Lab 6: Vector Semantics and Embeddings

Natural Language Processing

Informations

The chapter 6 in the *Speech and Language Processing* (Jurafsky & Martin) doesn't have its own exercises, so these exercises are created by ChatGPT.

Overviews

– **Subject**: Natural language processing

– **Topic**: Vector Semantics, TF-IDF, PPMI, Word Embeddings

– **Durations**: 2 lectures x 90 minutes

– **Tools**: Python (NumPy, scikit-learn), Jupyter Notebook or Colab

Objectives

After finishing this assignments, students can:

- Understand and use cosine similarity, PMI, PPMI, TF-IDF
- Build co-occurrence matrix
- Use TF-IDF and PPMI to represent words and documents
- train word embeddings for analogy calculate similarity
- Compare TF-IDF and PPMI in the classification problems

Pre-lab Reading

Chapter 6 in *Speech and Language Processing* (Jurafsky & Martin)

- 6.1 Vector Semantics
- 6.3 TF-IDF
- 6.6 PMI and PPMI
- 6.7 Embeddings via Matrix Factorization

Exercises

I. Theory

Ex 1. Proof these equations

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}_{\text{pos}}} = [\sigma(\boldsymbol{c}_{\text{pos}} \cdot \boldsymbol{w}) - 1]\, \boldsymbol{w} \tag{6.35}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}_{\text{neg}}} = [\sigma(\boldsymbol{c}_{\text{neg}} \cdot \boldsymbol{w})]\, \boldsymbol{w} \tag{6.36}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = [\sigma(\boldsymbol{c}_{\text{pos}} \cdot \boldsymbol{w}) - 1]\, \boldsymbol{c}_{\text{pos}} + \sum_{i=1}^{k} [\sigma(\boldsymbol{c}_{\text{neg}_i} \cdot \boldsymbol{w})]\, \boldsymbol{c}_{\text{neg}_i} \tag{6.37}$$

Ex 2. Compute PMI/PPMI. for 3 documents:

Doc1: `"dogs bark loudly"`
Doc2: `"cats meow softly"`
Doc3: `"dogs and cats play"`

- Build a co-occurences matrix with context window size = 1 (remoce stopwords)
- Compute PMI and PPMI for pairs of words: (”dogs”, ”bark”), (”cats”, ”meow”), (”dogs”, ”cats”)

Ex 3. Cosine similarity. Given 3 word vectors (normalized):

$$\text{king} = [0.7, 0.1, 0.3], \quad \text{queen} = [0.69, 0.12, 0.31], \quad \text{man} = [0.5, 0.09, 0.4]$$

- Compute cosine similarity between: (king, queen) and (king, man)
- Analyze why king is nearer queen than man in the vector space

Ex 4. Answers the questions:

1. what are the mathematical differences between PPMI and TF-IDF mathematically and their representing objectives?
2. when does PPMI cannot represent well?
3. Advantages and disadvantages of TF-IDF comparing to word embeddings?

II. Practice

Ex 5. TF-IDF + Cosine Similarity

- Pick 5 documents
- Compute TF-IDF vector for each documents
- Find the most similar pair of documents (cosine similarity)
- Print 3 words has the highest TF-IDF for each documents

Ex 6. PPMI Matrix + Visualization

- Create a co-occurances matrix for small documents
- Compute PPMI matrix
- Use SVD to reduce the dimension to 2 dimension
- Plot the words to the 2D spaces (matplotlib)

III. Advanced – Pratical applications

Ex 7. Word Analogies with GloVe

- download GloVe (glove.6B.100d.txt)
- Compute:

$$\text{king} - \text{man} + \text{woman} \approx ?$$

$$\text{paris} - \text{france} + \text{italy} \approx ?$$

- Print top 5 similar words

Ex 8. Classify with TF-IDF vs PPMI

- Get binary documents (ex: positive/negative review)
- Represent each documents by:
  - a) TF-IDF vector
  - b) PPMI vector (the mean of PPMI for each word)
- Use Logistic Regression to classify
- Comparing Accuracy and F1-score