Google Similarity Distance

Presented by:
Akshay Kumar
Pankaj Prateek

Are these similar?

- Number ‘1’ vs. color ‘red’
- Number ‘1’ vs. ‘small’
- Horse vs. Rider
- True vs. false
- ‘Monalisa’ vs. ‘Virgin of the rocks’

We need some universal similarity measure!!!

Information Distance

- *E(x, y) :* Given two strings x and y, the length of the shortest binary program, in the reference universal computing system, such that the program computes output y from input x, and also output x from input y is known as the information distance between x and y
- Up to a negligible logarithmic additive term,

$$E(x, y) = K(x, y) - \min\{K(x), K(y)\}$$

Information Distance

- Determines the distance between two strings minorizing the ***dominant*** feature in which they are similar.
- Not a good measure
  - If two small strings differ by an ID which is large compared to their lengths, then the strings are not similar. However, if two very large strings differ by the same distance, they are very similar

Normalized Information Distance

- To ensure that the ID expresses similarity between two strings, normalize it over the length of the strings

$$NID(x,y) = \frac{K(x,y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}$$

- Kolmogorov complexity is **uncomputable**!!!

Normalized Compression Distance

- To counter the problem of uncomputability of K(x), it's replaced by C(x) where C is some compression technique

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Normalized Google Distance

- NGD : a way of expressing NCD
- Premise :
  - Number of web pages indexed by Google is so vast that is actually approximates the actual relative frequency of various search terms
  - Invariant to growing size of Google database
- Uses probability of search terms to define similarity distance

Normalized Google Distance

S = set of Google search terms

Ω = set of webpage indexed by Google ; M=|Ω|

Assumption: All web pages have equal probability

- Event = subset of Ω
- A search term x defines an event **x** which is the set of webpages containing the word x
- L : Ω → [0,1] be a uniform mass probability distribution such that

$$L(x) = \frac{|\boldsymbol{x}|}{M};\ L(x \cap y) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{M}$$

Normalized Google Distance

- If we consider each event as code-word for encoding the Google Distance then Kraft's Inequality is violated since some webpage can have more than 1 search terms.
- The solution is to normalize.

$$NGD(x,y) = \frac{\max\{\log f(x),\ \log\ f(y)\}\ - \log f(x,y)}{\log N\ - \min\{\log f(x), \log f(y)\}}$$

where f(x) denotes the number of pages containing f(x)
and f(x, y) denotes the number of pages containing both x & y

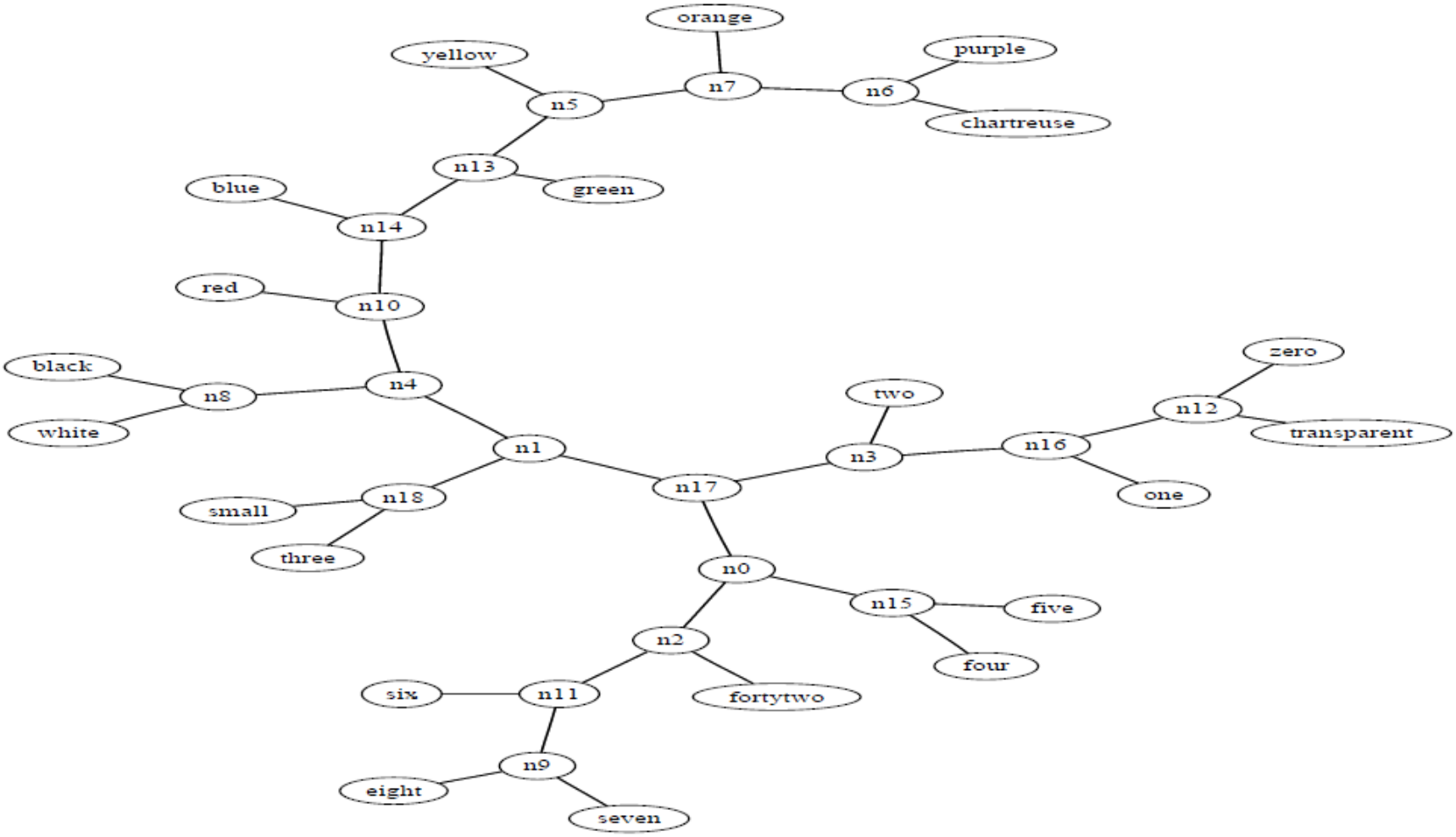
Normalized Google Distance

$$NGD(x,y) = \frac{\max\{\log f(x),\ \log\ f(y)\}\ - \log f(x,y)}{\log N\ - \min\{\log f(x), \log f(y)\}}$$

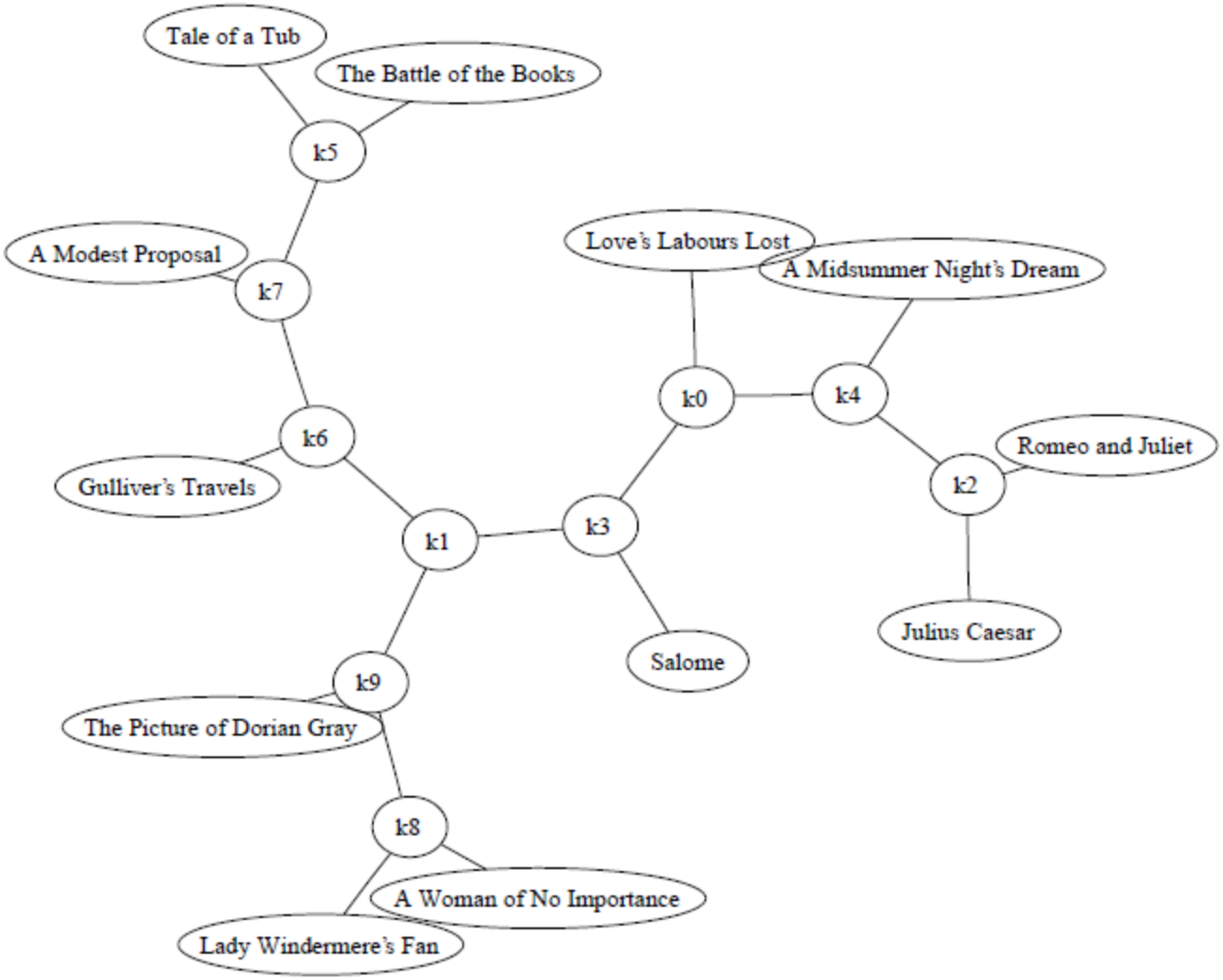
Properties :

1) $NGD(x,y) \geq 0$
2) $x = y\ or\ (x \neq y\ and\ f(x) = f(y) = f(x,y) > 0) \Rightarrow NGD(x,y) = 0$
3) $f(x) = 0 \Rightarrow f(x,y) = 0\ \forall\ y$
   Hence NGD(x, y) = ∞/∞ = 1 (by definition)
4) $NGD(x,x) = 0$
5) $NGD(x,y) = NGD(y,x)$
6) NGD is scale invariant

Example


orange
yellow
purple
n7
n5
n6
chartreuse
n13
blue
green
n14
red
n10
zero
black
n4
two
n12
n8
transparent
white
n1
n3
n16
n17
one
small
n18
three
n0
n15
five
n2
four
six
n11
fortytwo
n9
eight
seven

Example


Tale of a Tub
The Battle of the Books
k5
A Modest Proposal
k7
Love's Labours Lost
A Midsummer Night's Dream
k0
k4
k6
Romeo and Juliet
Gulliver's Travels
k2
k1
k3
Julius Caesar
Salome
k9
The Picture of Dorian Gray
k8
A Woman of No Importance
Lady Windermere's Fan

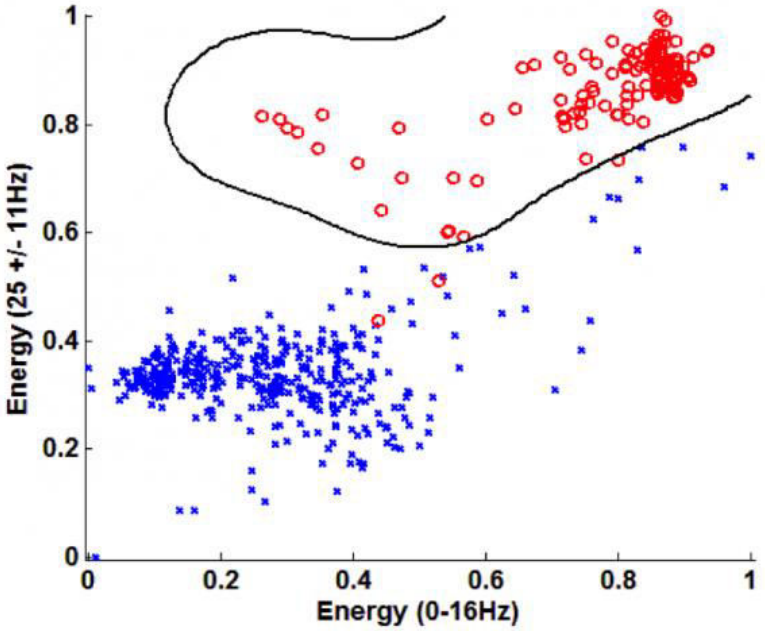
Errors/Shortcomings

- Not applicable to small datasets.
- Definition of N is still erroneous, the NGD can still be greater than 1 (Kraft's inequality is violated)
- Ignored the number of occurrence of a term on a page
- Page count analysis ignore the position of a word in a page
  - Even though 2 words appear on the same page, they may not be related.
- Page count of a polysemous word might contain a combination of all its senses, e.g. - Apple.
- The probability of occurrence of every page is taken to be the same (it is **NOT** the same, depends on the page rank)
- Given the scale and noise on the WWW, some words might just occur on a page arbitrarily (by random chance)

Suggestions

- Can use snippets which are returned along with the web searches to determine the similarity of the words
  - Only the snippets of the top ranking pages can be processed efficiently and no guarantee exists that all the information we need to measure the semantic similarity is contained in those snippets
- Combining information from the web-searches with the information obtained from other databases (Wikipedia, WordNet etc.) to find the similarity measure.

SVM – NGD Learning



- A set of n-dimensional data points along with their classes i.e. (x_i, y_i). Only 2 classes taken here for brevity.
- Divide this set into two parts : training set and testing set
- Determine a dividing curve to differentiate the points of two classes using learning set.
- Validate the results by testing on testing set.

Image Source: http://groups.csail.mit.edu/ddmg/drupal/sites/default/files/images/2.preview.png

SVM – NGD Learning

- Set of training words
- Set of anchor words of cardinality n(much smaller than training words set)
- Convert each of training word into a n-dimensional vector whose i^{th} dimension is NGD between that word and i^{th} anchor word
- Train it using SVM

Training Data

Positive Training	(21 cases)			
11	13	17	19	2
23	29	3	31	37
41	43	47	5	53
59	61	67	7	71
73				
Negative Training	(22 cases)			
10	12	14	15	16
18	20	21	22	24
25	26	27	28	30
32	33	34	4	6
8	9			
Anchors	(5 dimensions)			
composite	number	orange	prime	record

Testing Results

	Positive tests	Negative tests
Positive Predictions	101, 103, 107, 109, 79, 83, 89, 91, 97	110
Negative Predictions		36, 38, 40, 42, 44, 45, 46, 48, 49

Accuracy 18/19 = 94.74%

NGD Translation

English	Spanish
tooth	diente
joy	alegria
tree	arbol
electricity	electricidad
table	tabla
money	dinero
sound	sonido
music	musica

Fig. 7. Given starting vocabulary

English	Spanish
plant	bailar
car	hablar
dance	amigo
speak	coche
friend	planta

Fig. 8. Unknown-permutation vocabulary

English	Spanish
plant	planta
car	coche
dance	bailar
speak	hablar
friend	amigo

Fig. 9. Predicted (optimal) permutation

References

- Cilibrasi, Rudi L., and Paul MB Vitanyi. "The Google similarity distance." *Knowledge and Data Engineering, IEEE Transactions on* 19.3 (2007): 370-383.
- Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka. "Measuring semantic similarity between words using web search engines." *www* 7 (2007): 757-766.
- http://en.wikipedia.org/wiki/Normalized_Google_distance