Approach and Procedure Followed:

Step 1: I have downloaded the corpus from sematic scholar. total size 36gb, with 40 individual corpus files. Each of these corpus files have articles of semantic scholar in json format with one article per line.

Step 2: To filter the corpus articles for our keywords, we use two fields for each corpus article.

1. entities - an array which contains topics/important words/terms in the article

2. abstract - the abstract of the research paper. we can look in the abstract for our keywords/Search words ["pre frontal cortex", "posterior parietal cortex", "pfc", "prefrontal cortex", "posterior parietal cortex", "ppc"] for each article we first look for keywords in entities:

- if any are found we download the article

-if none are found in entities then we search for keywords in abstract, if any are found here we download the article

the article pdf is downloaded only if download link is available in the article json

some of the download links are behind a paywall/ need site login. I can't access these and will be downloaded as html instead.

I have to filter these html files from the downloaded docs and remove them.

The final downloaded docs are all in pdf format [Which has been attached in the Github Repo/Google Drive]

Step 3: I have extracted the text from all these downloaded docs using pdfminer library

Step 4: I have done the count on the occurrences of each keyword in the extracted text

Step 5: I apply tf_idf transsformer from sklearn to generate tf_idf matrix for the data

Step 6: Did Latent Semantic Analysis

Step 7: Did RNN LSTM for finding relation between the two regions of Brain

Shiva, I have completed and attached the output files as well. Please let me know your valuable feedbacks.

Regards,

Bharath Kumar N