CSCE 771: Computer Processing of Natural Languages
Prof. Biplav Srivastava, Fall 2022

Quiz 2 / Instructions

- This is a programming quiz. Code has to be submitted in a directory of your GitHub called "Quiz2" with sub-dir for code, data and doc. Code will have your source code, data will have any input or output generated, and doc will have a .pdf of this file (called Quiz2-CSCE771-answers.pdf) along with any answers
- Complete quiz by 9:00 am on Monday, Oct 3, 2022 by sending an email to biplav.s@sc.edu confirming completing the quiz and attaching your Quiz2-CSCE771-answers.pdf.
- Total points = 50 + 20 + 30 = 100
- Obtained =

Student Name: Guangyi Chen

Question 1: Contextual word embedding and TF-IDF
[5 + 5 + 20 + 10 + 10 = 50 points]

(a) What is the benefit of using a counting based vector representation like TF-IDF over a learning based representation like Word2Vec? [5 points]

TF-IDF can be used to a document but Word2Vec must be used to each word individually so it makes Word2Vec more memory intensive.
TF-IDF is easy to apply than Word2Vec.
Word2Vec needs a very large corpus for training .TF_IDF can work with smaller datasets as well.

(b) What are the advantages of character-based representation like fasttext over word-based representation like Word2Vec? [5 points]

Fasttext treats each word as n grams which makes better word embeddings for rare words.
But Word2Vec cannot handle the words they haven't seen before.

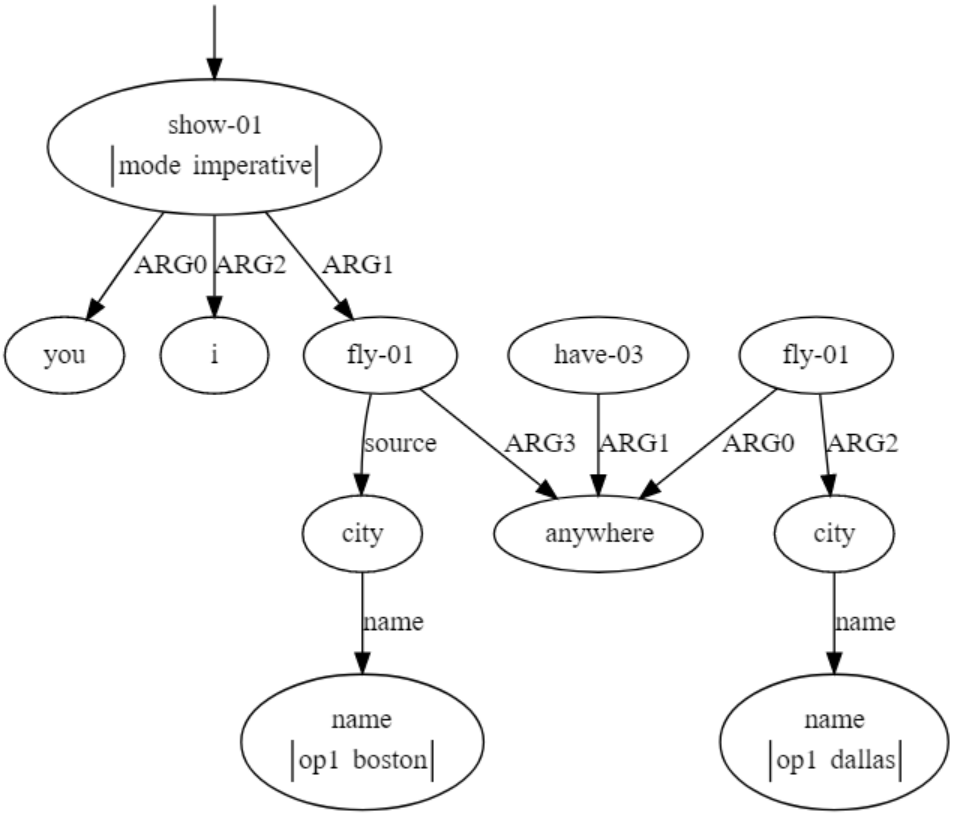
(c) In sample-code/l13-llm-quiz folder in course github, you will find a file called “projs.txt” containing the list of projects in the course. Do the following:

(i) Consider each line as a document and represent words in TF-IDF. [20 points]
(ii) Identify your project name and identify all projects similar to yours. Use a cosine similarity of 0.9 [10 points]
(iii) Identify clusters of projects along the same theme, based on similarity of project names. (Hint: you just have to reuse your code from (ii) above) [10 points]

Question 2: Semantics
[8 + 12 = 20 points]

Consider text = “show me flights from Boston to anywhere that has flights to Dallas”

(a) Using the online AMR tool at http://amparser.coli.uni-saarland.de:8080/, find the AMR structure of the example text. Paste it below.



(b) The AMR refers to specific variant of **show**, **fly** and **have**. Use pennbank and show the predicate, its arguments and its meaning. Use a propbank visualizer like https://verbs.colorado.edu/verb-index/index.php.

Predicate: *show*

Roleset id: show.01 , *cause to see*, Source: , vncls: , framnet:

Predicate: *fly*

Roleset id: fly.01 , *fly through the air, travel via air, fly in a flock.,*, Source: ,

vncls: , framnet:

Predicate: *have*

Roleset id: have.01 , *auxiliary*, Source: , vncls: , framnet:

Question 3: Word2Vec
[10 + 10 + 10 = 30 points]

(a) Take your latest resume (must be more than 1 page). Create a word2vec representation for it using genism and print statistics of embeddings.

(b) Visualize the embedding using PCA.

(c) Now create and visualize the embedding of the projects listed in the file - sample-code/l13-llm-quiz/projs.txt.