Text_similarity

General setup and libraries:

Load helper scripts:

```
#Preprocessing script
source(sprintf("%sScripts/Computations/prepro_text.R", mainpath))
#Similarity matrix script
source(sprintf("%sScripts/Computations/setup_sim_matrix.R", mainpath))
#Text similarity script
source(sprintf("%sScripts/Computations/text_similarity.R", mainpath))
```

Load text files:

```
#Text1
text1 = "PG31878_tokens"
text1_raw = read.csv(paste0(mainpath,"Scripts/Example/",text1,".txt"), sep = "", header = FALSE)
text1 = paste(text1_raw$V1)
text1 = iconv(enc2utf8(text1),sub="byte")
text1 = list(text1[nchar(text1)>0])
#Text2
text2 = "PG40983_tokens"
text2_raw = read.csv(paste0(mainpath,"Scripts/Example/",text2,".txt"), sep = "", header = FALSE)
text2 = paste(text2_raw$V1)
text2 = iconv(enc2utf8(text2),sub="byte")
text2 = list(text2[nchar(text1)>0])
```

Set up similarity matrix:

```
net="combined_clics3based"
beta = 0.8
lower_th = 0.5
sim = setup_sim_matrix(mainpath, net, beta, lower_th)
```

Preprocess texts:

Compute text similarity:

```
#Sample 2000 words from each text
set.seed(1)
text1 = sample(text1, 2000)
set.seed(2)
text2 = sample(text2, 2000)
#Apply similarity algorithm
similarity = suppressMessages(text_sim(mainpath, text1, text2, sim))
print(similarity)
```

```
## [1] 6.029815
```