SATHYABAMA INSTITUTE OF SCIENCE & TECHNOLOGY

SCHOOL OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SCSA 2604 NATURAL LANGUAGE PROCESSING LAB

LAB 4: SEMANTIC ANALYSIS

AIM: To perform Semantic Analysis using Gensim

PROCEDURE:

Semantic analysis is a broad area in NLP. This program demonstrates semantic analysis by leveraging pre-trained word vectors using Word2Vec from Gensim. It utilizes word embeddings to find words similar to each word in the provided sentences.

Library Installation: Ensure the necessary libraries (Gensim and NLTK) are installed.

Library Import: Import the required libraries (gensim for word vectors and nltk for tokenization).

Pre-trained Word Vectors: Load pre-trained word vectors (Word2Vec) using Gensim's api.load() method.

Sample Sentences: Define sample sentences for semantic analysis.

Tokenization: Break down the sentences into individual words using NLTK's word_tokenize() method.

Semantic Analysis: Iterate through each word in the tokenized sentences and:

> Check if the word exists in the pre-trained Word2Vec model.
>
> If the word exists, find similar words using the most_similar() method from the word vectors model.
>
> Display or store the similar words for each word in the sentence.
>
> If the word doesn't exist in the pre-trained model, indicate that it's not present.

The following algorithm outlines the steps involved in performing semantic analysis using pre-trained word vectors (Word2Vec) in Python, demonstrating how to find similar words foreach word in the provided sentences based on the loaded word vectors.

ALGORITHM:

1. Install Necessary Libraries: Install Gensim and NLTK libraries (!pip install gensim, !pip install nltk).
2. Import Libraries: Import required libraries: gensim for word vectors and nltk for tokenization.
3. Download Pre-trained Word Vectors: Download pre-trained word vectors (Word2Vec) using Gensim's api.load() method.
4. Define Sample Sentences: Create sample sentences for semantic analysis.
5. Tokenization: Tokenize the sentences into words using NLTK's word_tokenize() method.
6. Semantic Analysis with Word Vectors: Iterate through each tokenized sentence.
   For each word in the sentence:
   Check if the word exists in the pre-trained Word2Vec model.
   If the word exists:
   Find words similar to the current word using word_vectors.most_similar(word).
   Display or store the similar words.
   If the word doesn't exist in the model:
   Print a message indicating that the word is not in the pre-trained model.

PROGRAM:

```
# Install necessary libraries
!pip install gensim
!pip install nltk

# Import required libraries
import gensim.downloader as api
from nltk.tokenize import word_tokenize

# Download pre-trained word vectors (Word2Vec)
word_vectors = api.load("word2vec-google-news-300")

# Sample sentences
sentences = [
   "Natural language processing is a challenging but fascinating field.",
   "Word embeddings capture semantic meanings of words in a vector space."
]
```

```
# Tokenize sentences
tokenized_sentences = [word_tokenize(sentence.lower()) for sentence in sentences]

# Perform semantic analysis using pre-trained word vectors
for tokenized_sentence in tokenized_sentences:
    for word in tokenized_sentence:
        if word in word_vectors:
            similar_words = word_vectors.most_similar(word)
            print(f"Words similar to '{word}': {similar_words}")
        else:
            print(f"'{word}' is not in the pre-trained Word2Vec model.")
```

OUTPUT:

```
Requirement already satisfied: gensim in /usr/local/lib/python3.10/dist-packages (4.3.2)
Requirement already satisfied: numpy>=1.18.5 in /usr/local/lib/python3.10/dist-packages
(from gensim) (1.23.5)
Requirement already satisfied: scipy>=1.7.0 in /usr/local/lib/python3.10/dist-packages (from
gensim) (1.11.3)
Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.10/dist-packages
(from gensim) (6.4.0)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk)
(8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk)
(1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages
(from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk)
(4.66.1)
[==================================================]          100.0%
1662.8/1662.8MB downloaded
Words similar to 'natural': [('Splittorff_lacked', 0.636509358882904), ('Natural',
0.58078932762146), ('Mike_Taugher_covers', 0.577259361743927), ('manmade',
```

0.5276211500167847), ('shell_salted_pistachios', 0.5084421634674072), ('unnatural', 0.5030758380889893), ('naturally', 0.49992606043815613), ('Intraparty_squabbles', 0.4988228678703308), ('Burt_Bees_®', 0.49539363384246826), ('causes_Buxeda', 0.4935200810432434)]

Words similar to 'language': [('langauge', 0.7476695775985718), ('Language', 0.6695356369018555), ('languages', 0.6341332197189331), ('English', 0.6120712757110596), ('CMPB_Spanish', 0.6083104610443115), ('nonnative_speakers', 0.6063109636306763), ('idiomatic_expressions', 0.5889801979064941), ('verb_tenses', 0.58415687084198), ('Kumeyaay_Diegueno', 0.5798824429512024), ('dialect', 0.5724600553512573)]

Words similar to 'processing': [('Processing', 0.7285515666007996), ('processed', 0.6519132852554321), ('processor', 0.636760413646698), ('warden_Dominick_DeRose', 0.6166526675224304), ('processors', 0.5953895449638367), ('Discoverer_Enterprise_resumed', 0.5376213192939758), ('LSI_Tarari', 0.520267903804779), ('processer', 0.5166687369346619), ('remittance_processing', 0.5144169926643372), ('Farmland_Foods_pork', 0.5071728825569153)]

Words similar to 'is': [('was', 0.6549733281135559), ("isn'ta", 0.6439523100852966), ('seems', 0.634029746055603), ('Is', 0.6085968613624573), ('becomes', 0.5841935276985168), ('appears', 0.5822900533676147), ('remains', 0.5796942114830017), ('is', 0.5695518255233765), ('makes', 0.5567088723182678), ('isn_`_t', 0.5513144135475159)]

'a' is not in the pre-trained Word2Vec model.

Words similar to 'challenging': [('difficult', 0.6388775110244751), ('challenge', 0.5953003764152527), ('daunting', 0.569800615310669), ('tough', 0.5689979791641235), ('challenges', 0.5471934676170349), ('challenged', 0.5449535846710205), ('Challenging', 0.5242965817451477), ('tricky', 0.5236554741859436), ('toughest', 0.5169045329093933), ('diffi_cult', 0.5010539889335632)]

Words similar to 'but': [('although', 0.8104525804519653), ('though', 0.7285684943199158), ('because', 0.7225914597511292), ('so', 0.6865807771682739), ('But', 0.6826984882354736), ('Although', 0.6188263297080994), ('Though', 0.6153667569160461), ('Unfortunately', 0.6031029224395752), ('Of_course', 0.593142032623291), ('anyway', 0.5869061350822449)]

Words similar to 'fascinating': [('interesting', 0.7623067498207092), ('intriguing', 0.7245113253593445), ('enlightening', 0.6644250154495239), ('captivating', 0.6459898352622986), ('facinating', 0.6416683793067932), ('riveting', 0.6324825286865234), ('instructive', 0.6210989356040955), ('endlessly_fascinating', 0.6188612580299377), ('revelatory', 0.6170244216918945), ('engrossing', 0.6126049160957336)]

Words similar to 'field': [('fields', 0.5582526326179504), ('fi_eld', 0.5188260078430176), ('Keith_Toogood', 0.49749255180358887), ('Mackenzie_Hoambrecker', 0.49514278769493103), ('Josh_Arauco_kicked', 0.48817265033721924), ('Nick_Cattoi', 0.4863145053386688), ('Armando_Cuko', 0.4853871166706085), ('Jon_Striefsky', 0.48322004079818726), ('kicker_Nico_Grasu', 0.47572532296180725), ('Chris_Manfredini_kicked', 0.47327715158462524)]

```
'.' is not in the pre-trained Word2Vec model.

Words similar to 'word': [('phrase', 0.6777030825614929), ('words', 0.5864380598068237),
('verb', 0.5517287254333496), ('Word', 0.54575115442276), ('adjective',
0.5290762186050415), ('cuss_word', 0.5272089242935181), ('colloquialism',
0.5160348415374756), ('noun', 0.5129537582397461), ('astrology_#/##/##',
0.5039082765579224), ('synonym', 0.49379870295524597)]

'embeddings' is not in the pre-trained Word2Vec model.

Words similar to 'capture': [('capturing', 0.7563897371292114), ('captured',
0.7155306935310364), ('captures', 0.6099075078964233), ('Capturing',
0.6023245453834534), ('recapture', 0.5498639941215515), ('Capture', 0.5493018627166748),
('nab', 0.4941576421260834), ('Captured', 0.45745959877967834), ('apprehend',
0.4357919692993164), ('seize', 0.4338296055793762)]

Words similar to 'semantic': [('semantics', 0.6644964814186096), ('Semantic',
0.6464474201202393), ('contextual', 0.5909127593040466), ('meta', 0.5905876755714417),
('ontology', 0.5880525708198547), ('Semantic_Web', 0.5612248778343201), ('semantically',
0.5600483417510986), ('microformat', 0.5582399368286133), ('inferencing',
0.5541478991508484), ('terminological', 0.5533202290534973)]

Words similar to 'meanings': [('grammatical_constructions', 0.594986081123352), ('idioms',
0.5938195586204529), ('connotations', 0.5836683511734009), ('symbolic_meanings',
0.5806494951248169), ('meaning', 0.5785343647003174), ('literal_meanings',
0.5743482112884521), ('denotative', 0.5730364918708801), ('phrasal_verbs',
0.5697917342185974), ('contexts', 0.5609514713287354), ('adjectives_adverbs',
0.5569407343864441)]

'of' is not in the pre-trained Word2Vec model.

Words similar to 'words': [('phrases', 0.7100036144256592), ('phrase', 0.6408688426017761),
('Words', 0.6160537600517273), ('word', 0.5864380598068237), ('adjectives',
0.5812757015228271), ('uttered', 0.5724518299102783), ('plate_umpire_Tony_Randozzo',
0.5642045140266418), ('expletives', 0.5539036989212036), ('Mayor_Cirilo_Pena',
0.553884744644165), ('Tele_prompter', 0.5441114902496338)]

Words similar to 'in': [('inthe', 0.5891957879066467), ('where', 0.5662435293197632), ('the',
0.5429296493530273), ('In', 0.5415117144584656), ('during', 0.5188906192779541), ('iin',
0.48737412691116333), ('at', 0.484235554933548), ('from', 0.48268404603004456),
('outside', 0.47092658281326294), ('for', 0.4566476047039032)]

'a' is not in the pre-trained Word2Vec model.

Words similar to 'vector': [('vectors', 0.750322163105011), ('adeno_associated_viral_AAV',
0.5999537110328674), ('bitmap_graphics', 0.5428463220596313), ('Sindbis',
0.5353653430938721), ('bitmap_images', 0.5318013429641724), ('signal_analyzer_VSA',
0.5276671051979065), ('analyzer_VNA', 0.5184376239776611), ('vectorial',
0.5084835886955261), ('nonviral_gene_therapy', 0.5036363005638123), ('shellcode',
0.5015827417373657)]
```

```
Words similar to 'space': [('spaces', 0.6570690870285034), ('music_concept_ShockHound',
0.5850345492362976), ('Shuttle_docks', 0.5566749572753906), ('Space',
0.5478203296661377), ('Soviet_Union_Yuri_Gagarin', 0.5417766571044922),
('Shuttle_Discovery_blasts', 0.5352603197097778), ('Shuttle_Discovery_docks',
0.534925103187561), ('Shuttle_Endeavour_undocks', 0.532420814037323),
('Shuttle_Discovery_arrives', 0.5323426723480225), ('Shuttle_undocks',
0.523307740688324)]

'.' is not in the pre-trained Word2Vec model.
```