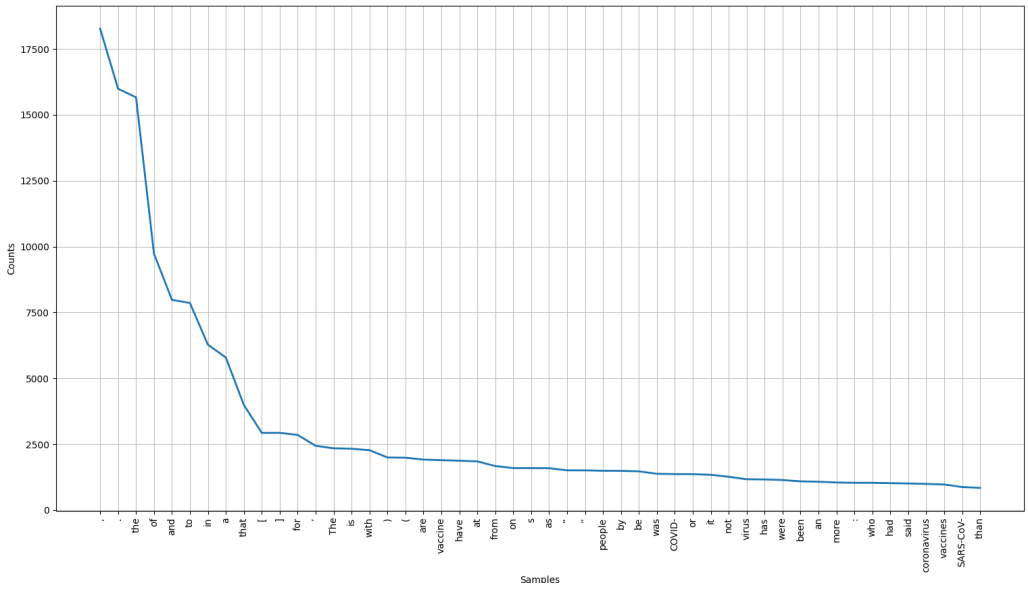
CL Project-1

-Eshika Khandelwal (2020114018)

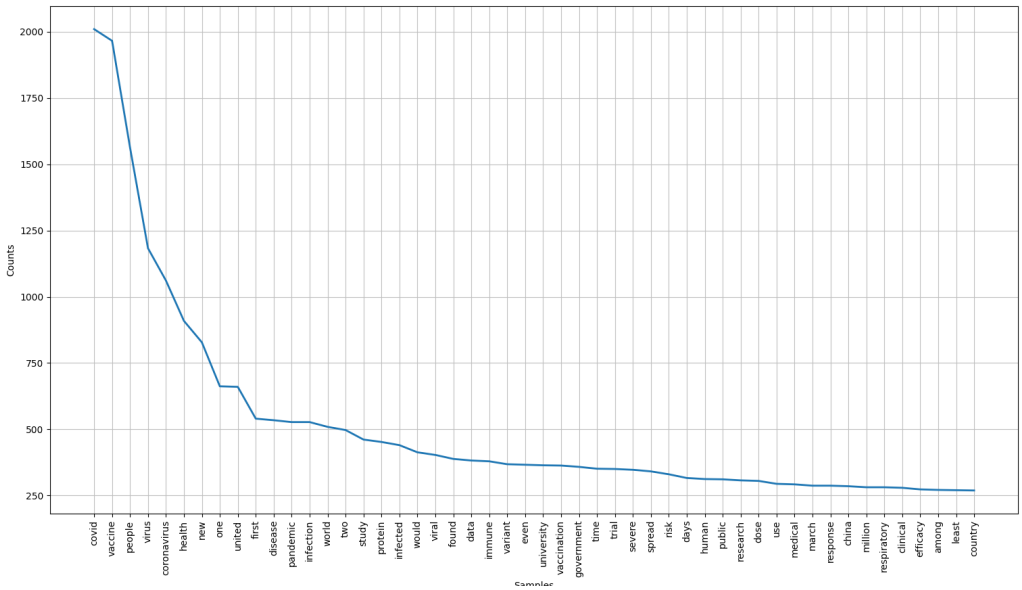
English Dataset

I have used around 200 news articles regarding 'Corona Virus' to form the dataset.

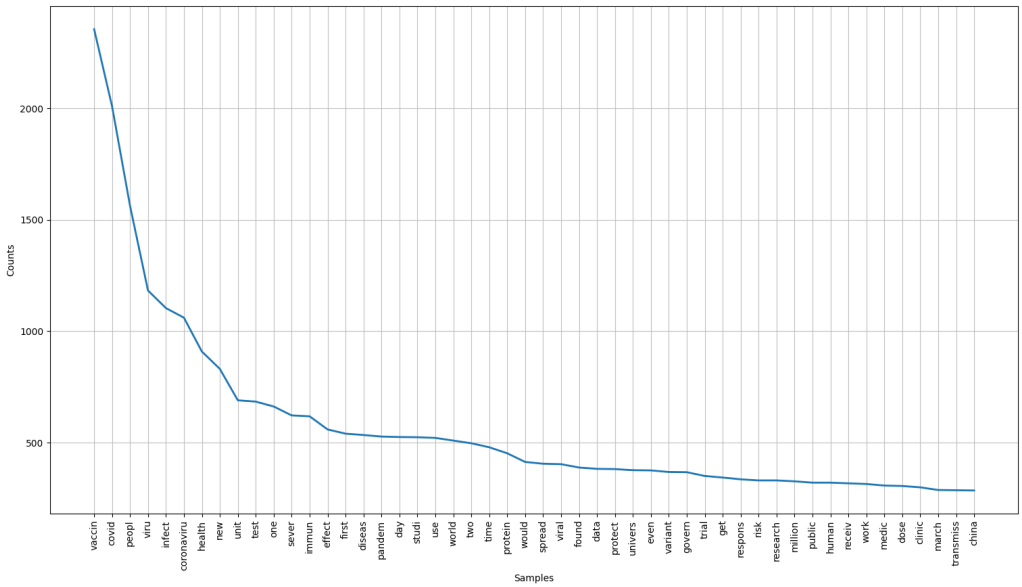
- Below is the inital Frequency graph, after just tokenizing the dataset



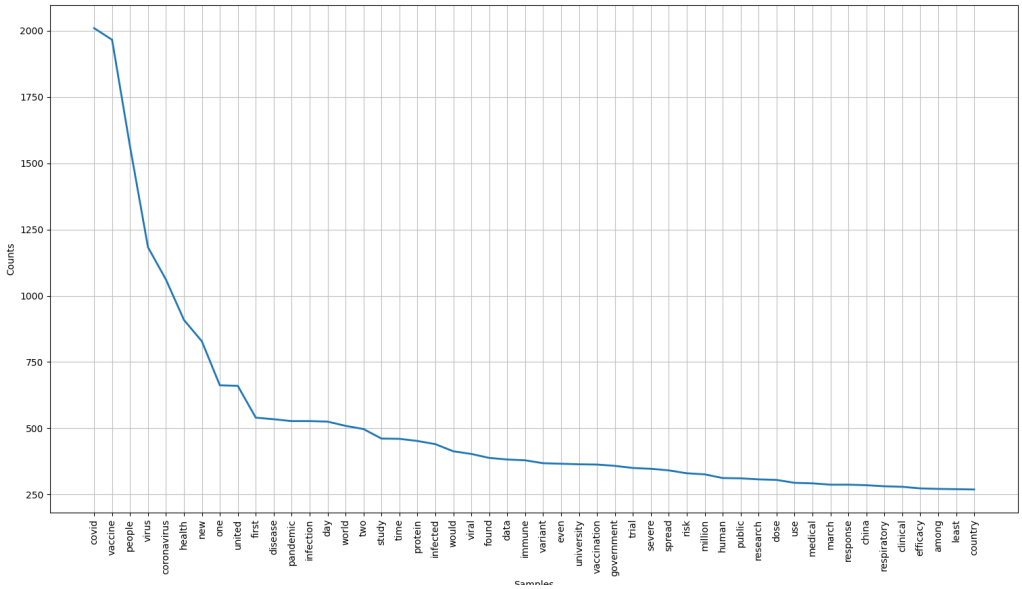
- We can see that a lot of irrelevant punctuations and stopwords are present in this stage.
- Inverted comma is shown to be the most used token, even a lot of stopwords like 'the', 'of', 'in' are present.
- After cleaning the data and removing punctuations and stopwords, the graph looks like this:



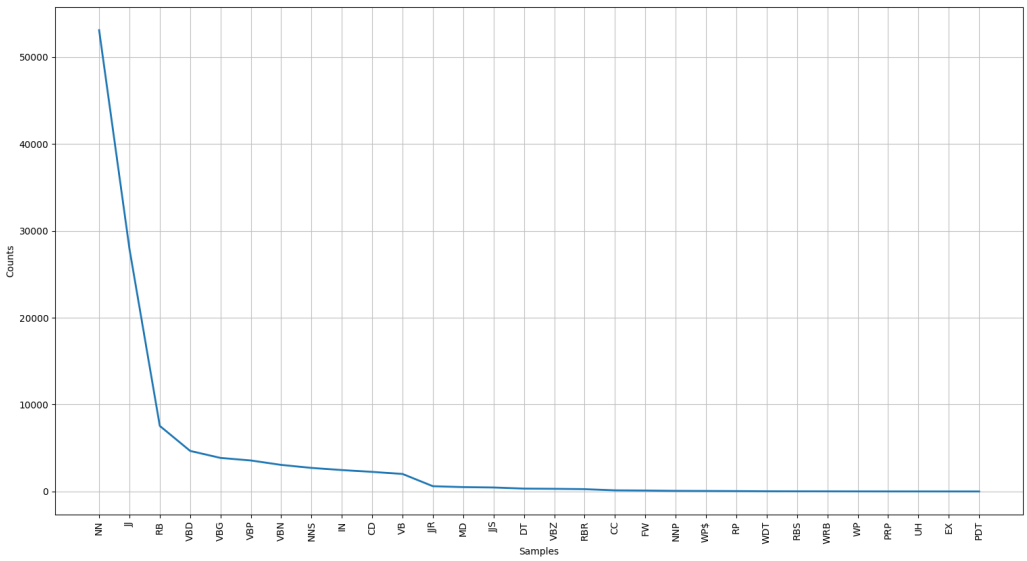
- Some extra words had to be added to the standard English dataset that NLTK provides as the pandemic gave rise to a lot of new terms that did not exist before like 'covid', 'pfizer, 'coronavirus' etc. This was done so that the relevant words aren't treated as random.
- Some additional stopwords are added too depending on the dataset crawled. For example: 'et al' was used a lot of times in the dataset as they were news articles and contained citations, these had to be removed. Apart from that, words like 'said', 'also', 'might' were also removed. These are very common words in news articles but are not relevant to our dataset.
- Clearly, the graph changes drastically.
- Now, we can see that 'covid' is the most used word followed by 'vaccines', which seems fair as all the recent covid articles talk about vaccines.
- Now, after stemming:



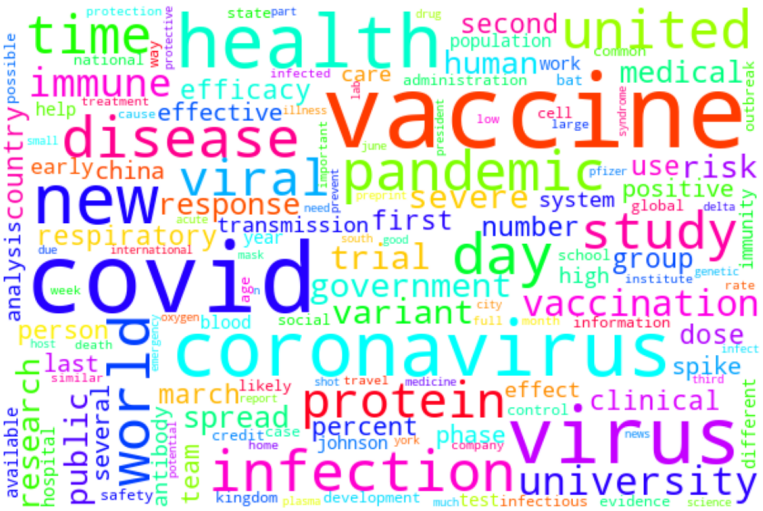
- The most used word now becomes vaccin. This is fair as I suppose all vaccines, vaccine, vaccinate, etc get stemmed to 'vaccin'. The count also considerably goes up. Wheras, no words can really be lemmatized to form 'covid'
- 'test' that wasn't amongst the most used words now is probably because words like 'tested', 'testing', 'tests' all got stemmed. Other stems like 'immun', 'univers' also move up due to similar reasons.
- Now, if we lemmatize the cleaned tokens instead of stemming them:



- The change between cleaned tokens and lemmatized tokens is not as drastic as stemmed. But the difference can be noticed in the ranges of the words in the center of the graph. The numbers go up consistently.
- The count of 'vaccine' goes up by a lot, but does not cross 'covid' in this case.
- Now if we see the POS tag frequency graph,



- We see that nouns and adjectives have the most frequency.
- Now, to make the wordcloud; Making the wordcloud of all the POS will not be as effective as just using nouns and adjectives.
- Hence, I have chosen to only include nouns followed by adjectives in my word cloud.
- Further, I have only chosen to include words that have been used more than 100 times. Hence, around 150 words have been included in the word cloud.
- This was done because any word used more than 100 seems to be relevant to the topic as a whole.
- The wordcloud made, looks like this:



- The frequencies in decreasing order:

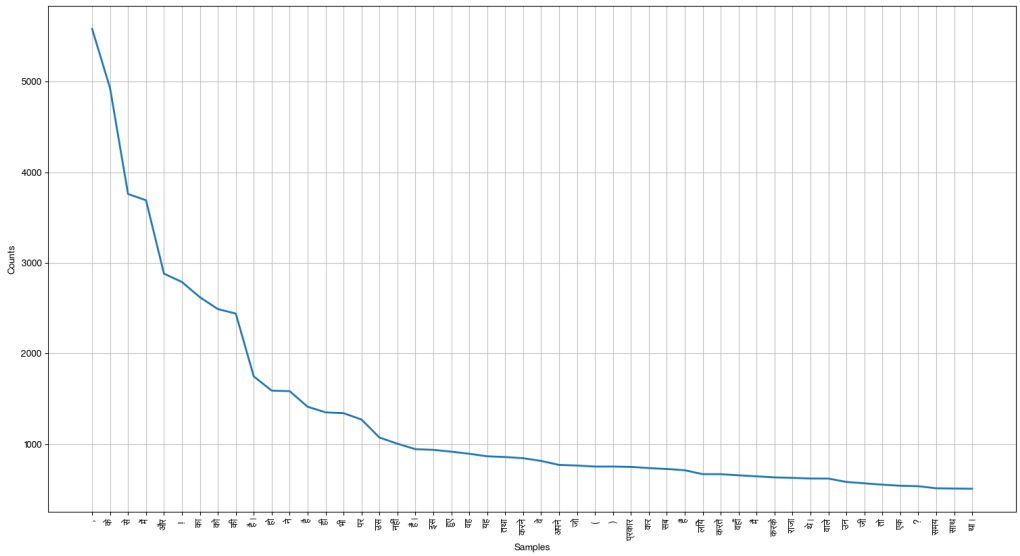
[('vaccine', 1913), ('covid', 1718), ('virus', 1063), ('coronavirus', 979), ('health', 909), ('new', 828), ('day', 525), ('disease', 520), ('infection', 520), ('world', 509), ('pandemic', 498), ('time', 460), ('united', 450), ('study', 423), ('viral', 402), ('protein', 384), ('university', 363), ('immune', 361), ('vaccination', 359), ('government', 358), ('variant', 358), ('trial', 350), ('severe', 325), ('public', 311), ('human', 311), ('risk', 311), ('research', 307), ('spread', 294), ('medical', 292), ('response', 287), ('respiratory', 280), ('clinical', 279), ('dose', 276), ('country', 269), ('efficacy', 268), ('group', 258), ('first', 256), ('use', 252), ('march', 248), ('number', 240), ('second', 240), ('percent', 235), ('team', 227), ('person', 226), ('positive', 224), ('spike', 224), ('effective', 221), ('phase', 220), ('last', 220), ('china', 217), ('several', 215), ('transmission', 212), ('effect', 212), ('antibody', 209), ('high', 209), ('system', 209), ('care', 207), ('early', 207), ('analysis', 207), ('population', 206), ('test', 206), ('available', 205), ('hospital', 202), ('help', 198), ('year', 198), ('johnson', 195), ('different', 191), ('work', 187), ('immunity', 187), ('blood', 180), ('development', 177), ('kingdom', 176), ('evidence', 175), ('age', 173), ('national', 172), ('likely', 168), ('cell', 168), ('outbreak', 167), ('social', 164), ('credit', 162), ('state', 161), ('information', 160), ('global', 159), ('case', 157), ('way', 157), ('possible', 156), ('control', 153), ('school', 152), ('common', 150), ('safety', 149), ('infectious', 147), ('administration', 147), ('bat', 145), ('protection', 145), ('institute', 144), ('week', 143), ('treatment', 140), ('death', 139), ('company', 139), ('month', 136), ('cause', 135), ('home', 132), ('infected', 132), ('low', 131), ('rate', 127), ('important', 127), ('travel', 126), ('full', 126), ('similar', 125), ('large', 125), ('international', 124), ('illness', 124), ('science', 124), ('emergency', 123), ('much', 123), ('part', 123), ('oxygen', 123), ('city', 122), ('preprint', 122), ('n', 121), ('drug', 119), ('due', 119), ('small', 116), ('potential', 115), ('genetic', 114), ('delta', 113), ('syndrome', 112), ('shot', 112), ('south', 111), ('medicine', 111), ('need', 109), ('news', 107), ('lab', 107), ('plasma', 107), ('infect', 105), ('pfizer', 105), ('host', 104), ('good', 103), ('june', 102), ('acute', 102), ('report', 102), ('third', 101), ('prevent', 101), ('mask', 101), ('president', 101), ('york', 100), ('protective', 100)]

As we can see, the noun + adjective word cloud works perfectly for this topic. All the above words perfectly sum up the pandemic.

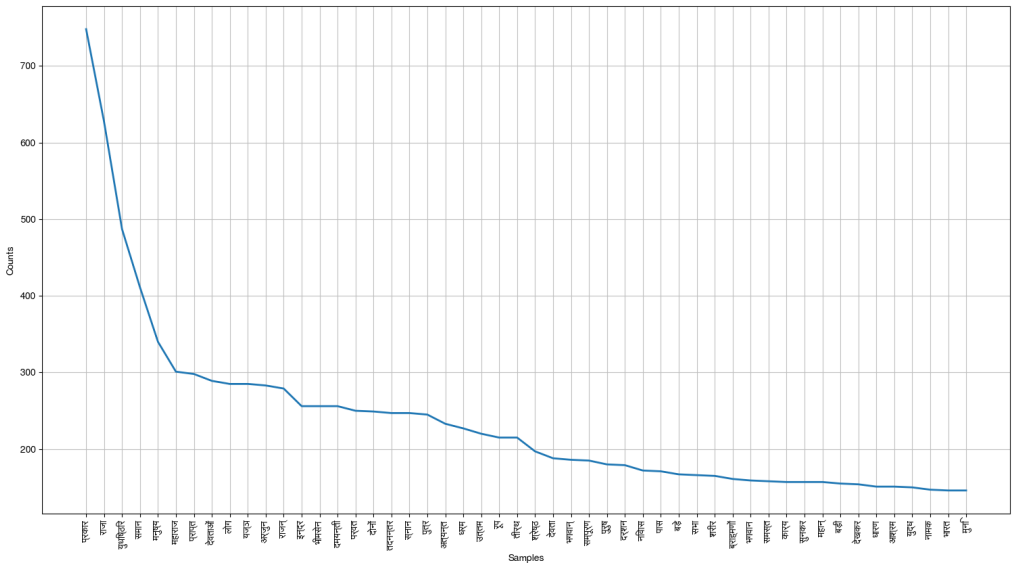
Hindi Dataset

The Hindi word cloud is about the Indian classic 'Mahabharata'. It comprises some of the incidences that took place during the beginning of 'Mahabharata'.

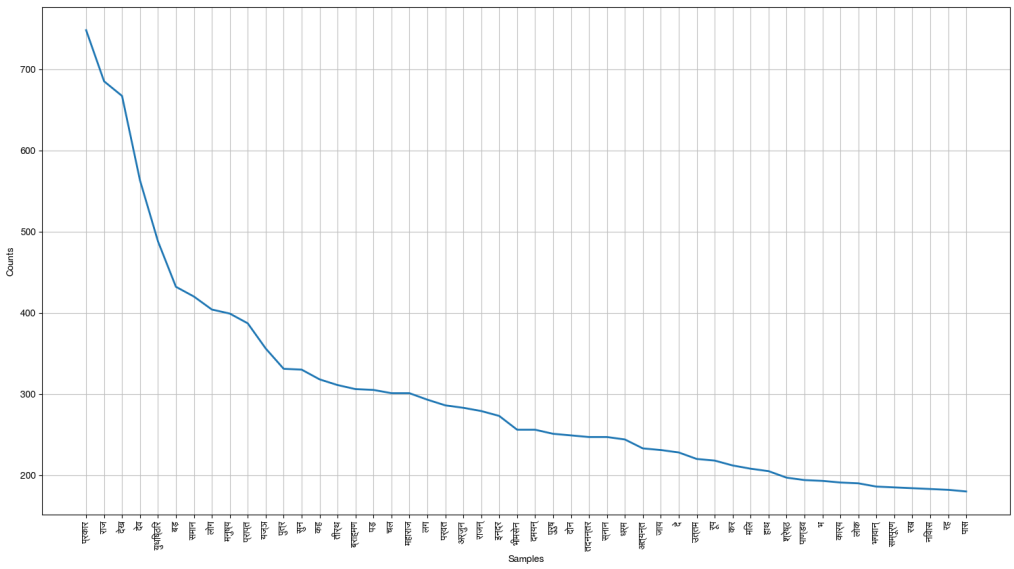
- Below is the graph of the inital tokens:



- We see a lot of punctuations - inverted commas and a lot of stopwords like 'का' 'की' 'है'.
- To clean the data many punctuations had to be added seperately, the string function did not account for '।' as a punctuation, which was problematic because it marks end of sentence in Hindi.
- Now after cleaning:



- The cleaned tokens depict the theme much better.
- A lot of extra stopwords had to be added, most of them were pronouns and general verbs. Like: मै, तुम, हम, थी etc.
- Right now the graph contain a lot of nouns and adjectives.
- Now when we apply stemming,



- We see that the graph changes drastically. The names on the x-axis change a lot. This is because in Hindi most of the verbs get stemmed, like:

  देखा, देखे, देखना, देखता etc all become देख

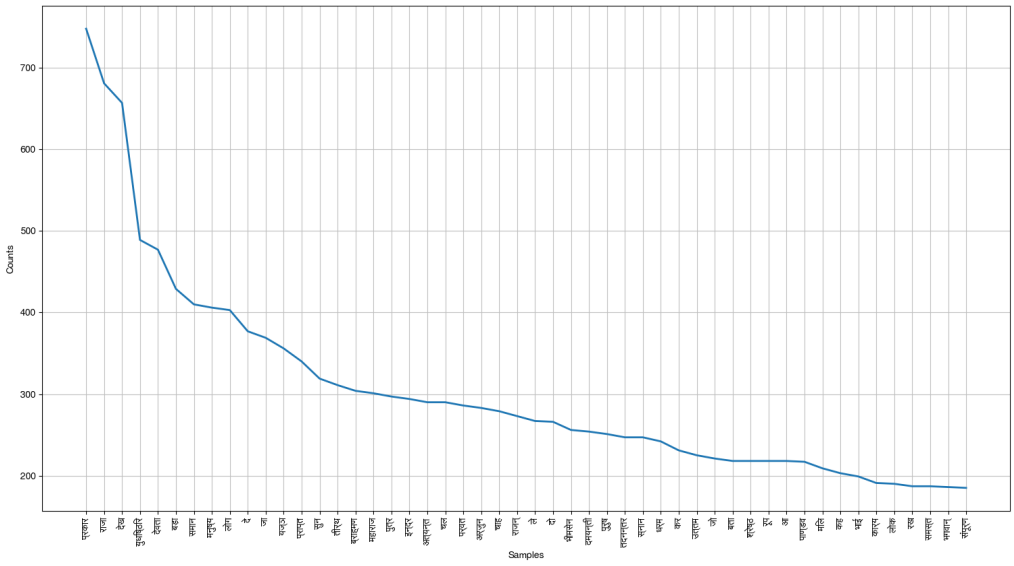
  कहता, कहती, कहना, कहा all become कह

  सुनना, सुनता, सुनती, सुना etc become सुन

  Similarly for the other verbs we can see in the graphs.

  These words were nowhere close to these frequencies before. In English such a drastic change was not noticed, but in Hindi most of the common verbs get stemmed

- Now, if we lemmatize,

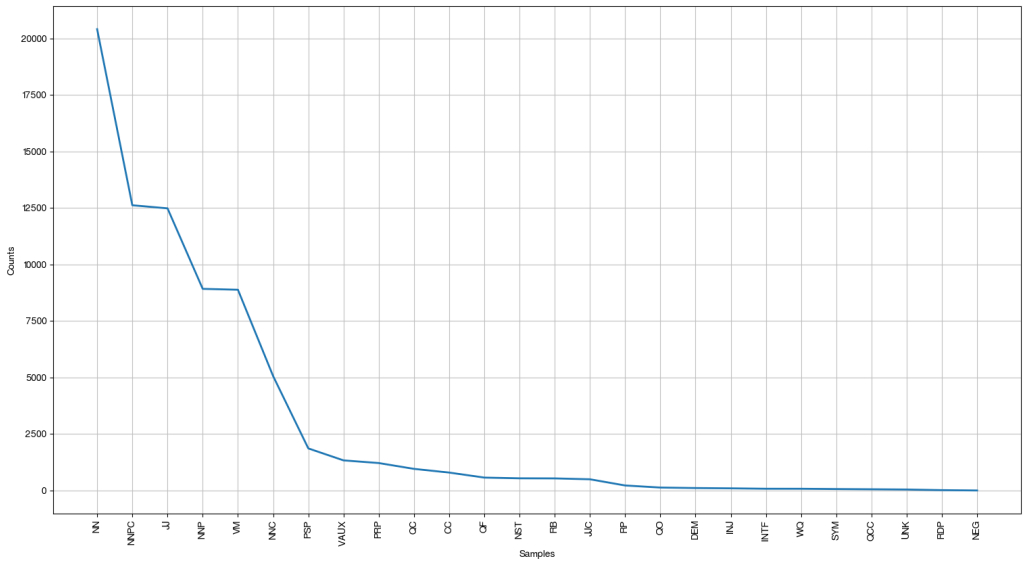


Again, we see a lot of verbs if we compare them to the cleaned tokens graph. The difference between the stemmed graph and this is that these are much more accurate. In stemming there was an instance where भाई got stemmed to भ; but in this, we can see that it remains भाई.

Example of a valid lemmatization:

आना,आने,आता,आती, आया all get lemmatized to आ

- Now let us look at the POS tag graphs



- We see that just like English nouns and adjectives have the highest frequency.

  When compared to English, verbs in Hindi are in a higher number.

  The number of pronouns are very less because most of these were removed while removing stopwords as they do not add to the theme but are used very frequently.
- Verbs, ordinals, cardinals, conjunctions, etc will not really describe the theme very well, so again we make a word cloud using nouns and adjectives.
- Around 150 words were used to make this wordcloud. The most common noun/adjective was of frequency 651 so it was only fair to include all words with atleast 1/10th this frequency.
- These frequencies are very less as compared to English because according to me, most of the high frequency words in Hindi are stopwords that do not add much to the theme.
- The wordcloud made, looks like this:

- The frequencies in decreasing order:

[('प्रकार', 651), ('राजा', 620), ('युधिष्ठिर', 487), ('मनुष्य', 323), ('महाराज', 301), ('प्राप्त', 298), ('देवताओं', 289), ('लोग', 284), ('अर्जुन', 283), ('यज्ञ', 277), ('राजन्', 269), ('इन्द्र', 256), ('भीमसेन', 256), ('दमयन्ती', 251), ('तदनन्तर', 247), ('स्नान', 246), ('पुत्र', 245), ('पर्वत', 235), ('धर्म', 224), ('उत्तम', 215), ('तीर्थ', 214), ('रूप', 204), ('समान', 200), ('श्रेष्ठ', 197), ('देवता', 187), ('भगवान्', 186), ('सम्पूर्ण', 185), ('दर्शन', 176), ('निवास', 171), ('पुरुष', 170), ('बड़े', 167), ('सभा', 166), ('शरीर', 164), ('ब्राह्मणों', 161), ('अत्यन्त', 160), ('भगवान', 158), ('बड़ी', 155), ('कार्य', 154), ('आश्रम', 151), ('धारण', 149), ('युद्ध', 149), ('नामक', 147), ('भारत', 146), ('मुनि', 146), ('महान्', 144), ('वीर', 142), ('ब्राह्मण', 140), ('नरेश', 140), ('नाम', 135), ('महात्मा', 133), ('समस्त', 130), ('दिव्य', 129), ('द्रौपदी', 124), ('पाण्डवों', 123), ('राजन', 122), ('महर्षि', 121), ('पवित्र', 121), ('क्रोध', 120), ('इच्छा', 120), ('लोगों', 119), ('पृथ्वी', 119), ('जनमेजय', 118), ('पालन', 118), ('लोमश', 118), ('देवराज', 117), ('श्रीकृष्ण', 116), ('हाथ', 113), ('आज्ञा', 111), ('शत्रुओं', 111), ('तपस्या', 107), ('भीम', 106), ('बड़ा', 106), ('भाई', 102), ('राक्षस', 100), ('सुन्दर', 100), ('विचार', 98), ('स्थान', 98), ('लोकों', 98), ('देश', 97), ('तीर्थों', 96), ('सूर्य', 95), ('शोभा', 94), ('कुबेर', 94), ('प्रसन्न', 93), ('मार्ग', 93), ('तेजस्वी', 93), ('वैशम्पायन', 91), ('धर्मराज', 91), ('भाँति', 91), ('भयंकर', 91), ('नदी', 91), ('रात', 91), ('सुशोभित', 89), ('भाइयों', 88), ('प्राप्ति', 88), ('सम्पन्न', 87), ('बातें', 85), ('यात्रा', 85), ('तेज', 84), ('विषय', 84), ('शोक', 83), ('वृक्ष', 82), ('समुद्र', 81), ('रक्षा', 80), ('हृदय', 80), ('ब्रह्मा', 80), ('वायु', 80), ('काल', 80), ('वृक्षों', 79), ('राज्य', 79), ('नरश्रेष्ठ', 78), ('प्रवेश', 78), ('अश्वमेध', 77), ('महान', 76), ('स्वर्गलोक', 76), ('बुद्धि', 76), ('भरतश्रेष्ठ', 75), ('प्रिय', 75), ('लोक', 75), ('राजेन्द्र', 75), ('योग्य', 74), ('सुख', 73), ('दिन', 73), ('दुःख', 73), ('प्रकट', 72), ('विशाल', 72), ('अग्नि', 72), ('तत्पश्चात्', 72), ('किया', 72), ('पिता', 71), ('पुरुषों', 71), ('यज्ञों', 71), ('पराक्रम', 71), ('अद्भुत', 70), ('पूर्ण', 70), ('बाहुक', 69), ('नाना', 68), ('वेग', 68), ('जरासंध', 68), ('रमणीय', 67), ('दान', 67), ('आकाश', 67), ('आश्रय', 67), ('परशुराम', 67), ('खड़े', 67), ('पितरों', 67), ('निश्चय', 66), ('पाण्डव', 66), ('मुख', 66), ('सरोवर', 66), ('कुन्तीनन्दन', 64), ('कर्म', 64), ('प्राणियों', 64), ('महाबली', 63), ('नष्ट', 63)]

As we can see, the noun + adjective wordcloud does justice to the theme Mahabharata.

More about the Word Cloud:

- To make the word cloud for both the data sets, a for loop was run on the POS tagged tokens to only append the chosen tags into a list.
- The reasoning behing the chosen tags has already been mentioned in the previous section.
- The Python Counter tool was then used on this list, and the most common words were used to make the wordcloud.
- Visualisation was done using the wordcloud package in Python and the matplotlib.