LECTURER: Nghia Duong-Trung

ARTIFICIAL INTELLIGENCE

TOPIC OUTLINE

- Course book: Artificial Intelligence_DLBDSEAIS01, provided by IU, myStudies
- Reading list provided by IU, myStudies
- The amount of slides content is based on the course book.
- Additional teaching materials:

https://github.com/duongtrung/IU-ArtificialIntelligenceCourse

DISCLAIMER

- This is the modified version of the IU slides.
- I used it for my lectures at IU only.

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

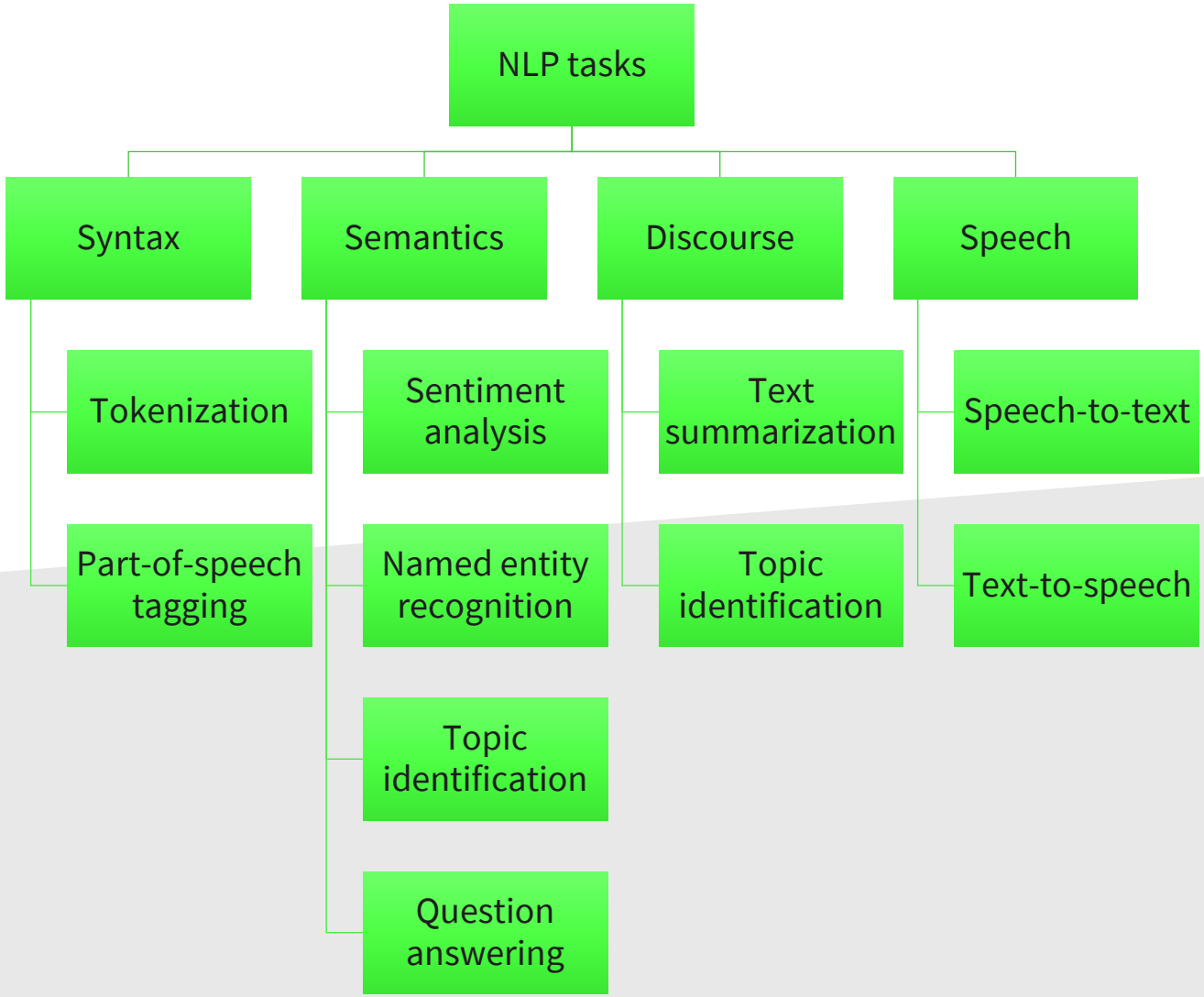


Image Source: Custom Depiction

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

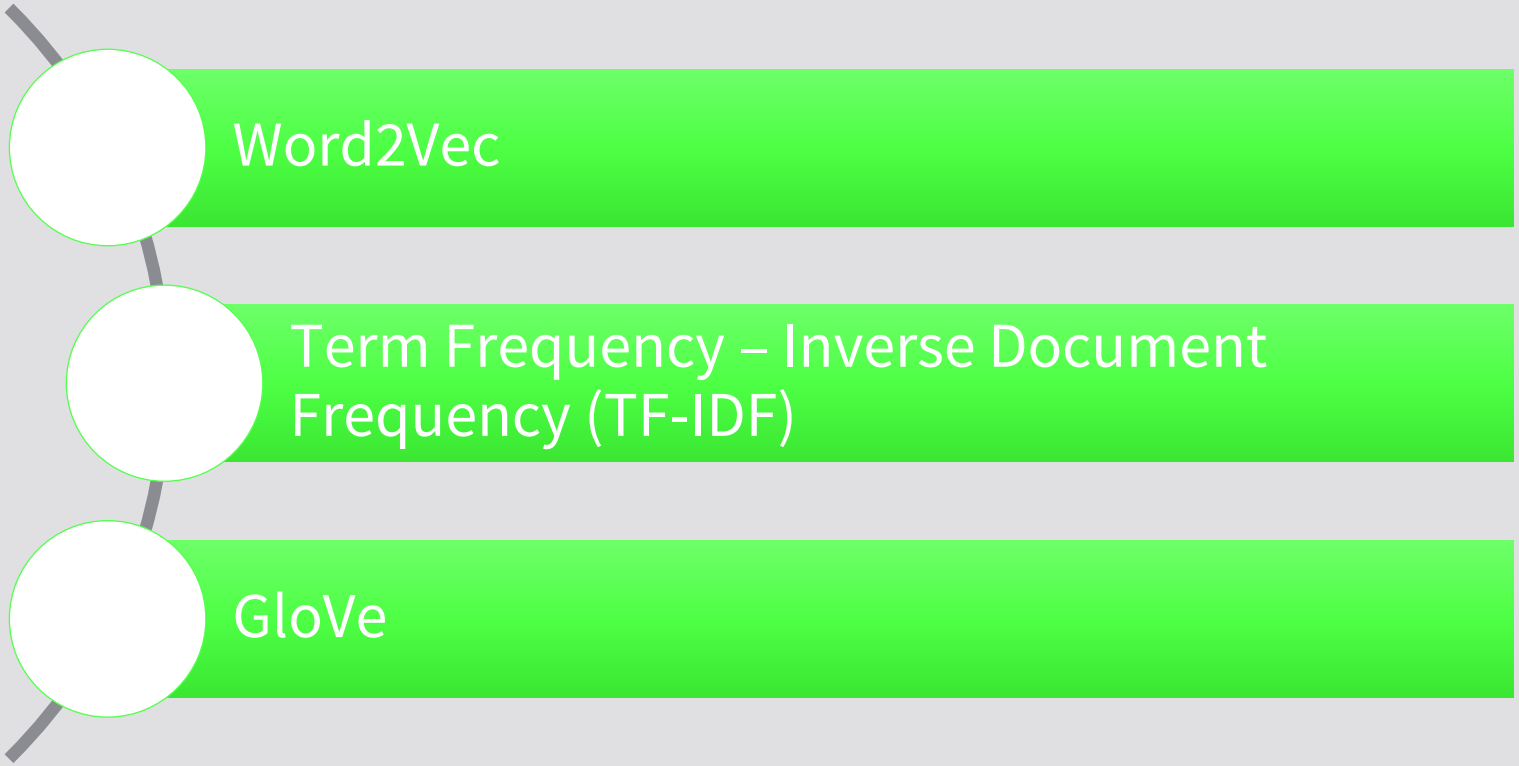
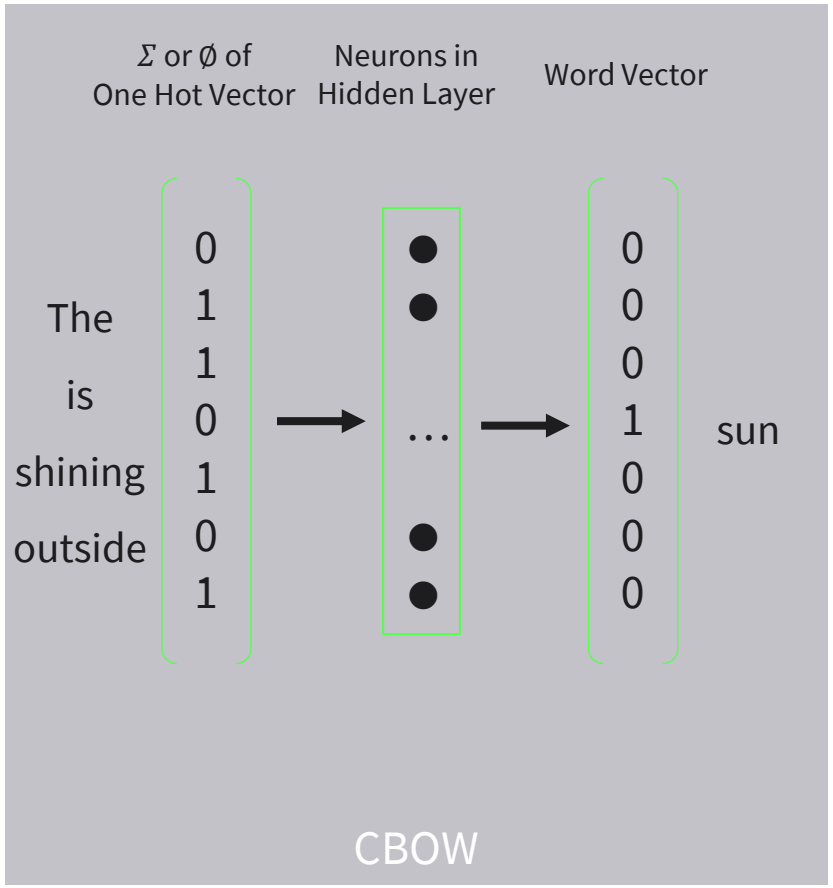


Image Source: Custom Depiction

WORD2VEC – CBOW VS. SKIP GRAM



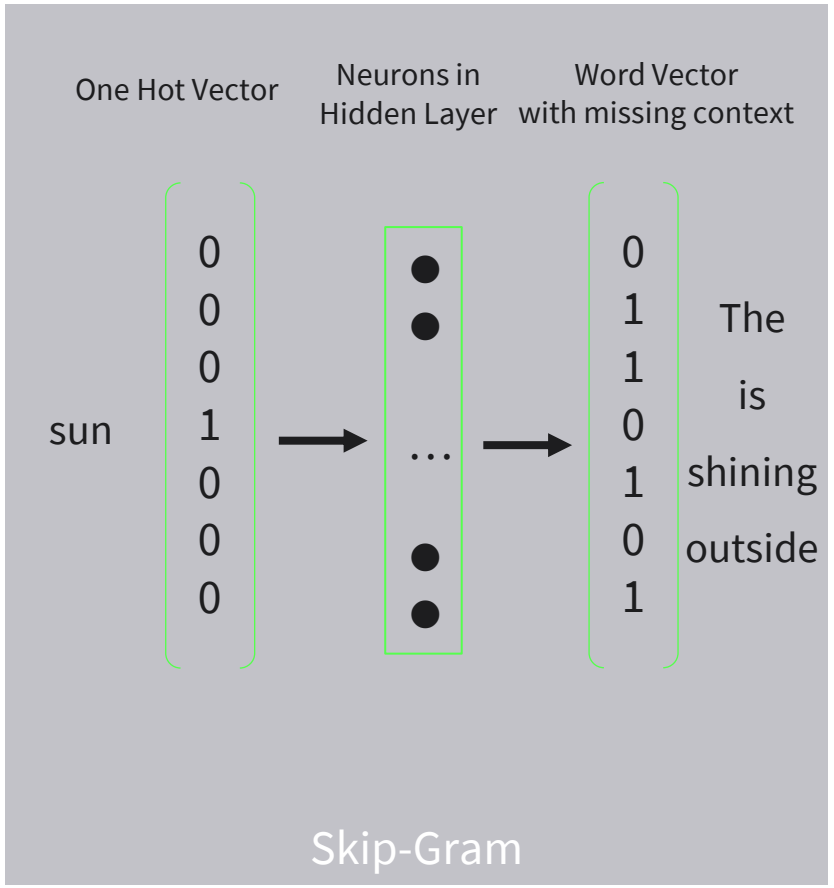


Image Source: Custom Depiction

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

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

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

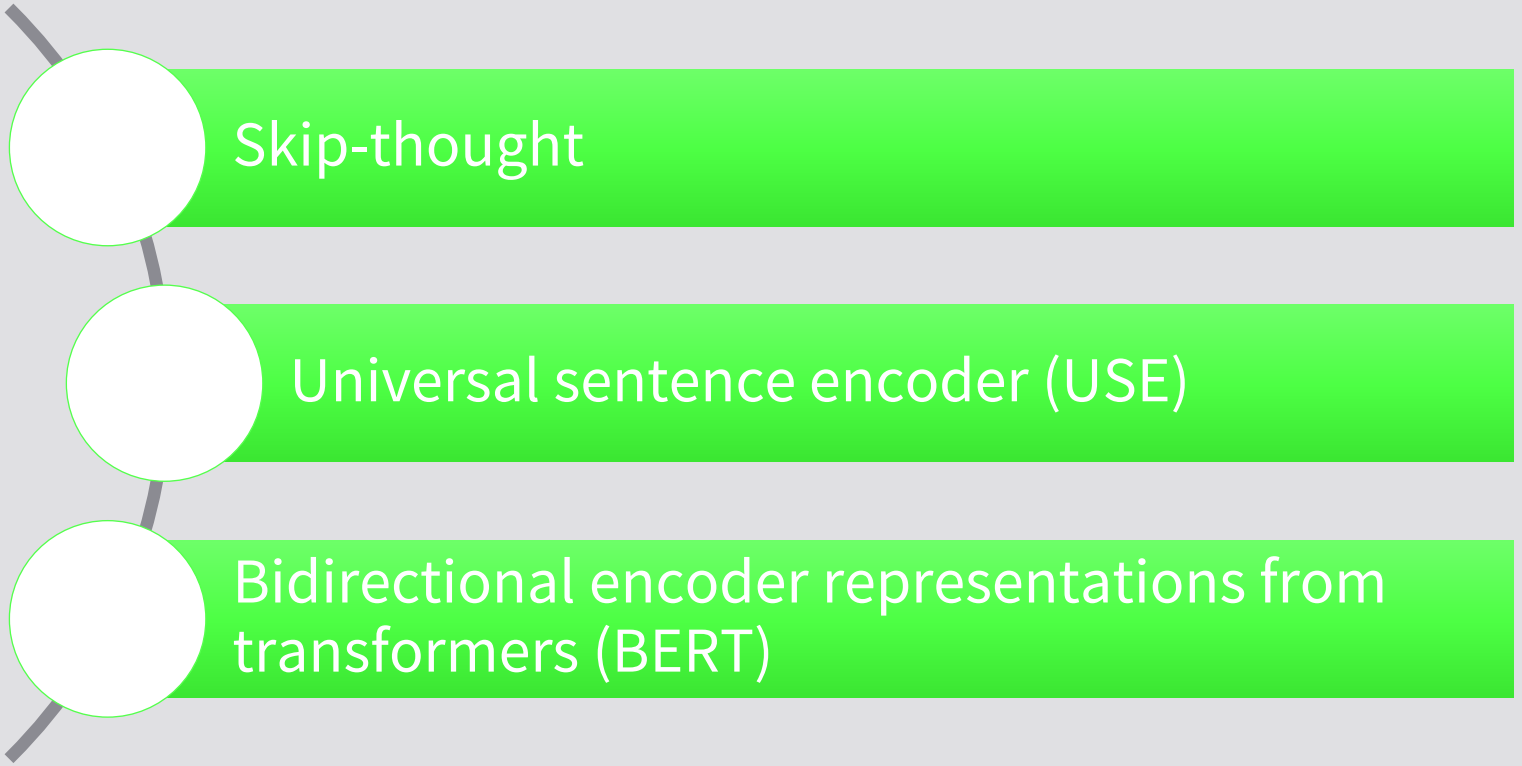


Image Source: Custom Depiction

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?

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?

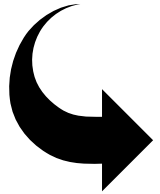
TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

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

[John, is, taller, than, Mary, and, Joe]



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

The question if Joe is taller than John can not be answered, as the structure of the sentence gets lost.

2. Term frequencies

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	0.6	0.25	0	0	0.15	0.02	0.08	0.01	0	0
Study	0.05	0	0.35	0.05	0	0	0.02	0	0.05	0.12
Cat	0	0.6	0	0.3	0.4	0.01	0.03	0.1	0	0.09

2. Document frequencies

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

→ DF(NLP, 6, 10) = 0.6

→ DF(Study, 6, 10) = 0.6

→ DF(Cat, 7, 10) = 0.7

→ IDF(NLP) = 0.737

→ IDF(Study) = 0.737

→ IDF(Cat) = 0.515

2. TF-IDF

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
NLP	0.442	0.184	0.000	0.000	0.111	0.015	0.059	0.007	0.000	0.000
Study	0.037	0.000	0.258	0.037	0.000	0.000	0.015	0.000	0.037	0.088
Cat	0.000	0.309	0.000	0.154	0.206	0.005	0.015	0.051	0.000	0.046

Studying something other than NLP: D3

Information about cats: D2

Darren does not like cats

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

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.

Solutions

1. Speech, discourse, syntax, semantics
2. The meaning gets lost
3. Word2Vec, TD-IDF, GloVe