THE ULTIMATE BEGINNERS GUIDE TO NATURAL LANGUAGE PROCESSING

COURSE CONTENT

- Part 1: Basics of natural language processing
- Part 2: Summarization, search, representation, and similarity
- Part 3: Sentiment analysis

PREREQUISITES

- Programming logic
- Basic Python programming
- Level: **beginners**

NATURAL LANGUAGE PROCESSING

Speech Transcription

Neural Machine Translation (NMT)

Chatbots

Q&A

Text Summarization

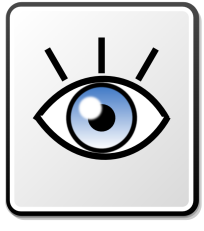


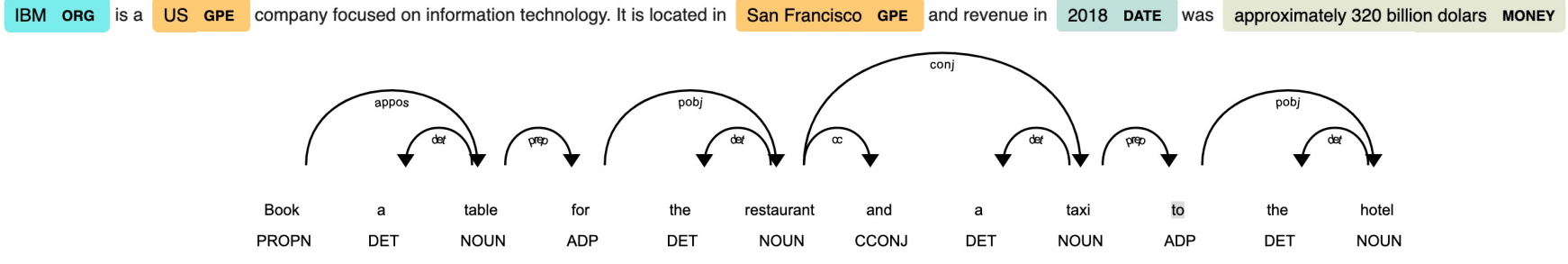
Image Captioning

Video Captioning

Sentiment analysis

PLAN OF ATTACK – BASIC NLP (SPACY)

1. Part-of-speech (POS)
2. Lemmatization and stemming
3. Named entity recognition
4. Stop words
5. Dependency parsing
6. Word similarity



PLAN OF ATTACK – SUMMARIZATION, SEARCH, REPRESENTATION AND SIMILARITY

1. Loading texts from the Internet
2. Most frequent words and wordcloud
3. Text summarization
4. Key word search
5. Bag of words and TF-IDF
6. Cosine similarity
7. Simulating a chatbot

machine-learning approach to language processing. [6] • 1990s: Many of the notable early successes on statistical methods in NLP occurred in the field of machine translation, due especially to work at IBM Research. These systems were able to take advantage of existing multilingual textual corpora that had been produced by the Parliament of Canada and the European Union as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding systems of government. However, most other systems depended on corpora specifically developed for the tasks implemented by these systems, which was (and often continues to be) a major limitation in the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data. • 2000s: With the growth of the web, increasing amounts of raw (unannotated) language data has become available since the mid-1990s. Research has thus increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web), which can often make up for the inferior results if the algorithm used has a low enough time complexity to be practical. In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing, due in part to a flurry of results showing that such techniques[7][8] can achieve state-of-the-art results in many natural language tasks, for example in language modeling,[9] parsing,[10][11] and many others. This is increasingly important in medicine and healthcare, where NLP is being used to analyze notes and text in electronic health records that would otherwise be inaccessible for study when seeking to improve care. [12] In the early days, many language-processing systems were designed by symbolic methods, i.e., the hand-coding of a set of rules, coupled with a dictionary lookup:[13][14] such as by writing grammars or devising heuristic rules for stemming. More recent systems based on machine-learning algorithms

the time not articulated as a problem separate from artificial intelligence. The premise of symbolic NLP is well-summarized by John Searle's Chinese room experiment: Given a collection of rules (e.g., a Chinese phrasebook, with questions and matching answers), the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it is confronted with. • 1950s: The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three

generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?". • 1970s: During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot

EXPERT ACADEMY

BAG OF WORDS

#1: This is the first document.
#2: This document is the second document.
#3: And this is the third one.
#4: Is this the first document?

```
[[0 1 1 1 0 0 1 0 1]
 [0 2 0 1 0 1 1 0 1]
 [1 0 0 1 1 0 1 1 1]
 [0 1 1 1 0 0 1 0 1]]
```

	and	document	first	is	one	second	the	third	this
#1	0	1	1	1	0	0	1	0	1
#2	0	2	0	1	0	1	1	0	1
#3	1	0	0	1	1	0	1	1	1
#4	0	1	1	1	0	0	1	0	1

TF-IDF (TERM-FREQUENCY – INVERSE DOCUMENT FREQUENCY

#1: This is the first document. #2: This document is the second document. #3: And this is the third one. #4: Is this the first document?

	and	document	first	is	one	second	the	third	this
	5	6			6		5		
#1		1	1	1			1		1
#2		2		1		1	1		1
#3	1			1	1		1	1	1
#4		1	1	1			1		1

TF = Number of times term *T* appears in the document / number of terms in the document

	and	document	first	is	one	second	the	third	this
#1		0.20	0.20	0.20			0.20		0.20
#2		0.33		0.16		0.16	0.16		0.16
#3	0.16			0.16	0.16		0.16	0.16	0.16
#4		0.20	0.20	0.20			0.20		0.20

TF-IDF (TERM-FREQUENCY – INVERSE DOCUMENT FREQUENCY

IDF = 1 + Log(Total number of documents / Number of documents term *T* appeared)

There are 4 documents
The term “document” appears in #1, #2 and #4

IDF = 1 + Log(4 / 3)

IDF = 1.28

There are 4 documents
The term “first” appears in #1 and #4

IDF = 1 + Log(4 / 2)

IDF = 1.69

#1: This is the first document. #2: This document is the second document. #3: And this is the third one. #4: Is this the first document?

TF-IDF (TERM-FREQUENCY – INVERSE DOCUMENT FREQUENCY

TF * IDF

	Sentence #1	Sentence #2	Sentence #3	Sentence #4
document	0.20 * 1.28 = 0.25	0.33 * 1.28 = 0.42	0	0.20 * 1.28 = 0.25
first	0.20 * 1.69 = 0.33	0	0	0.20 * 1.69 = 0.33

#1: This is the first document. #2: This document is the second document. #3: And this is the third one. #4: Is this the first document?

PLAN OF ATTACK – SENTIMENT ANALYSIS

1. Twitter dataset
2. Language detection
3. Sentiment analysis with NLTK
4. Introduction to classification and decision trees
5. Sentiment analysis with TF-IDF
6. Sentiment analysis with spaCy

CLASSIFICATION

Credit history	Debts	Properties	Anual income	Risk
Bad	High	No	< 15.000	High
Unknown	High	No	>= 15.000 a <= 35.000	High
Unknown	Low	No	>= 15.000 a <= 35.000	Moderate
Unknown	Low	No	> 35.000	High
Unknown	Low	No	> 35.000	Low
Unknown	Low	Yes	> 35.000	Low
Bad	Low	No	< 15.000	High
Bad	Low	Yes	> 35.000	Moderate
Good	Low	No	> 35.000	Low
Good	High	Yes	> 35.000	Low
Good	High	No	< 15.000	High
Good	High	No	>= 15.000 a <= 35.000	Moderate
Good	High	No	> 35.0000	Low
Bad	High	No	>= 15.000 a <= 35.000	High

Training

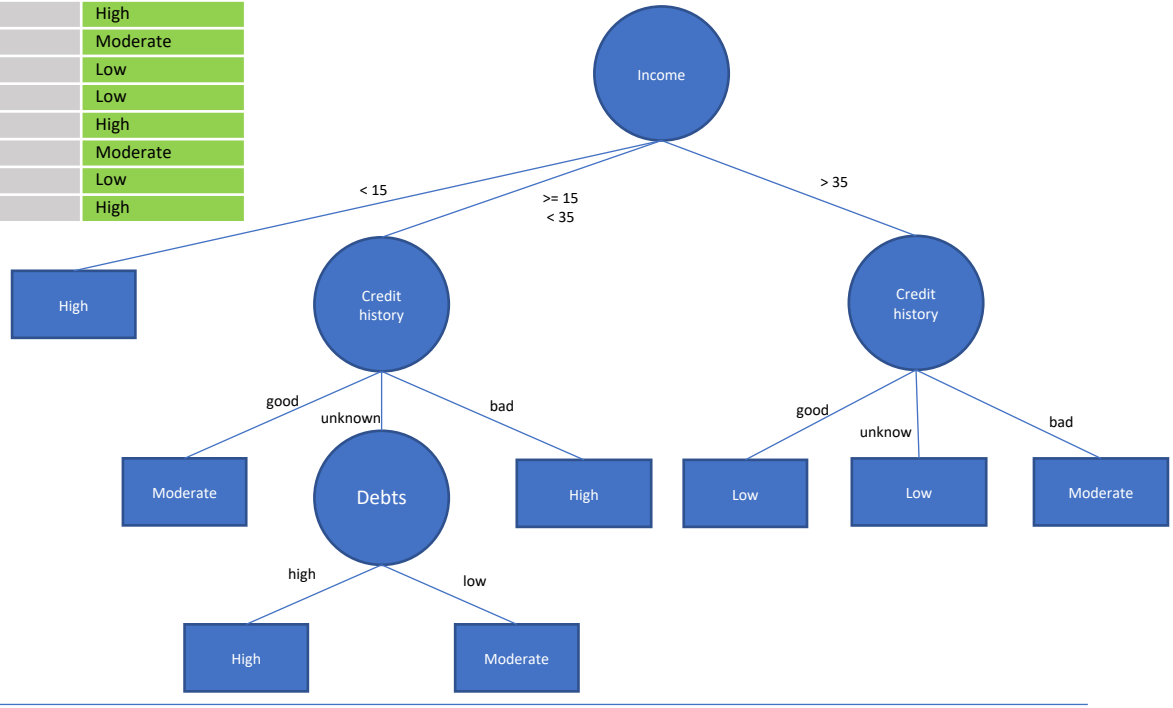
Credit history	Debts	Properties	Anual income
Bad	High	Yes	< 15.000
Unknown	High	Yes	< 15.000
Unknown	Low	No	>= 35.000
Good	High	Yes	>= 15.000 a <= 35.000

DECISION TREES

Credit history	Debts	Properties	Anual income	Risk
Bad	High	No	< 15.000	High
Unknown	High	No	>= 15.000 a <= 35.000	High
Unknown	Low	No	>= 15.000 a <= 35.000	Moderate
Unknown	Low	No	> 35.000	High
Unknown	Low	No	> 35.000	Low
Unknown	Low	Yes	> 35.000	Low
Bad	Low	No	< 15.000	High
Bad	Low	Yes	> 35.000	Moderate
Good	Low	No	> 35.000	Low
Good	High	Yes	> 35.000	Low
Good	High	No	< 15.000	High
Good	High	No	>= 15.000 a <= 35.000	Moderate
Good	High	No	> 35.0000	Low
Bad	High	No	>= 15.000 a <= 35.000	High

History = Good
Debts = High
Properties = No
Income = > 35

History = Bad
Debts = High
Properties = Yes
Income = < 15



EXPERT ACADEMY

TEXT SUMMARIZATION

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations.
- Steps
    1. Preprocessing the texts
    2. Word frequency
    3. Weighted word frequency
    4. Sentence tokenization
    5. Score for the sentences
    6. Order the sentences
    7. Generate the summary

1. PREPROCESSING THE TEXTS

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations

- artificial intelligence human like intelligence. study intelligent artificial agents. science engineering produce intelligent machines. solve problems intelligence. related intelligent behavior. developing reasoning machines. learn mistakes successes. artificial intelligence related reasoning everyday situations

2. WORD FREQUENCY

Word	Frequency
artificial	3
intelligence	4
human	1
like	1
study	1
intelligent	3
science	1
engineering	1
produce	1
machines	2
solve	1

Word	Frequency
agents	1
problems	1
related	2
behavior	1
developing	1
reasoning	2
learn	1
mistakes	1
successes	1
everyday	1
situations	1

3. WEIGHTED WORD FREQUENCY

Highest value: 4

Word	Frequency	Weigth
artificial	3	0.75
intelligence	4	1.00
human	1	0.25
like	1	0.25
study	1	0.25
intelligent	3	0.75
science	1	0.25
engineering	1	0.25
produce	1	0.25
machines	2	0.50
solve	1	0.25

Word	Frequency	Weight
agents	1	0.25
problems	1	0.25
related	2	0.50
behavior	1	0.25
developing	1	0.25
reasoning	2	0.50
learn	1	0.25
mistakes	1	0.25
successes	1	0.25
everyday	1	0.25
situations	1	0.25

4. SENTENCE TOKENIZATION

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations
- Tokenization
  - Artificial intelligence is human like intelligence.
  - It is the study of intelligent artificial agents.
  - Science and engineering to produce intelligent machines.
  - Solve problems and have intelligence.
  - Related to intelligent behavior.
  - Developing of reasoning machines.
  - Learn from mistakes and successes.
  - Artificial intelligence is related to reasoning in everyday situations

5. SCORE FOR THE SENTENCES

Sentence	Score (sum of weights)
Artificial **(0.75)** intelligence **(1.00)** is human **(0.25)** like **(0.25)** intelligence **(1.00)**.	3.25
It is the study **(0.25)** of intelligent **(0.75)** artificial **(0.75)** agents **(0.25)**.	2.00
Science **(0.25)** and engineering **(0.25)** to produce **(0.25)** intelligent **(0.75)** machines **(0.50)**.	2.00
Solve **(0.25)** problems **(0.25)** and have intelligence **(1.00)**.	1.50
Related **(0.50)** to intelligent **(0.75)** behavior **(0.25)**.	1.50
Developing **(0.25)** of reasoning **(0.50)** machines **(0.50)**.	1.25
Learn **(0.25)** from mistakes **(0.25)** and successes **(0.25)**.	0.75
Artificial **(0.75)** intelligence **(1.00)** is related **(0.50)** to reasoning **(0.50)** in everyday **(0.25)** situations **(0.25)**.	3.25

6. ORDER THE SENTENCES

Sentence	Score (sum of weights)
Artificial **(0.75)** intelligence **(1.00)** is related **(0.50)** to reasoning **(0.50)** in everyday **(0.25)** situations **(0.25)**.	3.25
Artificial **(0.75)** intelligence **(1.00)** is human **(0.25)** like **(0.25)** intelligence **(1.00)**.	3.25
It is the study **(0.25)** of intelligent **(0.75)** artificial **(0.75)** agents **(0.25)**.	2.00
Science **(0.25)** and engineering **(0.25)** to produce **(0.25)** intelligent **(0.75)** machines **(0.50)**.	2.00
Solve **(0.25)** problems **(0.25)** and have intelligence **(1.00)**.	1.50
Related **(0.50)** to intelligent **(0.75)** behavior **(0.25)**.	1.50
Developing **(0.25)** of reasoning **(0.50)** machines **(0.50)**.	1.25
Learn **(0.25)** from mistakes **(0.25)** and successes **(0.25)**.	0.75

7. GENERATE THE SUMMARY

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations

- Artificial intelligence is related to reasoning in everyday situations. Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents.