Natural Language Processing and Text Mining

Guide to AI and Data Courses · Part II: Branches of AI

Text preprocessing | Text representation | Language models | Transformers | Text mining | Neighboring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides:
https://sercankulcu.github.io/files/ai_courses/slides/Chapter_06_Natural_Language_Processing.html

INTRODUCTION

What Question Does It Ask?

> **How does a computer analyze, understand and generate human language?**

Natural language processing studies how computers analyze, understand and generate human language; text mining extracts information and patterns from large collections of text. At the heart of the course are **text representation**, which turns text into numbers, and language models that predict the next word.

At a glance

APPROACH linguistics, statistics and deep learning	**INPUT** text	**OUTPUT** label, translation, summary, answer
SUCCESS MEASURE accuracy, BLEU, human evaluation	**MATH LOAD** medium	**DATA NEEDS** large text corpora

A Short History

1954
Georgetown-IBM experiment: more than sixty Russian sentences are translated into English automatically.

1966
Weizenbaum's ELIZA imitates a therapist with simple pattern matching.

1990s
Statistical methods replace rule-based systems; IBM's translation models lead the way.

2013
Mikolov et al.'s word2vec embeds words as vectors that reflect meaning.

2017
The transformer architecture is introduced, followed by pretrained models such as BERT and GPT.

2022
Large language models reach a mass audience through chat interfaces.

THE COURSE

Core Concepts

Tokenization
Splitting text into word or subword units.

Morphology
Root and suffix analysis; critical in agglutinative languages like Turkish.

Bag of words
Representing text by word counts, ignoring order.

TF-IDF
High weight for words frequent in a document but rare in the corpus.

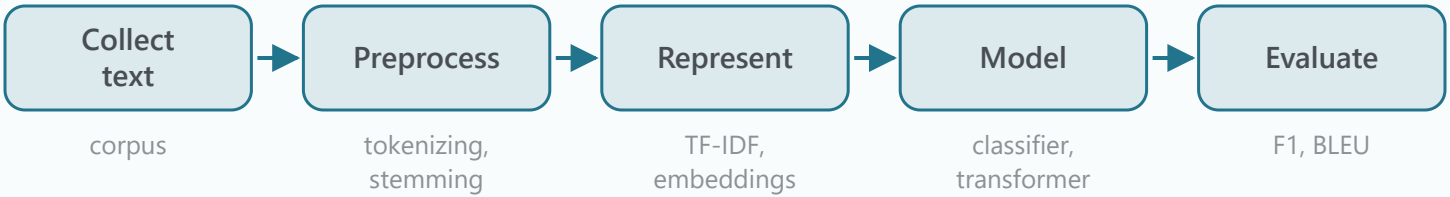
Word embeddings
Words with similar meanings lie close in vector space.

Language model
A model that predicts the probability of the next word.

Named entity recognition
Finding names of people, places and organizations in text.

Topic modeling
Finding hidden topics in documents with methods such as LDA.

A Typical Workflow



Example: An online store collects product reviews, tokenizes the text and analyzes suffixes, turns the reviews into vectors with a pretrained transformer, trains a positive/negative sentiment classifier and measures it with the F1 score. The result shows which topics the complaints cluster around.

THE COURSE

What Does the Course Cover?

1. Regular expressions, tokenization, normalization
2. Morphology: stemming and morphological analysis
3. N-gram language models and smoothing
4. Text classification: naive Bayes, logistic regression
5. Vector semantics: TF-IDF and word2vec
6. Part-of-speech tagging and hidden Markov models
7. Named entity recognition and sequence labeling (CRF)
8. Syntactic parsing: dependency trees
9. Language modeling with recurrent networks
10. Transformers, BERT and pretraining
11. Machine translation and evaluation with BLEU
12. Topic modeling (LDA) and information extraction

Skills you will gain

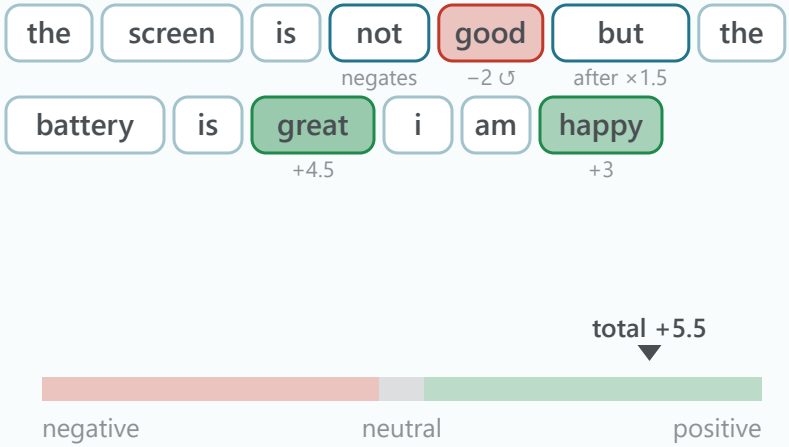
text processing with spaCy and NLTK · cleaning text with regular expressions · Turkish morphology with Zemberek · word embeddings with Gensim · fine-tuning a Hugging Face model · evaluating with F1 and BLEU

A typical semester; the order and coverage vary between universities.

TRY IT

Is the Customer Review Positive?

An online store gets thousands of reviews a day; negative ones should go straight to customer service. Pick a sample review or write your own; the model scores the words one by one.



5. The screen is not good but the bat...

The screen is not good but the battery is great, I

Apply negation (“not”, “n't”)

13 words, 3 of them carry sentiment.
Total score **+5.5** → decision: **positive**

The computer first splits the text into **words**, then gives each word a weight. A word like “not” or “didn't” flips the meaning of what follows; language processing has to teach the model this structure.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Artificial Intelligence	language understanding and generation
Programming	string handling, file reading and regular expressions in Python
Data Structures	dictionaries, hash tables and tries: for storing a vocabulary

CONTEXT

Are You Ready for This Course?

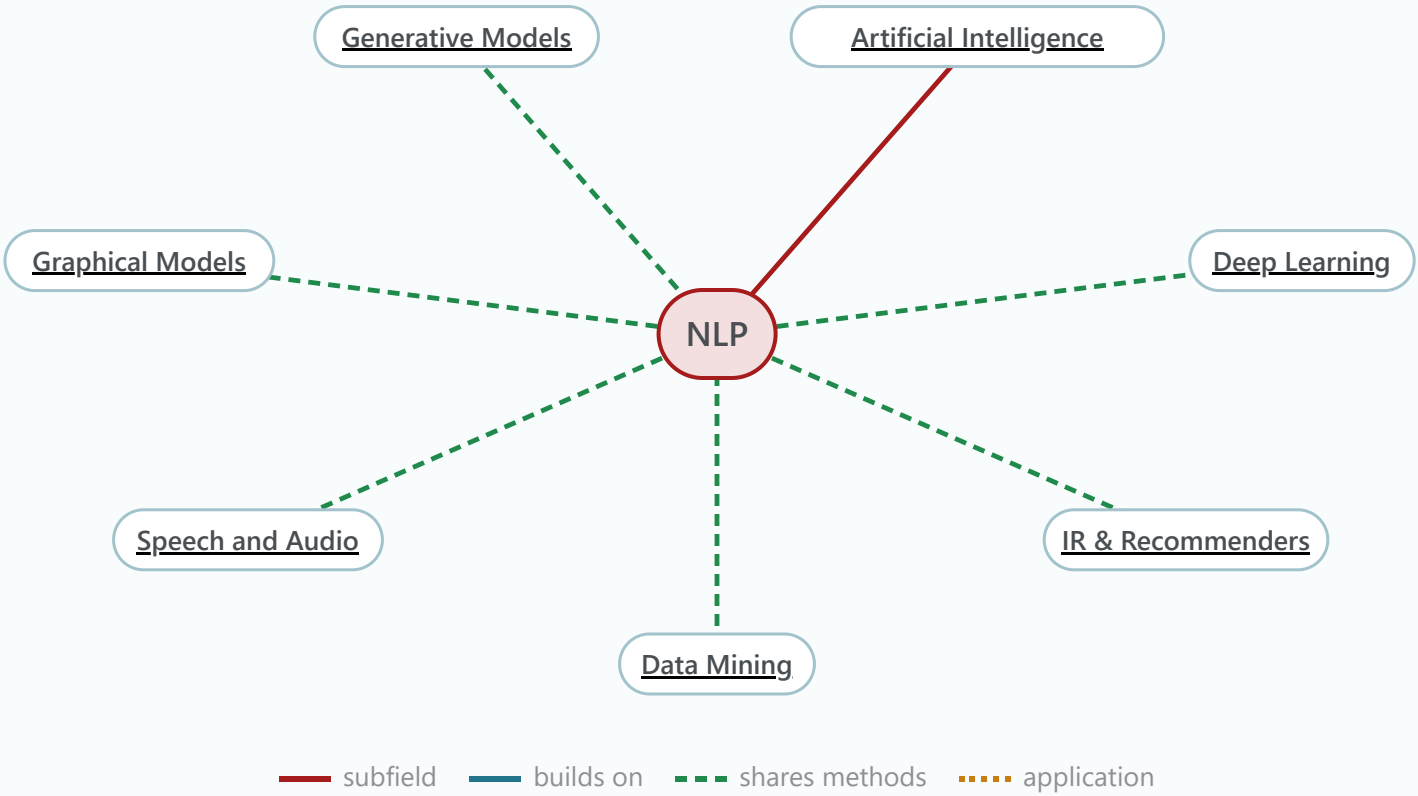
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you process strings in Python and search text with a regex?	An intro programming course, then a short regex tutorial
Can you count word frequencies using a dictionary (hash table)?	Hash tables from a data structures course
Can you explain conditional probability and Bayes' rule?	Probability and Statistics
Can you evaluate a classifier with precision, recall and F1?	Machine Learning

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



Overlapping Courses

Relation	Course	Shared ground
subfield	This course is a subfield of Artificial Intelligence	language understanding and generation
shares methods	Shares methods with Deep Learning	word embeddings, transformer architecture
shares methods	Shares methods with Probabilistic Graphical Models	hidden Markov model, conditional random field
shares methods	Shares methods with Generative Models	text generation
shares methods	Shares methods with Speech Recognition and Audio Processing	language model, speech to text
shares methods	Shares methods with Data Mining	text classification, topic modeling
shares methods	Shares methods with Information Retrieval and Recommender Systems	text representation, TF-IDF, semantic search

DIFFERENCES

Where It Differs (1/2)

Criterion	NLP	Speech and Audio	IR & Recommenders
Core question	How does a computer analyze, understand and generate human language?	How are words, speakers and events extracted from a sound wave?	How are the most relevant documents or products picked from millions of candidates?
Approach	linguistics, statistics and deep learning	acoustic signal and language model	matching and ranking
Input	text	audio waveform	query, documents, user history
Output	label, translation, summary, answer	text, speaker, sound event	ranked result list
Success measure	accuracy, BLEU, human evaluation	word error rate (WER)	precision, recall, NDCG
Math load	medium	medium-high	medium: vector spaces
Data needs	large text corpora	many hours of recordings	lots of interaction data

- **Speech and Audio:** Speech recognition goes from the sound wave to text; NLP analyzes meaning once the text is available.
- **IR & Recommenders:** Information retrieval fetches the documents most relevant to a query; NLP analyzes and generates the language inside them in detail.

DIFFERENCES

Where It Differs (2/2)

Criterion	NLP	Data Mining	Generative Models
Core question	How does a computer analyze, understand and generate human language?	How are hidden, useful patterns found in large piles of data?	How does a model generate new samples that look like real data?
Approach	linguistics, statistics and deep learning	discovery-driven analysis	learning the data distribution
Input	text	large tables or transaction data	huge, mostly unlabeled data
Output	label, translation, summary, answer	rules, clusters, outliers	new samples: images, text, audio
Success measure	accuracy, BLEU, human evaluation	interestingness, support, confidence	realism and diversity (FID etc.)
Math load	medium	medium	high
Data needs	large text corpora	large	very large

- **Data Mining:** Data mining looks for patterns in any kind of data; this course handles the structure specific to language (words, suffixes, sentences).
- **Generative Models:** Generative models produce all kinds of samples, including images and audio; NLP also covers understanding and analysis tasks.

PRACTICE

Applications

Machine translation
Automatic translation between languages.

Sentiment analysis
Finding the attitude in reviews and posts.

Question answering
Finding answers in documents or a knowledge base.

Summarization
Making a long text short and to the point.

Chat assistants
Systems that talk with users in natural language.

Document classification
Sorting news, petitions or emails by topic.

PRACTICE

Risks and Ethics

Hallucination

Language models fluently invent sources that do not exist; in the US, lawyers have filed briefs citing case law a chatbot made up.

Stereotypes in embeddings

Embeddings produce analogies like "man is to programmer as woman is to homemaker"; translation renders neutral pronouns by stereotype.

Low-resource languages

Models work best in English; on Turkish and languages with little data the same system makes noticeably more errors.

Surveillance through text

Emails, messages and posts are personal data under GDPR and similar laws; text mining can also be used to monitor and profile people.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

NLTK | spaCy | Hugging Face Transformers | Gensim | Stanza | Zemberek (Turkish)

Well-known datasets

Penn Treebank | CoNLL-2003 | SQuAD | GLUE | IMDb movie reviews | Wikipedia dumps

PRACTICE

Common Misconceptions

Common belief	In fact
The model understands text like a human.	Language models learn statistical patterns; fluent text is not always correct or consistent.
Any method built for English works for Turkish.	In agglutinative Turkish one root takes hundreds of forms; tokenization and morphology need separate care.
Large language models ended classic NLP.	TF-IDF, regular expressions and small models are still fast, cheap and enough for most jobs.
Text mining is just counting words.	Counting is the start; extracting entities, relations and topics requires language structure.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

NLP engineer | Computational linguist | Data scientist | Search and recommendation engineer | Researcher

After NLP you can move to spoken language with speech recognition, to semantic search with information retrieval, or to the inner workings of large language models with generative models. The Explainable AI and AI Ethics course covers the bias and reliability problems of language models.

NEXT STEPS

Further Reading

- Jurafsky and Martin, *Speech and Language Processing*, 2nd ed., Prentice Hall, 2009.
- Manning and Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
- Bird, Klein and Loper, *Natural Language Processing with Python*, O'Reilly, 2009.
- Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019.
- Vaswani et al., "Attention Is All You Need", *Advances in Neural Information Processing Systems* 30, 2017.

The End

Chapter 6: Natural Language Processing and Text Mining

← Previous chapter · Map of the field · Next chapter →