Summer Semester 2020 SNLP Assignment 2

Name: Awantee Deshpande
Id: 2581348
Email: s8awdesh@stud.uni-saarland.de

Name: Lakshmi Rajendra Bashyam
Id: 2581455
Email: s8laraje@stud.uni-saarland.de

1. Zipf's Law

1.1 Data Preprocessing

German
The top 10 most frequent words for the German language are 'der', 'die', 'und', 'in', 'von', 'im', 'das', 'den', 'des', 'mit'
Their average length is 2.8
Words like 'heimateinwohnermeldeamt', 'beeindruckender', 'entgegengesetzter' etc. occur once with an average length of 10.368376226624699

English
The top 10 most frequent English words are 'the', 'of', 'and', 'in', 'to', 'a', 'is', 'city', 'as', 'was'
Their average length is 2.4
Words like 'almond', 'outcross', 'observances', 'reoccupation' etc. occur only once with a combined average length of 7.975625625625626

Spanish
The top 10 most frequent Spanish words are 'de', 'la', 'el', 'en', 'y', 'del', 'los', 'que', 'a', 'se'
Their average length is 2.1
Words like 'defensivas', 'espectador', 'adolescencia' etc. occur once with a combined average length of 8.163870889602132

Hungarian
The top 10 most frequent Hungarian words are 'a', 'az', 'és', 'is', 'város', 'volt', 'de', 'egy', 'található', 'meg'
Their average length is 3.3
Words like 'jelentések', 'tatárvárosra', 'puritán', 'betekinteni' etc. occur once with a combined average length of 9.165741114701131

Turkish
The top 10 most frequent Turkish words are 've', 'bir', 'bu', 'en', 'olarak', 'büyük', 'ile', 'da', 'olan', 'de'
Their average length is 3.1
Words like 'partilerinin', 'istastistik', 'malzemesini' etc. occur once with a combined average length of 8.523685435086087

Analysis:
The most frequently occurring words in a corpus are what we generally term as 'stopwords'. These have no direct context in the domain but are used commonly in their respective texts. E.g. German articles like 'der', 'die', 'das', English words like 'has', 'it', 'and' etc. They are of short length but high frequency and one can observe that the average length of the top 10 frequent words across the corpora is not more than 4.
In case of the less frequently occuring words (in this case, occurring only once), the words are not commonly used words, nor do they have a very significant impact on the corpus domain. Across the different languages, we observe varying average lengths from 7 to 10 because of language specific characteristics. For example, words in German are compounded to form a new word leading to a longer length, as can be seen from the examples listed above.
A partial reasoning behind this can also be that considering drawing alphabets (or a space) at random from a uniform distribution, the probability of shorter words occurring would be much higher than that of longer words.

NOTE: We have preprocessed the text under the assumption that only punctuation is to be removed (and not other alphanumeric characters)

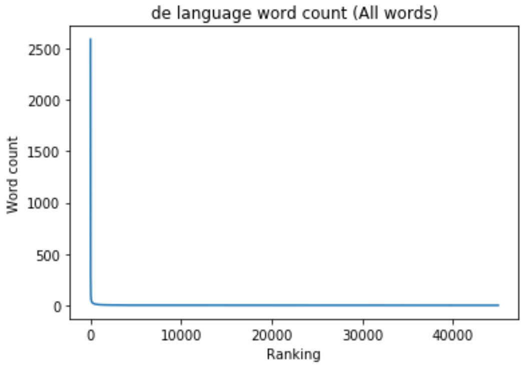
1.2 Plotting rank vs frequency

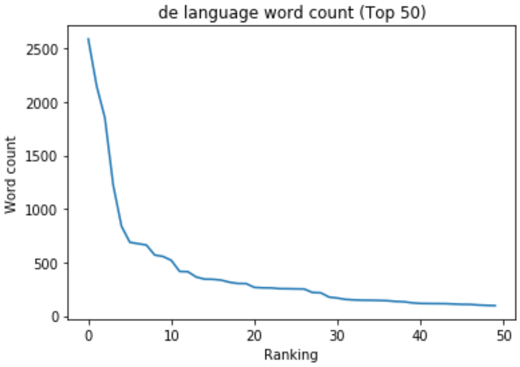
The data is normalised using the following formula

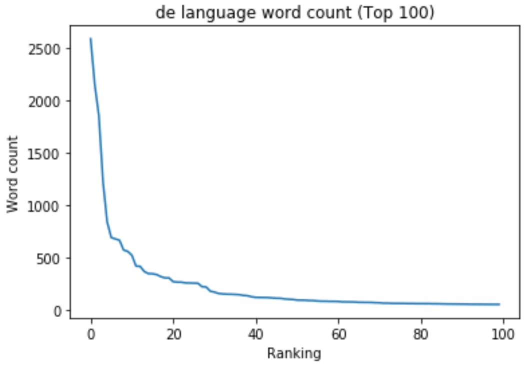
$$\text{normalised count} = \frac{\text{word frequency}}{\text{corpus size}} * \text{minimum corpus size}$$

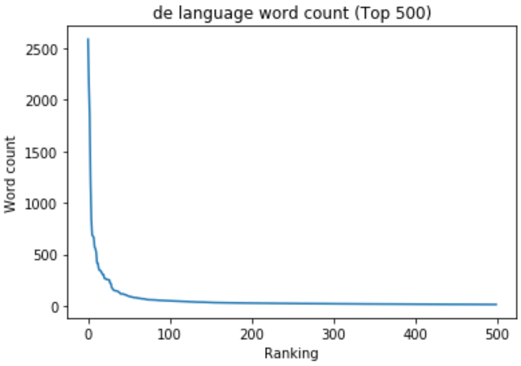
Then, for each language, these frequency counts are sorted in descending order (this order is their rank). The correspond Rank x Frequency graphs are plotted. Due to the vast size of the corpus and large differences in maximum and minimum frequencies, the graphs follow a steep curve. To make the plots more comprehensible, we make similar plots for the top 50, 100 and 500 ranked words.

German

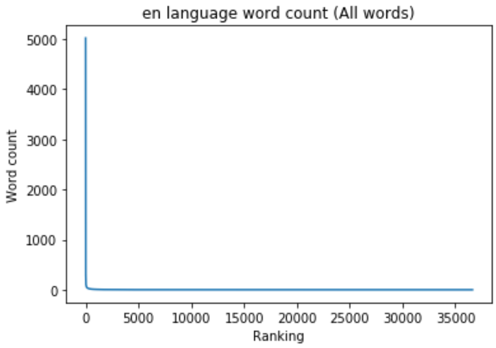


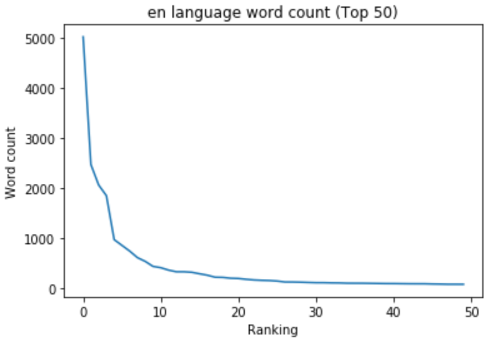


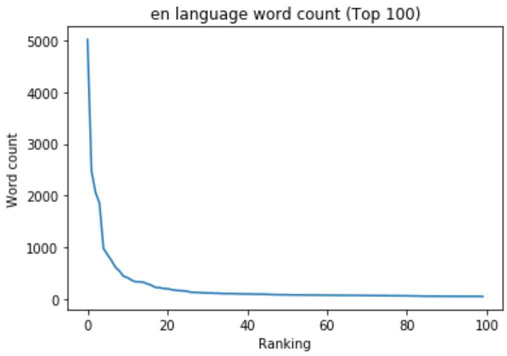


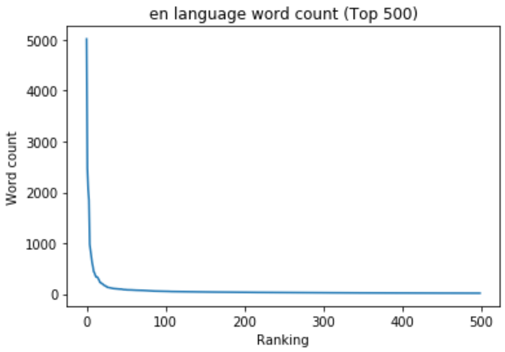


English

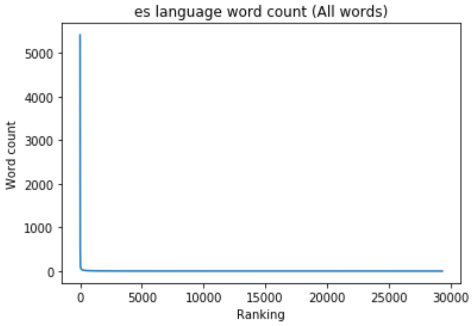


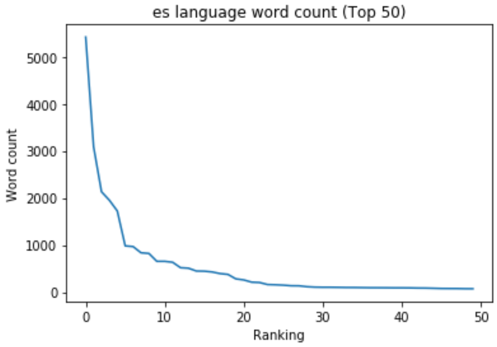


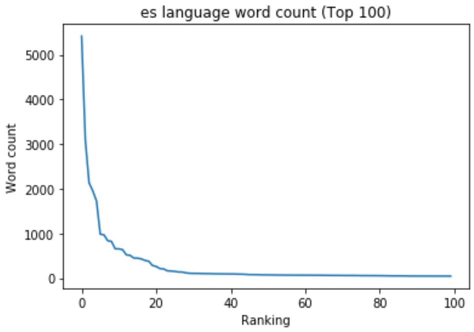


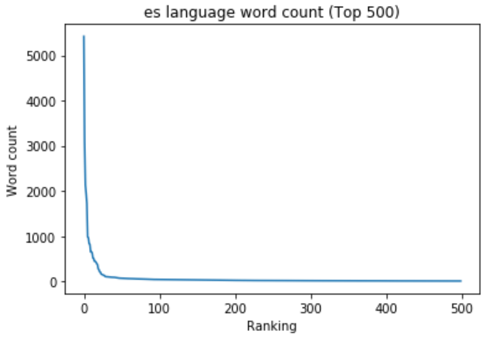


Spanish

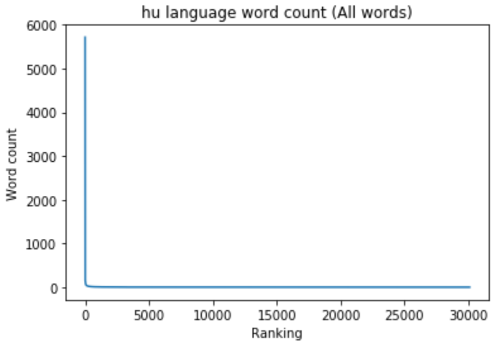


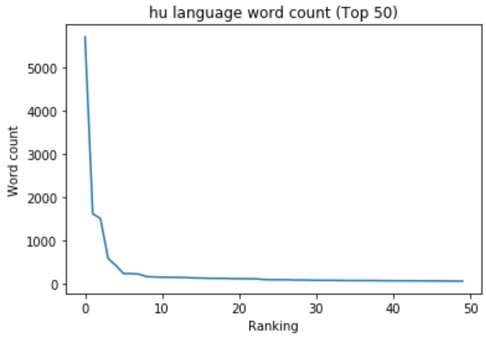


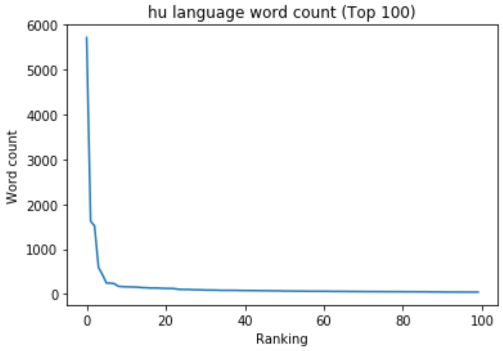


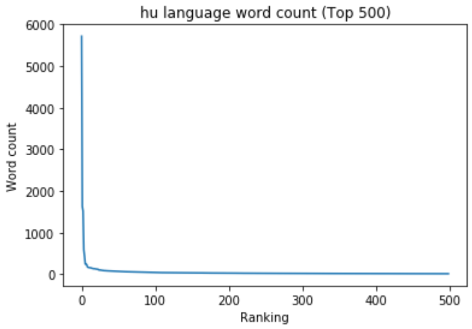


Hungarian

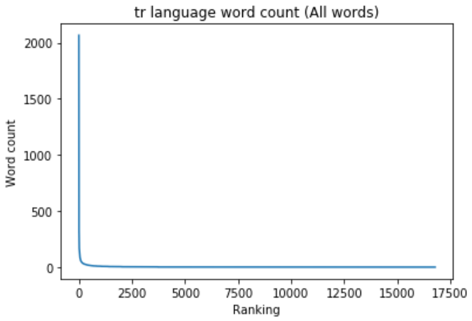


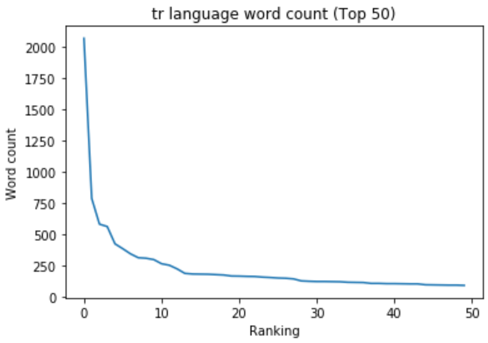


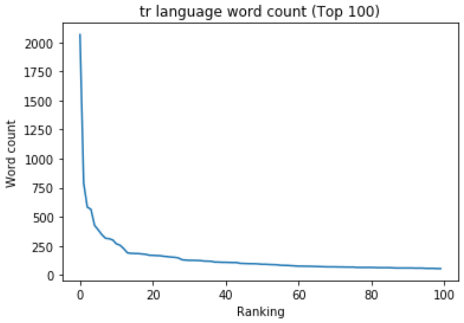


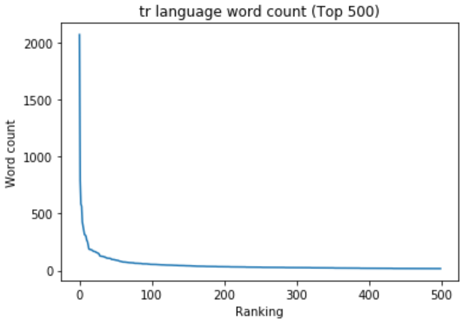


Turkish

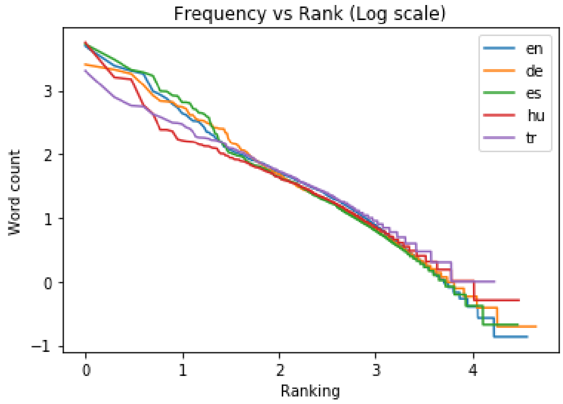








Analysis:
It can be observed from the plots that the languages follow Zipf's Law. Naturally, due to the differing vocabulary and semantics, the plots don't follow the exact same trend. Consider the same plots as above, but in a logarithmic scale for better visualisation. It is apparent that the languages not only follow the Zipf's law, but their plots are remarkably close to each others'.



2. Miller's Model

2.1 Random text generation

P(space) = 0.05
∴ P(alphabets) = 1 - 0.05 = 0.95
Since the probability distribution is uniform,
P(letter) = 0.95/26 = 0.03654

All letters of hello can occur equally probably. The probability of typing "hello" will be the probabilities of typing the letters 'h', 'e', 'l', 'l', and 'o' followed by a space. Thus, the probability is
P(h) * P(e) * P(l) * P(l) * P(o) * P(space)
= $(0.03654)^5$ * 0.05
= 3.25 x 10^{-9}

2.2 Random text based on statistics

The pseudocode for randomly generating sample words will be as follows (actual code in python file)

```
function output_word(length, prob_dist):
    Do not consider space in characters generated (but without changing the underlying distribution)
    word = `` //empty string
    for length times:
      randomly sample a character c from prob_dist
      append c to word
    return word
```

The corresponding python file also calculates the probability distribution of the underlying corpus using the formula

$$P(\text{char}) = \frac{\text{char count in corpus}}{\text{size of corpus}}$$

The corresponding probability of *hello* being generated is
P(h) * P(e) * P(l) * P(l) * P(o) * P(space) = $4.4638272591228335 * 10^{-8}$

3 Bonus: Frequency Analysis

Encoded text :
“PU JYFWAVNYHWOF H JHLZHY JPWOLY HSZV RUVDU HZ AOL ZOPMA JPWOLY PZ VUL VM AOL ZPTWSLZA HUK TVZA DPKLSF RU- VDU LUJYFWAPVU ALJOUPXBLZ. PA PZ H AFWL VM ZBIZAPABAPVU JPWOLY PU DOPJO LHJO SLAALY PU AOL WSHPUALEA PZ YLW- SHJLK IF H SLAALY ZVTL MPELK UBTILY VM WVZPAPVUZ KVDU AOL HSWOHILA. MVY LEHTWSL DPAO H SLMA ZOPMA VM AOYLL K DVBSK IL YLWSHJLK IF H HUK L DVBSK ILJVTL I. AOL TLAOVK PZ UHTLK HMALY QBSPBZ JHLZHY DOV BZLK PA PU OPZ WYPCHAL JVYYLZWVUKLUJL. IF NYHWOPUN AOL MYLXBLUJPLZ VM SLAA- LYZ PU AOL JPWOLYALEA HUK IF RUVDPUN AOL LEWLJALK KPZA- YPIBAPVU VM AOVZL SLAALYZ PU AOL VYPNPUHS SHUNBHNL VM AOL WSHPUALEA H OBTHU JHU LHZPSF ZWVA AOL CHSBL VM AOL ZOPMA IF SVVRPUN HA AOL KPZWSHJLTLUA VM WHYAPJBSHY ML- HABYLZ VM AOL NYHWO. AOPZ PZ RUVDU HZ MYLXBLUJF HUHS- FZPZ. MVY LEHTWSL PU AOL LUNSPZO SHUNBHNL AOL WSHPUALEA MYLXBLUJPLZ VM AOL SLAALYZ L, A (BZBHSSF TVZA MYLXBLUA) HUK X, G (AFWPJHSSF SLHZA MYLXBLUA) HYL WHYAPJBSHYSF KPZA- PUJAPCL. DPAO AOL JHLZHY JPWOLY LUJYFWAPUN H ALEA TB- SAPWSL APTLZ WYVCPKLZ UV HKKPAPVUHS ZLJBYPAF. AOPZ PZ ILJHBZL ADV LUJYFWAPVUZ VM ZOPMA H HUK ZOPMA I DPSS IL LXBPCHSLUA AV H ZPUNSL LUJYFWAPVU DPAO ZOPMA H + I. PU THAOLTHAPJHS ALYTZ AOL ZLA VM LUJYFWAPVU VWLYHAPVUZ BUKLY LHJO WVZZPISL RLF MVYTZ H NYVBW BUKLY JVTWVZPA- PVU.”

Since this text has been generated by simply replacing a character by another character with a one-one mapping, the approach to deciphering this would be to find the probability distribution of the encoded characters, map the characters of the source and the target language according to their sorted probability distribution (if ‘e’ occurs most in the source and is replaced by ‘k’ in the target, then k must occur most frequently in the target), and then replace the encoded text with the corresponding character from the source.
Following the above method results in the following text (code in python file):

TN DSFCARWSICHF I DIEOIS DTCHES ILOR JNRYN IO AHE OHTMA DTCHES TO RNE RM AHE OTGCLEOA INP GROA YTPELF JN- RYN ENDSFCATRN AEDHNTKUEO. TA TO I AFCE RM OUBOATAUATRN DTCHES TN YHTDH EIDH LEAAES TN AHE CLITNAEVA TO SEC- LIDEP BF I LEAAES ORGE MTVEP NUGBES RM CROTATRNO PRYN AHE ILCHIBEA. MRS EVIGCLE YTAH I LEMA OHTMA RM AHSEE P YRULP BE SECLIDEP BF I INP E YRULP BEDRGE B. AHE GEAHRP TO NIGEP IMAES QULTUO DIEOIS YHR UOEP TA TN HTO CSTXIAE DRSSEOCRNPENDE. BF WSICHTNW AHE MSEKUENDTEO RM LEAA- ESO TN AHE DTCHESAEVA INP BF JNRYTNW AHE EVCEDAEP PTOA- STBUATRN RM AHROE LEAAESO TN AHE RSTWTNIL LINWUIWE RM AHE CLITNAEVA I HUGIN DIN EIOTLF OCRA AHE XILUE RM AHE OHTMA BF LRRJTNW IA AHE PTOCLIDEGENA RM CISATDULIS ME- IAUSEO RM AHE WSICH. AHTO TO JNRYN IO MSEKUENDF INIL- FOTO. MRS EVIGCLE TN AHE ENWLTOH LINWUIWE AHE CLITNAEVA MSEKUENDTEO RM AHE LEAAESO E, A (UOUILLF GROA MSEKUENA) INP K, Z (AFCTDILLF LEIOA MSEKUENA) ISE CISATDULISLF PTOA- TNDATXE. YTAH AHE DIEOIS DTCHES ENDSFCATNW I AEVA GU- LATCLE ATGEO CSRXTPEO NR IPPTATRNIL OEDUSTAF. AHTO TO BEDIUOE AYR ENDSFCATRNO RM OHTMA I INP OHTMA B YTLL BE EKUTXILENA AR I OTNWLE ENDSFCATRN YTAH OHTMA I + B. TN GIAHEGIATDIL AESGO AHE OEA RM ENDSFCATRN RCESIATRNO UNPES EIDH CROOTBLE JEF MRSGO I WSRUC UNPES DRGCROTA- TRN.

While the logic behind this method is sound, the problem in this context is the lack of empirical data. The encoded distribution is too small to get realistic values. Some letters also have the same probability distributions leading to a confusion regarding what source characters to map them to. Only by swapping a few chosen letters (in this example, we replaced letters A, I, C, R, Y, J, S, F, D, M, V, K, X, W, Q) do we get the decoded text.

IN CRYPTOGRAPHY A CAESAR CIPHER ALSO KNOWN AS THE SHIFT CIPHER IS ONE OF THE SIMPLEST AND MOST WIDELY KN- OWN ENCRYPTION TECHNIQUES. IT IS A TYPE OF SUBSTITUTION CIPHER IN WHICH EACH LETTER IN THE PLAINTEXT IS REP- LACED BY A LETTER SOME FIXED NUMBER OF POSITIONS DOWN THE ALPHABET. FOR EXAMPLE WITH A LEFT SHIFT OF THREE D WOULD BE REPLACED BY A AND E WOULD BECOME B. THE METHOD IS NAMED AFTER JULIUS CAESAR WHO USED IT IN HIS PRIVATE CORRESPONDENCE. BY GRAPHING THE FREQUENCIES OF LETT- ERS IN THE CIPHERTEXT AND BY KNOWING THE EXPECTED DIST- RIBUTION OF THOSE LETTERS IN THE ORIGINAL LANGUAGE OF THE PLAINTEXT A HUMAN CAN EASILY SPOT THE VALUE OF THE SHIFT BY LOOKING AT THE DISPLACEMENT OF PARTICULAR FE- ATURES OF THE GRAPH. THIS IS KNOWN AS FREQUENCY ANAL- YSIS. FOR EXAMPLE IN THE ENGLISH LANGUAGE THE PLAINTEXT FREQUENCIES OF THE LETTERS E, T (USUALLY MOST FREQUENT) AND Q, Z (TYPICALLY LEAST FREQUENT) ARE PARTICULARLY DIST- INCTIVE. WITH THE CAESAR CIPHER ENCRYPTING A TEXT MU- LTIPLE TIMES PROVIDES NO ADDITIONAL SECURITY. THIS IS BECAUSE TWO ENCRYPTIONS OF SHIFT A AND SHIFT B WILL BE EQUIVALENT TO A SINGLE ENCRYPTION WITH SHIFT A + B. IN MATHEMATICAL TERMS THE SET OF ENCRYPTION OPERATIONS UNDER EACH POSSIBLE KEY FORMS A GROUP UNDER COMPOSIT- ION.