Assignment 1 209341U

Corpora taken from:

https://drive.google.com/file/d/0BxR2y9c3lt71d0VvM204NnB4ekk/view?usp=sharing

Github link: https://github.com/kanarupank/zipfs_law_test

Text Preparation:

1. Skip numbers
2. Skip words starting or ending with numbers
3. Skip empty string
4. Skip special characters
5. Remove special characters at the beginning or at the end of words
6. Did not skip English words, letters.

Item	Tamil	Sinhala
Number of words	249177	291637
Number of words after filtering	197752	240959
Number of unique words after filtering	33791	16772
Top 5 Frequency	மற்றும் - 8547 ஆம் - 6585 வேண்டும் - 4338 அல்லது - 4332 அவர் -3393	සඳහා - 3030 හා - 3029 යුතු - 2031 හෝ - 2014 ය - 1889
Top 5 Frequency of Frequencies	1 – 19396 2 - 5221 3 - 2223 4 - 1329 5 - 901	1 – 7316 2 – 2689 3 – 1301 4 – 855 5 - 548

Tamil Frequencies, descending

('மற்றும்', 2849), ('ஆம்', 2195), ('வேண்டும்', 1446), ('அல்லது', 1444), ('அவர்', 1131), ('திகதி', 1054), ('உத்தியோகத்தர்', 1018), ('கிழ்', 936), ('செயலாளர்', 834), ('தொடர்பாக', 733), ('மேற்படி', 651), ('மூலம்', 651), ('இலக்கம்', 651), ('பெற்றுக்', 608), ('ஒரு', 591), ('கௌரவ', 556), ('கல்வி', 550), ('பிரதேச', 492), ('இந்த', 480), ('அரச', 480), ('இலக்க', 479), ('உரிய', 467), ('தேதிய', 466), ('அபிவிருத்தி', 464), ('பணிப்பாளர்', 461), ('இலங்கை', 438), ('அரசாங்க', 426), ('நிதி', 411), ('யாது', 400), ('சேவை', 388), ('போது', 382), ('யாவை', 371), ('நடவடிக்கை', 366), ('திகதிய', 363), ('எண்ணிக்கை', 359), ('சட்டத்தின்', 358), ('பிரிவின்', 353), ('குறித்த', 349), ('கேட்பதற்கு', 346), ('விசேட', 345), ('முடியும்', 338), ('அமைச்சின்', 332)

…

…

Sinhala Frequencies, descending

('සඳහා', 3030), ('හා', 3029), ('යුතු', 2031), ('හෝ', 2014), ('ය', 1889), ('කරන', 1844), ('ඇති', 1776), ('එම', 1755), ('කළ', 1591), ('සහ', 1588), ('කිරීම', 1495), ('කර', 1475), ('ලබා', 1437), ('මෙම', 1386), ('ඒ', 1384), ('ලද', 1277), ('කටයුතු', 1224), ('විසින්', 1163), ('ද', 1156), ('වන', 1135), ('දින', 1107), ('නම්', 1094), ('සඳහන්', 1017), ('අංක', 984), ('ලේකම්', 932), ('වැනි', 899), ('කවරේද', 891), ('වෙත', 885), ('යටතේ', 875), ('අධ\u200dයාපන', 832),…

…

…

Tamil Frequency of Frequencies

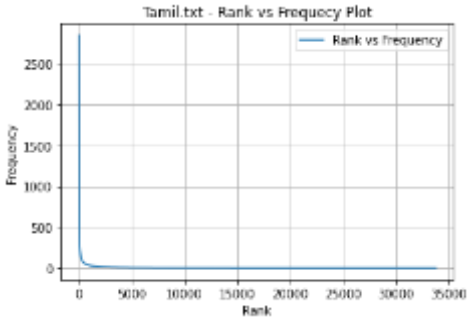
[(1, 7316), (2, 2689), (3, 1301), (4, 855), (5, 548), (6, 435), (7, 349), (9, 234), (8, 216), (10, 189), (11, 168), (12, 149), (13, 121), (14, 118), (15, 112), (16, 74), (17, 72), (18, 67), (19, 63), (21, 55), (20, 50), (28, 46), (23, 46), (22, 46), (25, 44), (24, 38), (27, 37), (26, 36), (30, 33), (34, 30), (33, 30), (31, 29), (35, 27), (32, 27), (40, 26), (29, 26), (46, 25), (48, 24), (37, 23), (43, 22), (36, 21), (39, 20), (38, 20), (51, 17), (49, 17), (47, 17), (41, 17), (69, 16), (42, 16), (56, 15), (44, 15), (50, 14), (45, 13), (60, 12), (58, 12), (83, 11), (74, 11), (65, 11), (59, 11), (55, 11), (81, 10), (72, 10), (62, 10), (103, 9), (87, 9), (85, 9), (77, 9), (63, 9), (57, 9), (54, 9), (95, 8), (73, 8), (61, 8), (52, 8), (161, 7), (144, 7), (80, 7), (75, 7), (71, 7), (70, 7), (67, 7), (66, 7), (64, 7), (130, 6), (114, 6), (110, 6), (108, 6), (99, 6), (90, 6), (89, 6), (79, 6), (53, 6), (147, 5), (112, 5), (111, 5), (107, 5), (88, 5), (82, 5), (78, 5), (68, 5), (335, 4), (153, 4), (145, 4), (137, 4), (133, 4), (127, 4), (126, 4), (123, 4), (120, 4), (119, 4), (106, 4), (104, 4), (100, 4), (98, 4), (97, 4), (92, 4), (91, 4), (84, 4), (76, 4), (262, 3), (257, 3), (203, 3), (182, 3), (181, 3), (179, 3), (172, 3), (168, 3), (162, 3), (152, 3), (149, 3), (143, 3), (142, 3), (141, 3), (136, 3), (135, 3), (125, 3), (113, 3), (105, 3), (102, 3), (96, 3), (86, 3), (616, 2), (575, 2), (521, 2), (406, 2), (382, 2), (364, 2), (354, 2), (329, 2), (271, 2), (260, 2), (258, 2), (236, 2), (234, 2), (230, 2), (217, 2), (215, 2), (210, 2), (208, 2), (205, 2), (198, 2), (195, 2), (194, 2), (189, 2), (188, 2), (180, 2), (178, 2), (176, 2), (171, 2), (170, 2), (164, 2), (163, 2), (160, 2), (158, 2), (157, 2), (151, 2), (150, 2), (146, 2), (138, 2), (132, 2), (129, 2), (122, 2), (121, 2), (118, 2), (117, 2), (109, 2), (101, 2), (3030, 1), (3029, 1), (2031, 1), (2014, 1), (1889, 1), (1844, 1), (1776, 1), (1755, 1), (1591, 1), (1588, 1), (1495, 1), (1475, 1), (1437, 1), (1386, 1), (1384, 1), (1277, 1), (1224, 1), (1163, 1), (1156, 1), (1135, 1), (1107, 1), (1094, 1), (1017, 1), (984, 1), (932, 1), (899, 1), (891, 1), (885, 1), (875, 1), (832, 1), (828, 1), (820, 1), (819, 1), (813, 1), (809, 1), (805, 1), (790, 1), (774, 1), (764, 1), (762, 1), (760, 1), (740, 1), (721, 1), (711, 1), (699, 1), (682, 1), (668, 1), (655, 1), (638, 1), (619, 1), (612, 1), (577, 1), (572, 1), (566, 1), (563, 1), (557, 1), (550, 1), (540, 1), (523, 1), (516, 1), (511, 1), (507, 1), (498, 1), (484, 1), (473, 1), (469, 1), (468, 1), (467, 1), (465, 1), (464, 1), (461, 1), (454, 1), (450, 1), (449, 1), (441, 1), (440, 1), (432, 1), (428, 1), (427, 1), (416, 1), (415, 1), (404, 1), (396, 1), (394, 1), (384, 1), (376, 1), (374, 1), (369, 1), (367, 1), (359, 1), (352, 1), (351, 1), (344, 1), (332, 1), (330, 1), (327, 1), (323, 1), (322, 1), (320, 1), (318, 1), (316, 1), (314, 1), (313, 1), (312, 1), (306, 1), (305, 1), (299, 1), (298, 1), (293, 1), (286, 1), (280, 1), (276, 1), (275, 1), (274, 1), (268, 1), (266, 1), (264, 1), (255, 1), (253, 1), (247, 1), (246, 1), (243, 1), (231, 1), (228, 1), (227, 1), (226, 1), (225, 1), (221, 1), (220, 1), (219, 1), (218, 1), (213, 1), (212, 1), (211, 1), (209, 1), (204, 1), (202, 1), (200, 1), (199, 1), (196, 1), (193, 1), (192, 1), (190, 1), (187, 1), (186, 1), (185, 1), (184, 1), (183, 1), (177, 1), (175, 1), (174, 1), (173, 1), (169, 1), (166, 1), (159, 1), (156, 1), (148, 1), (140, 1), (134, 1), (131, 1), (115, 1), (94, 1), (93, 1)]

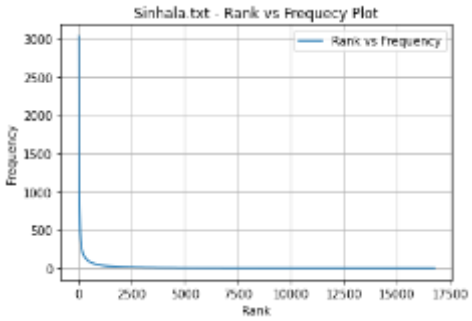
Sinhala Frequency of Frequencies

[(1, 7316), (2, 2689), (3, 1301), (4, 855), (5, 548), (6, 435), (7, 349), (9, 234), (8, 216), (10, 189), (11, 168), (12, 149), (13, 121), (14, 118), (15, 112), (16, 74), (17, 72), (18, 67), (19, 63), (21, 55), (20, 50), (28, 46), (23, 46), (22, 46), (25, 44), (24, 38), (27, 37), (26, 36), (30, 33), (34, 30), (33, 30), (31, 29), (35, 27), (32, 27), (40, 26), (29, 26), (46, 25), (48, 24), (37, 23), (43, 22), (36, 21), (39, 20), (38, 20), (51, 17), (49, 17), (47, 17), (41, 17), (69, 16), (42, 16), (56, 15), (44, 15), (50, 14), (45, 13), (60, 12), (58, 12), (83, 11), (74, 11), (65, 11), (59, 11), (55, 11), (81, 10), (72, 10), (62, 10), (103, 9), (87, 9), (85, 9), (77, 9), (63, 9), (57, 9), (54, 9), (95, 8), (73, 8), (61, 8), (52, 8), (161, 7), (144, 7), (80, 7), (75, 7), (71, 7), (70, 7), (67, 7), (66, 7), (64, 7), (130, 6), (114, 6), (110, 6), (108, 6), (99, 6), (90, 6), (89, 6), (79, 6), (53, 6), (147, 5), (112, 5), (111, 5), (107, 5), (88, 5), (82, 5), (78, 5), (68, 5), (335, 4), (153, 4), (145, 4), (137, 4), (133, 4), (127, 4), (126, 4), (123, 4), (120, 4), (119, 4), (106, 4), (104, 4), (100, 4), (98, 4), (97, 4), (92, 4), (91, 4), (84, 4), (76, 4), (262, 3), (257, 3), (203, 3), (182, 3), (181, 3), (179, 3), (172, 3), (168, 3), (162, 3), (152, 3), (149, 3), (143, 3), (142, 3), (141, 3), (136, 3), (135, 3), (125, 3), (113, 3), (105, 3), (102, 3), (96, 3), (86, 3), (616, 2), (575, 2), (521, 2), (406, 2), (382, 2), (364, 2), (354, 2), (329, 2), (271, 2), (260, 2), (258, 2), (236, 2), (234, 2), (230, 2), (217, 2), (215, 2), (210, 2), (208, 2), (205, 2), (198, 2), (195, 2), (194, 2), (189, 2), (188, 2), (180, 2), (178, 2), (176, 2), (171, 2), (170, 2), (164, 2), (163, 2), (160, 2), (158, 2), (157, 2), (151, 2), (150, 2), (146, 2), (138, 2), (132, 2), (129, 2), (122, 2), (121, 2), (118, 2), (117, 2), (109, 2), (101, 2), (3030, 1), (3029, 1), (2031, 1), (2014, 1), (1889, 1), (1844, 1),

(1776, 1), (1755, 1), (1591, 1), (1588, 1), (1495, 1), (1475, 1), (1437, 1), (1386, 1), (1384, 1), (1277, 1), (1224, 1), (1163, 1), (1156, 1), (1135, 1), (1107, 1), (1094, 1), (1017, 1), (984, 1), (932, 1), (899, 1), (891, 1), (885, 1), (875, 1), (832, 1), (828, 1), (820, 1), (819, 1), (813, 1), (809, 1), (805, 1), (790, 1), (774, 1), (764, 1), (762, 1), (760, 1), (740, 1), (721, 1), (711, 1), (699, 1), (682, 1), (668, 1), (655, 1), (638, 1), (619, 1), (612, 1), (577, 1), (572, 1), (566, 1), (563, 1), (557, 1), (550, 1), (540, 1), (523, 1), (516, 1), (511, 1), (507, 1), (498, 1), (484, 1), (473, 1), (469, 1), (468, 1), (467, 1), (465, 1), (464, 1), (461, 1), (454, 1), (450, 1), (449, 1), (441, 1), (440, 1), (432, 1), (428, 1), (427, 1), (416, 1), (415, 1), (404, 1), (396, 1), (394, 1), (384, 1), (376, 1), (374, 1), (369, 1), (367, 1), (359, 1), (352, 1), (351, 1), (344, 1), (332, 1), (330, 1), (327, 1), (323, 1), (322, 1), (320, 1), (318, 1), (316, 1), (314, 1), (313, 1), (312, 1), (306, 1), (305, 1), (299, 1), (298, 1), (293, 1), (286, 1), (280, 1), (276, 1), (275, 1), (274, 1), (268, 1), (266, 1), (264, 1), (255, 1), (253, 1), (247, 1), (246, 1), (243, 1), (231, 1), (228, 1), (227, 1), (226, 1), (225, 1), (221, 1), (220, 1), (219, 1), (218, 1), (213, 1), (212, 1), (211, 1), (209, 1), (204, 1), (202, 1), (200, 1), (199, 1), (196, 1), (193, 1), (192, 1), (190, 1), (187, 1), (186, 1), (185, 1), (184, 1), (183, 1), (177, 1), (175, 1), (174, 1), (173, 1), (169, 1), (166, 1), (159, 1), (156, 1), (148, 1), (140, 1), (134, 1), (131, 1), (115, 1), (94, 1), (93, 1)]

Plots with filtered text (without filtering also similar)





Comments on the Plots

Zipf's argument based on the equilibrium of speaker and hearer trying to minimize their efforts (the principle of least effort) suggests that the word usage is in a way that *rank*frequency* remains a constant (approximately).

The rank vs frequency graph according to above theory is expected to be a like an xy=c curve, more like a hyperbola.

Both graphs look alike, and neither are very much like xy=c.

To find the degree of the accordance with Zipf's law additional steps could to be taken, though no widely accepted steps exist in academic literature.

Based on simple frequency*rank calculation with first 25 frequent words and the very long tail with only rank increasing but not the frequency.

In my perspective,

- the law does not hold

- moreover, trying to prove Zipf's law using evidence would end up cherry-picking data, more on the pseudo-science side, only depending on affirmation trying to use data points that could serve as evidences, rather than providing genuine room for refutability.

- I will find Zipf's philosophy/principle of human beings (and the process of communication) are inclined toward 'least effort' limited to certain contexts only. There are numerous instances where effort is celebrated, boosted willingly (examples: sports, recreational, love and positive relationships...).

- I could not empathize much with Zipf's line of thought as in my observation as communication itself could be activity of pleasure, which could invoke a moving, temporal equilibrium. For instance, in my line of thinking, the speaker herself/himself is not actively looking for simple word pool and generate more meaning out it, rather both hearer and speaker are looking for a larger word pools with specific meanings. The process is not spiraling towards equilibriums but moving forward with speakers and hearers mutually sharing traits and finding themselves in each other.

- By and large the Zipf's law which claims to find the unifying principle (not limited to communication), in my opinion is not deep enough to provide interfaces/explanations for several available contexts.