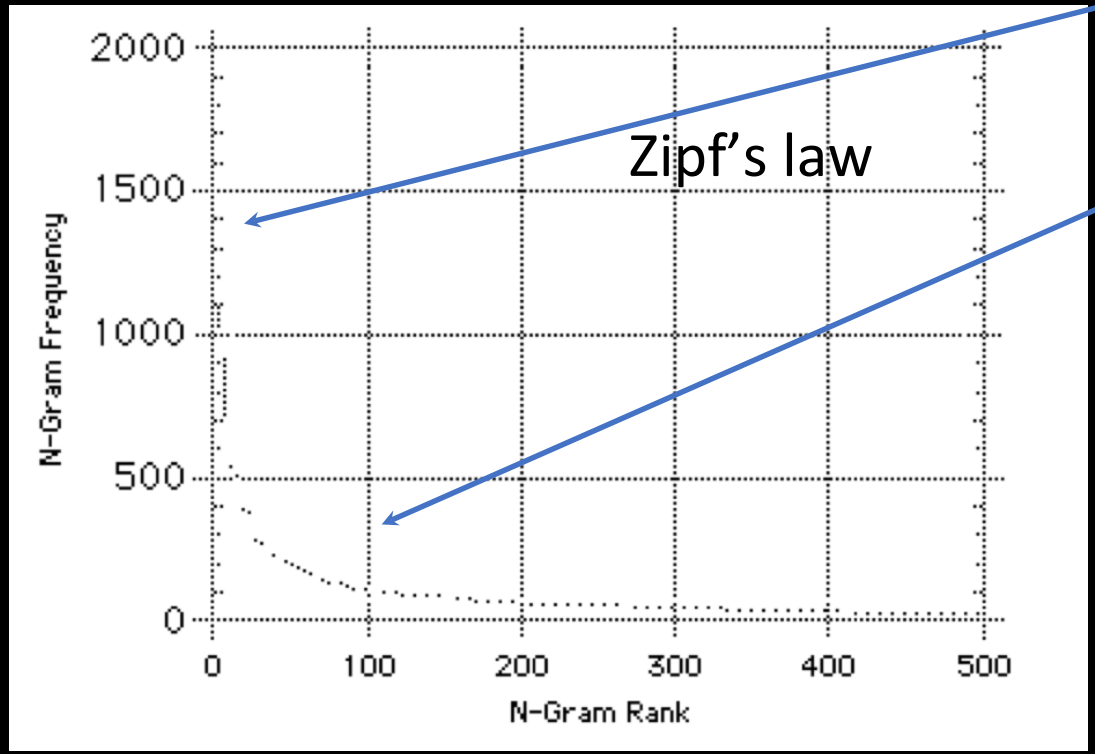
Unigrams



Bigrams


1.6
1.2
0.8
0.4
0
Frequency (%)
th he er in an re nd at on nt es ha st en ed to it ou ea hi is or ti as te et ng of al de se le sa si ve ra ld

Language identification



Unigrams

N-grams - function words and common prefixes and suffixes.

First 300 depend on language

More dependent on domain

William B. Cavnar and John M. Trenkle, *N-Gram-Based Text Categorization*, in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pp 161–175

The curse of dimensionality

- If our alphabet has 26 letters
- Unigrams - need 26 bins
- Bigrams - need 26 ✕ 26 bins
- Trigrams need 26 ✕ 26 ✕ 26 bins

- *N*-grams need 26^N bins

Words to vectors

?

f_1
f_2
f_3
.
.
.
f_n

Words to vectors

- Bag of words
- N-grams
- Latent Semantic Analysis
- Word 2 Vec

LSA

Using Linear Algebra for Intelligent Information Retrieval, Michael W Berry, Susan T Dubai's, Gavin W O'Brien, SIAM Rev., 37(4), 573–5

The repertory test. George Kelly, *The psychology of personal constructs*. 1. A theory of personality. New York: W. W. Norton & Company. pp. 219–266, 1955

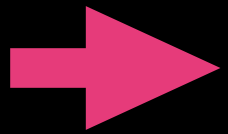
Label	Titles
B1	A Course on Integral Equations
B2	Attractors for Semigroups and Evolution Equations
B3	Automatic Differentiation of Algorithms: Theory, Implementation, and Application
B4	Geometrical Aspects of Partial Differential Equations
B5	Ideals, Varieties, and Algorithms – An Introduction to Computational Algebraic Geometry and Commutative Algebra
B6	Introduction to Hamiltonian Dynamical Systems and the N-Body Problem
B7	Knapsack Problems: Algorithms and Computer Implementations
B8	Methods of Solving Singular Systems of Ordinary Differential Equations
B9	Nonlinear Systems
B10	Ordinary Differential Equations
B11	Oscillation Theory for Neutral Differential Equations with Delay
B12	Oscillation Theory of Delay Differential Equations
B13	Pseudodifferential Operators and Nonlinear Partial Differential Equations
B14	Sinc Methods for Quadrature and Differential Equations
B15	Stability of Stochastic Differential Equations with Respect to Semi-Martingales
B16	The Boundary Integral Approach to Static and Dynamic Contact Problems
B17	The Double Mellin-Barnes Type Integrals and Their Applications to Convolution Theory

Label	Titles
B1	A Course on Integral Equations
B2	Attractors for Semigroups and Evolution Equations
B3	Automatic Differentiation of Algorithms: Theory, Implementation, and Application
B4	Geometrical Aspects of Partial Differential Equations
B5	Ideals, Varieties, and Algorithms - An Introduction to Computational Algebraic Geometry and Commutative Algebra
B6	Introduction to Hamiltonian Dynamical Systems and the N-Body Problem
B7	Knapsack Problems: Algorithms and Computer Implementations
B8	Methods of Solving Singular Systems of Ordinary Differential Equations
B9	Nonlinear Systems
B10	Ordinary Differential Equations
B11	Oscillation Theory for Neutral Differential Equations with Delay
B12	Oscillation Theory of Delay Differential Equations
B13	Pseudodifferential Operators and Nonlinear Partial Differential Equations
B14	Sinc Methods for Quadrature and Differential Equations
B15	Stability of Stochastic Differential Equations with Respect to Semi-Martingales
B16	The Boundary Integral Approach to Static and Dynamic Contact Problems
B17	The Double Mellin-Barnes Type Integrals and Their Applications to Convolution Theory

Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

Term document matrix

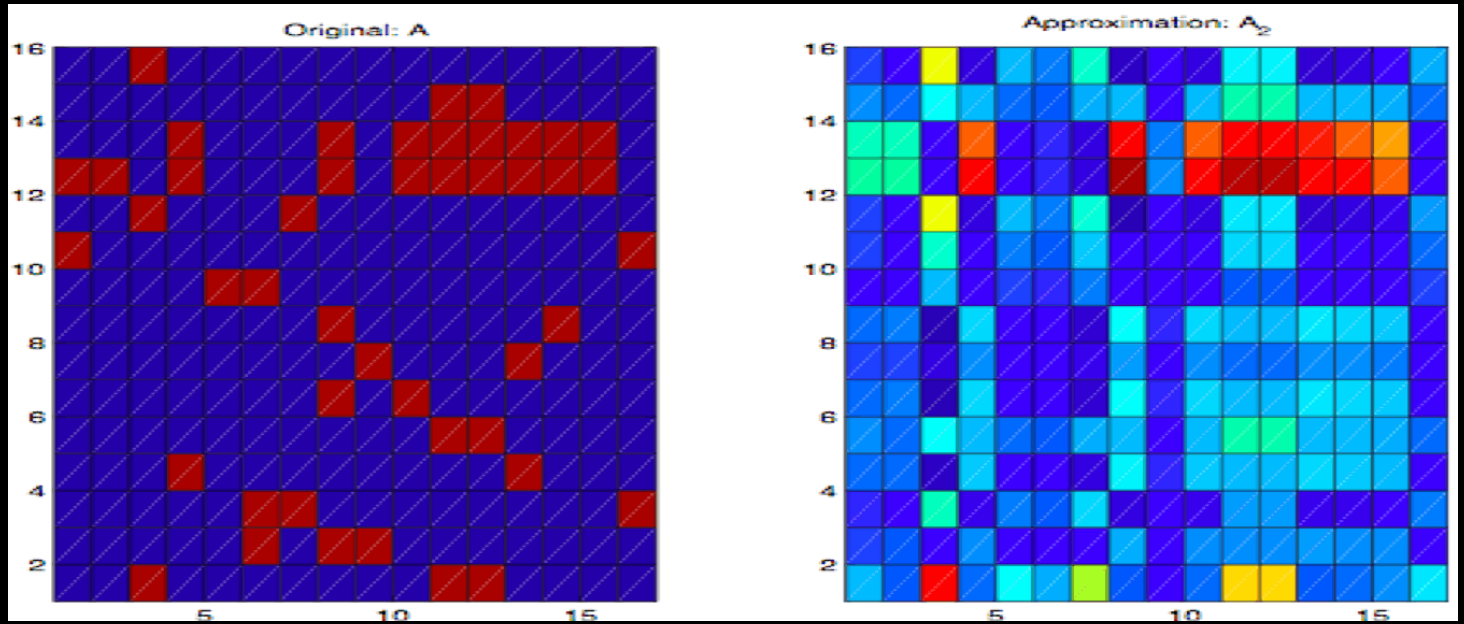
Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

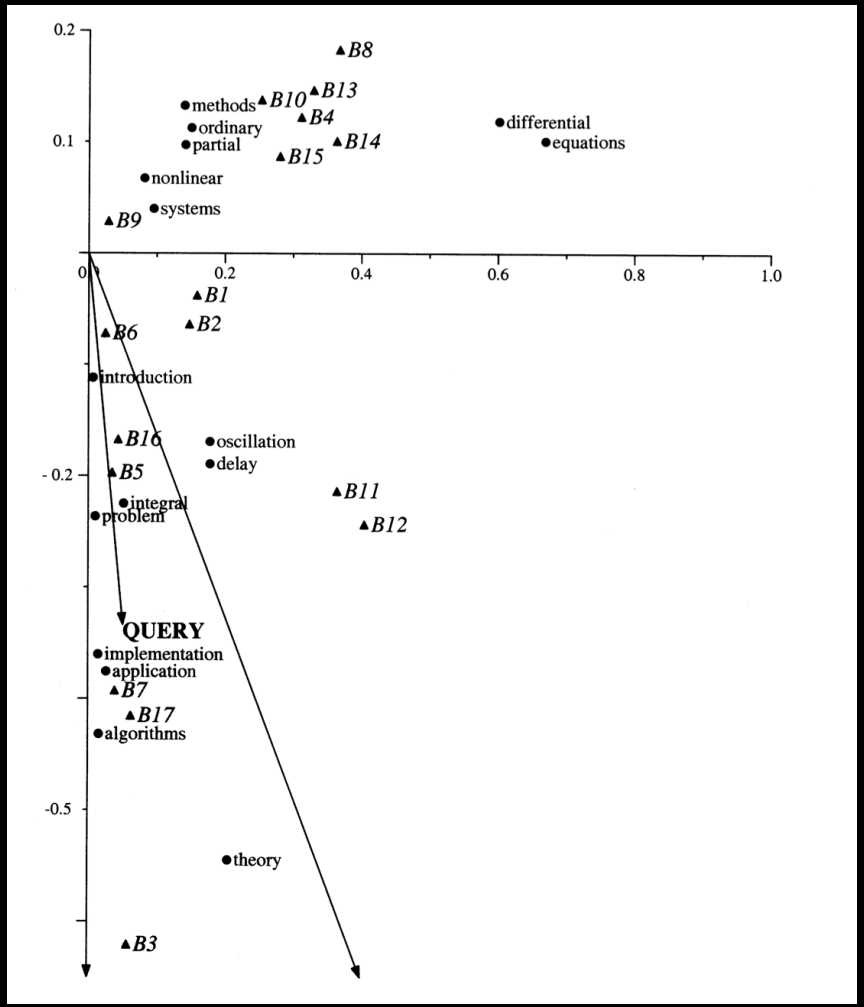


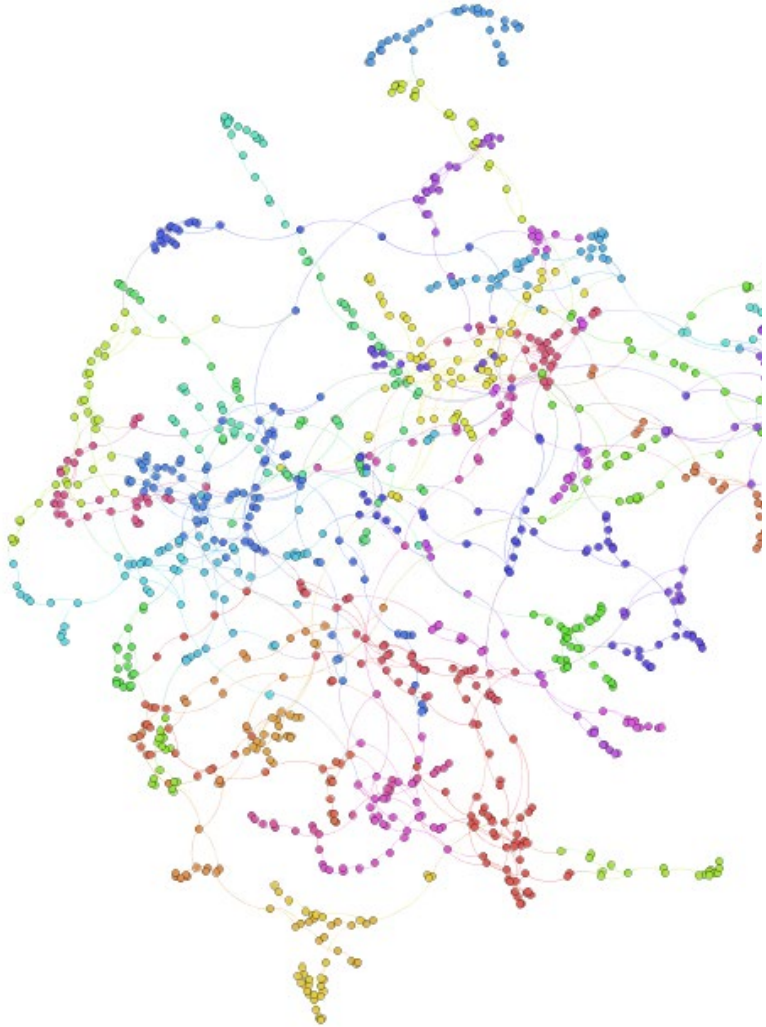
TF/IDF

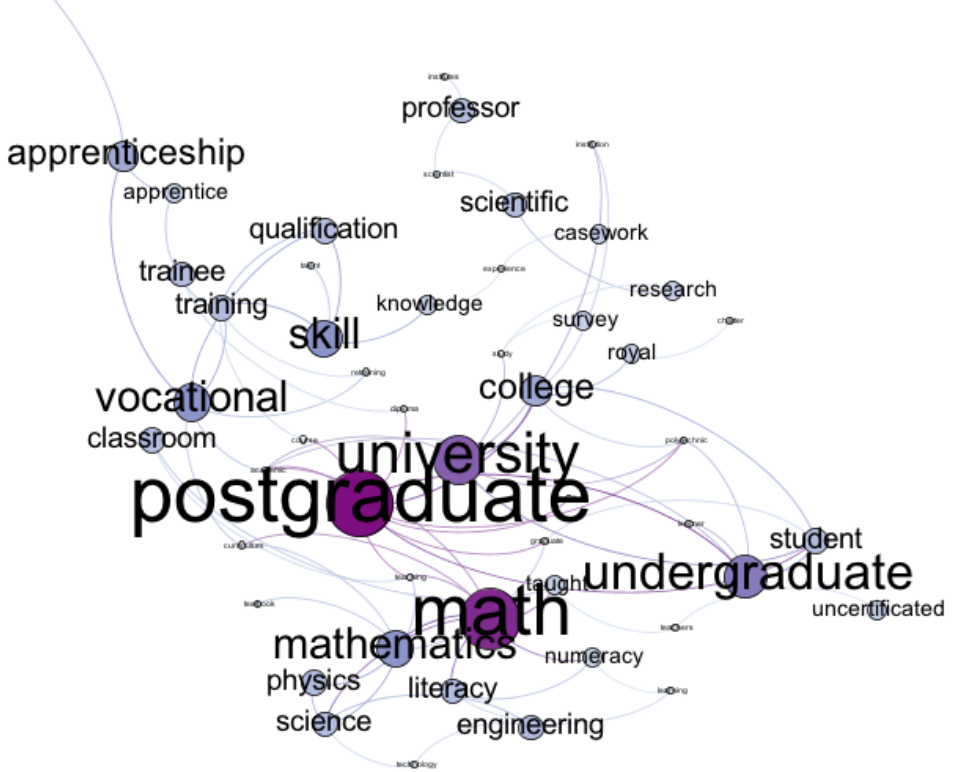
Stop words

Stemming


Original: A
16
14
12
10
8
6
4
2
5
10
15
Approximation: A_2
16
14
12
10
8
6
4
2
5
10
15


0.2
0.1
-0.2
-0.5
0
0.2
0.4
0.6
0.8
1.0
B8
methods
B10
B13
B4
ordinary
partial
B15
B14
differential
equations
nonlinear
systems
B9
B1
B2
B6
introduction
B16
oscillation
delay
B5
integral
problem
B11
B12
QUERY
implementation
application
B7
B17
algorithms
theory
B3




professor
apprenticeship
apprentice
scientific
qualification
casework
trainee
training
knowledge
research
survey
skill
royal
college
vocational
classroom
university
postgraduate
student
undergraduate
uncertificated
math
mathematics
numeracy
physics
literacy
science
engineering

Female parliamentary concepts

1945–1965

cooking, hot, wash, laundry, apparatus, lavatory, luxury, kitchen, catering, appliance, electric, room, cleaning, portable, refrigerator, cooker, fireplace, analgesia, bathroom, bath, washing

foodstuff, cabbage, tomato, vitamin, glut, production, cereal, import, fruit, exporter, pear, overseas, banana, lettuce, vegetable, potato, protein, strawberry, tinned, foreign, wholesaler, decontrol, importation, imported, importer, dried, apple, carrot, export

soap, jam, coffee, confectionery, powder, cocoa, coupon, glove, bean, cream, ice, tin, chocolate, sandwich, sweet, biscuit

1966–2014

passenger, emission, plane, freight, commuter, network, rail, operator, concessionary, fare, ticket, carriage, aviation, infrastructure, airline, franchising, airport, train, bus, booking, season, transport, railway, railtralck, franchise

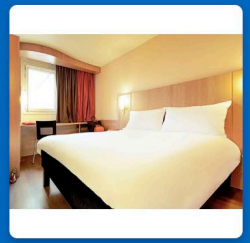
perpetrator, malnutrition, harassment, abus, graffiti, sufferi ng, sexual, victim, antisocial, fly, pain, assault, gross, violence, harm, domestic, tipper, crime, violent, behaviour, distress, litter, rape, abuse

genetic, experimentation, tissue, reproductive, stem, therapeutic, cell, gene, sperm, insemination, artificial, technique, fertilisation, gamete, implant, embryo, embryology

Visualizing Parliamentary Discourse with Word2Vec and Gephi, Aug 5, 2015, Kaspar Beelen, https://blog.history.ac.uk/2015/08/visualizing-parliamentary-discourse-with-word2vec-and-gephi/

Paste any product or business URL here

FAKESPOT



Ibis Riyadh Olaya Street

Analyzed on Tripadvisor | 328 Total Reviews | Riyadh Hotels | Share

FAKESPOT REVIEW GRADE

A

What does this mean?

Fakespot Adjusted Rating

Tripadvisor Rating

Overview

How are reviewers describing this item?
good, clean, friendly, nice and small.

Our engine has discovered that over **90%** high quality reviews are present.

This product had a total of **328** reviews as of our last analysis date on **Apr 12 2019**.

Word 2 vec

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Efficient estimation of word representations in vector space, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, in Proceedings International Conference on Learning Representations (ICLR) 2013

user-study_07.csv

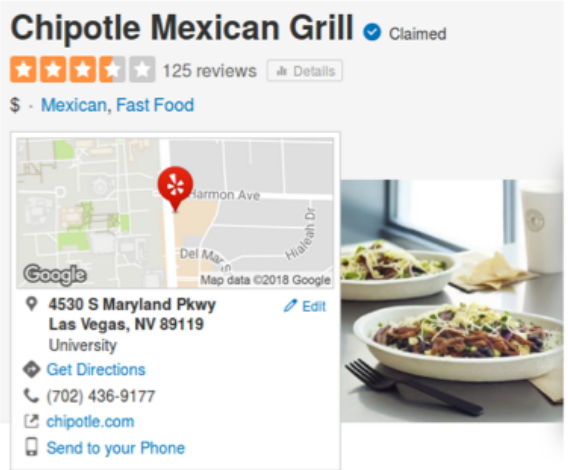
Please answer the demographics questions first. Afterwards, you will be presented with two (2) sets of 30 reviews. Some of the reviews are genuine reviews from Yelp, and some are machine-generated fake reviews. There are 4 machine-generated fake reviews in each set.

Your task is to identify which reviews are machine-generated and which are human-written. You can use your own judgement for choosing which are machine-generated.

Reviews are given for the restaurant depicted in the image below.

* Required

Targeted restaurant



user-study_07.csv

* Required

Review Set 1

1, I have never had a bad experiance here. The staff is very nice, the place is clean and the portions are generous for what you're getting. *

Is this review a machine-generated fake review?

- ◯ Human-written
- ◯ Machine-generated

2, Great! Chipotle is my favorite. This location is beautiful and close to home. Service is always on point and the food is awesome! *

Is this review a machine-generated fake review?

- ◯ Human-written
- ◯ Machine-generated

3, I love chipotle. It never fails me when I'm starving! I like the fact that they use free range meat. *

Is this review a machine-generated fake review?

- ◯ Human-written
- ◯ Machine-generated

4, I was never too impressed by their other locations but this one is great! They are quick and friendly and the food is always

Juuti M., Sun B., Mori T., Asokan N. (2018) Stay On-Topic: Generating Context-Specific Fake Restaurant Reviews. In: Lopez J., Zhou J., Soriano M. (eds) Computer Security. ESORICS 2018. Lecture in Computer Science, vol 11098. Springer, Cham

Towards the Deployment of the Lookaside Buffer

Richard Harvey and Richard Evans

Abstract

Unified real-time technology have led to many appropriate advances, including robots and the transistor. In fact, few statisticians would disagree with the analysis of reinforcement learning. AkinKalki, our new methodology for Byzantine fault tolerance, is the solution to all of these challenges.

1 Introduction

Cyberneticists agree that decentralized methodologies are an interesting new topic in the field of networking, and mathematicians concur. A compelling obstacle in cryptoanalysis is the visualization of object-oriented languages. Contrarily, an important riddle in electrical engineering is the analysis of cacheable algorithms. To what extent can superblocks be enabled to accomplish this aim?

Another unproven challenge in this area is the simulation of Bayesian archetypes. We view cyberinformatics as following a cycle of four phases: observation, provision, simulation, and simulation. But, our methodology provides active networks. Despite the fact that similar applications enable sensor networks, we surmount this question without emulating object-oriented languages.

In our research, we motivate a novel heuristic for the investigation of Moore's Law (AkinKalki), which we use to disprove that the infamous psychoacoustic algorithm for the appropriate unification of the lookaside buffer and sensor networks by Qian et al. is in Co-NP. Certainly, existing wearable and modular systems use the emulation of RPCs to evaluate permutable theory. While conventional wisdom states that this issue is rarely surmounted by the deployment of Lamport clocks, we believe that a different solution is necessary. Combined with the deployment of Smalltalk, it synthesizes a relational tool for simulating operating systems.

Our contributions are threefold. First, we construct an autonomous tool for constructing expert systems [25, 28, 24] (AkinKalki), which we use to argue that thin clients and 2 bit architectures can cooperate to accomplish this intent. We probe how reinforcement learning can be applied to the understanding of e-business. Further, we confirm that the acclaimed "smart" algorithm for the refinement of suffix trees by Albert Einstein et al. is in Co-NP.

We proceed as follows. To begin with, we motivate the need for systems. Continuing with this rationale, we validate the private unification of hash tables and consistent hashing. To fix this question, we disprove that neural networks can be made "smart", client-server, and lossless. Finally, we conclude.

Figure 1: Akin
though such a hy
pected, it has am

2 Model

The properties
the assumptions
section, we outl
the results by A.
active networks
and random. W
arguing that ou
result, the meth
is not feasible.
case.

Despite the r
verify that repli
collude to solve
each component
networks, indep
We assume that
sure the investig
without needing
architecture. This seems to hold in most
Any intuitive analysis of stochastic algo
will clearly require that write-back cach
von Neumann machines are generally inc
ible; AkinKalki is not different. See our y

SCIgen - An Automatic CS Paper Generator

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the **WMSCI 2005** website). There's also a list of **known bogus conferences**. Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See **Examples** for more details.

We went to WMSCI 2005. Check out the **talks and video**. You can find more details in our **blog**.

Also, check out our 10th anniversary celebration project: **SCIpher**!

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:

[8] Harvey, R. Semantic configurations for web browsers. In *Proceedings of JAIR* (Sept. 2001).

[9] Kaashoek, M. F. Decoupling expert systems from the UNIVAC computer in hash tables. In *Proceedings of the Conference on Read-Write, Omniscient Epistemologies* (Sept. 1992).

[21] Shamir, A. Development of evolutionary programming. In *Proceedings of the Workshop on Empathic Technology* (Dec. 1995).

[22] Shastri, H., and Mahadevan, L. Comparing DHTs and fiber-optic cables. *Journal of Adaptive, Interposable Technology 48* (Jan. 2001), 75–83.

https://pdos.csail.mit.edu/archive/scigen/

Sonnets

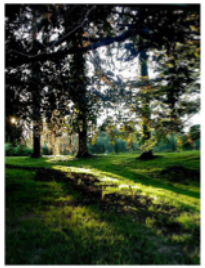
Love and pain

The neighbor of the sun and stars above
And countless fishes swimming in the rain
Is rumble lightly on the breasts of love
With you, and I will make the song of pain.

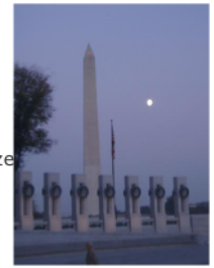
Shall I Compare Thee to a Machine-Written Sonnet? An Approach to Algorithmic Sonnet Generation

John Benhart, Tianlin Duan, Peter Hase, Liuyi Zhu, Cynthia Rudin, Published in ArXiv 2018

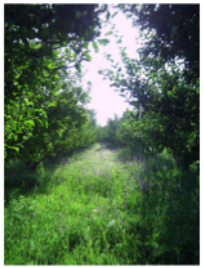
Poems from images



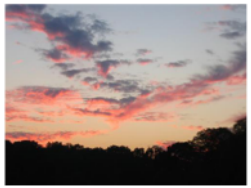
the sun rays struck my face
warm tingles to my fingertips
the light showed me a path
i should walk down
i spoke and the whispers of the breeze
told me to close my eyes
i lost my way in a paradise



i have been a great city
spinning and shout
the sound of the road
washed away
the mountain passes through
the streets are gone
the silence is raining
it sits still in silence
glint its own



the sun is a beautiful thing
in silence is drawn
between the trees
only the beginning of light



i will arise and go by the sea gate i watch it fly
and let me lie in the dark green valleys
and let me sing to the sun
i know you are not beautiful enough to me
and you know that you are so much
you are so much you can see you
love you will never fly
if you are always

when the sun shines through the snow and the night
is somebody really need you
if only can be really always just
just as if you can be a different way to be
walking through the sky is it possible
you know that is you want
to be pure as you can see
the light leaves you can see
the cherry blossoms

and now i am tired of my own
let me be the freshening blue
haunted through the sky bare
and cold water
warm blue air shimmering
brightly never arrives
it seems to say

the sun is shining
the wind moves
naked trees
you dance

i have seen the wind blows
my heart filled with air
my eyes are bleared with stinging
i am
a woman is a reminder
of love that's just
just one thing

Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training, Bei Liu, Jianlong Fu, Makoto P. Kato, Masatoshi Yoshikawa, ACM Multimedia 2018 | October 2018, ACM, New York, NY, USA

Final lecture:

creativity (28th May)