2019-02-14 Lab 5

Contents

Exercise: Review Annotation

- □ Open today's notebook
- □ Discuss use of NLTK's tokenizer and its relationship to annotation – specifically, part-of-speech.
  - □ Why does NLTK's parser produce better results? What does this imply about the POS algorithm?

Exercise: Create BOW Vector Space Model

- □ Discuss method of converting the tokens table from our F2 model to a vector space
  - □ What are our choices (hyperparameters) in making this conversion?
- □ Convert the BOW data frame into a document-term matrix data frame.
  - □ What happens in the process? What is added?

Exercise: Compute Term Frequencies and Weights

- □ Discuss method of computing term frequencies.
  - □ How can we confirm that we have created a probability distribution?
- □ Discuss methods of computing weights for term frequencies.
- □ What values can we add to the vocabulary table?

Exercise: Find and Compare Significant Words

- □ Compare results of weighting methods by inspecting top words in the vocabulary table.
- □ Does the experimental TFTH method help us at all?

Exercise: Compare Chapters

- □ Compare each chapter to every other chapter using both euclidean and cosine measures.
  - □ How do generate a set of pairs to compare?
- □ Discuss how the metrics are computed in Pandas.
- □ Discuss how the measures differ.

Exercise: Compare Two Interesting Words

- □ Go back to our list of top words.
- □ What are two top words that might be useful to compare?
- □ Compare these words by plotting their distribution over the novel.
    - □ How is the form of comparison different from what is expected by traditional information retrieval methods?

Homework

- □ Rewrite today's notebook to work with Jane Austen's *Persuasion*.
    - □ Don't worry about answering the discussion questions in the notebook – they are for in-class use.
- □ In your rewrite, write a function that can create a bag of words from a tokens table that takes the following arguments:
    - □ The tokens data frame to use.
    - □ Choice of OHCO container, e.g. which "bag" to use.
    - □ Choice of item to count, e.g. terms or stems.
- □ Based on the application of your weighting metrics, find two top words to compare, using the methods in the notebook.
- □ Submit your notebook or Spyder file to the appropriate Collab assignment.

Files

- DS5559_VSM.ipynb