Corpora analyzed in our work

January 11, 2019

We not only studied real corpora in this work, but also generated fake ones to clarify the origin of scaling structure. The novels, proses, and poems included by the former are listed in the reference below. As for the fake corpora there are three kinds:

1 Paper generator

We use SCIgen[25] to generate random Computer Science research paper.

2 N-gram generator

Based on Sinica Corpus[26], we construct the data set of Chinese words. The N-gram generator[27] groups N random Chinese characters into one word. After building the word bank, the user can arrange these words to obey Zipf's distribution or non-Zipf distribution such as Gaussian.

3 Mixed text

A text composed of multiple books by different authors is also regarded as a fake corpus. We made two different mixed texts: Newspaper and Mixed. The former is the collection of Chinese Time News over the past 10 years, while the latter is the mixed text of three books[14, 8, 4].

References

[1] Jeremy Stribling, Max Krohn, Dan Aguayo. SCIgen - An Automatic CS Paper Generator. Retrieved from https://pdos.csail.mit.edu/archive/scigen/ (2017)

[2] Academia Sinica. Sinica Corpus. Retrieved from http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm (2017)

[3] Shan-Jyun Wu. Fake Script Vietnamese. Retrieved from http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm (2017)

[4] The standard corpus

[5] Herman Melville. *Moby-Dick*. Urbana, Illinois: Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/2701 (2008)

[6] J. R. R. Tolkien. *The Hobbit*. ISBN: 0618260307 (2002)

[7] Wu Jingzi 吳敬梓. *The Scholar* 儒林外史. Retrieved from https://zh.wikisource.org/zh-hant/%E5%84%92%E6%9E%97%E5%A4%96%E5%8F%B2 (2017)

[8] Mo Yan 莫言. *Frog* 蛙. ISBN: 9787532136766 (2009)

[9] S. T. Tsai *et al.* Acoustic Emission from Breaking a Bamboo Chopstick. Phys. Rev. Lett. **116**, 035501 (2016)

[10] B. P. Abbott *et al.* Observation of Gravitational Waves from a Binary Black Hole Merger. Phys. Rev. Lett. **116**, 061102 (2016)

[11] T. Maillart, D. Sornette, S. Spaeth, G. Von Krogh. Empirical Tests of Zipf's law Mechanism In Open Source Linux Distribution. Phys. Rev. Lett. **101**, 218701 (2008)

[12] Yu Guangzhong 余光中. *Selected Poetry of Yu Guangzhong* 余光中詩選. ISBN：9576742730 (2006)

[13] Xu Zhimo 徐志摩. Selected poems. Retrieved from http://w3.loxa.com.tw/fxp6033/poet01.htm (2017).

[14] Lung Yingtai 龍應台. 野火集. ISBN: 9576070589 (1985)

[15] Haruki Murakami 村上春樹. Kafka on the Shore (Chinese version) 海邊的卡夫卡. ISBN:2966622172 (2017)

[16] Jin Yong 金庸. *Demi-Gods and Semi-Devils* 天龍八部. ISBN: 9789573256748 (1996)

[17] Jin Yong 金庸. *Mandarin Duck Blades* 鴛鴦刀. ISBN: 9789573256748 (1996)

[18] Chang Show-Foong 張曉風. *Zhang Xiaofeng Prose Collection* 張曉風散文集. ISBN: 9574441407 (2004)

[19] Ni Kuang 倪匡. Wisely 衛斯理系列-第二種人. ISBN:9576455936 (1995)

[20] Zheng Chou-yu 鄭愁予. Selected poems. Retrieved from http://w3.loxa.com.tw/fxp6033/poet03.htm (2017).

[21] J.K. Rowling, Mary GrandPré. *Harry Potter and the Sorcerer's Stone.* ISBN:0439554934 (1997)

[22] J.K. Rowling, Mary GrandPré. *Harry Potter and the Chamber of Secrets.* ISBN:0439064864 (1999)

[23] J.K. Rowling, Mary GrandPré. *Harry Potter and the Prisoner of Azkaban.* ISBN:043965548X (2004)

[24] J.K. Rowling, Mary GrandPré. *Harry Potter and the Goblet of Fire.* ISBN:0439139600 (2002)

[25] J.K. Rowling, Mary GrandPré. *Harry Potter and the Order of the Phoenix.* ISBN:0439358078 (2002)

[26] J.K. Rowling, Mary GrandPré. *Harry Potter and the Half-Blood Prince.* ISBN:0439785960 (2006)

[27] J.K. Rowling, Mary GrandPré. *Harry Potter and the Deathly Hallows.* ISBN:0545010225 (2007)

[28] Done by S.T. Tsai. News collection from Chinese times, Apple daily, The Liberty Times, etc. (2016, 2017)