Google Samankaltaisuusetäisyys

Timo Sand

Tietojenkäsittelytieteen laitos
Helsingin Yliopisto
Helsinki

14. Marraskuu, 2013

Johdanto

- Internet on miljoonien käyttäjien täyttämä tietokanta
- Melkein jokainen mahdollinen aihe on katettu
- Sisällöt keskimäärin huonolaatuisia
- Aiheitten matala-arvoisa approksimaatioita

- GSD on yleinen menetelmä hyödyntää tätä huonolaatuista, ilmaista tietoa
- Maailman suurin semanttinen sahköinen tietokanta
- Hakukone palauttaa sivunumeroarvion mille tahansa hakukyselylle

Jäsentely

- Esimerkki
- Algoritmin perusta
- Samankaltaisuuden googlaus
- Sovellukset ja kokeilut
- Kertaus

Esimerkki

- Google haku sanoille "horse"ja "rider"
- Kirjataan sivujen lukumäärät, yksittäin ja yhdessä
- Maaliskuu 2007 $NGD('horse','rider') \approx 0.443$
- Marraskuu 2013 $NGD('horse','rider') \approx 0.233$
- Asteikko $0 - 1$

Algoritmin Perusta: Kolmogorov-kompleksisuus

Kolmogorov-kompleksisuus merkkijonosta x on lyhimmän tietokoneohjelman pituus, joka tuottaa merkkijonon x. Jokaista olemassaolevaa pakkausalgoritmiä kohtaan meillä on $K(x) \leq$ pakatun x:n pituus.

Normalisoitu informaatioetäisyys

Informaatioetäisyys

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

Normalisoitu informaatioetäisyys

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

Normalisoitu pakkausetäisyys

- NID ei ole laskettava, joten sitä ei voi tosielämässä käyttää.
- Pakkausalgoritmeillä (C) voi approksimoida Kolmogorov-kompleksisuuksia.
- $C(x)$ kuvastaa merkkijonon x pakattua versiota

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

Google-jakauma

- Yksittäisten hakujen joukko S
- Indeksöityjen internetsivujen joukko Ω
- $M = |\Omega|$, joukon Ω mahtavuus
- Yksittäisiä hakuja $|S|$
- Yhdistettyjä hakuja $\binom{|S|}{2}$
- Määritellään $N = \sum_{\{x,y\} \subseteq S} |x \cap y|$
- $N \geq M$

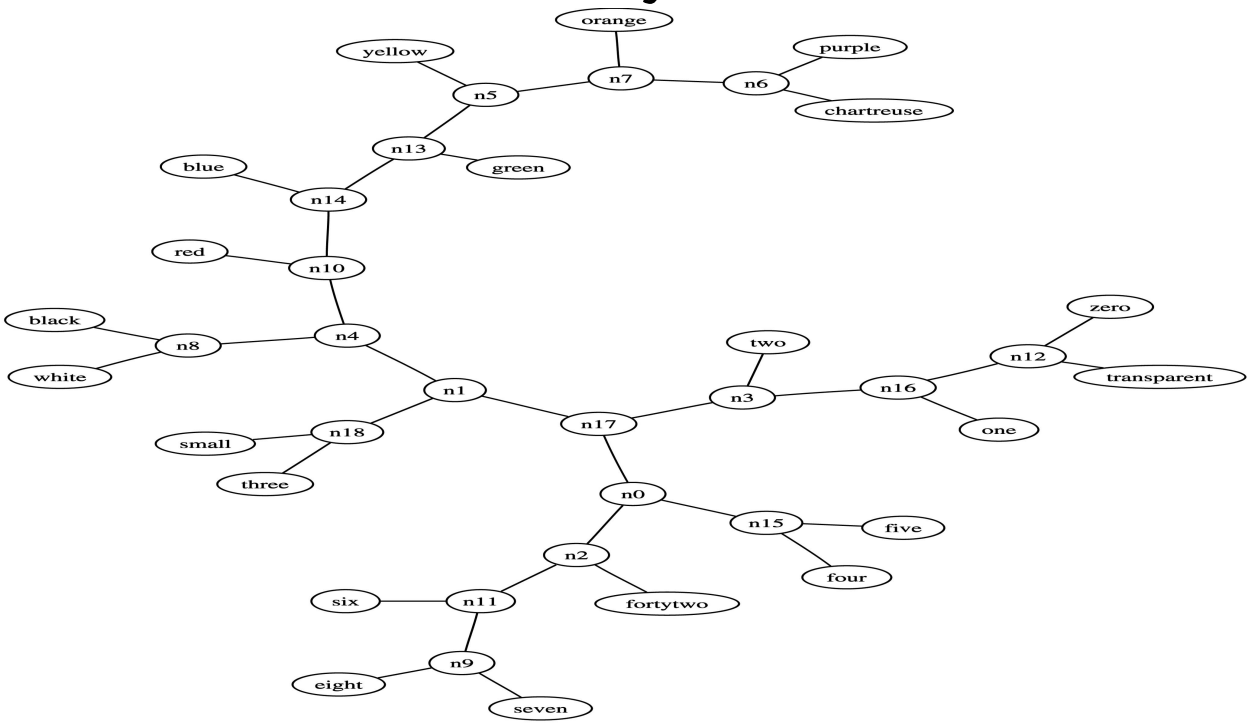
Google samankaltaisuusetäisyys

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

$f(x)$ ilmaisee sivujen lukumäärän, jokta sisältävät x:n, ja $f(x,y)$ ilmaisee sivujen lukumäärän, jotka sisältävät x:n ja y:n, Googlen mukaan.

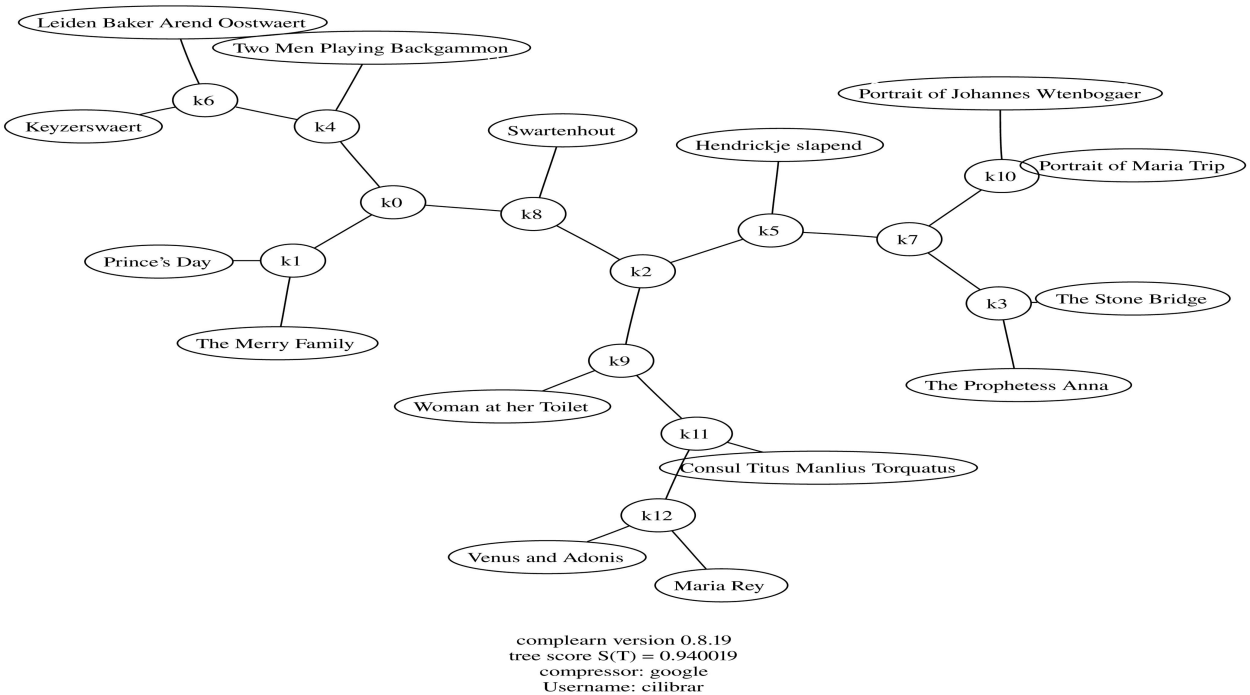
Sovellukset: Hierarkinen klusterointi

Hakusanoina: värit, numerot ja muutama konstikas sana



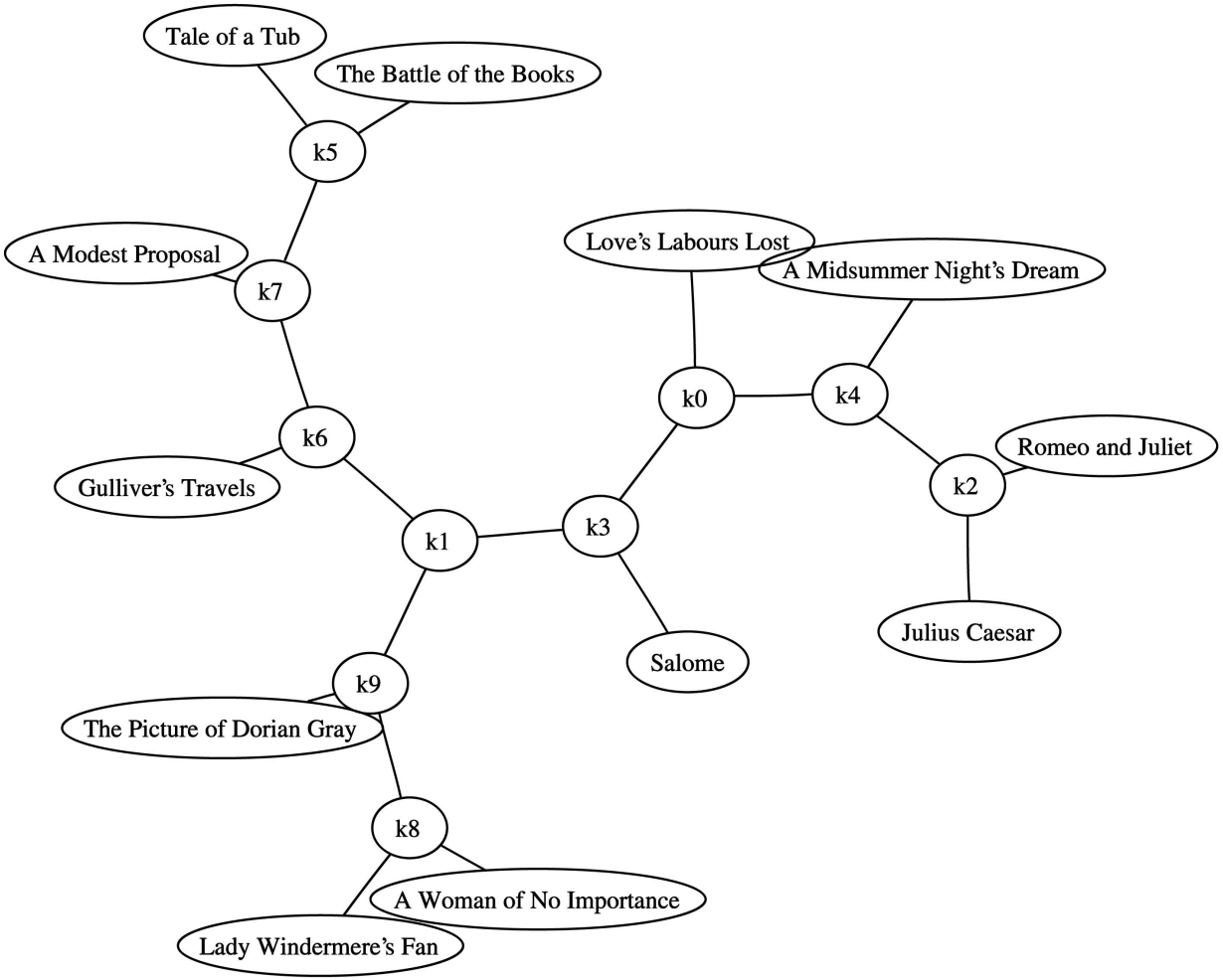
Sovellukset: Hollantilaisia maalareita

Hakusanoina 15 Steenin, Rembrandtin ja Bolin maalausta.



Sovellukset: Englantilaisia kirjailijoita

Hakusanoina Shakespearen, Oscar Wilden ja Jonathan Swiftin kirja



complearn version 0.8.19
tree score S(T) = 0.940416
compressor: google
Username: cilibrar

Kertaus

Kertaus

- NID
- NCD
- NGD

Kysymyksiä?