CS563 - NLP

(Read all the instruction carefully and adhere to them.)

Assignment - 3: Word and Sense Embedding

Deadline: 31st March 2019 **Date: 25th March 2019**

Identify the Most Frequent Sense of a word using Word Embeddings.

Input: A word w.
Output: Most frequent sense of the given word w.

Approach: Solve the problem by constructing the sense embedding. A sense embeddings are similar to word embeddings which are low dimensional real valued vectors. The steps are as follows:

1. Create the sense-bag for each sense of a word by extracting the context words from the WordNet such as
    - Content words in the gloss,
    - Content words in the example sentence,
    - synset members of the hypernymy synsets,
    - synset members of the hyponymy synsets,

    Note: Use nltk.wordnet package for the creation of sense-bag.
2. Construct the sense embeddings by taking the average of word embeddings of each word in the sense-bag. Take any pre-trained word-embedding model (e.g., Google News word2vec or CommonCrawl GloVe embedding)
3. Identify the most frequent sense, of a given word w, by computing the cosine similarity between the word embedding and sense embeddings of the word w. The sense having maximum cosine similarity value should be returned as the most frequence sense.
4. Plot the word embedding and sense embeddings for word w in 2D space. (Use PCA to reduce the dimensionality of embeddings)

Reference:
https://pdfs.semanticscholar.org/b58e/477022d79562ce1c5e76218bb328c8fb7c3c.pdf

Submission guidelines:

- Please adhere to following guidelines while submitting your assignment.
- Please submit your assignment **on or before the deadline**.
- Compress all your files **(Input / Output / Codes / Analysis)** in zip file. It should be named as **Roll1_Roll2_Roll3-Assignment-#.zip**
- Please submit your assignment on "https://goo.gl/gCMwfV**".**