CORE147

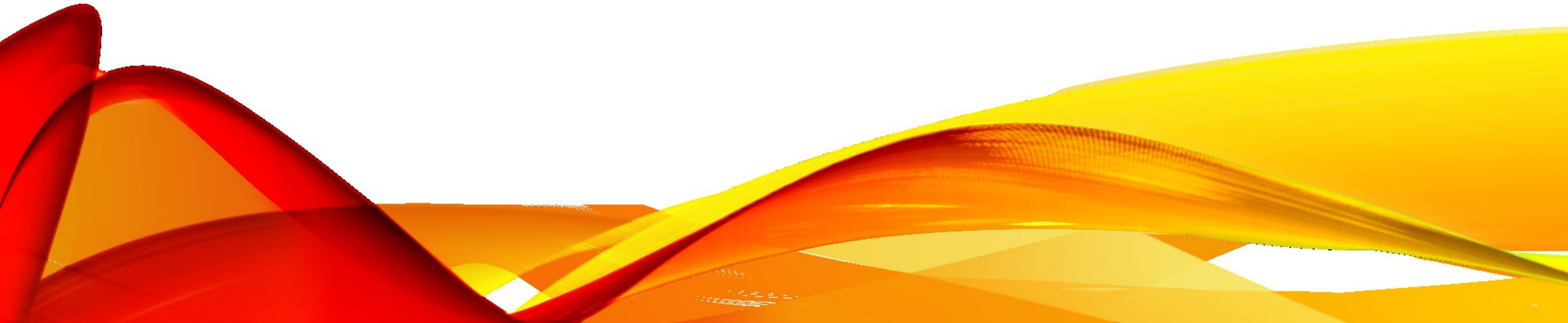
DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

TEXTOVÁ DATA A PŘEDZPRACOVÁNÍ

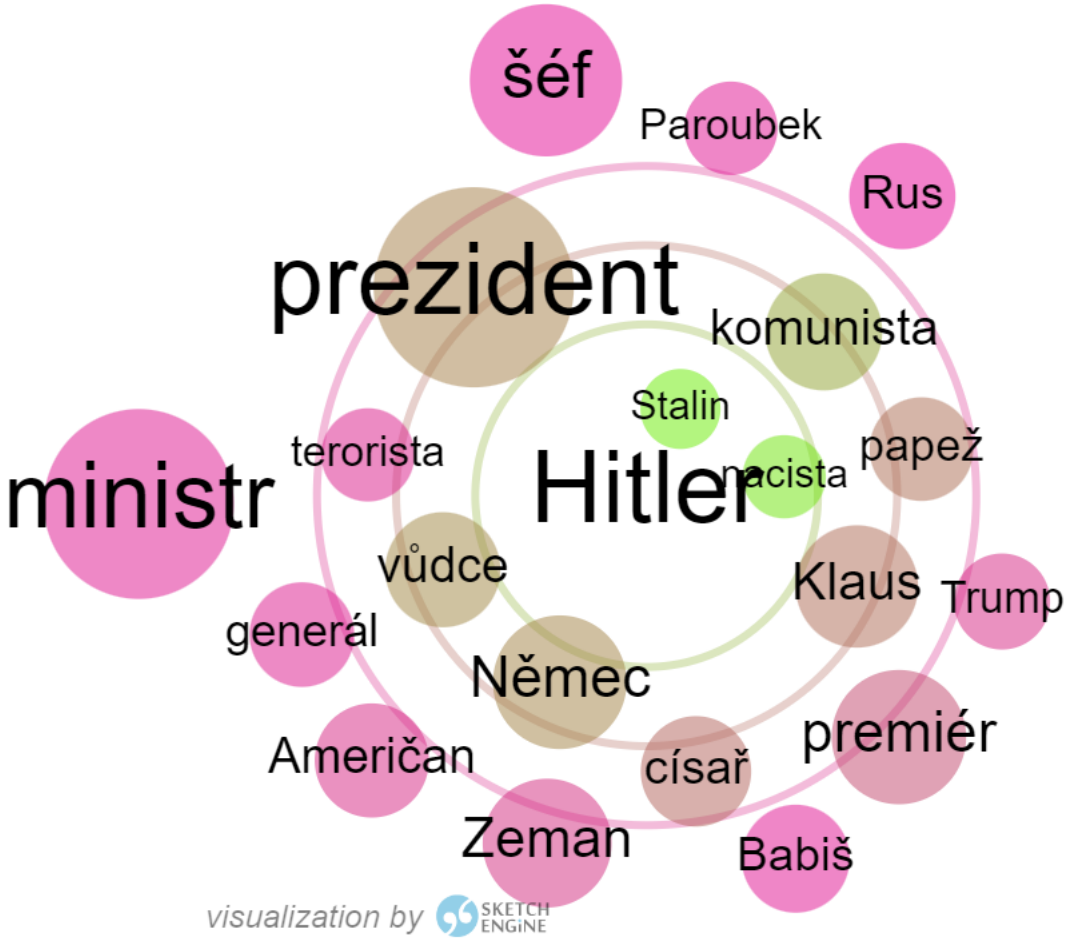
Text data preprocessing



TEXTOVÁ DATA

- **Korpus** = soubor textů, anotací a metadat
- **Texty** = souvislé, vyprodukované lidmi, případně jako výsledek přepisu mluveného slova
- **Anotace** = označkování jako výsledek analýzy textu, např. segmentace na slova a věty
- **Metadata** = informace o jednotlivých dokumentech

zmrazit ús
esvědčeni
rů! Třicátél
měnit. O tu
marskou r
e. A při výs
a Háchy u
u s Němci
čně. Křest

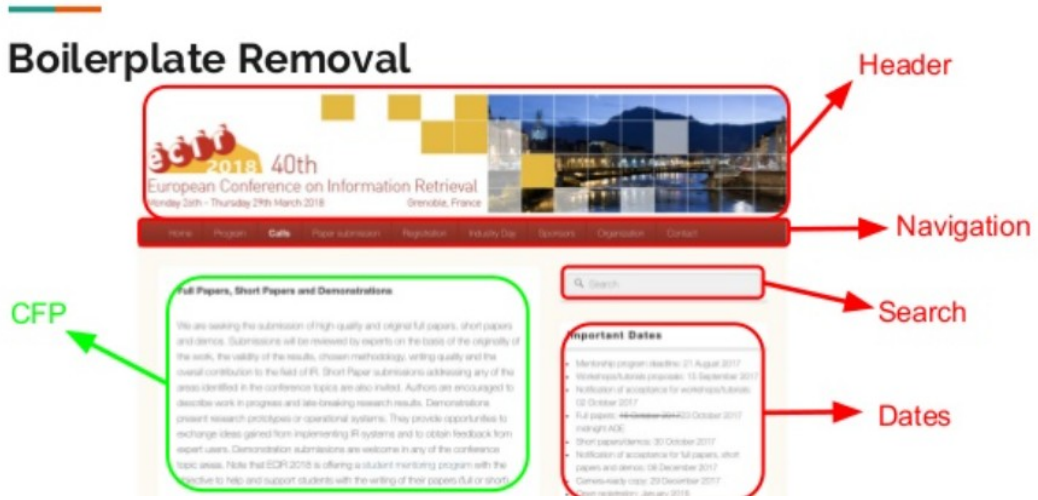


PŘEDZPRACOVÁNÍ TEXTŮ: AKTIVITY

1. Extrakce textu
2. Sjednocení kódování
3. Čištění: odstranění nežádoucího obsahu (kdo pozná, co je nežádoucí)
4. Rozdělení na kapitoly, sekce, odstavce, nadpisy, perexy, ...
5. Rozdělení na věty
6. Ponechat velká a malá písmena jako v originále?
7. Tokenizace
8. Stemming/lemmatizace
9. POS-tagging (určování slovních druhů a dalších gramatických kategorií)
10. Odstranění stop slov

EXTRAKCE TEXTU

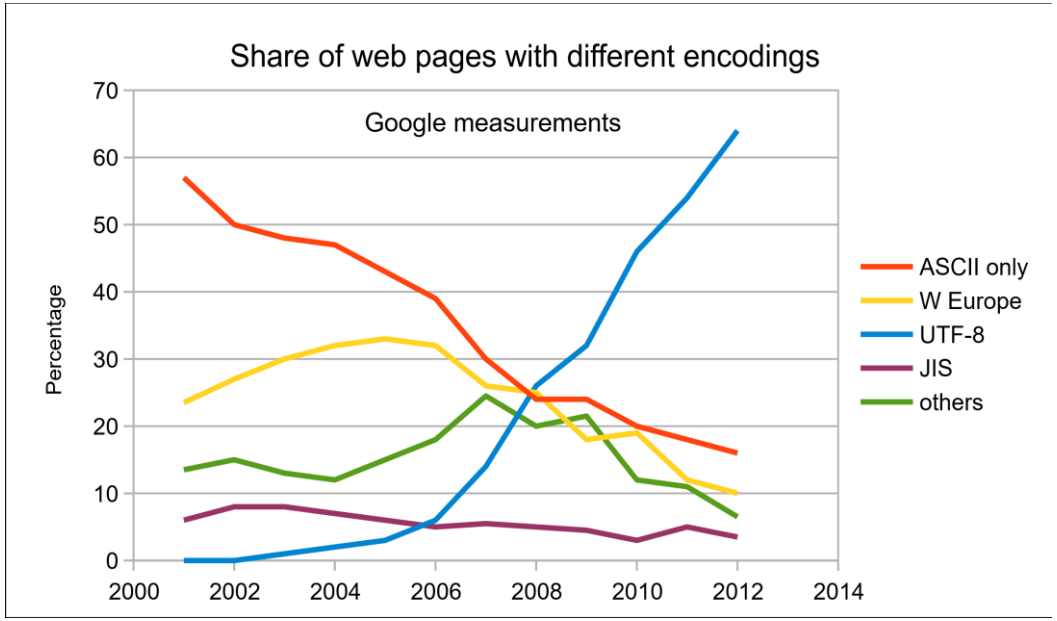
- OCR
- Konverze podle formátu
  - Apache Tika (PDF, Office)
  - pdf2text (PDF)
  - justText (HTML)
- Specializované (netextové) formáty
  - XML
  - Databáze



https://www.slideshare.net/eickhoff/web2text-deep-structured-boilerplate-removal

ZNAKOVÁ SADA A KÓDOVÁNÍ
CHARACTER SET AND ENCODING

- Tradiční znaková sada ASCII
  - 7bit = 128 znaků
- Unicode = univerzální znaková sada
- **Kódování = kód bajtů, které odpovídají znaku ve znakové sadě**
- Pro češtinu existovalo více než 5 různých kódování (zmatek)
- Od začátku století začíná převažovat UTF-8
  - 1,112,064 znaků
  - 1-4 bajtů/znak



ČIŠTĚNÍ DAT

Odstranění šumu:

- V případě OCR pixely omylem rozpoznané jako znaky

Odstranění nadbytečné informace:

- Čísla stran
- Záhlaví, zápatí?
- Dělení slov?

Kdy a jak čištění dělat?

- Záleží na cíli projektu
- Replikovatelná procedura je lepší

- Rozložení dokumentu (layout)
  - Pozice
  - Mezery
  - Velikost písma
- Význam textu (implicitní)
  - Jména osob = autoři
  - Bloky: nadpis, perex, abstrakt, popis pod obrázkem...

ORIGINAL ARTICLE

Correlation between tau phosphorylation sites and tangle morphology in Alzheimer's disease

Masao SHIMAZAKI,[1,2] Hiroyuki NAKANO[1,3] and Katsuji KOBAYASHI[1]

[1]Department of Psychiatry and Neurobiology, Kanazawa University Graduate School of Medical Sciences, Takaramachi, Kanazawa, [2]Department of Neuropsychiatry, Hokuriku National Hospital, Nobusue, Nanto, Toyama and [3]Department of Psychiatry, Takaoka City Hospital, Takaramachi, Takaoka, Japan

Correspondence: Dr Masao Shimazaki, Department of Psychiatry and Neurobiology, Kanazawa University Graduate School of Medical Sciences, 13-1, Takaramachi, Kanazawa 920-8641, Japan. Email: simazaki@hokuriku.hosp.go.jp

Received 03 September 2004; accepted 14 January 2005.

Abstract

Background: To study the relationship between phosphorylation sites of the tau protein and changes in the configuration of neurofibrillary tangles (NFT) in brains with Alzheimer's disease (AD).

Methods: Seven brains from patients with non-familial AD and three control brains were examined. NFT were immunolabeled with five monoclonal antibodies against phosphorylated serine (pSer) and threonine (pThr): AT180 targets pThr231; AT8 targets pSer202 and pThr205; HT7 targets pSer159–163; Tau2 targets pSer101; and Tau5 targets the central region of the tau sequence. Tau-labeled NFT (tNFT) were grouped into pretangles (p-NFT), intracellular tangles (i-NFT) and extracellular ghost tangles (e-NFT) according to their cytological features. Gallyas-stained NFT (gNFT) were regarded as the NFT population of AD. The cerebral regions examined included the cornu ammonis (CA), entorhinal cortex, anterior cingulate cortex, inferior parietal lobe, temporal neocortex, occipital cortex and lateral prefrontal lobe. The first three regions were grouped into the limbic cortex and the others into the association cortex.

Results: p-NFT were observed with all tau labeling, with the p-NFT/tNFT percentage ranging from 14% to 19%, and densities of p-NFT and i-NFT were higher in the limbic cortex than in the association cortex. The p-NFT/tNFT ratio in Tau2 labeling was the lowest compared with the other tau labeling. e-NFT showed an irregular density distribution across both cerebral regions and different tau labeling, and the densities of p-NFT and e-NFT were correlated with that of tNFT. In contrast, e-NFT were correlated with p-NFT in AT180 and Tau5 labeling.

Conclusion: The higher densities of p-NFT and i-NFT in the limbic cortex compared to the association cortex are compatible with the regional extension of NFT. pSer101 is closely associated with the late process of p-NFT generation, and pSer202 and pThr205 are presumed to be triggers for the generation of p-NFT. Phosphorylation of pSer101 detected by Tau2 plays an important role in the formation of e-NFT and p-NFT.

Key words: *Alzheimer's disease, AT180, AT8, extracellular ghost tangle, HT7, intracellular tangle, pretangle, tau protein, Tau2, Tau5.*

INTRODUCTION

Neurofibrillary tangles (NFT) form the core of the pathological substrates as well as senile plaques (SP) in Alzheimer's disease (AD). NFT are composed of paired helical filaments (PHF) caused by the abnormal and excessive phosphorylation of the microtubule-associated protein, tau.[1–4] NFT occur initially in the neurons of the entorhinal cortex, extend to the limbic cortex and finally become distributed in the association cortex.[5] Three types of NFT can be recognized morphologically: pretangles (p-NFT), intracellular tangles (i-NFT) and extracellular tangles (e-NFT). The

ROZDĚLENÍ NA VĚTY
SENTENCE SEGMENTATION

- Sentence splitter
  - ".", "!", "?" jsou velmi mnohoznačné
  - Jednoduchá pravidla nejsou 100%
- Existující software
  - NLTK
  - SpaCy
  - Syntok
- Segmentace textu bez interpunkce (např. přepis mluveného slova)
  - DeepSegment

I saw Mr. Smith.

I have a 5.4 m carpet.

The winner of the Jeopardy! contest was …

WTF? is an amazing book

I am Batman I live in Gotham

UPPERCASE/LOWERCASE

- Na jednu stranu užitečné:
  - „Strom“ i „strom“ znamenají totéž
- Na druhou stranu ztráta informace:
  - Autor textu si kvůli něčemu malá a velká písmena vybral
  - Ale jaká je ta informace?

teXT caN EasIlY Be CoNveRtEd frOM uPPeRCaSe or LoWeRcAse To stUDlycAPs.

TOKENIZACE
TOKENIZATION

- CJK (Chinese-Japanese-Korean) – nepoužívají mezery – tokenizace je obtížná
- Ostatní jazyky – poměrně snadný úkol...
- ... až na interpunkci
- ... and data, e-maily, IP adresy, ... ☺

はじめまして。ズザナと言います。

The quick brown fox jumped over a lazy dog.

The, quick, brown, fox, jumped, over, a, lazy, dog, .

There, is, no, place, like, 192.168.0.0, .

I, did, n't, do, that, !

STEMMING

Nalezení kmenu slova **stem** (inflectional root):

- Odstranění předpon (prefixů)
  - Nejpovedenější - vedenější
- Odstranění přípon (sufixů)
  - veden
- Odstranění koncovek
  - ved
- Sjednocení morfologických variant
  - vést - ved

Výhoda

Seskupení slov se společným kmenem k sobě.

Stemming je ***nezávislý na kontextu***

- *lepší = lepš*
- *ženu = žen*
- *hnát = hná*

LEMMATIZACE
LEMMATIZATION

Základní tvar slova

- Konvence: první pád, jednotné číslo, u sloves infinitiv
- Řada nejasností
  - Zápor: Nejsem = být/nebýt?
  - Stupňování: Lepší = dobrý?
  - Pomnožná slova (pluralia tantum): tepláky = tepláky?

Stemming

Seskupení slov se společným kmenem k sobě.

Lemmatizace

Seskupení různých gramatických tvarů téhož slova k sobě.

*Lemmatizace je **kontextově závislá**:*

Ženu = žena? Hnát?

Červenej = červený? Červenat?

ZNAČKOVÁNÍ
TAGGING

- Morfologická analýza = určení lemmatu a gramatických kategorií
- Morfologická analýza je mnohoznačná

like	**flies**
Verb	Verb
Prep	Noun

- V kontextu se (často) zjednoznační.

Time flies like an arrow.

Fruit flies like a banana.

Tagger

- První taggery založené na pravidlech
- V současnosti založené na neuronových sítích

Tagging

- Pro řadu úloh stačí slovní druh
- Některé taggery poskytují i další informace

Známé taggery

Stanford POS Tagger, TreeTagger, SpaCy

ODSTRANĚNÍ STOP SLOV
STOPWORD REMOVAL

- Česká stop slova: a, v, se, na, je, že, o, s, z, do, i, to, k, ve, pro, za, by, ale, si, po, jako, podle, od, jsem, tak, jsou, které, který

- English stopwords: the, of, and, a, in, to, it, is, to, was, I, for, that, you, he, be, with, on, by, at, have, are, not, this, ‘s, but, had, they, ... too

Představ si to, že tak v pátek jsem si to koupila.

představ, pátek, koupila

Nick likes to play football, however, he is not too fond of tennis.

Nick, likes, play, football, however, fond, tennis

PŘEDZPRACOVÁNÍ: OBECNÁ DOPORUČENÍ

Reprodukovatelný postup:

- Co nejvíc automatizace (hotový software)
- Tam, kde to nestačí, tak skripty (závislé na datech)
- Tam, kde je potřeba ruční zásah, uložit i nedotčenou verzi
- U ručních zásahů / anotací mít dokument, který postup popisuje (např. Anotační manuál)

DALŠÍ ČTENÍ

- Matthew Mayo: ***A General Approach to Preprocessing Text Data***. KDnuggets, 2017. https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html
- Kavita Ganesan: ***All you need to know about text preprocessing for NLP and Machine Learning***, KDnuggets, 2019. https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html
- Text Analysis Portal for Research. http://tapor.ca/
- Methodica Commons: Digital Text Methods. http://methodi.ca/