universidade de aveiro
theoria poiesis praxis

Introdução à Ciência de Dados

João Lourenço Marques: jjmarques@ua.pt

Processamento de linguagem natural (NLP)

Data Science/Ciência de Dados para Ciências Sociais

NLP /Análise de Conteúdo

1. Conceitos
2. Tipos de utilização
3. Abordagens e Fases
4. Processo
5. Estruturação de informação textual
6. Text mining
7. …

Quando aparecer o símbolo “#” tem hiperlink

Data Science/Ciência de Dados

1 Conceitos

Análise de conteúdo

"the longest established ***method of text analysis*** *among the set of empirical methods of social investigation" (p.55)* Titscher et al. (2000)

…contudo, não parece existir, ainda, uma consensualização sobre o significado deste tipo de métodos

Concentram-se em **aspetos direta e claramente quantificáveis** do **conteúdo do texto**

(frequências absolutas e relativas de palavras por texto ou por uma qualquer fração desse texto)

The Use of Qualitative Content Analysis in Case Study Research

Tipicamente (nos métodos clássicos) analisa-se o texto:

Análise conceptual

- determina a existência e a frequência dos conceitos num texto

Frequência de palavras

(simplesmente) contagem de ocorrências de palavras (ou conjunto de palavras), número de vezes que aparecem num texto, assim como a sua proximidade com outras palavras ou expressões.

Os termos podem ser explícitos ou implícitos

#Text Analyzer

Análise relacional

- relações entre conceitos num texto

Codificação de texto

ideias e temas expressos por muitas palavras transformadas numa frase comum – dá para perceber como evolui o discurso

#WebQDA

Abordagem/análise
Qualitativa ou quantitativa?

A tribo dos da investigação:

Quantitativa (so-called "QUANs")
Qualitativa (so-called "QUALs")

cientistas soft
não científico,
exploratório
subjetivo

mixed method approaches (triangulation)
(e.g. BRYMAN, 2004; CRESWELL, 2003; JICK, 1979)

Análise de conteúdo

…técnicas que permitem estudar (indiretamente) o comportamento humano através da análise das suas "comunicações"

- Texto estruturado (documentos, artigos, etc.).
- Texto não estruturado (transcrições, entrevistas, grupos focais, conversas, memorandos)
- Gravações áudio (entrevistas, debates, conversas, música)
- Gravações de vídeo (filmes, documentários, …)
- Outros (gráficos, imagens, pinturas, desenhos visuais)

analisar **grandes quantidades de dados** de forma encontrar informação (útil e relevante) de uma forma sistemática.

…identificando padrões e associações

tendências no uso de palavras, estrutura sintática, etc.

Podem ser utilizadas muitas técnicas, incluindo:

- **contagem de palavras**
  procurar temas e palavras-chave interessantes para explorar - remover palavras triviais

- **modelação**
  processo exploratório de **procura de padrões/associações** de temas - localizar "partes" que podem ser interpretadas como "tópicos" lógicos - como as palavras estão ligadas

- **análise dos sentimentos**
  identifica o conteúdo emocional das palavras.
  Na sua abordagem mais básica rotulam-se as palavras como negativas e positivas.
  Para realizar esta análise, deve ser dada ao computador uma lista de palavras já codificadas como positivas e negativas (campo em desenvolvimento).

- ....

#Why You Should Do Text Analysis in Python (Even if You Don't Want to)

Data Science

1. O conceito

(em sentido lato)

Supervisionados

Utilizam dados de treino para generalizar padrões

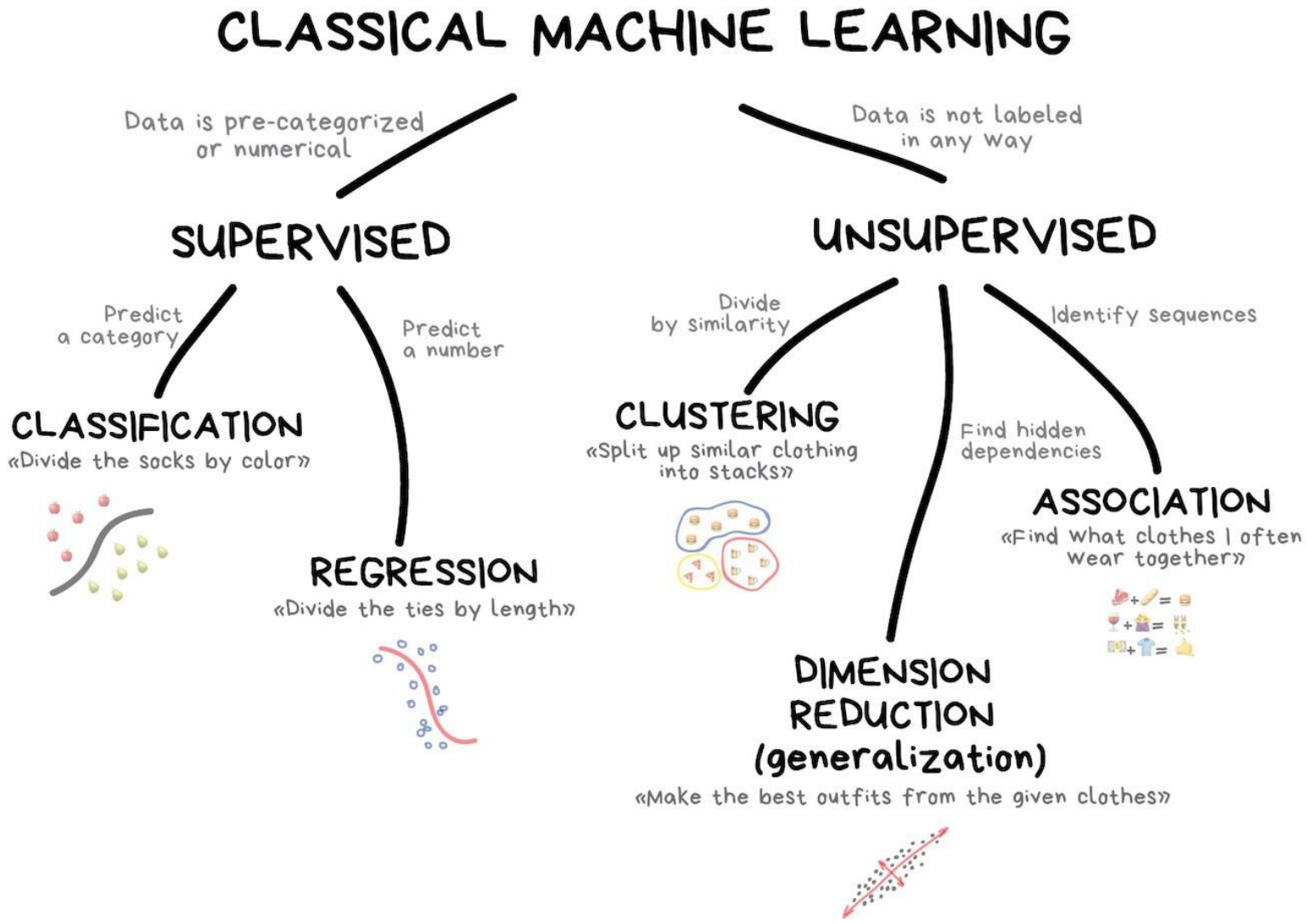
Não-supervisionados

regras genéricas de algoritmos que se aplicam diretamente

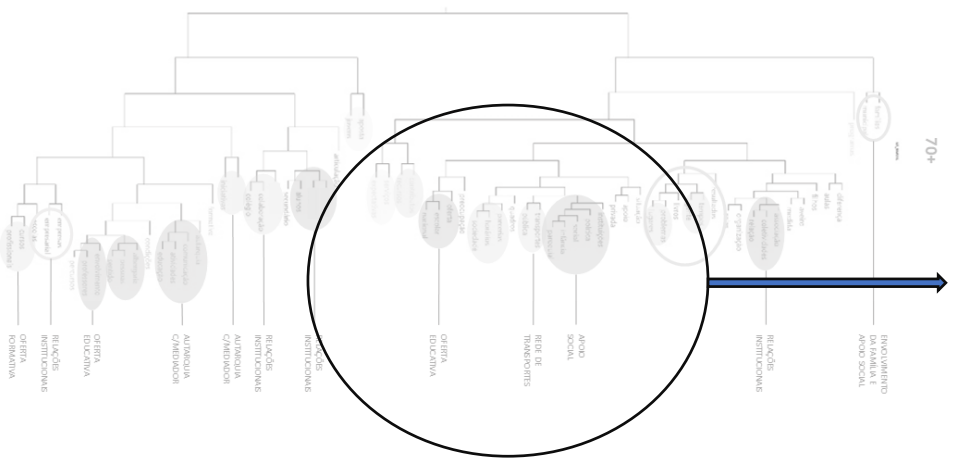
Data Science

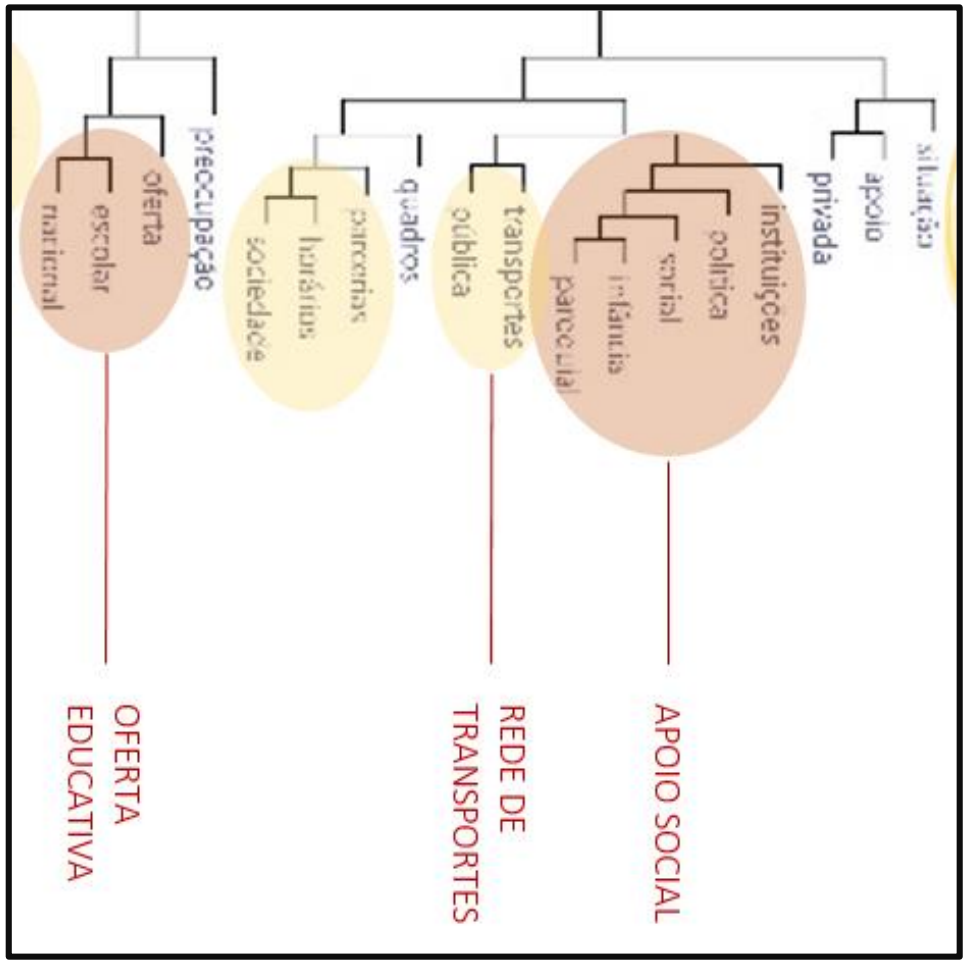
1. O conceito

(em sentido lato)



Classificação





Envolvimento da família e apoio social	Oferta Educativa e Formativa	Rede de Transportes
• Necessidade de maior interação entre Escola–Pais-CM • Respostas da rede solidária	• Diversificação da oferta formativa • Articulação insuficiente entre a oferta formativa e as necessidades da população e do tecido empresarial • Redefinição da rede escolar, nomeadamente no EPE e no ES	• Desajustamento entre a rede de transportes e as necessidades dos alunos/munícipes • Assunção de uma dimensão estratégica ao nível da rede de transportes

universidade de aveiro
theoria poiesis praxis

Data Science

1. O conceito

(em sentido lato)

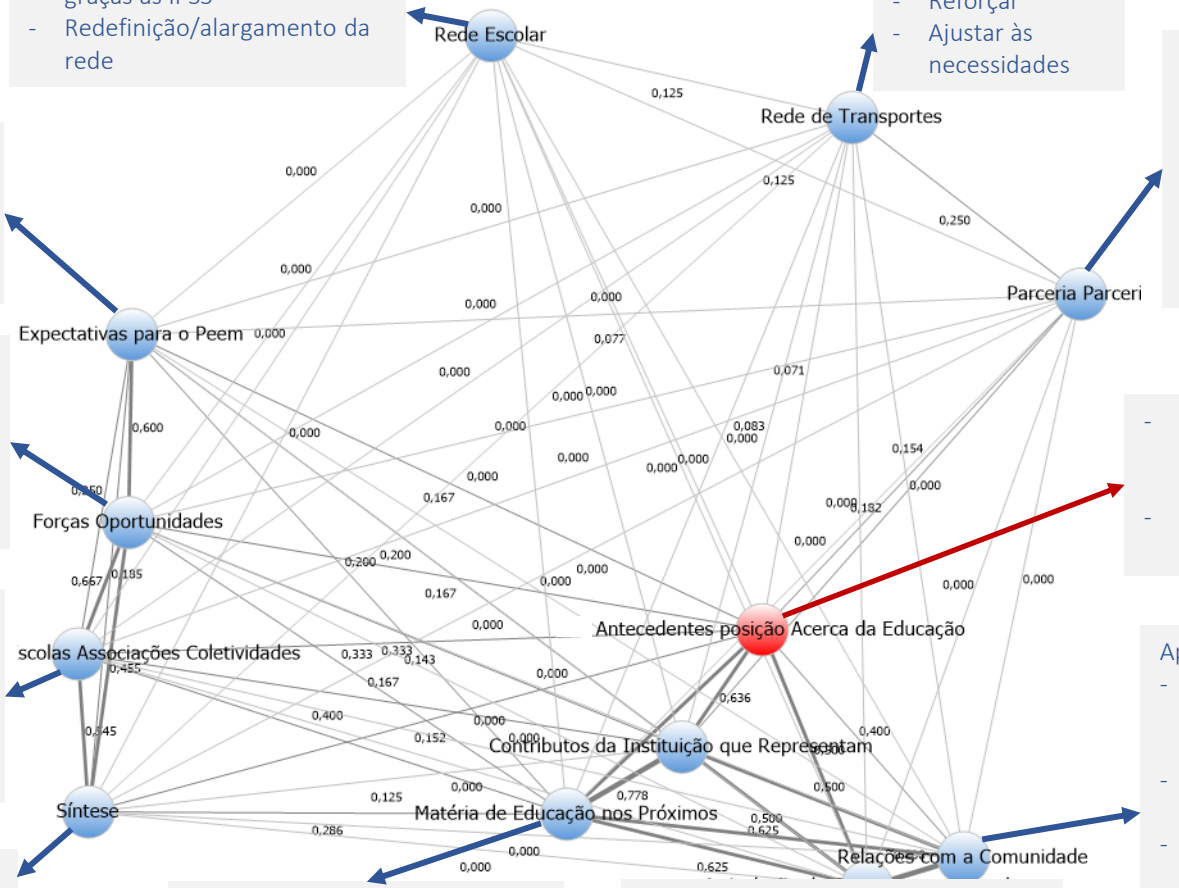
Associação

- Alargamento do EPE muito graças às IPSS
- Redefinição/alargamento da rede

- Reforçar
- Ajustar às necessidades

São necessárias + parcerias entre:
- INSTITUIÇÕES
- ASSOCIAÇÕES
- TECIDO EMPRESARIAL
- AGENTES EDUCATIVOS
- AUTARQUIA
- ESCOLA

- Reforço do trabalho conjunto
- Maior envolvimento
- Repensar o ensino à escala local



- Escola de Artes
- Práticas de integração e formação de alunos com NEE
- Oportunidades de formação

- Perceção positiva da realidade educativa do concelho
- Tirar partido das redes de proximidade

- Algum distanciamento entre agrupamentos
- Fraca interação entre escolas e associações culturais, desportivas, empresariais e de jovens

Aposta:
- Na construção da identidade do Município/Munícipes;
- Na resolução de problemas sociais
- No reforço do papel de coordenação do Município
- No fomento de esforços comuns

- Retrato geral favorável

- Falta de interação regular entre Escola, Pais e Autarquia
- Aposta premente na concertação entre agentes educativos
- Reforço das parcerias entre os diferentes agentes educativos, com a CM como mediadora

Aposta:
- Na interação entre os diferentes agentes educativos
- Na participação alargada na elaboração do plano de atividades
- Na articulação entre o PEEM, os Projetos Educativos das escolas e o Programa Municipal de Educação

Data Science

1. O conceito

(a redundância semântica)

- TEXT MINING
- *TEXT ANALYTICS*
- *TEXT ANALYSIS*
- *CONTENT ANALYSIS*
- *CONTEXTUAL CONTENT*
- *SENTIMENT ANALYSIS*
- BIBLIOMETRIC

- ***NLP***

Text Mining
… processo de extrair informações úteis de grandes volumes de texto - através da aplicação de algoritmos de aprendizagem de máquina, estatísticas e técnicas de processamento de linguagem natural (NLP).

Text Analytics
…semelhante ao Text Mining, mas geralmente é mais focado na análise estruturada de textos para obter informações específicas, como sentimentos, categorias temáticas

Text Analysis
…termo mais genérico que se refere à análise de texto para extrair informações que pode ser feito manualmente ou através de métodos automatizados e pode incluir tanto Text Mining quanto Text Analytics.

Content Analysis
Content Analysis é uma técnica de pesquisa qualitativa e quantitativa que envolve a categorização e interpretação de aspectos do conteúdo de um texto, como temas, ideias e padrões.

Contextual Content
…refere-se ao conteúdo que é entendido em relação ao seu contexto. Isso pode incluir o ambiente em que o conteúdo é apresentado, o público-alvo, ou outros fatores que afetam como o conteúdo é interpretado.

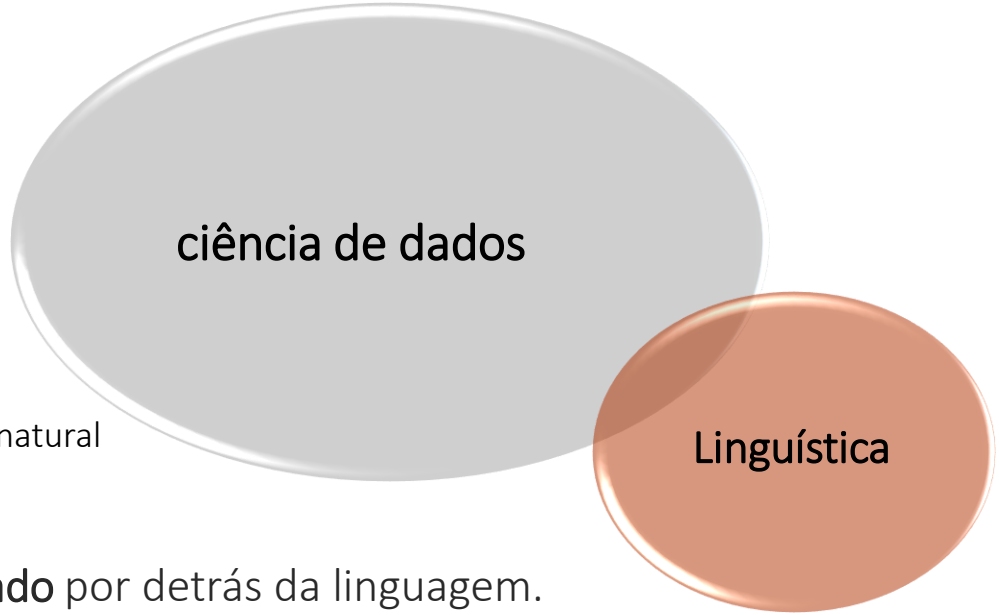
Sentiment Analysis
…o uso de técnicas de processamento de linguagem natural, estatísticas ou aprendizagem de máquina para identificar e extrair opiniões subjetivas ou sentimentos de um texto.

Bibliometric
… é o estudo quantitativo de publicações científicas e literárias. Isso inclui a análise de citações, padrões de publicação e outras métricas para avaliar o impacto ou a relevância de um trabalho acadêmico.

…e a **NLP** é uma área da **Ciência de dados** (inteligência artificial)
que lida com **a linguagem humana** (tanto escrita como a falada)

O processamento de linguagem natural

(*Natural Language Processing* - NLP)



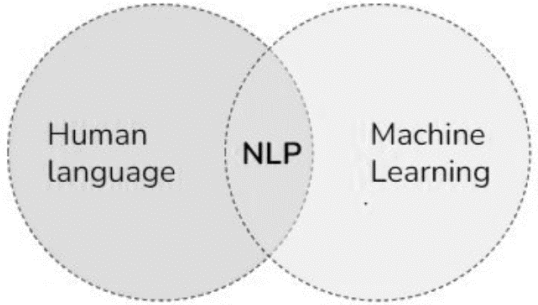
O objetivo **é automatizar a leitura**, **interpretação** e **compreensão da linguagem humanas -** linguagem natural

…preocupa-se com a **compreensão do significado** por detrás da linguagem.

… desenvolvimento e aplicação de ferramentas, técnicas e algoritmos para processar e compreender a linguagem natural dados,

…a NLP tem tudo a ver com a forma como os computadores funcionam com a linguagem humana.

é um ramo da inteligência artificial (I.A.) que lida com a interação entre máquinas e línguas humanas.

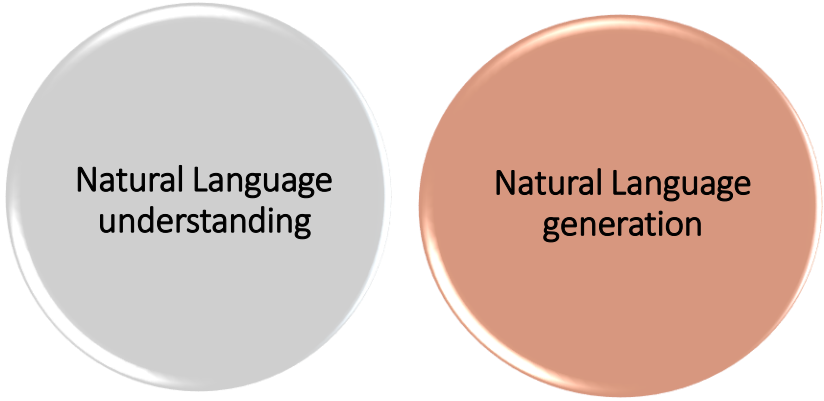


(Sarkar, 2019)

(NLP)

Natural Language Processing

(como os computadores compreendem, interpretam e usam a linguagem humana)



(capacidade dos computadores compreenderem texto e discursos humanos)

(capacidade dos computadores gerarem textos e discursos)

(NLP)

Natural Language Understanding

Words
↓
Morphological Analysis
↓
Morphologically analyzed words *(another step: POS tagging)*
↓
Syntactic Analysis
↓
Syntactic Structure
↓
Semantic Analysis
↓
Context-independent meaning representation
↓
Discourse Processing
↓
Final meaning representation

Natural Language Generation

↓
Meaning representation
Utterance Planning
↓
Meaning representations for sentences
Sentence Planning and Lexical Choice
↓
Syntactic structures of sentences with lexical choices
Sentence Generation
↓
Morphologically analyzed words
Morphological Generation
↓
Words
↓

Data Science/Ciência de Dados

2 Tipos de utilização

NLP refere-se a um conjunto de **tarefas que ajudam os sistemas computacionais** a **processar** e a **aprender** a utilização de línguas faladas por humanos

… tarefas de processamento de linguagem, tais como

Classificação de textos:

compreender os sentimentos a partir de um conjunto e palavras ou identificar documentos semelhantes com base no texto dos documentos

Geração de texto:

gerar um parágrafo de texto que descreva uma dada imagem

Resumir textos:

extrair ideias-chave a partir de um grande corpo de texto

Tradução de texto

traduzir de uma língua para outra

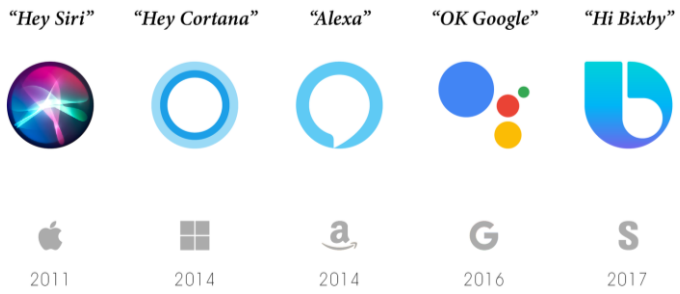
Processamento de texto para voz e vice-versa

conversação com sistemas treinados (robots) e obtêm uma resposta “adequada” às suas perguntas

NLP nas nossas vidas

- Pesquisa de informação (Google encontra resultados relevantes e similares).
- Tradução automática (Google Translate e DeepL Translator traduz de uma língua para outra).
- Simplificação de Texto (Rewordify, Smmry or Reddit's autotldr simplifica o significado das frases).
- Sentiment Analysis (Hater News dá-nos o sentimento do utilizador).
- Filtros de Spam (classificação)
- A dactilografia preditiva (Pesquisa Google prevê os resultados da pesquisa do utilizador).
- Auto-Correção (Google/Grammarly correção de palavras mal escritas).
- Reconhecimento ótico de caracteres (OCR)
- Reconhecimento de voz (Google WebSpeech or Vocalware).
- Resposta a perguntas (*IBM Watson's* Trivial answers *to a query*).
- Natural Language Generation (Geração de texto a partir de dados de imagem ou vídeo data.)
- Smart Assistants
- Aplicações de IA generativa (CHATGPTs…)

  (sistemas de inteligência artificial projetados para criar conteúdo novo e original)



https://www.expert.ai/blog/natural-language-processing-examples/
https://www.wonderflow.ai/blog/natural-language-processing-examples

Data Science/Ciência de Dados

3 Abordagem e Fases

Tipos de abordagem

Abordagem indutiva

abordagem exploratória em que não sabemos que padrões ou conteúdos vamos descobrir,
por isso vamos de mente aberta.

Abordagem dedutiva

implica começar com uma ideia vamos tentar encontrar nas fontes de informação como ela aparece

Fases

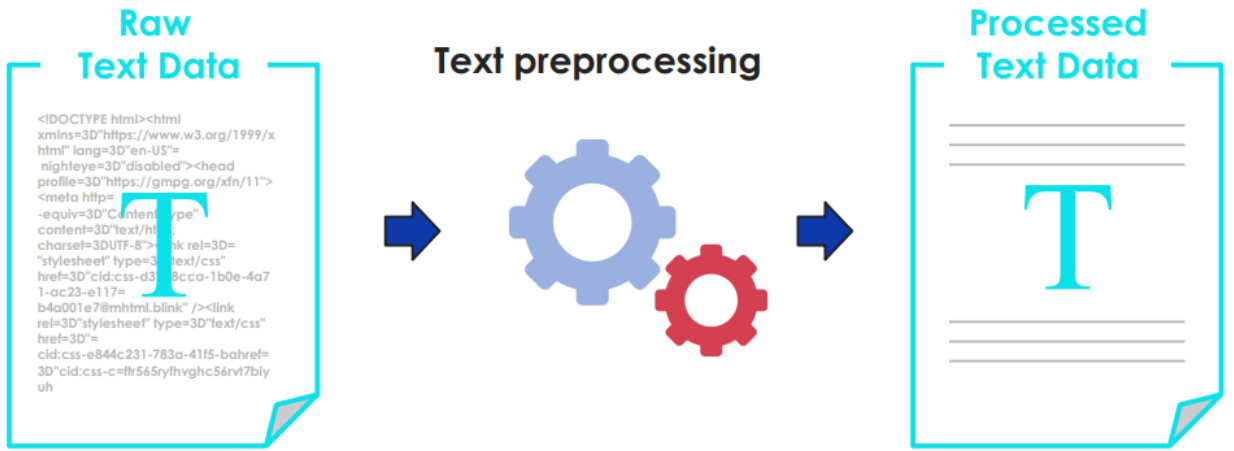

Definição do problema
perguntas a serem respondidas
Recolha de dados
seleção das fontes de informação
Redução dos dados
Seleção, Simplificação, transformação
Resultados
Matrizes, gráficos, redes
Conclusões
Padrões, associações e significados
Fases

Data Science/Ciência de Dados

4 Processo

Pré Processamento

É o processo de transformação dos dados de texto em bruto e não limpos numa forma analisável para o modelo



Modelo

Se treinarmos o modelo em dados não estruturados que não tenham sido pré-processados, modelo pode perder aprendizagens de informações importantes

Diferentes métodos de extração de características documentais

técnicas básicas

...

técnicas avançadas de Processamento de Linguagem Natural

#Guide to deal with Text Data

1.Basic feature extraction using text data

1. Number of words
2. Number of characters
3. Average word length
4. Number of stopwords
5. Number of special characters
6. Number of numerics
7. Number of uppercase words

2.Basic Text Pre-processing of text data

1. Lower casing
2. Punctuation removal
3. Stopwords removal
4. Frequent words removal
5. Rare words removal
6. Spelling correction
7. Tokenization
8. Stemming
9. Lemmatization

3.Representing text

1. N-grams
2. Term Frequency
3. Inverse Document Frequency
4. Term Frequency-Inverse Document Frequency (TF-IDF)
5. Bag of Words
6. Sentiment Analysis
7. Word Embedding

Pre-processamento dos dados/Etapas da indexação do texto

1.Basic Text pre-processing

1. Remove punctuation (1.1)
2. Lowercasing (2.1)
3. Stopwords (1.2)
4. Tokenization (segmentação) (1.3)
5. Stemming (1.4.1)
6. Lemmatization (1.4.2)

Normalização de texto

2. Representing text

7. Vectorizing Data: Bag-Of-Words (2.1.1) → (1-Gram)
8. Vectorizing Data: N-Grams (2.1.2)
9. Term Frequency-Inverse Document Frequency (TF-IDF) (2.2)
10. Part of Speech (POS) tagging (2.3)
11. Sentiment analysis (3.6)

Natural Language Processing(NLP) for Machine Learning
NLP - Text Preprocessing and Text Classification (using Python)

Pre-processamento dos dados/Etapas da indexação do texto

(Natural Language Toolkit)NLTK: NLTK is a popular open-source package in Python.
Rather than building all tools from scratch, NLTK provides all common NLP Tasks.

Installing NLTK

Type !pip install nltk in the Jupyter Notebook or if it doesn't work in cmd type conda install -c conda-forge nltk. This should work in most cases.

Install NLTK: http://pypi.python.org/pypi/nltk

Importing NLTK Library

In [*]:
```
import nltk
nltk.download()
```

O comando nltk.download() é usado para baixar pacotes adicionais para a biblioteca NLTK (Natural Language Toolkit) em Python

1. Remove punctuation and special characters

A pontuação pode fornecer um contexto gramatical a uma frase … **., ?, ! , #, $, &, @,**

Mas para no processos de NLP o que conta, num primeiro momento é número de palavras e não o contexto, não acrescenta valor, por isso removemos todos os caracteres especiais.

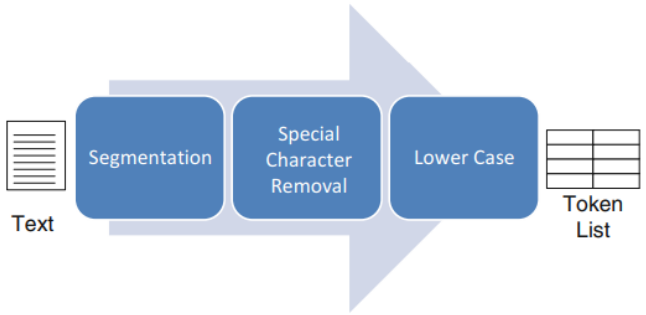
Por exemplo: Exemplo: "Olá, mundo!" → "Olá mundo"

In [3]:
```
1 import string
2
3 text = "Olá, mundo!"
4 text = text.translate(str.maketrans('', '', string.punctuation))
5 text
```

Out[3]: 'Olá mundo'

2. Lowercasing

… converter todas as palavras em minúsculas para garantir que as ocorrências repetidas da mesma palavra em casos diferentes continuam a ser tratadas como a mesma palavra.



In [4]:
```
1 text = "Hello, World! This is an example."
2 text_lower = text.lower()
3
4 print(text_lower)
```

```
hello, world! this is an example.
```

3. Stop-Word Removal

…refere-se ao processo de remoção de *stop words* - partes do texto com pouca informação
Por exemplo: *Apple or orange is fine for me-> apple, orange, fine*

Using natural language processing, we make use **of the** text data available across **the** internet **to** generate insights **for the** business. **to** make **this** huge amount **of** data usable **for a** natural language processing task, we use text preprocessing.

No entanto, a remoção incorreta de palavras pode até alterar o significado do nosso texto
'O filme não era mau'
Se removermos a palavra 'não', a frase passa a ser: 'O filme era mau', o que altera completamente o significado da frase original.

Uma forma de evitar este problema é escolher manualmente a lista de palavras a retirar

```
1 from nltk.corpus import stopwords
2
3 stop_words = set(stopwords.words('portuguese'))
4 text = "Eu gosto de café"
5 text = ' '.join([word for word in text.split() if word.lower() not in stop_words])
6 print(text)
```

```
gosto café
```

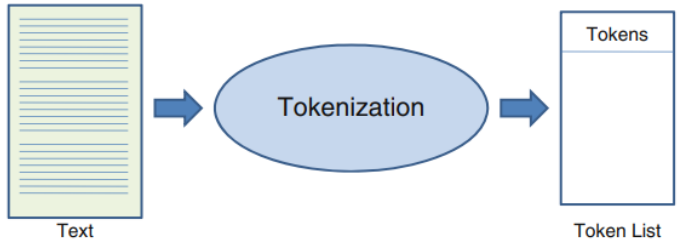
4.Tokenization (segmentação)

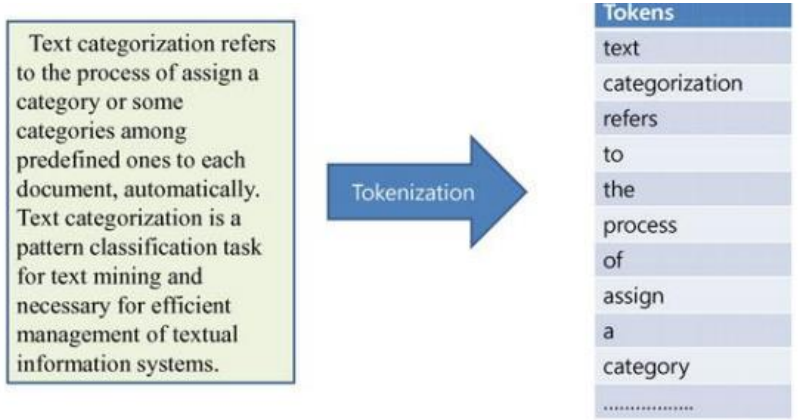
o processo de segmentar um texto ou textos em unidades (tokens)

- *Word Tokenization*
- *Sentence Tokenization*

- quebrar uma frase em palavras ou um texto em frases,

Por exemplo:






```
1 from nltk.tokenize import word_tokenize
2
3 text = "Eu gosto de café"
4 tokens = word_tokenize(text)
5 print(tokens)
```

```
['Eu', 'gosto', 'de', 'café']
```

5. Stemming

…cada palavra é reduzida à sua raiz (stem) - Reduz o *corpus* de palavras
Os algoritmos de *stemming* têm um conjunto de regras para decidir como fazer os cortes (#RSLP Stemmer.)

Por exemplo: *remoção de sufixos, "ing", "ly", "s",*
(Alguns motores de busca já tratam as palavras com o mesmo tronco/raiz, como sinónimos.)

Likes
Liked
Liking
Likely
[INFLECTIONS]

Like
[STEM]

Stemmers mais usados :
- Porter Stemmer (soft)
- Snowball Stemmer (soft)
- Lancaster Stemmer (hard)

Word	Suffix	Lancaster	Snowball	Porter
Connecting	- ing	Connect	Connect	Connect
Connected	- ed	Connect	Connect	Connect
generous	- erous	Gen	Generous	Gener
generously	- erously	Gen	Generous	Gener

```
1 from nltk.stem import RSLPStemmer
2
3 stemmer = RSLPStemmer()
4 stemmed_word = stemmer.stem("correndo")
5 print(stemmed_word)
```

```
corr
```

5. Stemming

As limitações do Stemming podem ser agrupadas em duas categorias principais:

- Overstemming
- Understemming

Overstemming

Universe		**Univers**
Universal	→	**Univers**
University		**Univers**
Universities		**Univers**

Understemming

Universe		**Univers**
Universal	→	**Universa**
University		**Univers**
Universities		**Universit**

6. Lemmatization

…processo de conversão de uma palavra para a sua forma base: deriva a forma canónica ("lemma") de uma palavra

analisa uma palavra e associa-a ao seu lema utilizando dicionários

Word	Lemmatization Output	Stemming Output
Changes	Change	Changes
Changing	Change	Changing
Multiplied	Multiply	Multiplied
Multiplier	Multiply	Multiplier

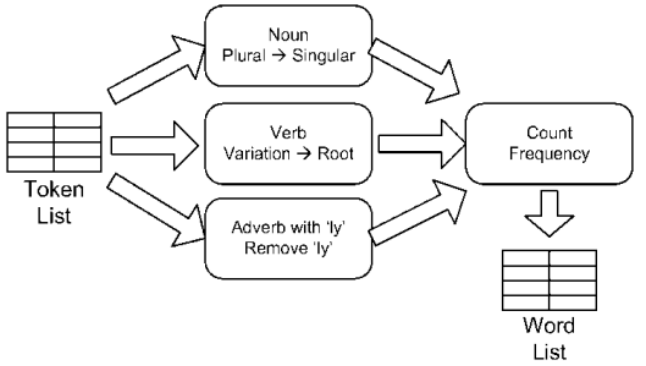
Por exemplo:

	Original	nltk_stemmer	spacy_lemma
0	amigos	amig	amigo
1	amigas	amig	amigo
2	amizade	amizad	amizade
3	carreira	carr	carreira
4	carreiras	carr	carreira

abordagem baseada em dicionários: que permite fazer uma análise morfológica da raiz da palavra

```
1 from nltk.stem import WordNetLemmatizer
2
3 lemmatizer = WordNetLemmatizer()
4 lemmatized_word = lemmatizer.lemmatize("changing", pos="v")
5 print(lemmatized_word)
```

```
change
```



7. Vectorizing Data: **Bag-Of-Words** ou **1 gram**

… (forma mais simples de representação de texto em números)

ou *CountVectorizer* -representa as palavras num vetor, sinalizando quais estão no texto.

Dá um resultado de 1 se estiver presente na frase e 0 se não estiver presente.

Por exemplo:

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Tem-se vetores de comprimento-6 para cada *documento*!

- *the cat sat*: [1, 1, 1, 0, 0, 0]
- *the cat sat in the hat*: [2, 1, 1, 1, 1, 0]
- *the cat with the hat*: [2, 1, 0, 0, 1, 1]

O modelo BoW não retém informações sobre a gramática das frases nem sobre a ordem das palavras no texto

Sobre BoW:
https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp

```
1 from sklearn.feature_extraction.text import CountVectorizer
2 import pandas as pd
3 vectorizer = CountVectorizer()
4 X = vectorizer.fit_transform(["gato cão", "gato"])
5
6
7 df = pd.DataFrame(X.toarray(), columns=vectorizer.get_feature_names_out())
8 print(df)
```

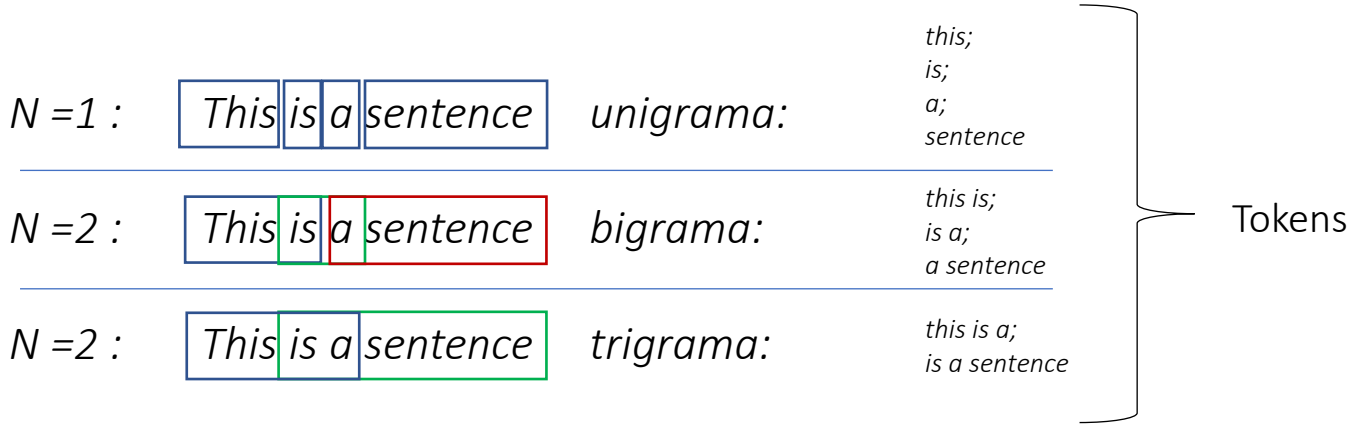
```
   cão  gato
0    1     1
1    0     1
```

8. Vectorizing Data: N-Grams

… são todas as combinações de palavras ou letras adjacentes à distância de n, que se podem encontrar no texto.

N-gramas com n=1 –> unigramas; bigramas (n=2), trigramas (n=3) etc

Por exemplo:

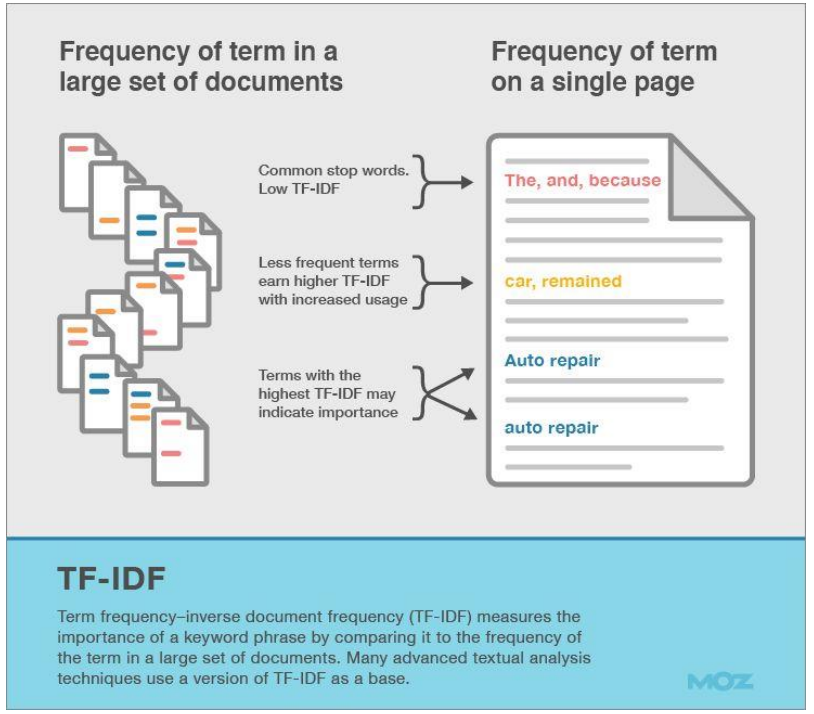



```
vectorizer = CountVectorizer(ngram_range=(2,2))
X = vectorizer.fit_transform(["gato da rua preto"])

df = pd.DataFrame(X.toarray(), columns=vectorizer.get_feature_names_out())
print(df)

```

```
   da rua  gato da  rua preto
0       1        1          1
```

9. Term Frequency-Inverse Document Frequency (TF-IDF)

Métrica para **ponderar tokens individuais** - reflete a importância de uma palavra num *corpus* (conjunto de textos) As palavras que são frequentes num documento mas não são em outros documentos tendem a ter uma pontuação elevada.



O TF-IDF pondera as palavras pela frequência de aparecimento num documento, mas penaliza as palavras que aparecem em muitos documentos.

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

Frequência do termo = número de ocorrências de i em j (contagem do termo) / (contagem total de palavras no documento)

Frequência inversa de documentos = log (número documentos) / (documento contendo palavras-chave)

```python
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(["gato cão", "gato", "gato da rua"])

df = pd.DataFrame(X.toarray(), columns=vectorizer.get_feature_names_out())
print(df)

```

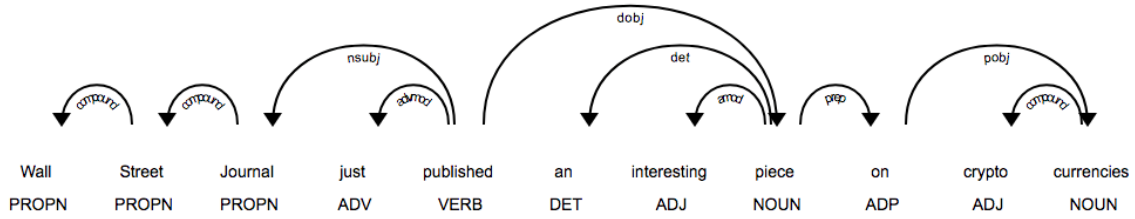
```
        cão        da      gato       rua
0  0.861037  0.000000  0.508542  0.000000
1  0.000000  0.000000  1.000000  0.000000
2  0.000000  0.652491  0.385372  0.652491
```

10. Part of Speech (POS) tagging

…categorização das palavras (corpus) fazendo corresponder o termo lexical a uma determinada parte da frase - **nível mais baixo de análise sintática.**

usam-se "tags" para representar as categorias (tais como substantivos, verbos, pronomes, advérbios, conjunção, adjetivos, interjeição..)

Por exemplo:



Como pode ser feito?

- pesquisa em dicionários
- análise morfológica
- etiquetagem

```
# Exemplo em inglês usando NLTK
import nltk

text = nltk.word_tokenize("The cat ran")
nltk.pos_tag(text)

```

```
[('The', 'DT'), ('cat', 'NN'), ('ran', 'VBD')]
```

Noun (Naming word)
Pronoun (Replaces a noun)
Verb (Action Word)
Adjective (Qualifies noun)

Adverb (Describe a word)
Preposition (Shows relationship)
Conjunction (Joining word)
Interjection (Expressive word)

10. Part of Speech (POS) tagging

- Noun (person, place or thing)
  - Singular **(NN)**: dog, fork
  - Plural **(NNS)**: dogs, forks
  - Proper **(NNP, NNPS)**: John, Springfields
  - Personal pronoun **(PRP)**: I, you, he, she, it
  - Wh-pronoun **(WP)**: who, what
- Verb (actions and processes)
  - Base, infinitive **(VB)**: eat
  - Past tense **(VBD)**: ate
  - Gerund **(VBG)**: eating
  - Past participle **(VBN)**: eaten
  - Non 3rd person singular present tense **(VBP)**: eat
  - 3rd person singular present tense: **(VBZ)**: eats
  - Modal **(MD)**: should, can
  - To **(TO)**: to (to eat)
- Adjective (modify nouns)
  - Basic **(JJ)**: red, tall
  - Comparative **(JJR)**: redder, taller
  - Superlative **(JJS)**: reddest, tallest
- Adverb (modify verbs)
  - Basic **(RB)**: quickly
  - Comparative **(RBR)**: quicker
  - Superlative **(RBS)**: quickest
- Preposition **(IN)**: on, in, by, to, with
- Determiner:
  - Basic **(DT)** a, an, the
  - WH-determiner **(WDT)**: which, that
- Coordinating Conjunction **(CC)**: and, but, or,
- Particle **(RP)**: off (took off), up (put up)

Categorias de classes abertas (substantivos, verbos, adjetivos, advérbios)
Categorias de classes fechadas (preposições, determinantes, pronomes, conjunções, ...)

11. Sentiment Analysis

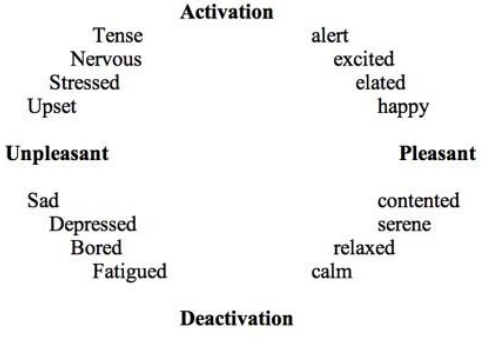
Opinion mining
Sentiment mining
Subjectivity detection

Impressões, não factos

Estudo computacional de sentimentos, opinião, avaliações, atitudes, afetos, emoções, subjetividade, etc.,

Text = Reviews, blogs, discussions, news, comments, feedback ….

expressos em texto



Duas abordagens

- **Baseada no léxico:** contagem do número de palavras positivas e negativas num determinado texto – a maior contagem será o sentimento do texto.

- **Machine Learning**: modelos de classificação treinados para um conjunto de dados pré-rotulados de positivo, negativo, e neutro - ou mais do que isso…:

```
# Usando uma biblioteca como TextBlob (em inglês)
from textblob import TextBlob

text = "I love NLP"
blob = TextBlob(text)
blob.sentiment
```

```
Sentiment(polarity=0.5, subjectivity=0.6)
```

Data Science/Ciência de Dados

5 Estruturação de informação textual

As bases de dados de texto são, em geral, **semi-estruturadas**

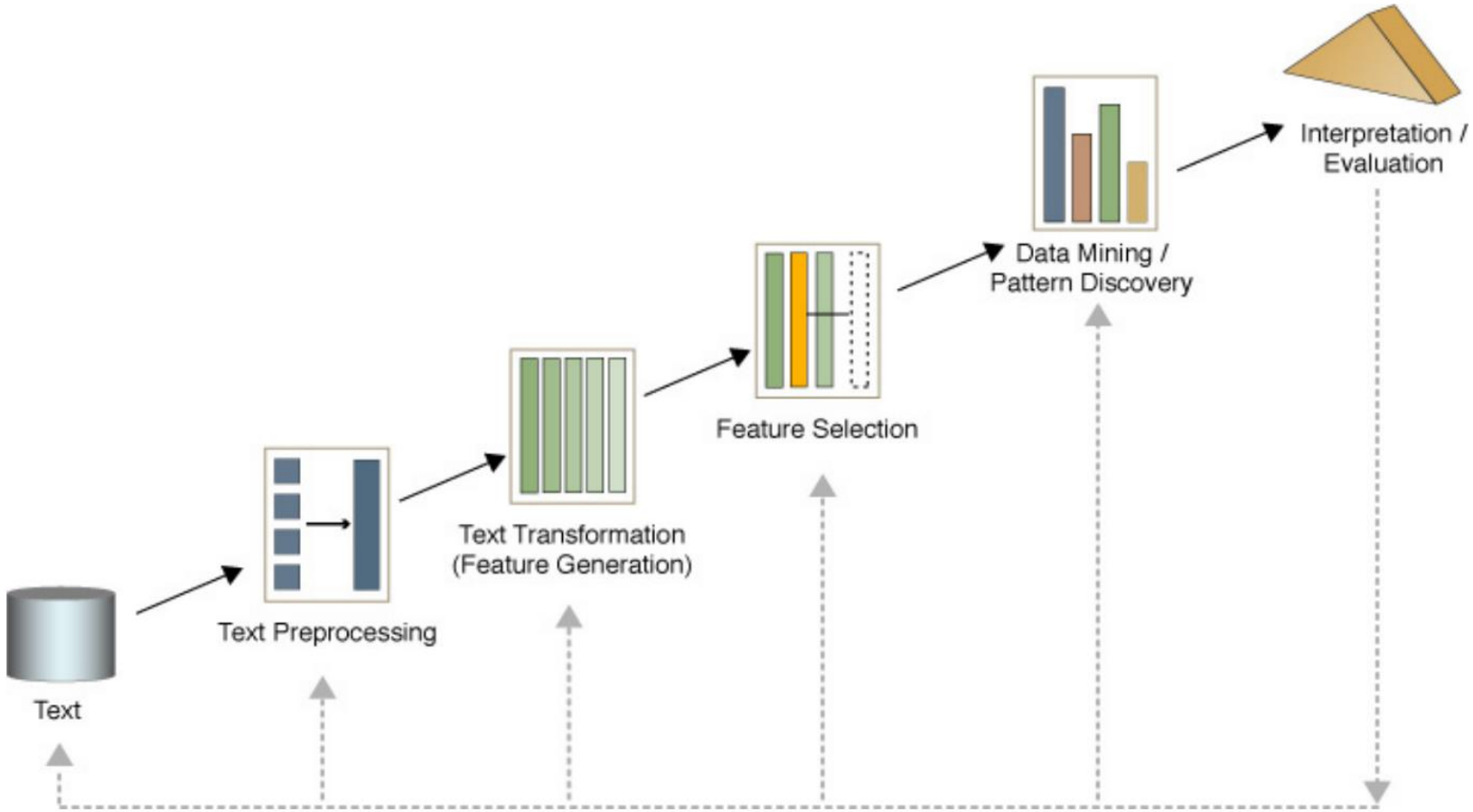
- *Title*
- *Author*
- *Publication Date*
- *Pages*
- *Category*

Atributos/valores estruturados

- *Abstract*
- *Content*

Dados não-estruturados

Processamento de texto



Remove punctuation
Tokenization
Stopwords
Stemming
Lemmatization

Bag-Of-Words
N-Grams
TF-IDF

Contagens simples
Estatísticas

Classificação
Clusterização
Associações

Análise de resultados

Estruturação de informação textual

Como **representar o conteúdo** de um documento para que seja **analisado computacionalmente**?

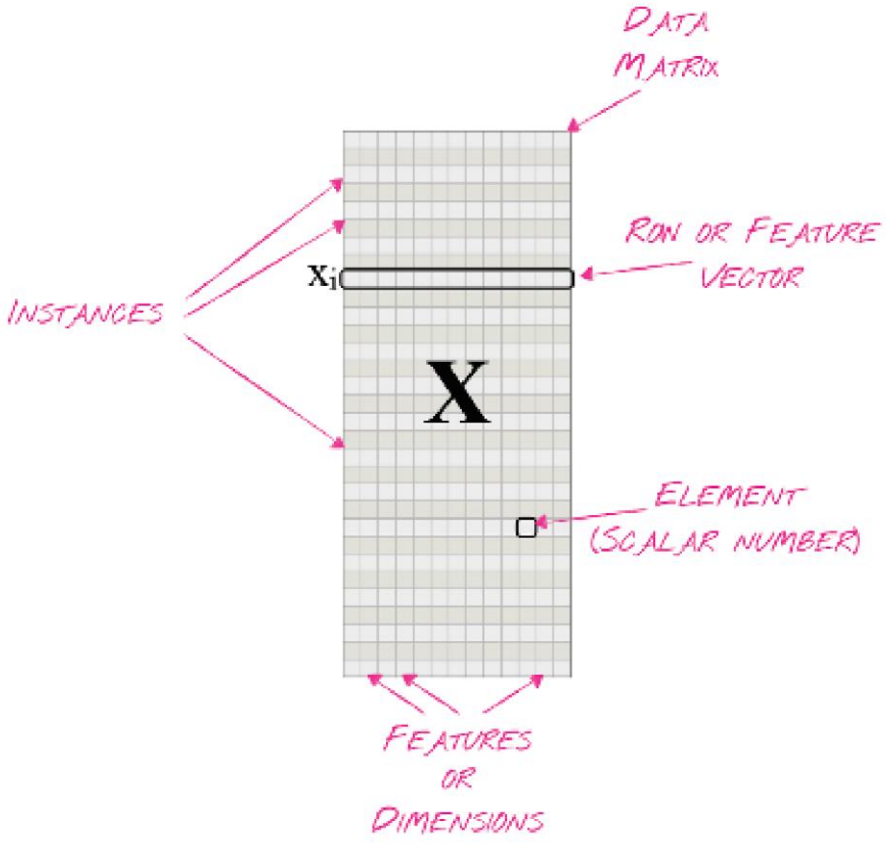
(Precisamos de colocar a informação de forma a que
os computadores possam compreender e manipular)

Usam-se estatísticas para adicionar uma dimensão numérica ao texto não estruturado

Representando os conteúdos num conjunto de atributos - com **base em vetores**

(listas de números com várias propriedades úteis
que os tornam fáceis de trabalhar)

Uma **representação documental** tem como **objetivo** perceber do que é que documento trata



(existem muitos métodos para analisar dados estruturados - já vimos alguns)

(Uma abordagem possível)

Cada entrada descreve um documento

Atributo **descrever se um termo aparece ou não no documento**

Exemplo: Tabela de valores boleanos

Nominal

Características (features) - n-gramas?

ID do document/texto/parágrafo?

	Termos				
	Data	**Science**	**Social**	**University**	…
Documento 1	1	1	1	0	…
Documento 2	1	1	0	1	…
…	…	…	…	…	

Vector de um documento

(Outra abordagem)

Cada entrada descreve um documento

Os atributos **representam a frequência com que um termo aparece no documento**

Exemplo: Tabela de frequência de termos (relativos)

Numérico

	Termos				
	Data	**Science**	**Social**	**University**	…
Documento 1	5	7	4	0	…
Documento 2	3	2	0	2	…
…	…	…	…	…	

(ainda outra abordagem)

…por vezes um termo é mencionado mais vezes em documentos mais longos

Assim, utilizam-se as **frequências relativas (% do documento):**

Exemplo: Nº de ocorrências/ Nº de palavras no documento

Numérico

	Termos				
	Data	**Science**	**Social**	**University**	**…**
Documento 1	0.005	0.04	0.007	0	…
Documento 2	0.04	0.05	0	0.04	…
…	…	…	…	…	

(ainda mais outra…)

Se um termo é frequente em muitos documentos, não tem poder discriminatório

Ponderação TF-IDF: dar maior peso a termos que são raros

TF: frequência dos termos no documento (peso dos termos frequentes)
IDF: frequência de termo inverso

Exemplo: Nº de ocorrências/ Nº de palavras no documento

Numérico

	Termos				
	Data	**Science**	**Social**	**University**	**…**
Documento 1	0.012	0.013	0.3	0	…
Documento 2	0.014	0.016	0	0.5	…
…	…	…	…	…	

Data Science/Ciência de Dados

6 Text mining

Classificação de documentos

Agrupamento de documentos

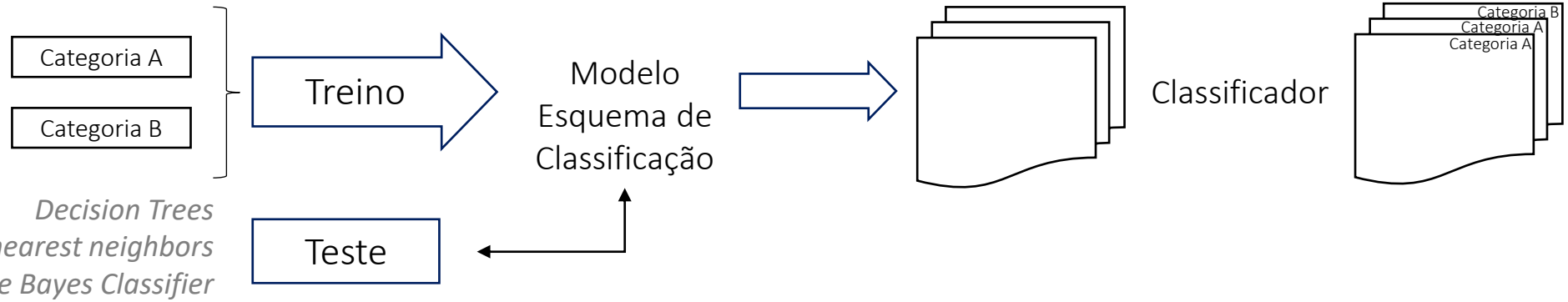
Regras de associação baseadas em palavras-chave

Classificação

Supervisionados

Utilizam dados de treino para generalizar padrões

A “mão do homem” classifica um conjunto de documentos
conjunto de dados de formação
Induza um modelo de classificação



Decision Trees
K-nearest neighbors
Naïve Bayes Classifier
Neural networks,
Support Vector Machines

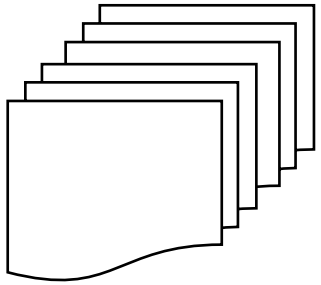
	Termos					
	Data	**Science**	**Social**	**University**	**…**	***Nível de Relevância***
Documento 1	0.012	0.013	0.3	0	…	*Relevante*
Documento 2	0.014	0.016	0	0.5	…	*Não relevante*
…	…	…	…	…	…	

Agrupamento

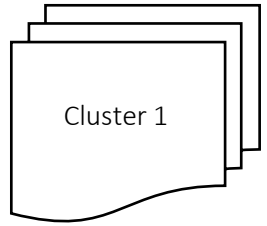
Não-supervisionados

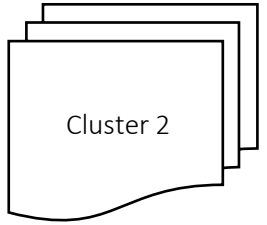
regras genéricas de algoritmos que se aplicam diretamente

Encontrar grupos de documentos semelhantes



Métodos de Clusterização





Métodos de partição:
k Means

Métodos Hierárquicos:
Aglomerativos ou Divisivos

	Termos					
	Data	**Science**	**Social**	**University**	**…**	***Classificação***
Documento 1	0.005	0.001	0.007	0	…	*?*
Documento 2	0.04	0.03	0	0.04	…	*?*
…	…	…	…	…		

Não-supervisionados

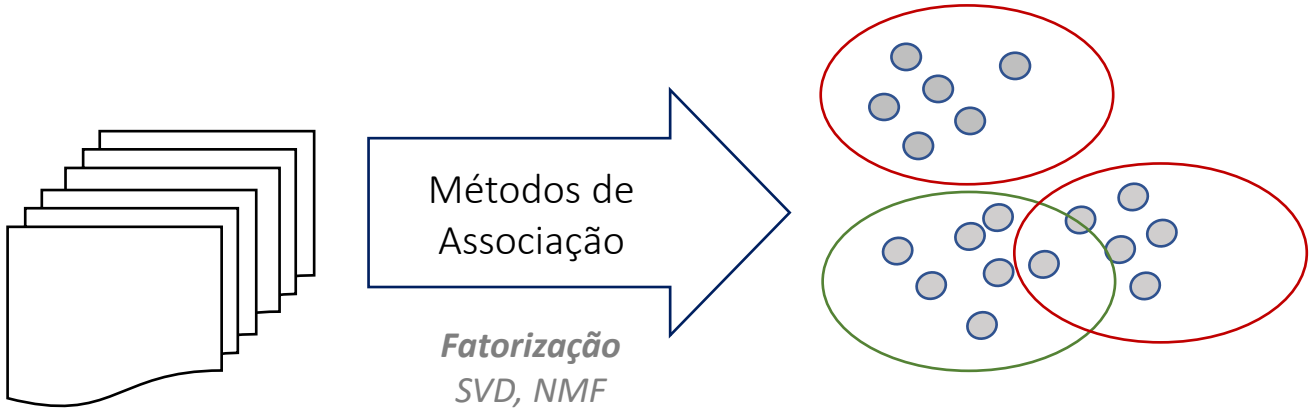
regras genéricas de algoritmos que se aplicam diretamente

Associação

Encontrar grupos de termos (Redução de dimensionalidade)

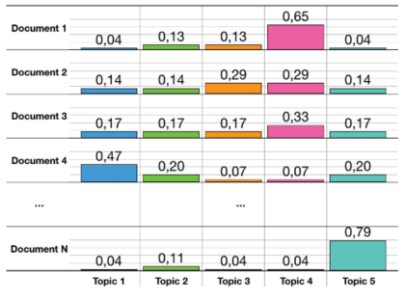
Descobrir padrões temáticos ocultos através de regularidades estatísticas

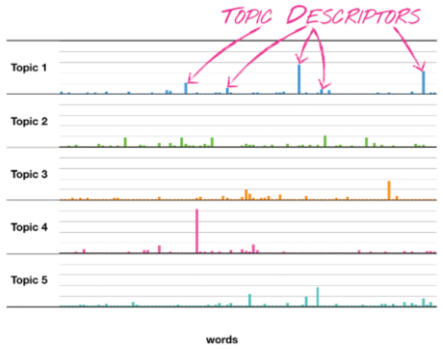
Objectos ≡ n gramas traduzidas por dimensões latentes



Fatorização
SVD, NMF

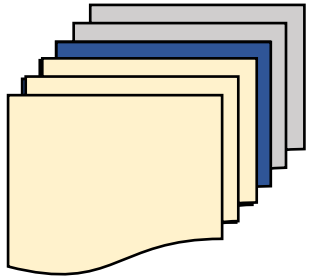
Topic models
LSA, NMF, LDA





Grupo de palavras por tópicos

Grupos de documentos por tópicos



Data Science/Ciência de Dados

7 … em suma

Ferramentas

Desafios

Bibliotecas (populares) NLP

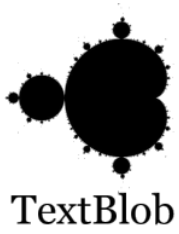
NLTK

NLTK - the Natural Language Toolkit - is a suite of open source Python modules, data sets, and tutorials supporting research and development in Natural Language Processing.

spaCy

spaCy is a free open-source library for Natural Language Processing in Python. It features Named Entity Recognition, Part of Speech tagging, dependency parsing, word vectors and more.

TextBlob



TextBlob is a Python library for processing textual data. It provides a simple API* for diving into common Natural Language Processing tasks.

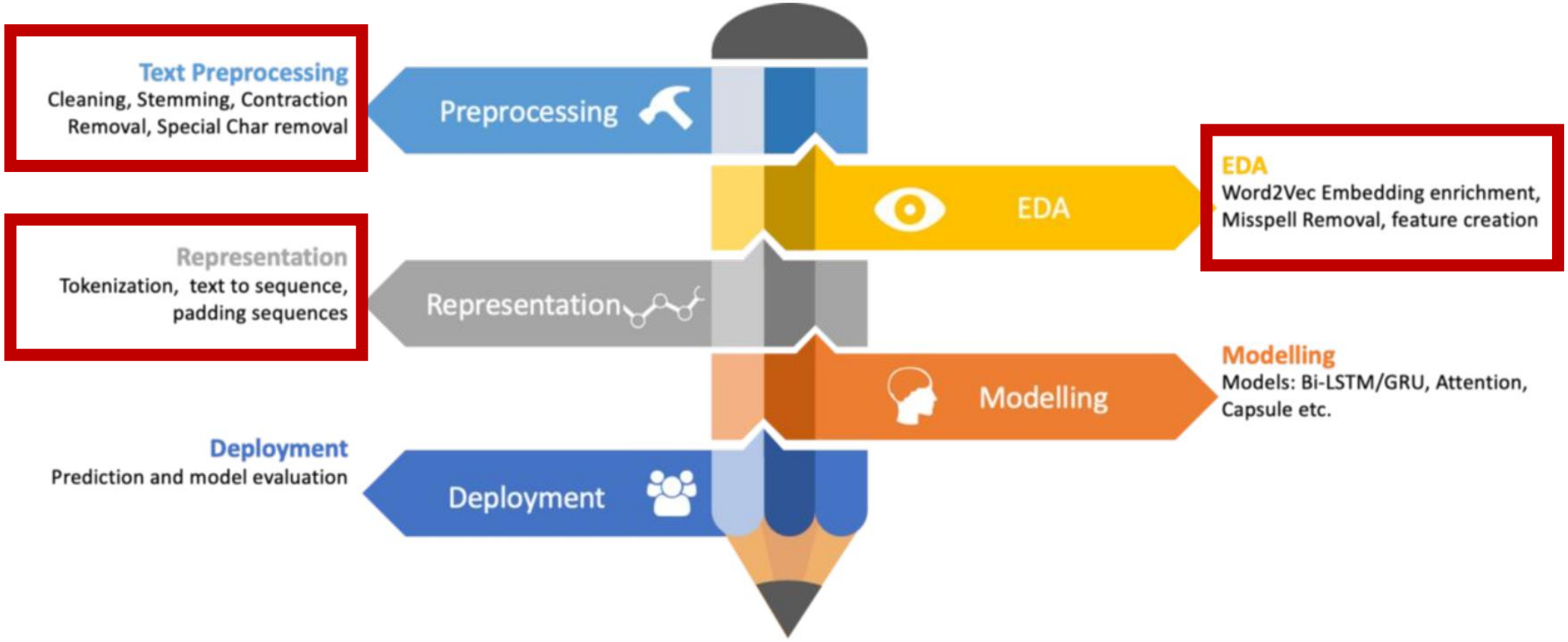
RASA

RASA is an open-source ML framework to automate text and voice-based conversations.

Hugging Face

Hugging Face is a platform that provides the community with application programming interface to access and use state-of-the-art pre-trained models available from the Hugging Face hub.

NLP Pipeline



Desafios

Palavras de contexto contextuais, frases e homónimos - As mesmas palavras e frases podem ter significados diferentes de acordo com o contexto de uma frase e de muitas palavras.
Por exemplo:

Ironia e Sarcasmo e Ambiguidade - Os seres humanos geralmente usam palavras e frases que podem ser positivas ou negativas de acordo com o dicionário, mas que na realidade significam o oposto.

Coloquialismos e gíria - Frases informais, expressões, expressões idiomáticas e linguagem específica de uma cultura apresentam inúmeros problemas para os modelos de NLP

Linguagem específica de uma área ou setor - Diferentes empresas e indústrias utilizam frequentemente linguagens muito diferentes para as suas actividades e operações.

Desafios

Contextual meaning and homonyms

Some words and phrases are based on the context of the sentence.
For example: Can I **run** something past you real quick or the house is looking really **run** down.

Ambiguity

- Lexical ambiguity
- Semantic ambiguity
- Syntactic ambiguity

Evolving Language

Although human languages are rule-based to some level, these rules have many exceptions and are prone to change over time.

Tone of Voice

The intonation of voice can vary from region to region and dialect to dialect, which can make it more difficult for machines to understand.

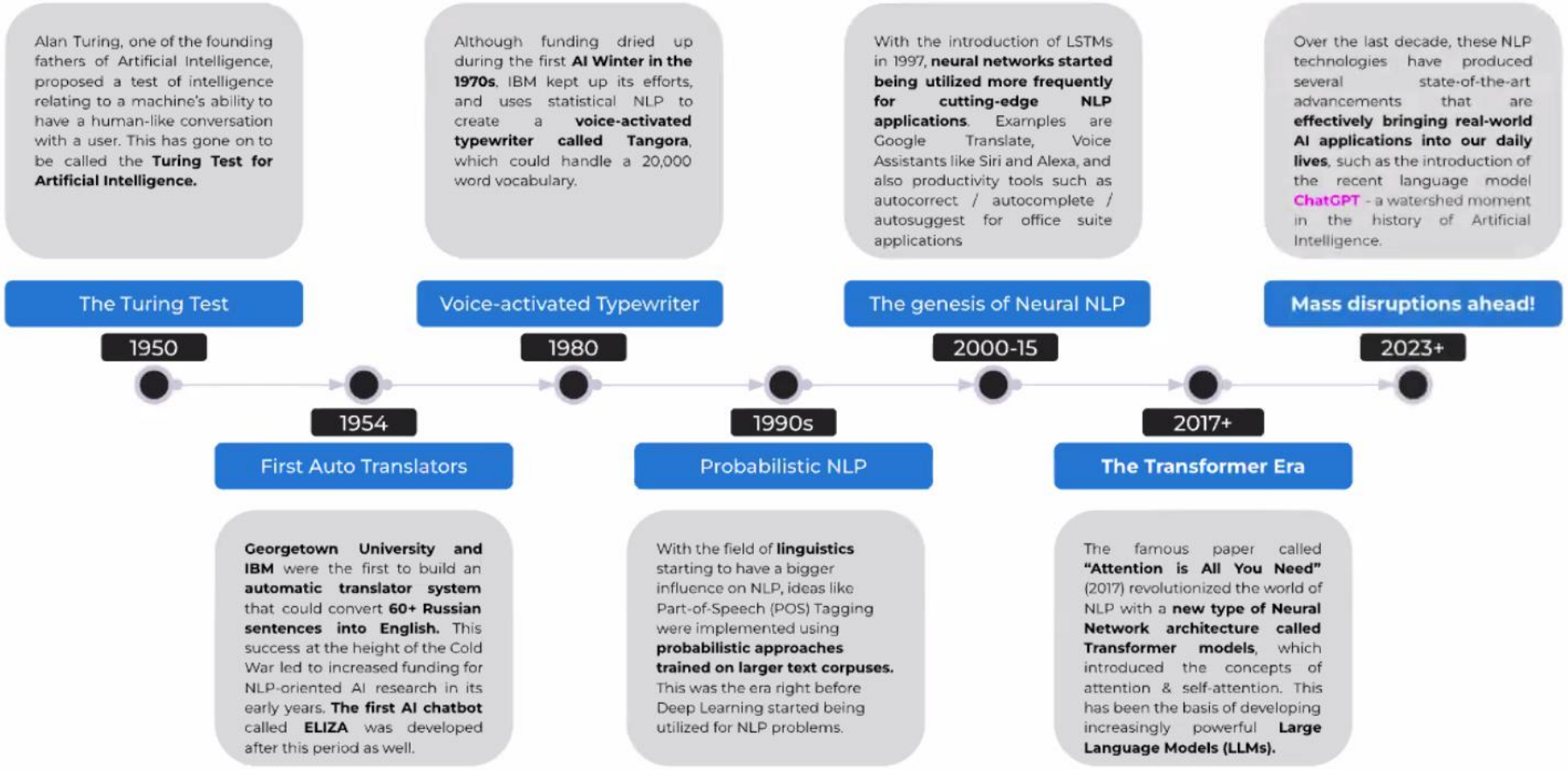
Precision

Human speech is ambiguous, and linguistic structures can be impacted by various factors such as slang, regional dialects, social context or colloquial usage of words.

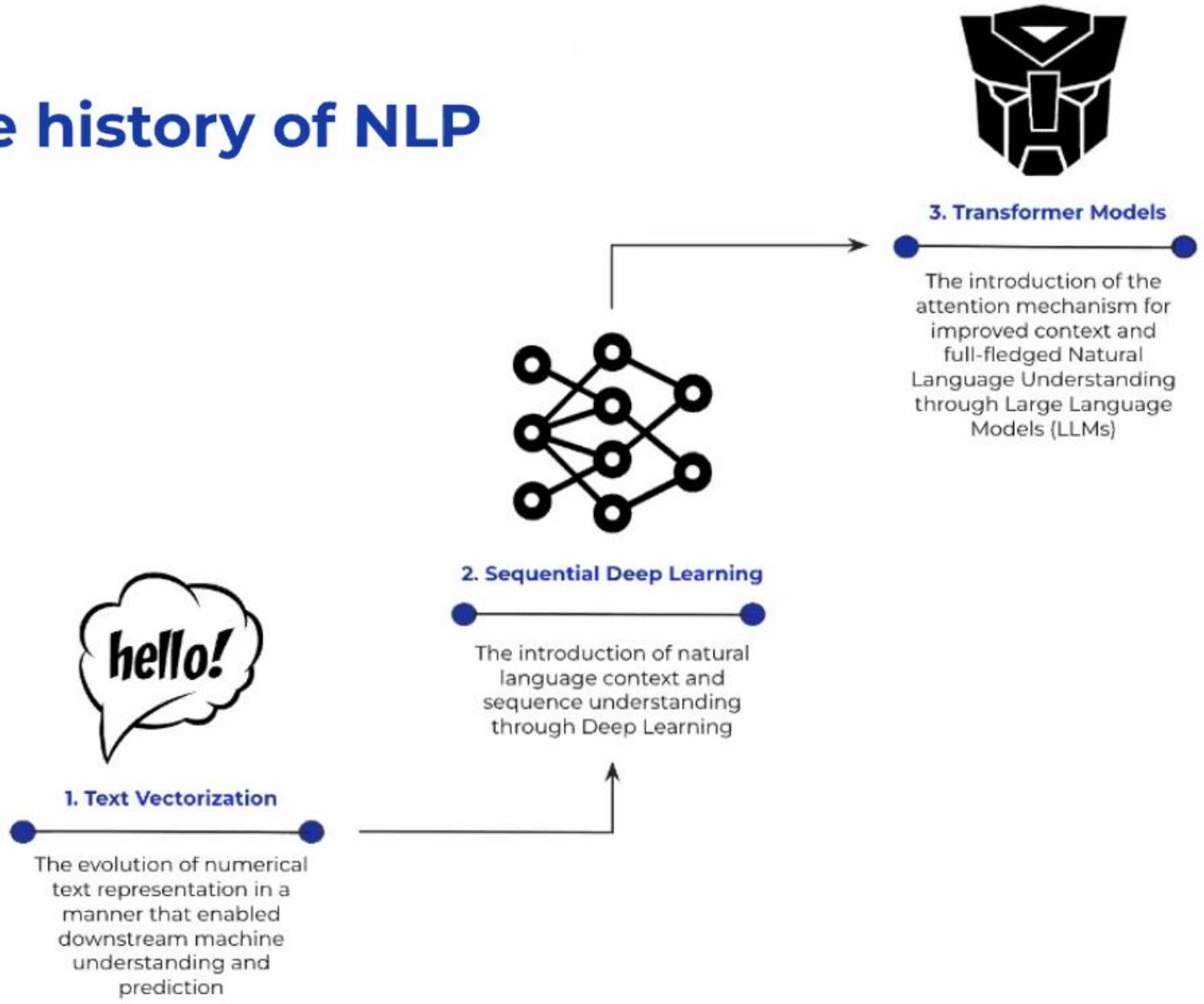
Errors & abbreviations

Errors in text such as incorrect spelling, incorrect use of tenses along with the use of abbreviations and slangs like LOL and ROFL make NLP tasks more challenging for computers.

The history of NLP



The history of NLP



Data Science

Referências

[Text Analysis]

https://www.nltk.org/