Foundations of Natural Language Processing

Yansong Feng
fengyansong@pku.edu.cn

Wangxuan Institute of Computer Technology
Peking University
https://yansongfeng.github.io/

February 18, 2025

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- website: course.pku.edu.cn

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- website: course.pku.edu.cn

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- Exams: Mid-term (written)
- website: course.pku.edu.cn

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- Exams: Mid-term (written)
- website: course.pku.edu.cn

Tips

- Coding: You have to be able to code with python.

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- Exams: Mid-term (written)
- website: course.pku.edu.cn

Tips

- Coding: You have to be able to code with python. Otherwise, learn it from today!

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- Exams: Mid-term (written)
- website: course.pku.edu.cn

Tips

- Coding: You have to be able to code with python. Otherwise, learn it from today! With the help of Large Language Models.

Welcome to FNLP@PKU

This course is designed for runs in the second semester each year

- Instructor: Yansong Feng
- Teaching Assistants: Chen Zhang, Mingxu Tao, Zirui Wu, Jiuheng Lin, Kangcheng Luo
- Lectures: 10:10AM Wednesday/Thursday, 203, Sci. Bld.
- Office Hours: By appointment
- Language: Chinese (Notes in English)
- Assignments: 1 short report
- Labs: solve about 3 tasks individually
- Warning: No Plagiarism!
- Exams: Mid-term (written)
- website: course.pku.edu.cn

Tips

- Coding: You have to be able to code with python. Otherwise, learn it from today! With the help of Large Language Models.
- NNs: Not necessary, but would be superb if you already know

Scoring

- Assignments:
  - a short report on challenging LLMs (instructions available soon)
- Labs: implement straightforward methods to solve tasks
  - simple text classifiers
  - slightly complex tasks with combinations of classifers
  - Solving tasks with the help fo LLMs
- Midterm Exam:
  - Written Examination
- In-class Participation:
  - quizzes, in-class discussions

$$\text{A+L} : \text{ME} : \text{IP} \approx 60 : 30 : 10$$

Scoring

- Assignments:
  - a short report on challenging LLMs (instructions available soon)
- Labs: implement straightforward methods to solve tasks
  - simple text classifiers
  - slightly complex tasks with combinations of classifers
  - Solving tasks with the help fo LLMs
- Midterm Exam:
  - Written Examination
- In-class Participation:
  - quizzes, in-class discussions
  - bonus
    - voluntary presentations ($\sim$5 mins) if you want to share with us
    - $+1\sim3\%$

$$\text{A+L} : \text{ME} : \text{IP} \approx 60 : 30 : 10$$

References

- Book: Speech and Language Processing. Dan Jurafsky and James H. Martin, 3rd Edition, https://web.stanford.edu/ jurafsky/slp3/
- Papers: Research papers, Lecture notes, etc

More Resources

- Jurafasky and Manning: https://www.coursera.org/course/nlp
- M. Collins: https://www.coursera.org/course/nlangp
- ACL Anthology: http://aclweb.org/anthology/
- arXiv: http://arxiv.org/list/cs.CL/recent
- Conferences/Journals to Look at: ACL, NAACL, EMNLP, CL, TACL, AIJ, JAIR, T-PAMI, COLING, EACL, NeurIPS, ICML, AAAI, IJCAI...

References

- Book: Speech and Language Processing. Dan Jurafsky and James H. Martin, 3rd Edition, https://web.stanford.edu/ jurafsky/slp3/
- Papers: Research papers, Lecture notes, etc

More Resources

- Jurafasky and Manning: https://www.coursera.org/course/nlp
- NLP with DL: http://cs224n.stanford.edu/
- ACL Anthology: http://aclweb.org/anthology/
- arXiv: http://arxiv.org/list/cs.CL/recent
- Conferences/Journals to Look at: ACL, NAACL, EMNLP, CL, TACL, AIJ, JAIR, T-PAMI, COLING, EACL, NeurIPS, ICML, AAAI, IJCAI...

Questions

- How many of you guys have taken **Probability and Statistics**?
  - probability, joint probability, conditional probability, distribution, ...
  - Bayes, Gaussian, Bernoulli, ...
  - prior, posterior, likelihood, expectation, ...
  - .....

Questions

- How many of you guys have taken **Probability and Statistics**?
  - probability, joint probability, conditional probability, distribution, ...
  - Bayes, Gaussian, Bernoulli, ...
  - prior, posterior, likelihood, expectation, ...
  - .....

Questions

- How many of you guys have taken **Probability and Statistics**?
  - probability, joint probability, conditional probability, distribution, ...
  - Bayes, Gaussian, Bernoulli, ...
  - prior, posterior, likelihood, expectation, ...
  - .....
- Still remember what you have learned in the past?
  - Subject, Object, Predicate, Grammar, ...
  - .....

Questions

- How many of you guys have taken **Probability and Statistics**?
  - probability, joint probability, conditional probability, distribution, ...
  - Bayes, Gaussian, Bernoulli, ...
  - prior, posterior, likelihood, expectation, ...
  - .....
- Still remember what you have learned in the past?
  - Subject, Object, Predicate, Grammar, ...
  - .....
- How many of you guys have used **neural networks models** before?
  - TensorFlow, PyTorch, ...
  - numpy, scikit-learn, ...

FNLP v.s. EMNLP

FNLP

- for sophomore

EMNLP

- for junior/senior students

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding

EMNLP

- for junior/senior students

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding
- **what and why**

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding
- **what and why**

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data
- what and how

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding
- **what and why**
- a bit harsh assessment
  - assignments
  - labs/projects
  - exams

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data
- what and how

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding
- **what and why**
- a bit harsh assessment
  - assignments
  - labs/projects
  - exams

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data
- what and how
- more flexible assessment
  - assignments
  - presentations
  - projects

FNLP v.s. EMNLP

FNLP

- for sophomore
- linguistic and algorithmic foundations of NLP
  - phenomena, concepts
  - empirical models, algorithms
  - corpora, applications
  - conceptual understanding
- **what and why**
- a bit harsh assessment
  - assignments
  - labs/projects
  - exams

EMNLP

- for junior/senior students
- formal models/algorithms for NLU/NLG
  - concepts, tasks
  - SOTA models, algorithms
  - corpora, applications
  - play with data
- what and how
- more flexible assessment
  - assignments
  - presentations
  - projects

Overlap in syllabus but with different expectations

Syllabus

- Lexical: Word Sense Disambiguation : Classification
- Lexical: Sequence Models: POS Tagging, Word Segmentation, NER
- Semantics: Word Representations
- Language Modeling: Smoothing, back-off, Neural LM and beyond
- Neural solutions - 1
- Information Retrieval
- Syntactic: Tree Structures and Grammars: CFG, PCFG, Dependency Parsing
- Semantics: Compositional Semantics
- Applications and Neural solutions - 2
- Neural solutions - 3 (Large Language Models)
- Special issues: bias, ethics, safety, etc.
- ...

A New Syllabus

+ What can we do for Large Language Models?

- Lexical: Word Sense Disambiguation : Classification
- Lexical: Sequence Models: POS Tagging, Word Segmentation, NER
- Semantics: Word Representations
- Language Modeling: Smoothing, back-off, Neural LM and beyond
- Neural solutions - 1
- Information Retrieval
- Syntactic: Tree Structures and Grammars: CFG, PCFG, Dependency Parsing
- Semantics: Compositional Semantics
- Applications and Neural solutions - 2
- Neural solutions - 3 (Large Language Models)
- Special issues: bias, ethics, safety, etc.
- ...

A New Syllabus

- Lexical: Word Sense Disambiguation : Classification
- assignment 1
- Lexical: Sequence Models: POS Tagging, Word Segmentation, NER
- lab 1
- Semantics: Word Representations
- Language Modeling: Smoothing, back-off, Neural LM and beyond
- Neural solutions - 1
- Information Retrieval
- lab 2
- Syntactic: Tree Structures and Grammars: CFG, PCFG, Dependency Parsing
- Semantics: Compositional Semantics
- *mid-term* exam
- Applications and Neural solutions - 2
- lab 3
- Neural solutions - 3 (Large Language Models)
- Special issues: bias, ethics, safety, etc.