Project 2 - Implementation of HAL Algorithm

1. Implement the HAL algorithm in Python and estimate the word-vectors for the dataset COVID19 or with any of your language corpus of choice. Make sure that the size of the vocabulary is around 8000 words.
2. Use SVD to reduce word vector size to 50
3. Select ten words from the vocabulary and list their syntactic (associative) and semantic similarities. Use Cosine distance to measure the similarities

Wiki dump link to several languages

1. Bengali - https://dumps.wikimedia.org/bnwiki/latest/bnwiki-latest-pages-articles.xml.bz2
2. Hindi - https://dumps.wikimedia.org/hiwiki/latest/hiwiki-latest-pages-articles.xml.bz2
3. Kannada - https://dumps.wikimedia.org/knwiki/latest/knwiki-latest-pages-articles.xml.bz2
4. Malayalam - https://dumps.wikimedia.org/mlwiki/latest/mlwiki-latest-pages-articles.xml.bz2
5. Marati - https://dumps.wikimedia.org/mrwiki/latest/mrwiki-latest-pages-articles.xml.bz2
6. Tamil - https://dumps.wikimedia.org/tawiki/latest/tawiki-latest-pages-articles.xml.bz2
7. Telugu - https://dumps.wikimedia.org/tewiki/latest/tewiki-latest-pages-articles.xml.bz2

Covid19 Dump

https://drive.google.com/file/d/1PVA5IR4bAn7RulAeuSgxRBNMtmRGXaUm/view?usp=sharing