ITMO UNIVERSITY

NLP Basic and Selected Topics

A Practical and Easy Introduction to Selected Topics

Aug 31st, 2019

Overview of the Unit Today

1) **Applications of NLP / Introduction (30min)**

2) Practical Basic NLP (NLTK / pythainlp) (45min)

3) Modern NLP with ML/DL (45min)

4) Example: Word Similarity and WordNet (30min)

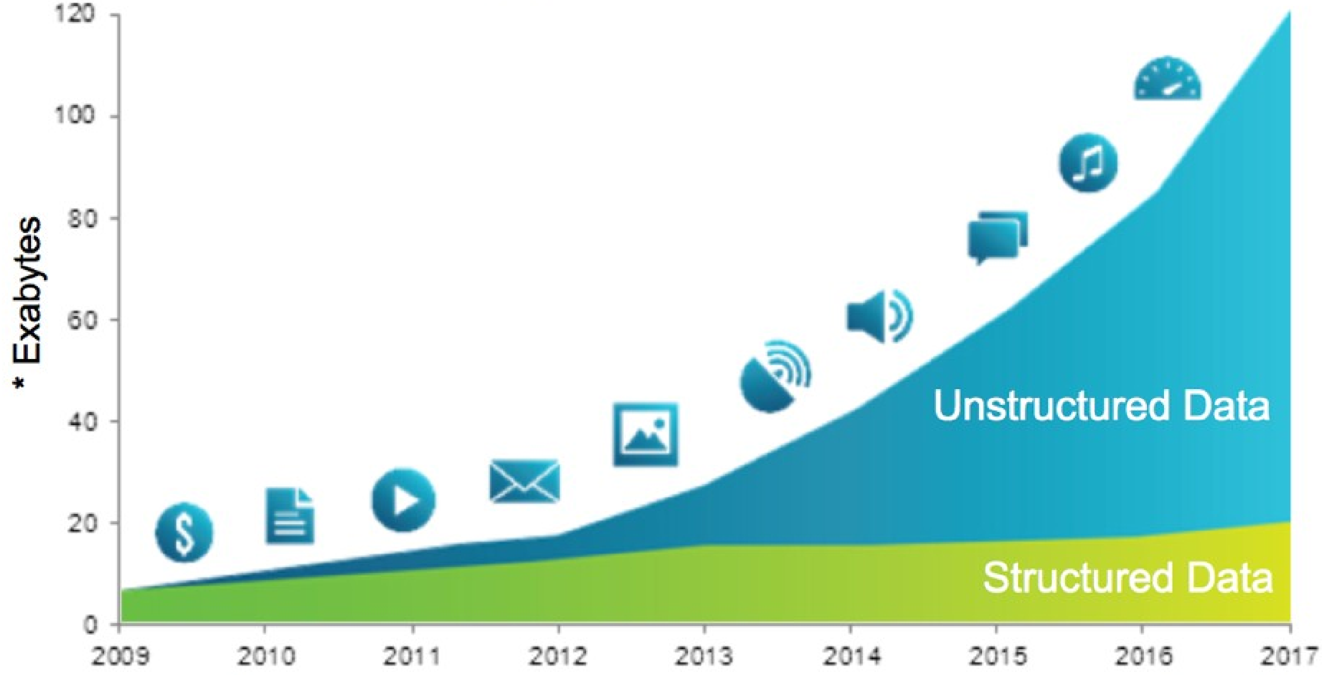
5) Modern NLP with fastAI / flair (30min)

Applications – Motivation. NLP used for …

- ✓ Sentiment Detection
- ✓ Language Modeling
- ✓ Machine Translation
- ✓ Classification (eg. spam detection)
- ✓ …. many other tasks …

Data Growth

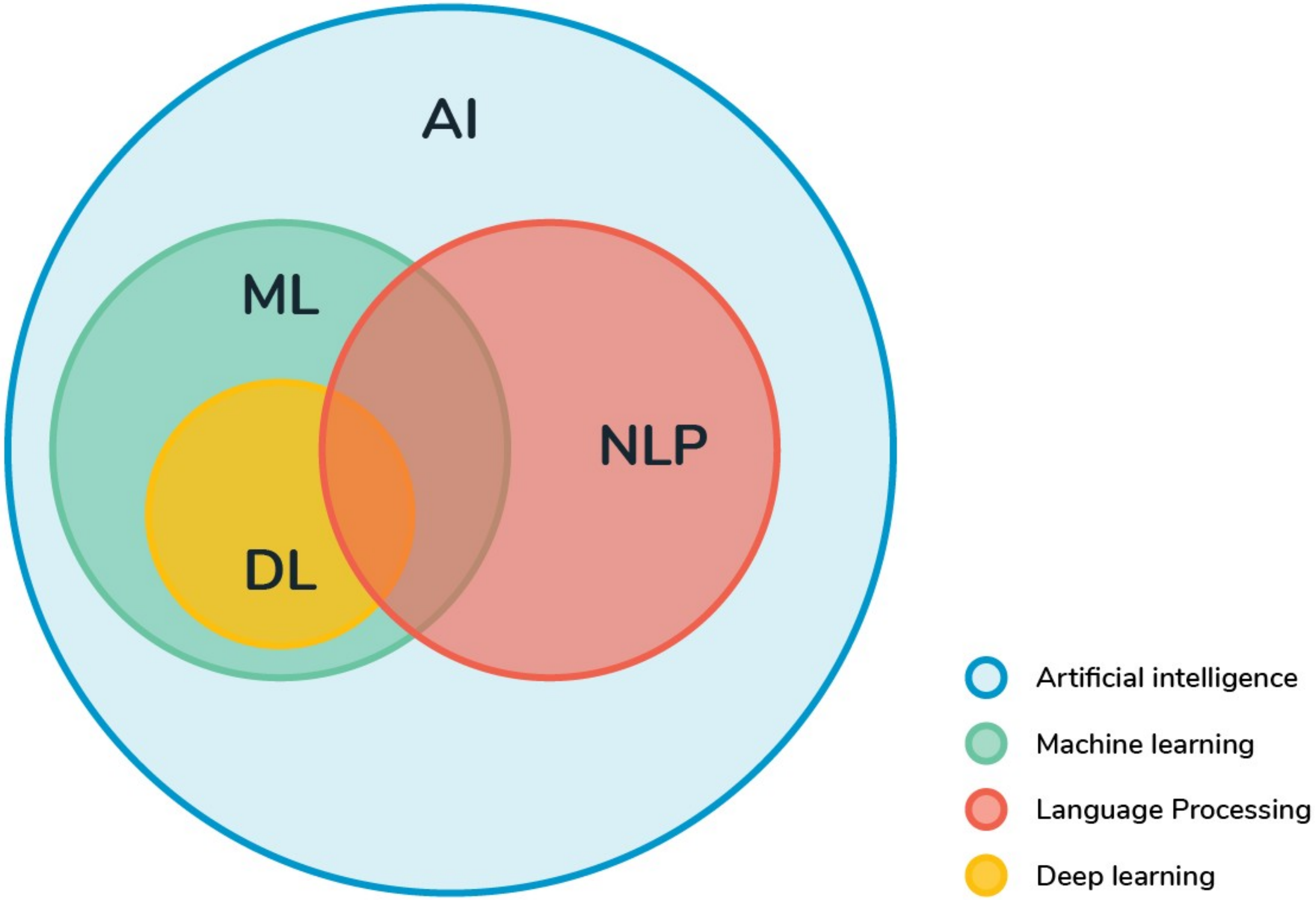
Problem - Traditional and Legacy Storage Designed for Transactional, Not Unstructured Data

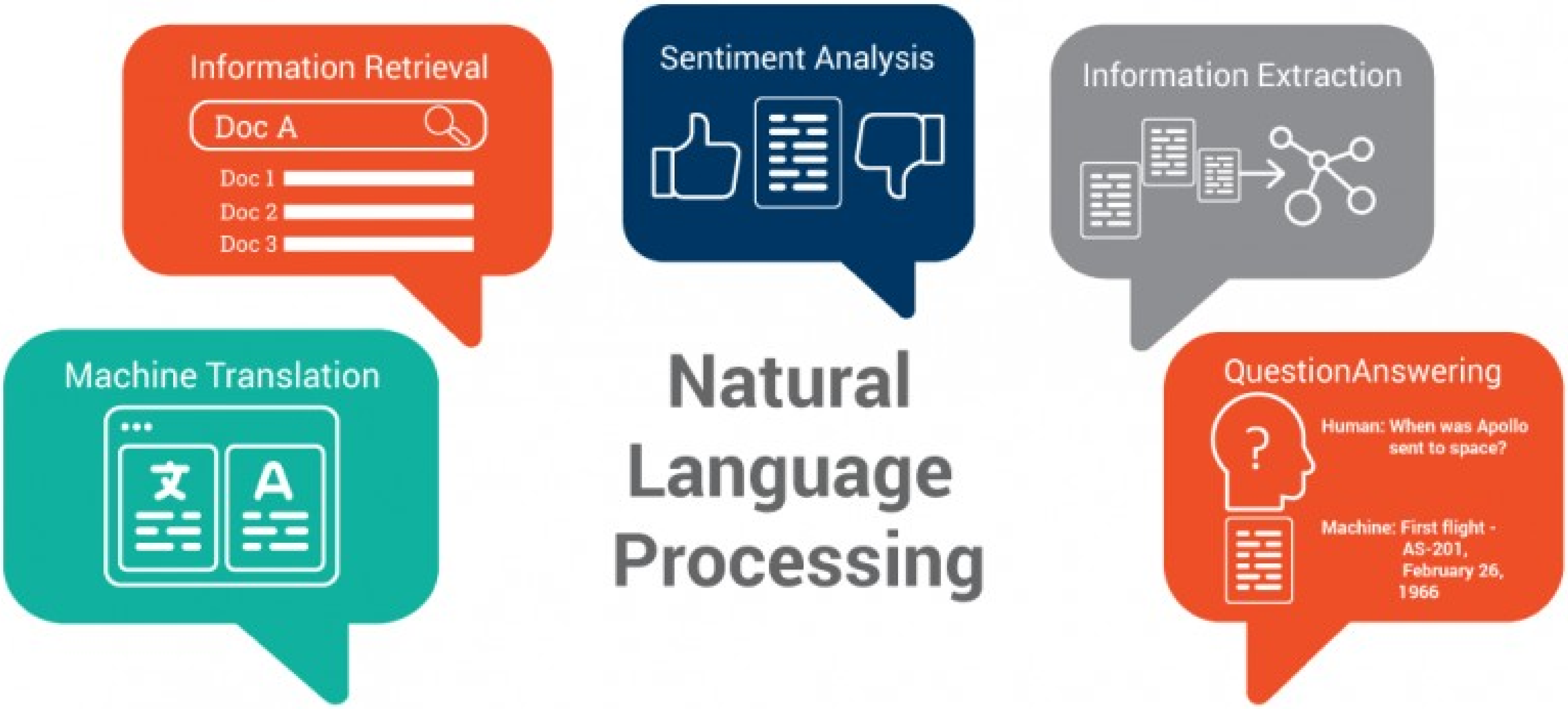


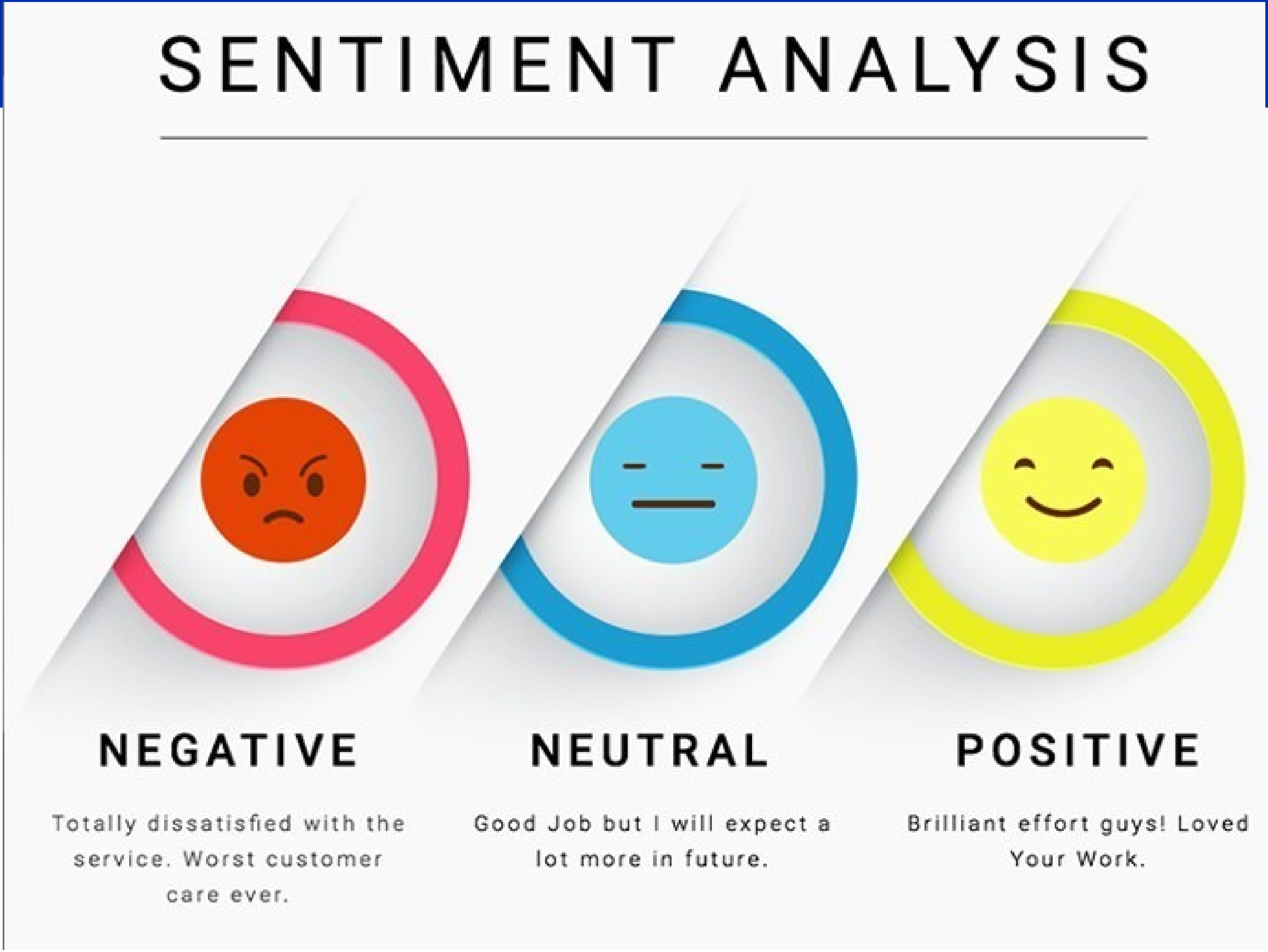
*1 exabyte = 1,000 petabytes =1 million terabytes = 1 billion gigabytes

Source:IDC

- Unstructured data growth of
- 60–80% per year
- creates Web-scale storage needs


AI
ML
NLP
DL
Artificial intelligence
Machine learning
Language Processing
Deep learning


Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
文
A
Natural Language Processing
QuestionAnswering
?
Human: When was Apollo sent to space?
Machine: First flight - AS-201, February 26, 1966



SENTIMENT ANALYSIS

NEGATIVE

Totally dissatisfied with the service. Worst customer care ever.

NEUTRAL

Good Job but I will expect a lot more in future.

POSITIVE

Brilliant effort guys! Loved Your Work.

Sentiment Detection

- **How would you do it?**
- … Ideas … be creative …

Sentiment Detection

- Most simple: Dictionary-based
- a) **create a dictionary** of positive and negative terms
- b) **count** how often these terms in your document

- **What is the problem with this method?**

Sentiment Detection

- **Problems of dictionary-based methods:**
- a) Negation
- b) “I heard, that this is a **good** movie and that this movie is **entertaining,** but I don’t think so.”
- c) Sarcasm.

Sentiment Detection

- **Solve the problems:**

  Negation: just flip polarity.

  But not perfect, because:

  ```
  "not good" != "bad"
  "not very good" != "very bad"
  ```

Sentiment Detection

- Problems b)

  b) “I heard, that this is a **good** movie
  or that this movie is **entertaining,**
  but I don’t think so.”

- **What needs to be done here?**

Sentiment Detection

- Problems b)

Dependency parser

Show SpaCy:

https://explosion.ai/demos/displacy

Language Modeling

- When you type on your phone, the app suggests you the current/next word to type. This is a language modeling (text generation) task.

- **How would you do this? How can it be implemented?**

Language Modeling
N-Gram Model

- ✓ Show on whiteboard for (Unigram, Bi-Gram, Tri-Gram)
- ✓ **What is the problem here?**
- ✓ Give super basic idea of **RNN**

Machine Translation

- ✓ Show Google translate
- ✓ **How would you do this? How can it be implemented?**

Machine Translation

- Idea 1: Dictionary-based
- What is the problem?

Machine Translation

- One problem:
  "I **walk**. The **walk** was long."
  "Ich **gehe**. Der **Spaziergang** war lange."

- Part-of-speech tagging

- **How could we do this?**

  https://parts-of-speech.info/

Machine Translation

- Another problem:
  "I bought a new **mouse**."
- Computer mouse? Animal?
- Meaning depends on the context of this sentence.
- **Task:** Word Sense Disambiguation (WSD)
- … We look at **WordNet** later

Machine Translation

- ✓ Another problem:
  Can you just translate word by word?

  ”The **kitchen floor**” → “Der **Kuechenboden**”

- ✓ **So:** simple word by word is never perfect.

- ✓ **Modern techniques:** Encoder – decoder neural architectures
  - Show on whiteboard
  - Translate to multiple languages

Last: Text Classification Example: Spam Detection

- ✓ Is this email **spam** or **ham**?
- ✓ **How to do this? How would a human do it?**

Last: Text Classification Example: Spam Detection

- Simple words occurrence maybe not enough.

- Volleyball → sport!
  ”In the movie, Tom Hanks talks to a volleyball and makes him his friend.”

Example: Spam Detection

- Is this email **spam** or **ham**?

- **How to do this? How does a human do it?**

  Given:

  you have 10000 spam,
  and 10000 ham emails.

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$