THINKFUL

Intro: Natural Language Processing

DATA SCIENCE

Warm Up

- What challenges do you anticipate encountering when analyzing text vs. numeric data?

Agenda

- What is Language?
- Language Concepts & Terminology
- What is Natural Language Processing?
- Why Natural Language Processing?
- NLP Concepts & Terminology
- Natural Language Processing Workflow
- NLP Applications
- Python NLP Tools

What Is Language?

- Language is unstructured data that has been produced by people to be understood by other people.
- Refers to both the *medium* through which people communicate with each other and also the *content* of that communication.
- While language is unstructured data, it is not random - it is governed by linguistic properties that make it understandable.
- Language is flexible, ambiguous, and dynamic. This makes it useful for communication but difficult to analyze.

Language Concepts & Terminology

- Lexicon - body of knowledge about a language including the forms, meanings, usage, categories, and relationships of words and phrases.
- Vocabulary - subset of words in a language that are used in a particular context.
- Grammar - system and structure of a language, consisting of syntax and morphology (and sometimes also phonology and semantics).

Language Concepts & Terminology

- Syntax - set of rules, principles, and processes that govern the structure of sentences in a given language.
- Morphology - the study of how words are formed, their structure, and their relationships to other words.
- Semantics - the study of meaning in language.

Language Concepts & Terminology

- Stem - form of a word to which affixes (prefix or suffix) can be attached.
  - Word: puppies
  - Stem: puppi
- Lemma - root or base form of a word.
  - Word: puppies
  - Lemma: puppy
- Part of speech - category to which a word is assigned in accordance with its syntactic functions (noun, verb, adjective, adverb, etc.).

What is Natural Language Processing

- Subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages.
- The computational processing, analysis, and modeling of language encoded as text.
- Fascinating and rapidly-evolving field and a high demand skill set in industry.

Why Natural Language Processing

- Natural language is one of the most untapped forms of data available today.
- Companies have lots of text data (emails, documents, chat logs, employee reviews, etc.).
- The Internet is full of user-generated data, the majority of which is text.
- The amount of text data is growing extremely quickly, and there will be a need for talent knowledgeable about how to process and analyze it.

NLP Concepts & Terminology

- Document - the basic unit of observation in NLP. Documents can be of varying lengths (entire books, chapters within a book, articles, reviews, tweets, etc.).
- Corpus - collection of related documents.
- Token - unit of text analysis, generally words and punctuation.

NLP Concepts & Terminology

- Stop words - words that are frequently encountered but hold relatively unimportant meanings (ex. the, and, I, we, my, etc.).
- Named entity - words (or a group of words) in a sentence that indicate a predefined category of objects (entities). These categories include names of people, organizations, locations, time expressions, quantities, monetary values, percentages, etc.
- Dependency tree - represents the grammatical structure and indicates the relationship between words.

Natural Lang. Processing Workflow

- Text Data Acquisition - getting text data from sources.
- Corpus Building and Storage - organizing and storing the data.
- Preprocessing - cleaning, tokenizing, and tagging the data.
- Text Analysis - exploring and analyzing the data to extract insights.
- Vectorization - engineering features and preparing the data for modeling.
- Modeling - applying machine learning algorithms to text data.
- Operationalization - deploying models and creating language-powered data products.

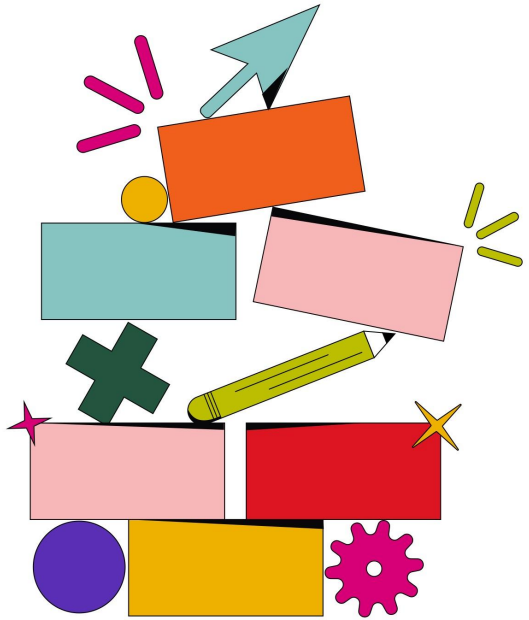
NLP Applications

- Spam Filtering
- Search Results
- Newsfeed Curation
- Sentiment Analysis
- Chatbots
- Digital Assistants (Siri, Alexa, Cortana, etc.)
- Language Detection
- Machine Translation
- Language Generation

Python NLP Tools

- Natural Language Toolkit (NLTK)
- Spacy
- Scikit-Learn
- Gensim
- Stanford CoreNLP
- BeautifulSoup
- Readability
- And more!

Questions?



Summary

- What language is and what important language concepts and terms we will need to know.
- What NLP is, why it's important, and what NLP concepts and terms we will need to know.
- What the steps in the NLP workflow are.
- Examples of real-world NLP applications.
- Overview of Python tools for NLP.

Assignment

Try to answer the following questions without looking up the answers:

1. What properties of language make it difficult to analyze?
2. What is the difference between a lexicon and a grammar?
3. What is the difference between syntax and morphology?
4. What is the difference between a word stem and a lemma?
5. Name 4 examples of parts of speech.
6. What are some examples of named entities?
7. What are the steps in the NLP workflow?
8. Provide 3 examples of real-world NLP applications.

Thank You

Introduction to Natural Language Processing

Warm Up

- What challenges do you anticipate encountering when analyzing text vs. numeric data?

High Level Agenda

- What is Language?
- Language Concepts & Terminology
- What is Natural Language Processing?
- Why Natural Language Processing?
- NLP Concepts & Terminology
- Natural Language Processing Workflow
- NLP Applications
- Python NLP Tools

What is Language?

- Language is unstructured data that has been produced by people to be understood by other people.
- Refers to both the *medium* through which people communicate with each other and also the *content* of that communication.
- While language is unstructured data, it is not random - it is governed by linguistic properties that make it understandable.
- Language is flexible, ambiguous, and dynamic. This makes it useful for communication but difficult to analyze.

Language Concepts & Terminology

- Lexicon - body of knowledge about a language including the forms, meanings, usage, categories, and relationships of words and phrases.

- Vocabulary - subset of words in a language that are used in a particular context.

- Grammar - system and structure of a language, consisting of syntax and morphology (and sometimes also phonology and semantics).

Language Concepts & Terminology

- Syntax - set of rules, principles, and processes that govern the structure of sentences in a given language.

- Morphology - the study of how words are formed, their structure, and their relationships to other words.

- Semantics - the study of meaning in language.

Language Concepts & Terminology

- Stem - form of a word to which affixes (prefix or suffix) can be attached.
  - Word: puppies
  - Stem: puppi

- Lemma - root or base form of a word.
  - Word: puppies
  - Lemma: puppy

- Part of speech - category to which a word is assigned in accordance with its syntactic functions (noun, verb, adjective, adverb, etc.).

What is Natural Language Processing?

- Subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages.

- The computational processing, analysis, and modeling of language encoded as text.

- Fascinating and rapidly-evolving field and a high demand skill set in industry.

Why Natural Language Processing?

- Natural language is one of the most untapped forms of data available today.
- Companies have lots of text data (emails, documents, chat logs, employee reviews, etc.).
- The Internet is full of user-generated data, the majority of which is text.
- The amount of text data is growing extremely quickly, and there will be a need for talent knowledgeable about how to process and analyze it.

NLP Concepts & Terminology

- Document - the basic unit of observation in NLP. Documents can be of varying lengths (entire books, chapters within a book, articles, reviews, tweets, etc.).

- Corpus - collection of related documents.

- Token - unit of text analysis, generally words and punctuation.

NLP Concepts & Terminology

- Stop words - words that are frequently encountered but hold relatively unimportant meanings (ex. the, and, I, we, my, etc.).

- Named entity - words (or a group of words) in a sentence that indicate a predefined category of objects (entities). These categories include names of people, organizations, locations, time expressions, quantities, monetary values, percentages, etc.

- Dependency tree - represents the grammatical structure and indicates the relationship between words.

Natural Language Processing Workflow

1. Text Data Acquisition - getting text data from sources.
2. Corpus Building and Storage - organizing and storing the data.
3. Preprocessing - cleaning, tokenizing, and tagging the data.
4. Text Analysis - exploring and analyzing the data to extract insights.
5. Vectorization - engineering features and preparing the data for modeling.
6. Modeling - applying machine learning algorithms to text data.
7. Operationalization - deploying models and creating language-powered data products.

NLP Applications

- Spam Filtering
- Search Results
- Newsfeed Curation
- Sentiment Analysis
- Chatbots
- Digital Assistants (Siri, Alexa, Cortana, etc.)
- Language Detection
- Machine Translation
- Language Generation

Python NLP Tools

- Natural Language Toolkit (NLTK)
- Spacy
- Scikit-Learn
- Gensim
- Stanford CoreNLP
- BeautifulSoup
- Readability
- And more!

Questions?

Recap

In this session, we covered:

- What language is and what important language concepts and terms we will need to know.
- What NLP is, why it’s important, and what NLP concepts and terms we will need to know.
- What the steps in the NLP workflow are.
- Examples of real-world NLP applications.
- Overview of Python tools for NLP.

Assignment

Try to answer the following questions without looking up the answers:

- What properties of language make it difficult to analyze?
- What is the difference between a lexicon and a grammar?
- What is the difference between syntax and morphology?
- What is the difference between a word stem and a lemma?
- Name 4 examples of parts of speech.
- What are some examples of named entities?
- What are the steps in the NLP workflow?
- Provide 3 examples of real-world NLP applications.