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: Sourin Dey

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]
Ans: TF-IDF calculates the frequency as well as importance of words while Word2Vec provides vector representation (word embedding) of a word which does not have interpretable representation. Word2Vec also requires further works of dimensionality reduction for its word representation. It is also expensive to train such a model.

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

Ans: Character-based representations provide infer meaning of unseen word. For example, if a word prestigious is absent in the corpus, it can still find meaning of it by finding meaning of the word prestige. Word2Vec would not be able to do that. Also, in different languages such as in German language, training Word2Vec is difficult. For example, tennis and tischtennis are similar but word2vec won’t recognize it while there will be overlapping n-grams while using fasttext.

(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]

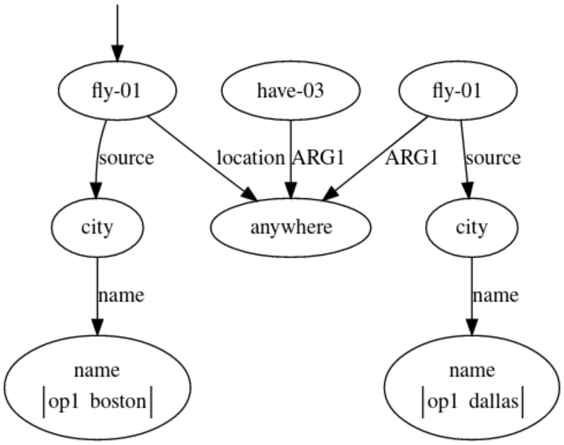
Ans: In Quiz2/code/ Quiz question 1.ipynb

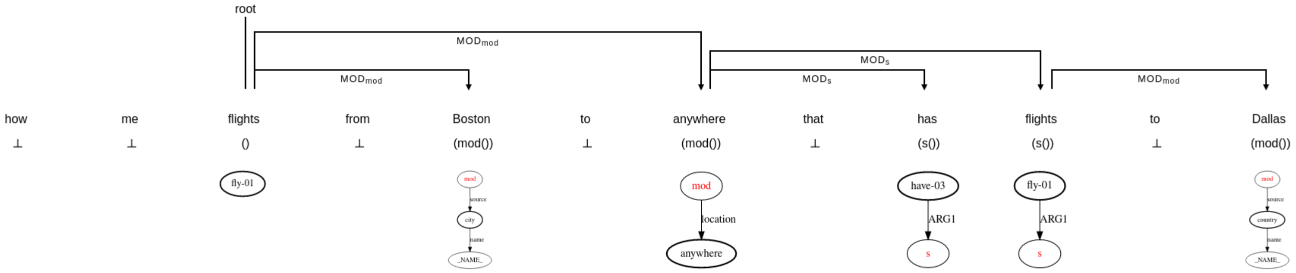
Question 2: Semantics

[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.

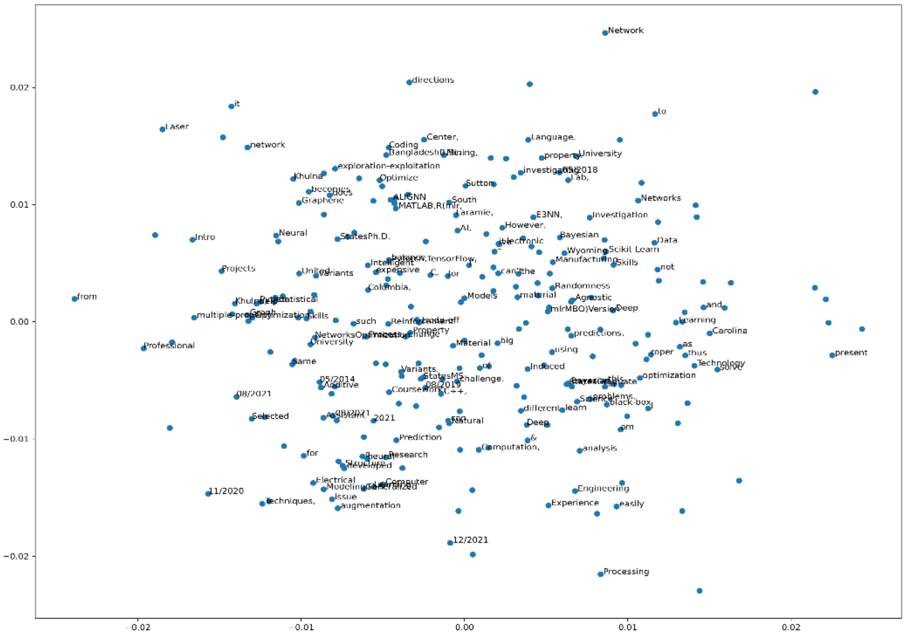
Ans: I could not solve this problem.

Question 3: Word2Vec
Ans: In Quiz2/code/ Quiz question 3.ipynb

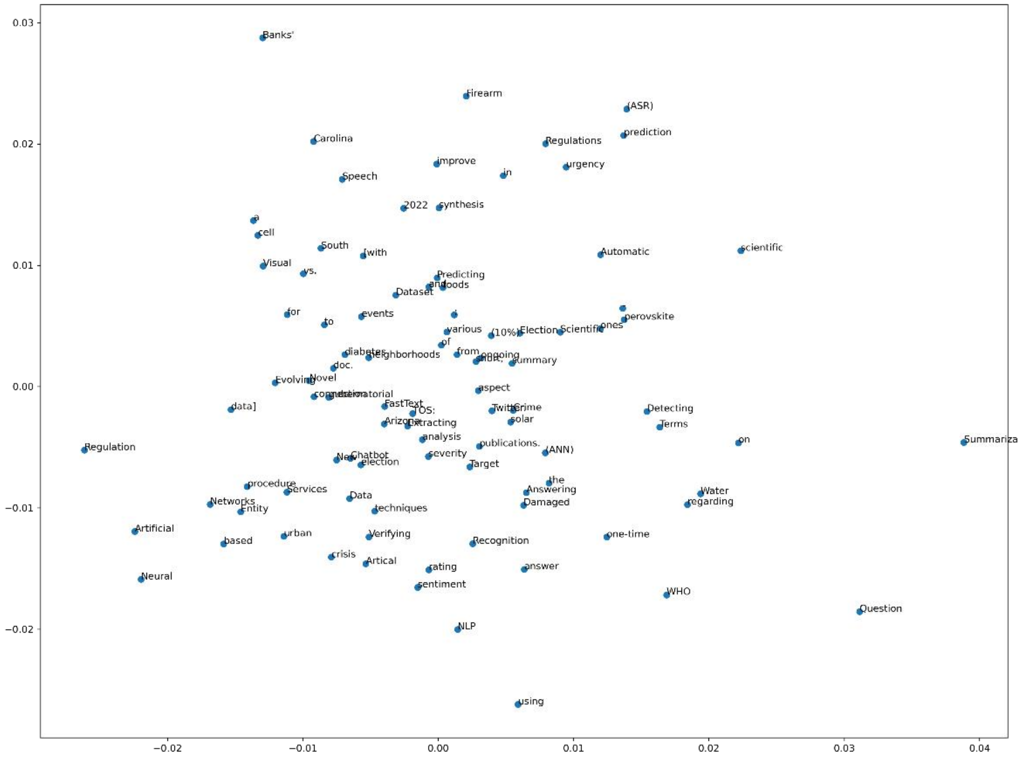
[10 + 10 + 10 = 30 points]

(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.


Banks'
Firearm
(ASR)
prediction
Carolina
Regulations
improve
urgency
in
Speech
synthesis
a
cell
South
[with
Automatic
scientific
Visual
vs.
Predicting
and
foods
Dataset
for
to
events
i
perovskite
various
(10%)
Election
Scientific
ones
of
diabetes
neighborhoods
from
ongoing
short,
summary
doc.
Evolving
Novel
aspect
combinatorial
data]
FastText
TOS:
Twitter
Crime
Detecting
Arizona
Extracting
solar
Terms
analysis
on
Summarization
Regulation
publications.
(ANN)
New
Chatbot
severity
election
Target
the
procedure
Services
Answering
Water
Data
Networks
Damaged
regarding
Entity
techniques
Artificial
urban
Verifying
one-time
based
Recognition
crisis
Artical
rating
answer
Neural
sentiment
WHO
Question
NLP
using