LECTURER: Nghia Duong-Trung

ARTIFICIAL INTELLIGENCE

TOPIC OUTLINE

History of AI	
Modern AI Systems	
Reinforcement Learning	
Natural Language Processing – Part 1	
Natural Language Processing – Part 2	5
Computer Vision	6

UNIT 5

NATURAL LANGUAGE PROCESSING

PART 2

STUDY GOALS

- Identify the typical tasks in NLP.

- Understand how to vectorize data, including
  - Bag-of-Words
  - Neural word vectorization techniques
  - Neural sentence vectorization techniques

EXPLAIN SIMPLY

1. What are the typical tasks in NLP?

2. How does Bag-of-Words work?

3. How can words and sentences be vectorized using neural models?

NLP TASKS

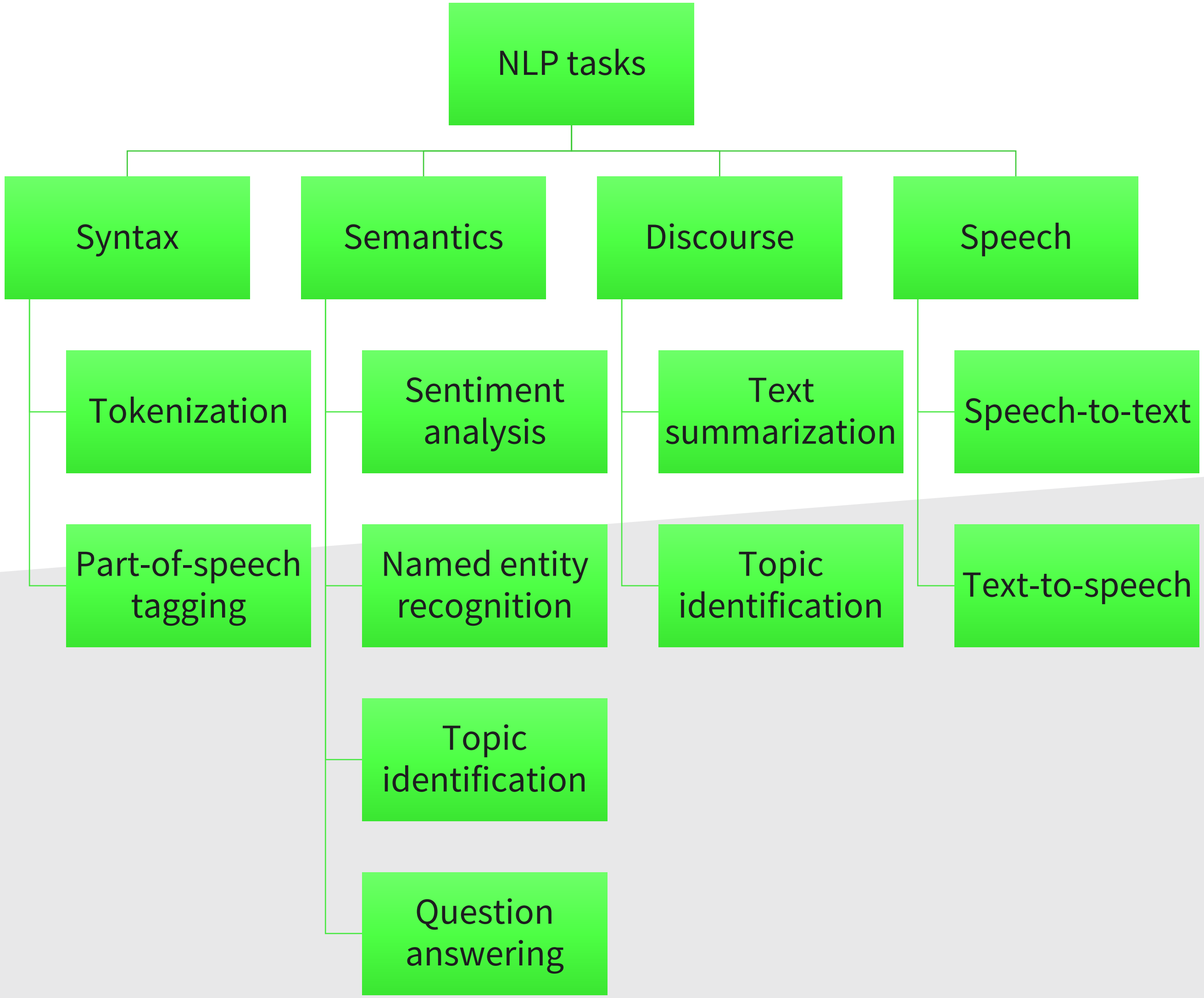


Image Source: Custom Depiction

VECTORIZING DATA – BAG-OF-WORDS

Darren loves dogs.

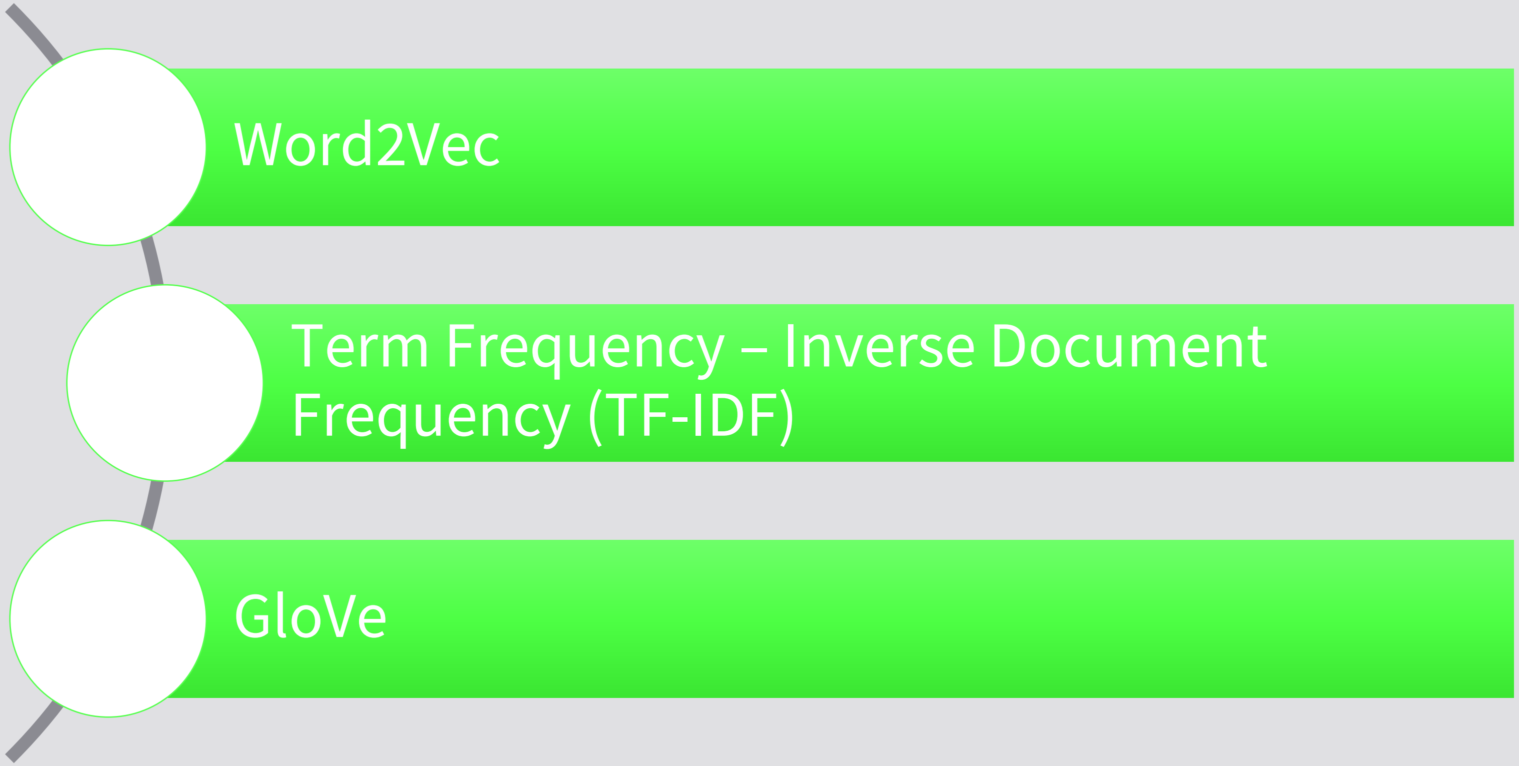
Darren does not like cats.

Cats are not like dogs.

Darren, loves, dogs, does, not, like, cats, are

[2, 1, 2, 1, 2, 2, 2, 1]

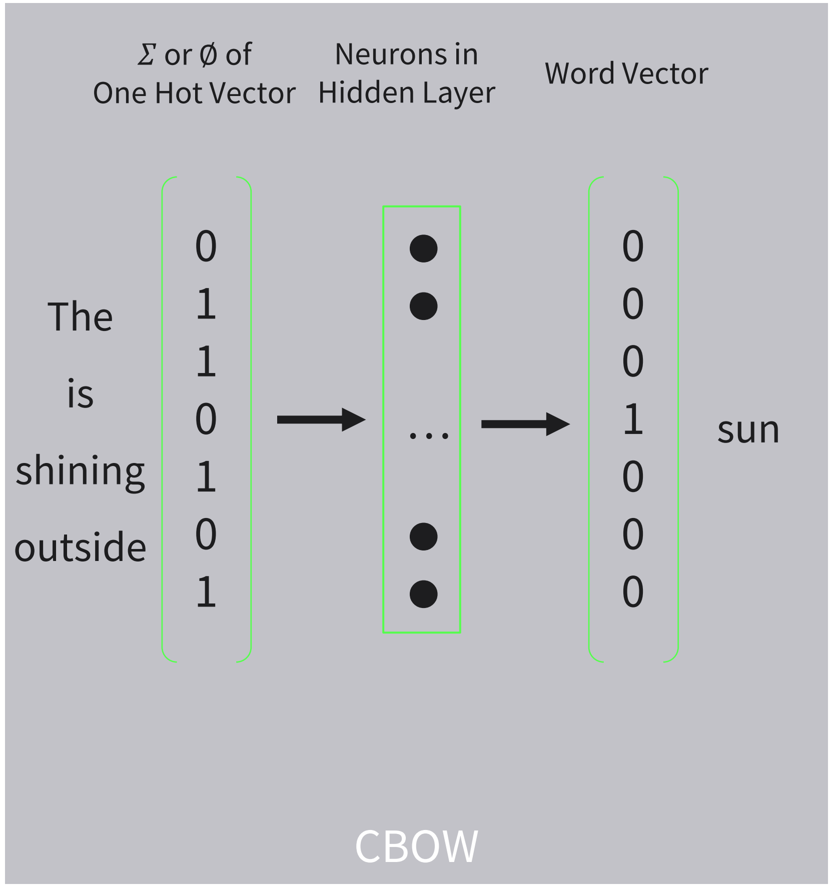
VECTORIZING DATA – WORD VECTORS



- Word2Vec
- Term Frequency – Inverse Document Frequency (TF-IDF)
- GloVe

Image Source: Custom Depiction

WORD2VEC – CBOW VS. SKIP GRAM



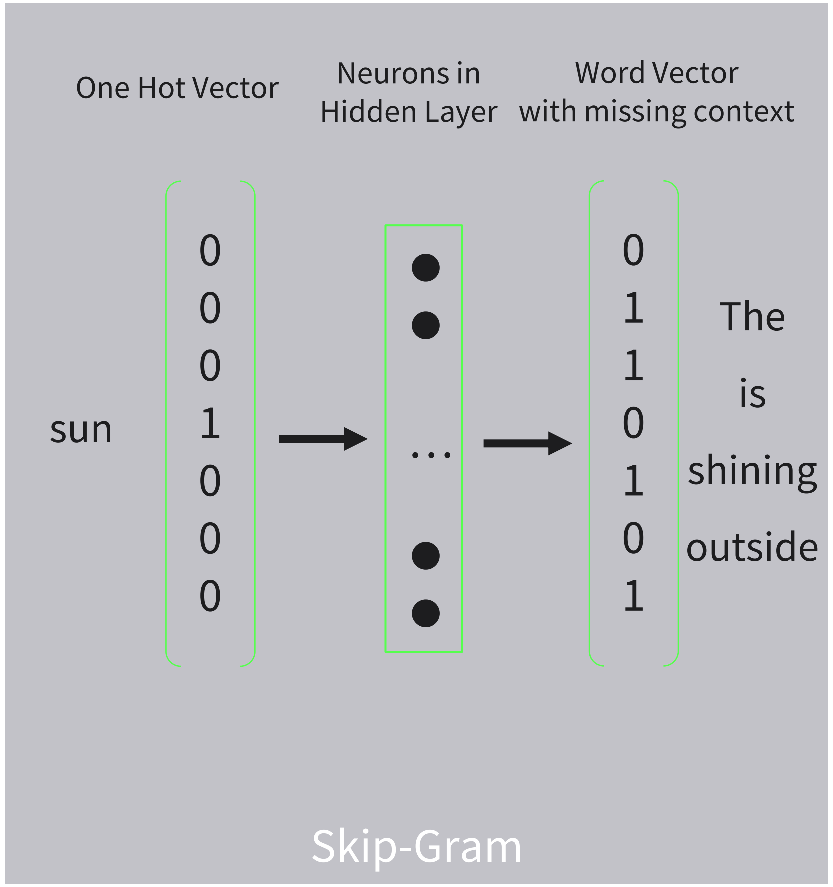


Image Source: Custom Depiction

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY

1

$$TF(t,d)$$

$$= \frac{number\ of\ occurences\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

2

$$DF(t,d,D)$$

$$= \frac{number\ of\ documents\ d\ containing\ t}{total\ number\ of\ documents\ D}$$

3

$$IDF(t)$$

$$= \log \frac{1}{DF(t,d,D)}$$

4

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

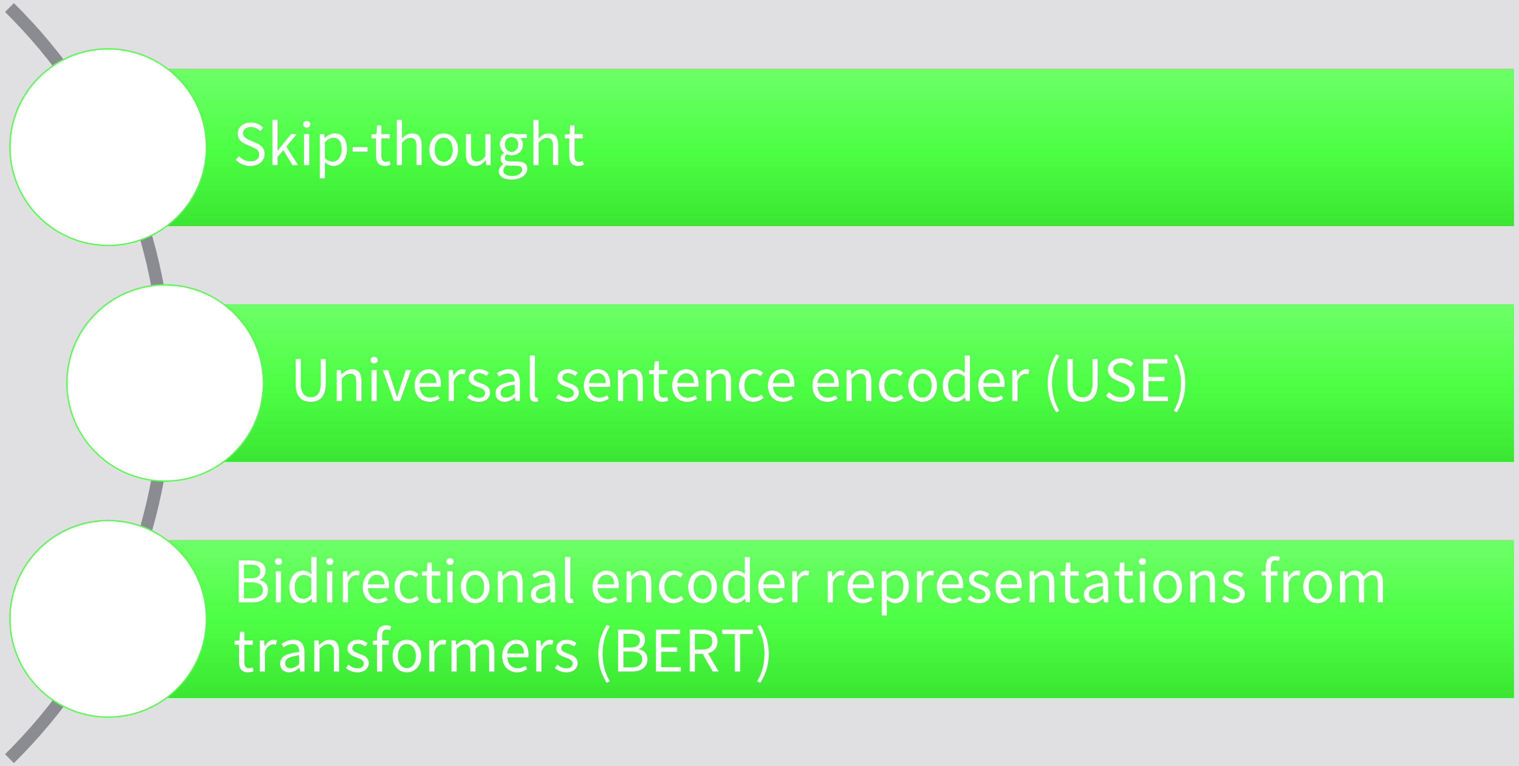
GLOVE – CO-OCCURENCE MATRIX

Darren does not like cats.

	Darren	Does	Not	Like	Cats
Darren	0	1	0	0	0
Does	1	0	1	0	0
Not	0	1	0	1	0
Like	0	0	1	0	1
Cats	0	0	0	1	0

Co-occurence matrix, window size = 1

VECTORIZING DATA – SENTENCE VECTORS



Skip-thought

Universal sentence encoder (USE)

Bidirectional encoder representations from transformers (BERT)

Image Source: Custom Depiction

REVIEW STUDY GOALS

- Identify the typical tasks in NLP.
- Understand how to vectorize data, including
  - Bag-of-Words
  - Neural word vectorization techniques
  - Neural sentence vectorization techniques

REFERENCE

- https://edumunozsala.github.io/BlogEms/jupyter/nlp/classification/embeddings/python/2020/08/15/Intro_NLP_WordEmbeddings_Classification.html
- https://medium.com/analytics-vidhya/basics-of-using-pre-trained-glove-vectors-in-python-d38905f356db

SESSION 5

TRANSFER TASK

TRANSFER TASK

1. Use the Bag-of-Words (BoW) approach to convert the following sentence into the corresponding vector representation:

John is taller than Mary and Mary is taller than Joe.

Now think about the question “Is John taller than Joe?” and discuss the shortcomings of the BoW approach.

TRANSFER TASK

2. In 10 documents, the words **NLP**, **study**, and **cat** have the following frequencies:

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	12	5	0	0	3	2	8	1	0	0
Study	1	0	7	1	0	0	2	0	5	12
Cat	0	12	0	6	8	1	3	10	0	9

Assume, that the D1-D5 contain 20 words. D6-D10 contain 100 words each.
Compute the TF-IDF for each term.

Which document will be returned if somebody wants to study something other than NLP?
Which document contains the most information about cats?

TRANSFER TASKS

Go back to the GloVe example sentence “Darren does not like cats.” How would the co-occurence matrix change for a window size of 2?

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. Name the four categories of NLP tasks.

2. How is the meaning of a text represented using the BoW model?

3. Name three methods for word vectorization.