Requirement:

Generate a stoplist using the unigram data. How would you choose your cutoff value? Briefly justify your choice and comment on the stoplist content.

Solution:

In this scenario, tf-idf can be used to define our stopwords, because tf-idf reflects how a word is important to a document in a corpus.

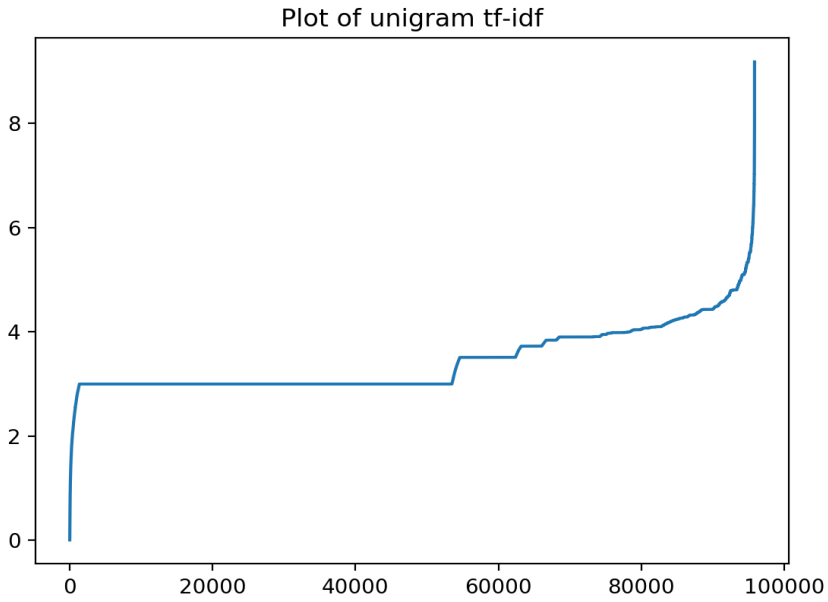
The tf-idf here is calculated by the formula below:

$$\text{tf-idf} = [1+\log10(ft)]*\log10(N/fd)$$

where ft represents for term frequency and fd represents for doc frequency, N = 1000.

The mechanism of this formula is that a big value is achieved by high term frequency and low doc frequency. When a term appears in more documents, the ratio inside log10(N/fd) approaches 1, making tf-idf closer to 0. (log10(ft) + 1) makes sure that those words appear only very little times will not make tf-idf a value closer to 0.

The sorted tf-idf plot:



Where we can see words with value less than 2 might be stopword. But we need to check the tf-idf list generated(to save the page space we only list the top 100 words and their tf-idf value)

As we can see from the list below, those words are mainly function words, although several of them are not because they are meaningful on their own. Therefore, it depends on our needs to choose the stoplist. A simple strategy is to choose words with tf-idf less than a value no largSer than 0.9 because meaningful words grows after 0.9.

```
('the', 0.008160033391613851)
('of', 0.01306735905827888)
('and', 0.02585349077691907)
('to', 0.040635136451843856)
('is', 0.043924364309242785)
('a', 0.04768890002478284)
('in', 0.04873121408164283)
('for', 0.12803416627229416)
('by', 0.13525671022278818)
('as', 0.1628165965989098)
('that', 0.19676171960829164)
('on', 0.20972427173788344)
('with', 0.21213381015052907)
('an', 0.22553818965304678)
('it', 0.24350299520735882)
('from', 0.25381894661429844)
('are', 0.26703820773877895)
('which', 0.3168827030217673)
('at', 0.3200984485034461)
('be', 0.3275806989153972)
('or', 0.34804651297161976)
('also', 0.34985620771430587)
('has', 0.3509434178992374)
('this', 0.3920330932880227)
('more', 0.4470546094934265)
('its', 0.472477503281175)
('have', 0.4745653093102944)
('was', 0.48839797184782635)
('such', 0.4925416841792625)
('other', 0.4954076907165948)
('can', 0.5374962778003846)
('not', 0.53992404586272)
('use', 0.5648265608426376)
('their', 0.5870833757020398)
('than', 0.5906775074177577)
('these', 0.6027749015329772)
('one', 0.606993878247564)
('but', 0.6183825996152418)
('into', 0.6376076143264732)
('been', 0.641321185241744)
('used', 0.6564885572601302)
('some', 0.6623852531212191)
('all', 0.6648142176485546)
```

```
('may', 0.7105011208318361)
('energy', 0.7180738825987807)
('they', 0.7318395609265185)
('there', 0.7370906321148404)
('most', 0.7409019959225058)
('many', 0.7507912157262693)
('over', 0.7629896909267042)
('between', 0.7671014330538347)
('when', 0.7785288295507062)
('new', 0.7851314370257749)
('through', 0.8343773640387541)
('include', 0.8353375166231801)
('only', 0.8386541535324409)
('well', 0.8846360153186985)
('were', 0.8854428608880832)
('however', 0.8881209849701867)
('will', 0.8952498215774871)
('including', 0.9059648935936229)
('both', 0.9086038117717703)
('where', 0.917084764683229)
('while', 0.9196402517893923)
('up', 0.9265970480428359)
('being', 0.9284254380616709)
('two', 0.934591876334683)
('time', 0.949953566154992)
('about', 0.9713210013909379)
('using', 0.9747159448137691)
('system', 1.0036139155978594)
('since', 1.0134976876129946)
('large', 1.018116867238486)
('under', 1.0329426152280525)
('united', 1.0344724069418643)
('would', 1.038167668409587)
('systems', 1.056427152913274)
('states', 1.0608451925514175)
('often', 1.0716300014864366)
('during', 1.071977837257048)
('any', 1.0770719292893653)
('high', 1.0864165546270552)
('world', 1.101920700728991)
('made', 1.104204641096231)
('development', 1.107063708223117)
('part', 1.1161078350917861)
('due', 1.128562961826739)
('because', 1.129386150739713)
('so', 1.1339955613238772)
('no', 1.1373601738645938)
('within', 1.140497883691604)
('then', 1.1455204767579643)
('much', 1.1470618327716757)
('example', 1.1500346380191488)
('power', 1.1523368148523818)
('number', 1.176115262341608)
('known', 1.1806023865194912)
```