Corpra used in our paper

May 31, 2018

There are two kinds of corpra used in our paper: Real and fake. The real corpra we used are included in reference. We will introduce the fake corpra in the following content.

1 Paper generator

We use SCIgen[12] to generate random Computer Science research paper.

2 N-gram generator

Based on Sinica Corpus, we build the table of Chinese words. The N-gram generator[14] selects N Chinese characters into one word. After build up a word list, the user can asks these word obey Zipf's distribution or nonZipf distribution like Gussian.

3 Mix text

A text that composed of different books written by different authors is also a fake corpus. We made two different mix text: Newspaper and Mix. The former is the collection of Chinese Time News over the past 10 years, while the latter is mixing of Ref.[11][8][5].

References

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

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

[3] J.K. Rowling, Mary GrandPré (Illustrator), Adolfo Muñoz García y Nieves Martín Azofra. *Harry Potter and the Prisoner of Azkaban*. ISBN: 9780439655484. (2004)

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

[5] Mo Yan. *Frog*. ISBN: 9787532136766. (2009)

[6] Einstein. *Relativität der Gleichzeitigkeit*. Retrieved from http://www.mahag.com/srt/1905.php (1905)

[7] S.T. Tsai. et al. Acoustic Emission from Breaking a Bamboo Chopstick. Phys. Rev. Lett. 116, 035501 – Published 19 January 2016

[8] Yu Guangzhong. *Selected Poetry of Yu Guangzhong*. ISBN：9576742730. (2006)

[9] Xu Zhimo. *Xu Zhimo - Selected Poems*. ISBN：9789861783840. (2016)

[10] Jin Yong. *Demi-Gods and Semi-Devils*. ISBN: 9789573256748. (1996)

[11] Chang Show-Foong. *Zhang Xiaofeng Prose Collection*. ISBN：9574441407. (2004)

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

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

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