MCA Assignment 3: Text Representation and Retrival

Sarthak Bhagat

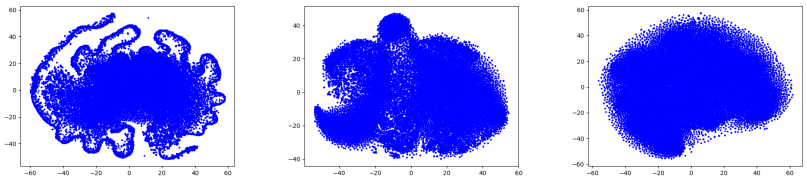
May 10, 2020

1 Work2Vec

Steps of the word2vec algorithm:

- Data preparation: Creating corpus and tokenizing words.
- Generating training data: Building the vocabulary and create dictionary that give id to word and word to id.
- Training: Pass latent word embeddings through forward pass of the network, calculate loss (NLL in our case), update weights of the network using gradient of the loss. Trained for 3 epochs as the loss converged.

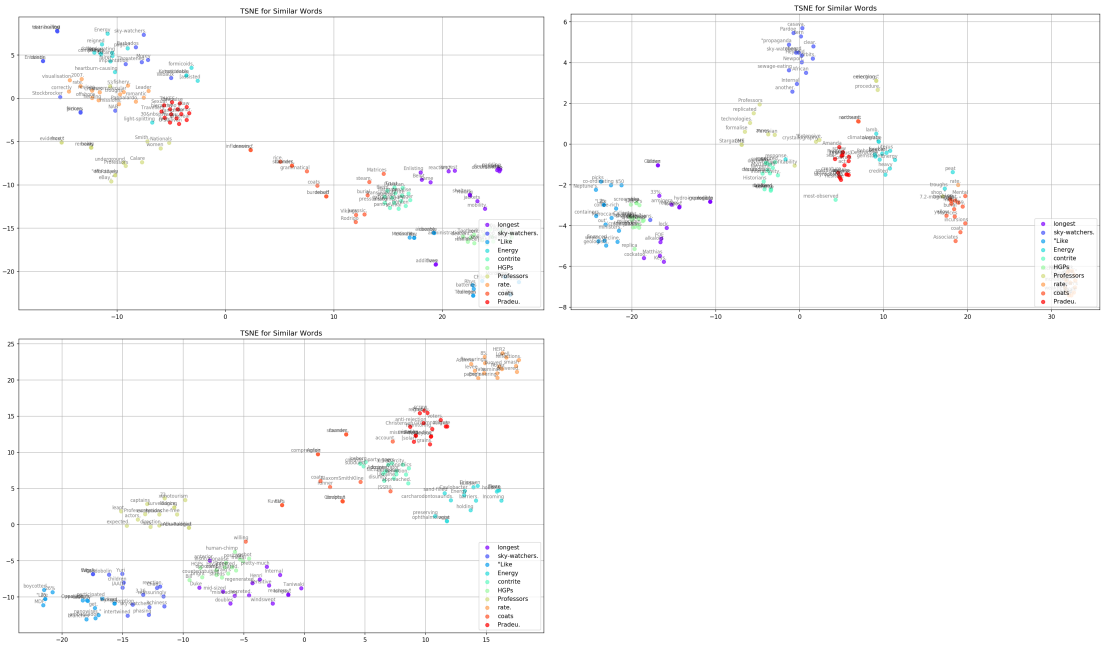
The loss log file is provided in form of a text file that depicts the convergence of the loss over iterations over each epoch.



(a) TSNE visualizations over epochs

As shown in the above figure, the manifold formed by the latent space points flattens out as the training continues to form a much generalised version (and low dimensional) of the initial high dimensional latent manifold. This would be much more generalised for the ensure data distribution.

For inference of the model, we also plot the similar word encodings using a TSNE plot over the epochs.



(a) TSNE visualizations for similar words over epochs

2 Document Retrieval using Query Expansion

Performance on Baseline: 49.22

Performance on Relevance Feedback: 60.06

Performance on Relevance Feedback using Query Expansion: 61.69

Performance Discussion. The performance of relevance feedback improved by a significant margin when we include the query expansion part. This is completely inline with what was expected as including query expansion into the model ensures that we consider all synonyms of a word (by choosing n closest words from the thesaurus). Doing this, we are enhancing the queries as words with similar meaning are considered strongly during the retrieval process. For example if we search pilot, words that are similar to it like plane should automatically be given more importance.

Performance Evaluation over Iterations: The performance of both models improves as the iterations proceed in the following way:

Method	Iteration 1	Iteration 2	Iteration 3
Relevance Feedback	55.6	58.5	60.0
Relevance Feedback with Query Expansion	52.7	59.8	61.7