Simply Diagram Flow Diachronic POS Tagger
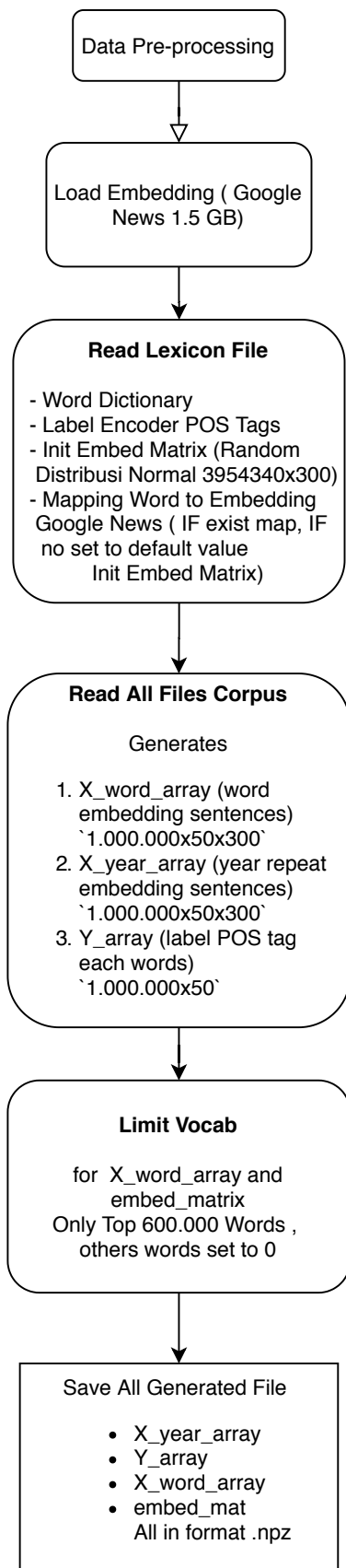
Step 1

Data Pre-processing

↓

Load Embedding (Google News 1.5 GB)

↓

Read Lexicon File

- Word Dictionary
- Label Encoder POS Tags
- Init Embed Matrix (Random Distribusi Normal 3954340x300)
- Mapping Word to Embedding Google News (IF exist map, IF no set to default value
  Init Embed Matrix)

↓

Read All Files Corpus

Generates

1. X_word_array (word embedding sentences) `1.000.000x50x300`
2. X_year_array (year repeat embedding sentences) `1.000.000x50x300`
3. Y_array (label POS tag each words) `1.000.000x50`

↓

Limit Vocab

for X_word_array and embed_matrix
Only Top 600.000 Words , others words set to 0

↓

Save All Generated File

- X_year_array
- Y_array
- X_word_array
- embed_mat
  All in format .npz

Step 2

LSTM Training

↓

Data Train Load Data Test Load

↓

Load Embed Matrix

↓

Create TF session

↓

Training Data LSTM , Feed Forward

↓

Calculation Accuracy Training set Calculation Accuracy Test set

↓

PCA linear Reduction Feature Matrix

↓

Calculate Average Perplexity for Training set and Test set and Visualization

↓

Find Sample Sentences for Inferencing Model