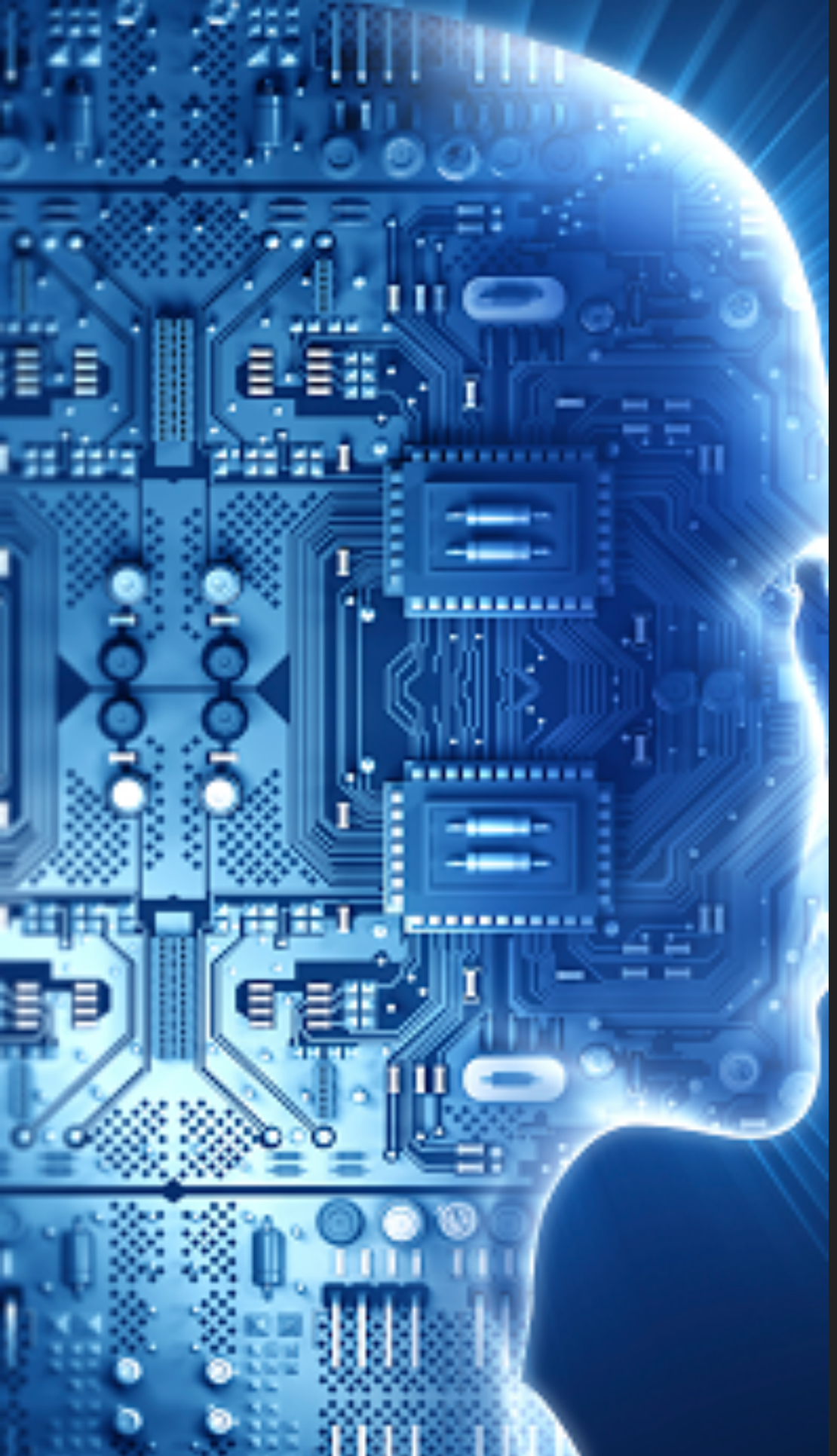


INTRODUCTION TO AI AND MACHINE LEARNING

SESSION #3

COURSE AGENDA

Session #1: Introduction to machine learning, concepts, basics, capabilities. Classification basics. ✓

Session #2: Feature engineering, data wrangling. Regression basics. ✓

Session #3: Working with textual data, text classification, NLP basics

Session #4: Introduction to neural networks, deep learning, image recognition

SESSION #3 AGENDA

SECTION 1

- Language interpretation
- Statistical models for text processing
- Information extraction methods

SECTION 2

- Case Study: Predicting SMS spam

NATURAL LANGUAGE PROCESSING

Goal: interpreting human language by a computer

Application areas:

- text classification (assigning text into categories, e.g. spam detection)
- machine translation
- chatbots
- sentiment analysis (deciding whether a text is positive, negative or neutral in nature)
- natural language generation (e.g. predicting next word, summarising text)
- speech recognition (sound -> text -> NLP)

LANGUAGE INTERPRETATION

“This is our second vacation at an Iberostar Grand and it delivers it spades! The service is like no other, quality food and a well maintained grand resort on an amazing clean white sand beach that you can walk for miles! It’s a small and quaint resort which means less people everywhere and no line ups. From check in to check out a fantasy experience”

“My wife and I stayed in this hotel from January 13-20 2020 and absolutely loved it. Obviously a little more expensive than other all inclusives in the area but more eating usthan worth it in my opinion. Staff was beyond belief, especially want to shoutout our butler Marggie, as well as Omar and Sylvester and pool bar waiter Tobia. Food was awesome, we always came back to a room with something beautiful or cool greeting us, and the grounds were immaculate”

LANGUAGE INTERPRETATION

Syntactic analysis

- grammar
- rules

Semantic analysis

- meaning
- context

Example: I went home before 11 a.m.

Example 1: these guys gave me nuts!

Example 2: hollow keyhole fly 32 papers

Example 3: “Service is down again.

Thank you Telekom! Thank you very much!!”

BAG OF WORDS AND N-GRAM MODELS

Text dataset = collection of documents (corpus)

Machines only understand numbers -> convert documents to numeric data

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

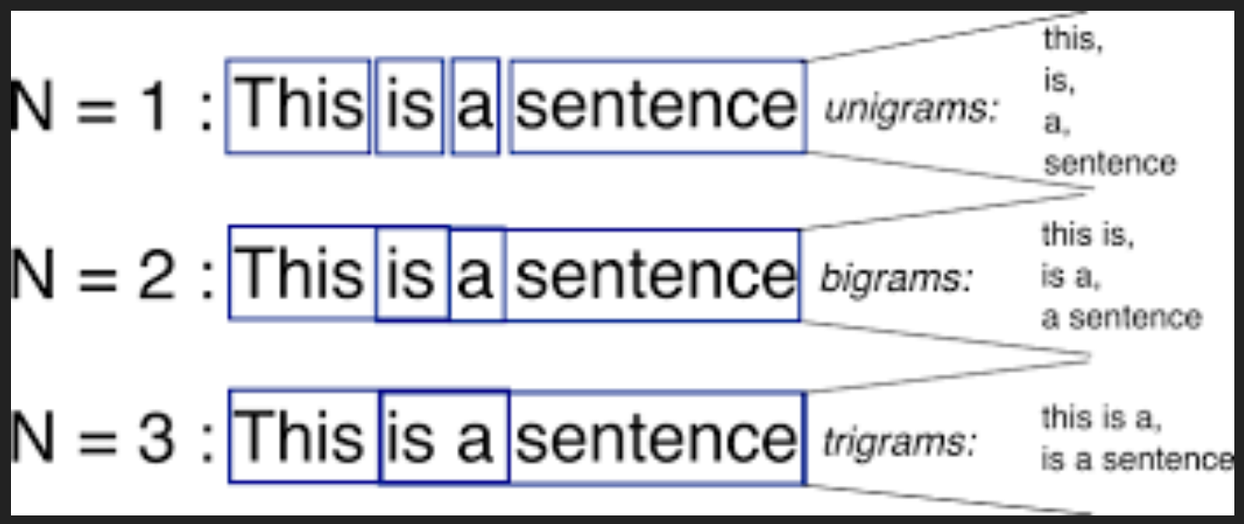
Bag of words

- describe the occurrence of words in the documents
- defines a vocabulary of known words
- measures the presence of known words

Every document in the corpus becomes a binary vector

BAG OF WORDS AND N-GRAM MODELS

Bag of words is a special case for n-gram models (unigram model)



How this can be used:

- measure document similarity
- topic modeling
- finding dominant words

Advantages:

- easy to understand and implement
- flexible, easy to customize

Drawbacks:

- semantic, e.g. meaning, is discarded
- manage huge vocabulary

TF-IDF REPRESENTATION

Problem with simple bag of words:

highly frequent words can dominate document => may suppress rare, but informational words

Term frequency (TF) = term count per document / total terms per document

Inverse Document Frequency (IDF): log (no. of documents / no. of documents with term)

Example:

The hotel was great. The restaurant in the hotel was fine.

TF1: 'The hotel was great' => The: 0.25, hotel: 0.25, was:0.25, great: 0.25

TF2: 'The restaurant in the hotel was fine.' => The:0.14, restaurant:0.14, in:0.14, the:0.14,

hotel:0.14, was:0.14, fine: 0.14

IDF1: the: log(2/2) = 0, hotel: log(2/2)=0, was: log(2/2)=0, great: log(2/1)=0.69

TFIDF1: the: 0.25*0=0, hotel: 0.25*0=0, was: 0.25*0=0, great: 0.25*0.69=0.17

TEXT PREPROCESSING

Based on the project needs you may want to do some text preprocessing:

- normalizing case (Apple -> apple)
- remove punctuation and whitespace (‘this is cooool!!!\n want some more\n\n’)
- removing stop words (I, is, the, how, about, etc.)
- stemming (studies, studying, studied -> stud)
- lemmatization (went, gone, going -> go)

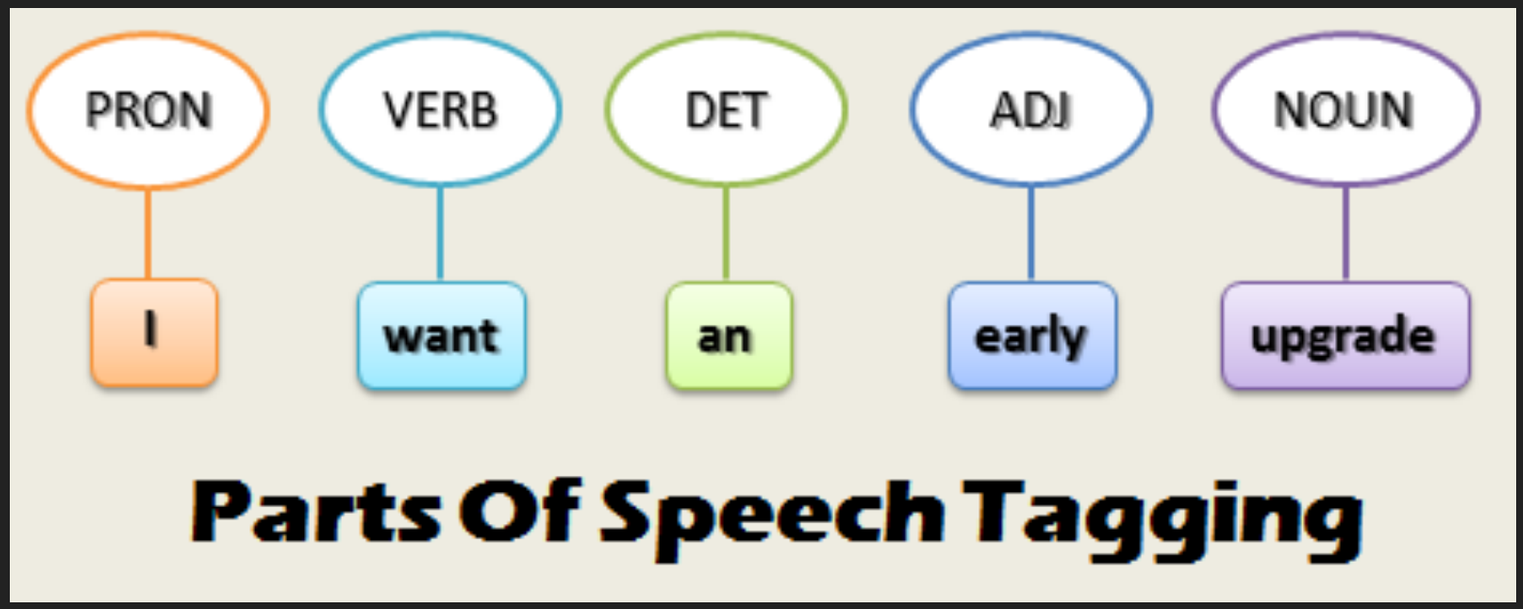
INFORMATION EXTRACTION

Beyonds statistical models the are several methods to extract further information from text

- part of speech tagging
- dependency parsing
- named entity recognition
- sentiment analysis

PART OF SPEECH TAGGING (POS TAGGING)

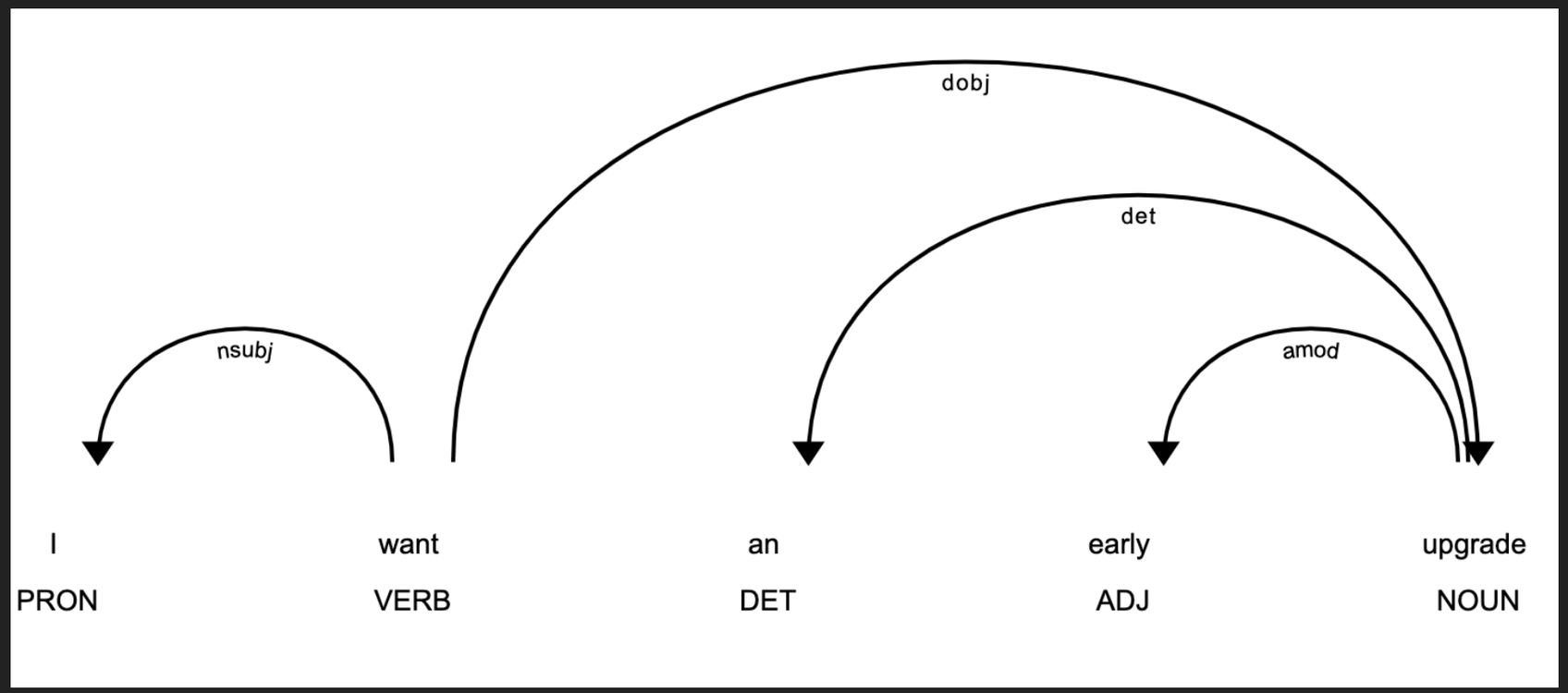
Goal: Identifying grammatical elements in a text



Source: https://www.thinkinfi.com/2018/10/extract-custom-entity-using-nltk-pos.html

DEPENDENCY PARSING

Goal: Extracting relations between grammatical elements in a sentence



Further reading: https://nlp.stanford.edu/software/dependencies_manual.pdf

NAMED ENTITY RECOGNITION

Goal: locate and classify named entities in a text into pre-defined categories

Figure 1: An example of NER application on an example text

SENTIMENT ANALYSIS

Goal: determine sentiment and/or objectivity from text.

SUMMARY

We learnt today:

- how humans and machines interpret natural languages
- statistical models (bag of words, TF-IDF)
- how we can extract more information from text by various methods

HOMEWORK

THE ALTERED STEAM REVIEW DATASET

Can you predict whether a user is recommended a computer game or not based on his/her review only?

Dataset description: https://www.kaggle.com/luthfim/steam-reviews-dataset