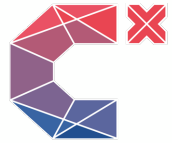


CORRELAID

GOOD CAUSES. BETTER EFFECTS.

Text Data Visualisations

CorrelAid X Austria

Tilman Kerl

Who Am I?

- Tilman Kerl
- Master Data Science @ TU Wien
- Computer Science @ Uni Konstanz
- CorrelAid since 2019
- Thesis on Visual Analytics for Transformer models

- I own very fashionable hats
- (and I don’t know where to put my hands during photos)

Scope & Agenda for today

Text & Language

- Why is it important?
- What is it?
- Why is it difficult to work with?

Visualizing Text

- What is possible
- Feature Extraction

Code Examples

Who are you?

Slido: What is your background?

http://bit.ly/3FanvjI

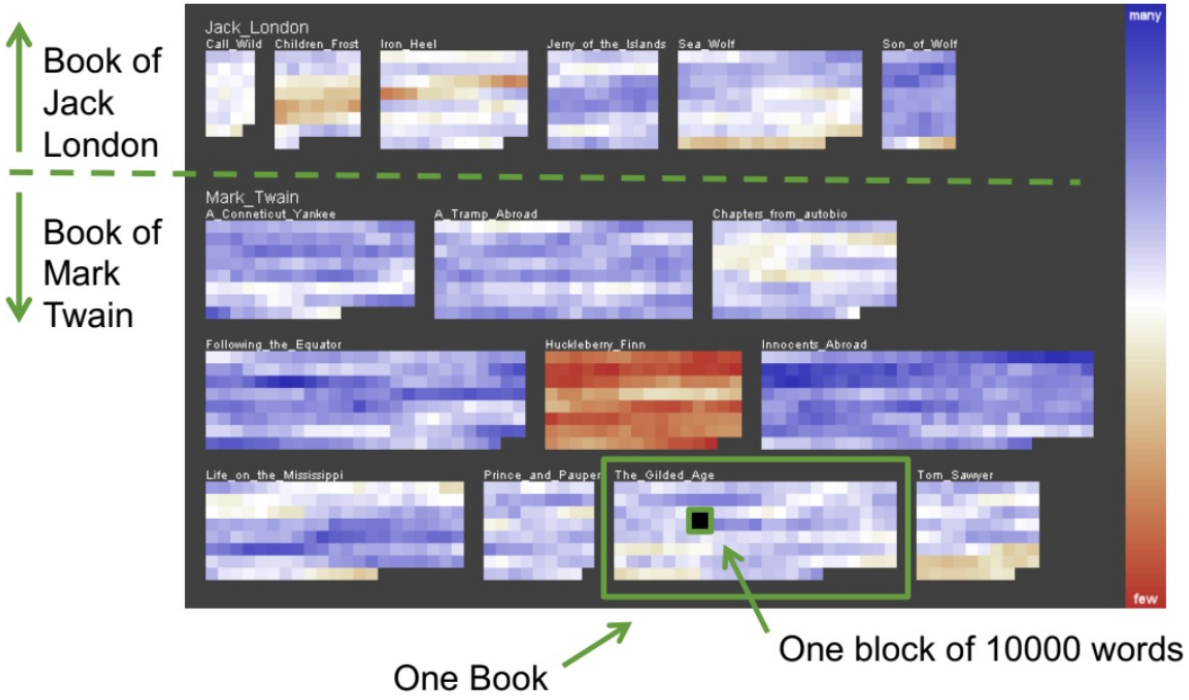
Part 1

What is Language and Text?

Why is text data important?

- Text data is everywhere: Reviews, Social Media, Newspaper, Chats, Political documents
- We communicate with text
- There is (most of the times) a lot of information in text data about
  - what we want
  - how we feel
  - how we think
  - potential plagiasm
  - ….

Example Use Case: Literature Fingerprinting



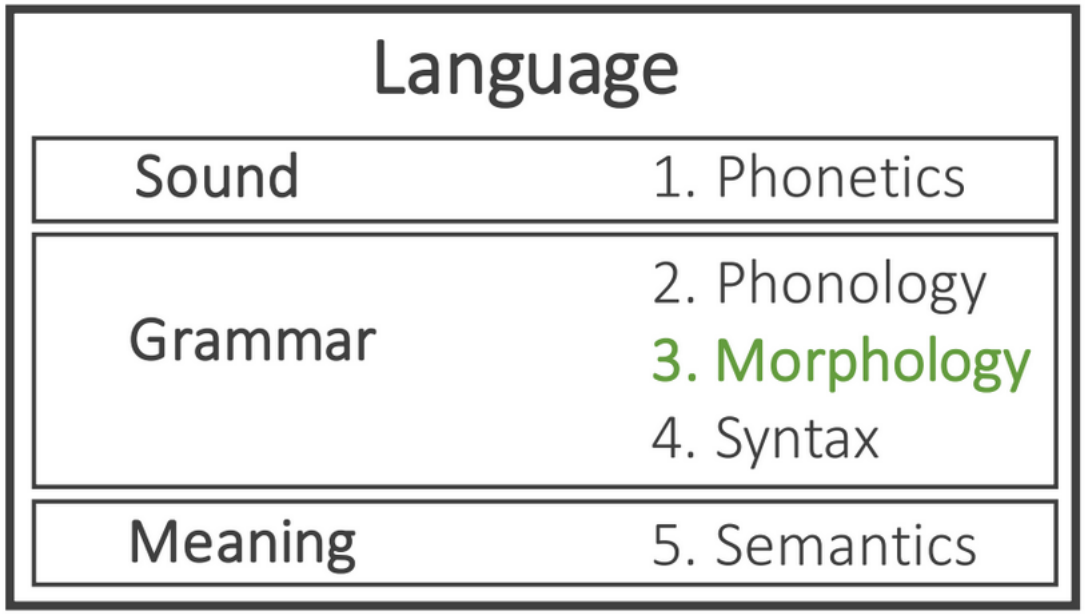
- characterize a text and a writing style
- assign and obtain “fingerprints” of an author
- Authorship Attribution
- Plagiarism Checks

- We can spot differences between the two authors
- Something special about “The Adventures of Huckleberry Finn”
  → Reason unclear, maybe Ghost Writer?

See: Literature Fingerprinting: A New Method for Visual Literary Analysis: https://ieeexplore.ieee.org/abstract/document/4389004

What is text?

What is language?



Morphology

- The study of the way words are built up from smaller meaning units

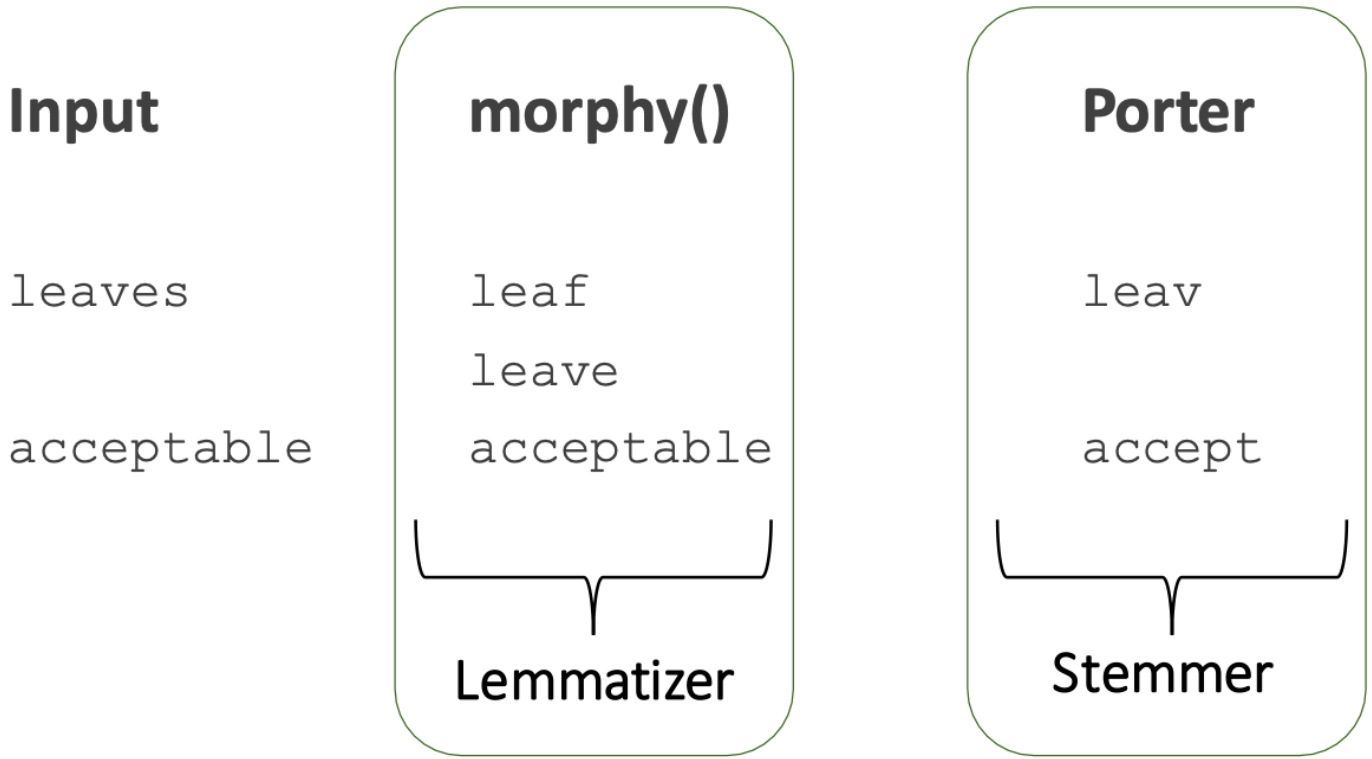
Morphemes

- The smallest meaningful unit in the grammar of a language
- Root, Stem, Lemma

Stemming & lemmatization

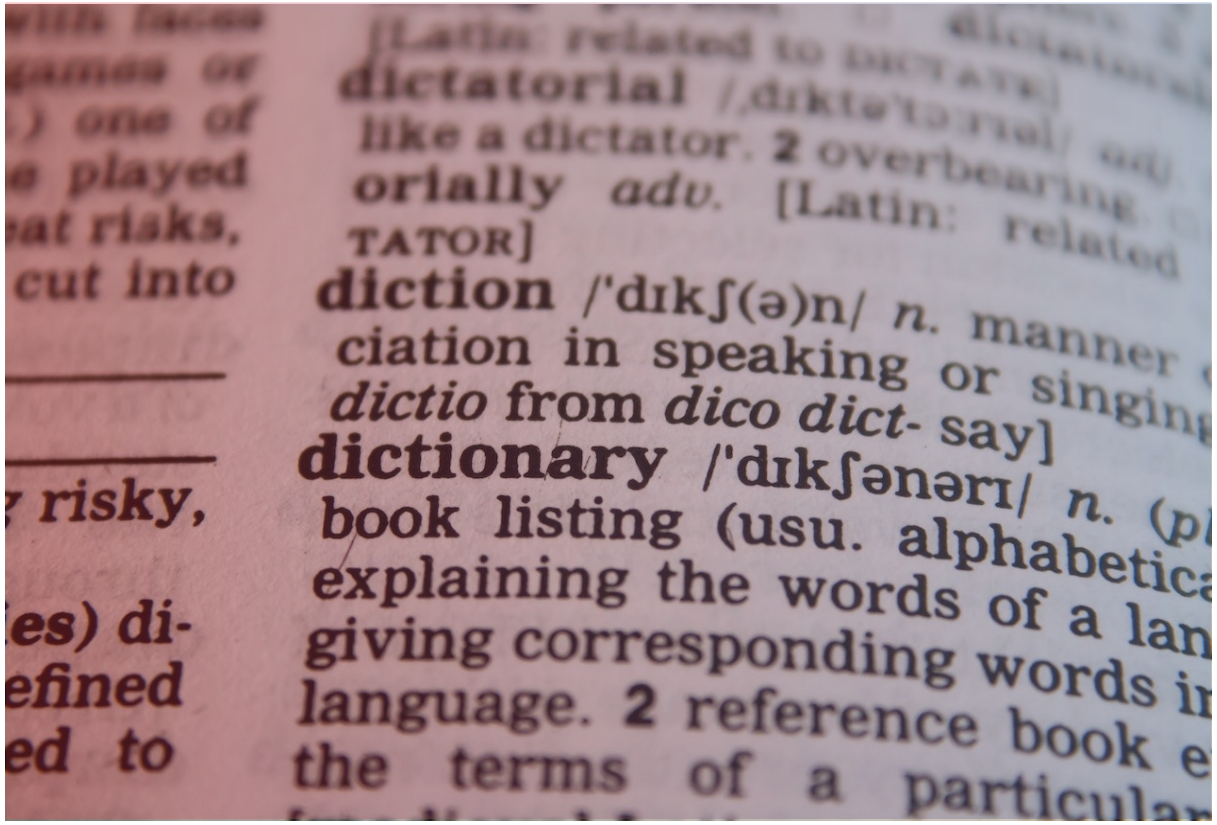
- Different Approaches to the problem
- morphy() vs. Porter Stemmer

morphy() vs Porter Stemmer



* Slide from the Computational Methods for Document Analysis by Prof. Daniel A. Keim, University Konstanz (2019)

Why is text so difficult to work with?



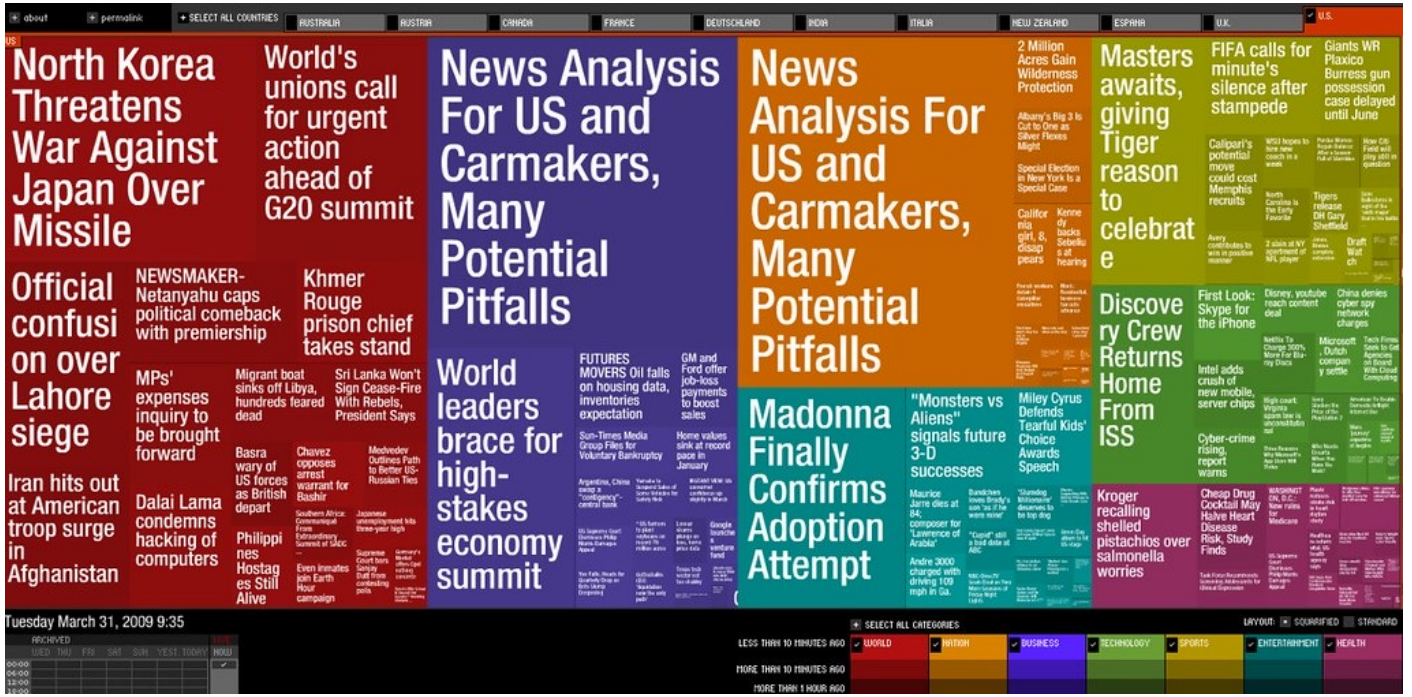
- Language has multiple layers which are difficult to understand for machienes
- Working, worked, works refer to the same action but are different words (stemming)
- Human concepts as sarcasm and irony are hard to grasp, even for humans sometimes
- How we percive text and spoken language is also influenced by our mood (“Just relax”)
  --> the same sentence or word can have multiple meanings
- Punctuation matters (Let’s eat (,) grandpa)
- Context matters
- We have 100+ languages

Part 2

Visualizing Text

How can we visualize text data?

- Text itself has limits on how we can visulize it
- Some basic visualizaitons include:
  - Keywords over time
  - Wordclouds
  - Newsmaps (Treemap)
- Most visualisations need extracted features



What features can be extracted?

Basics informations

- Term frequency (see Zipfs-Law)

Token-relationships

- POS-Tags
- Dependencies
- Co-occurence

Token/Phrase simililarity

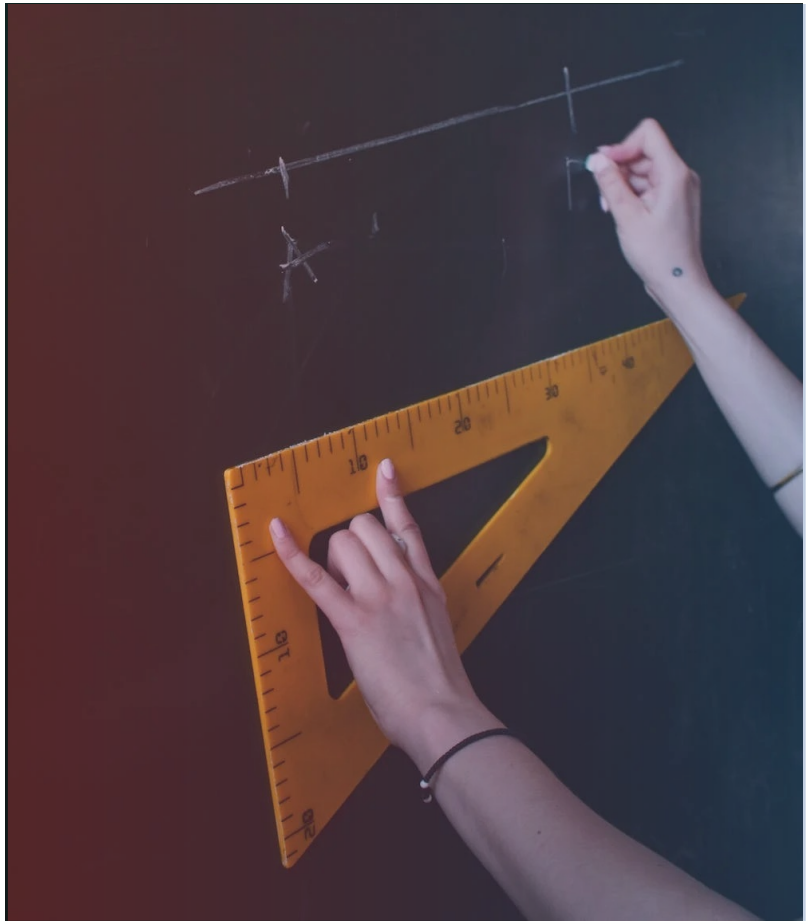
- Levenstein distance
- Embedding distance

Word Embeddings

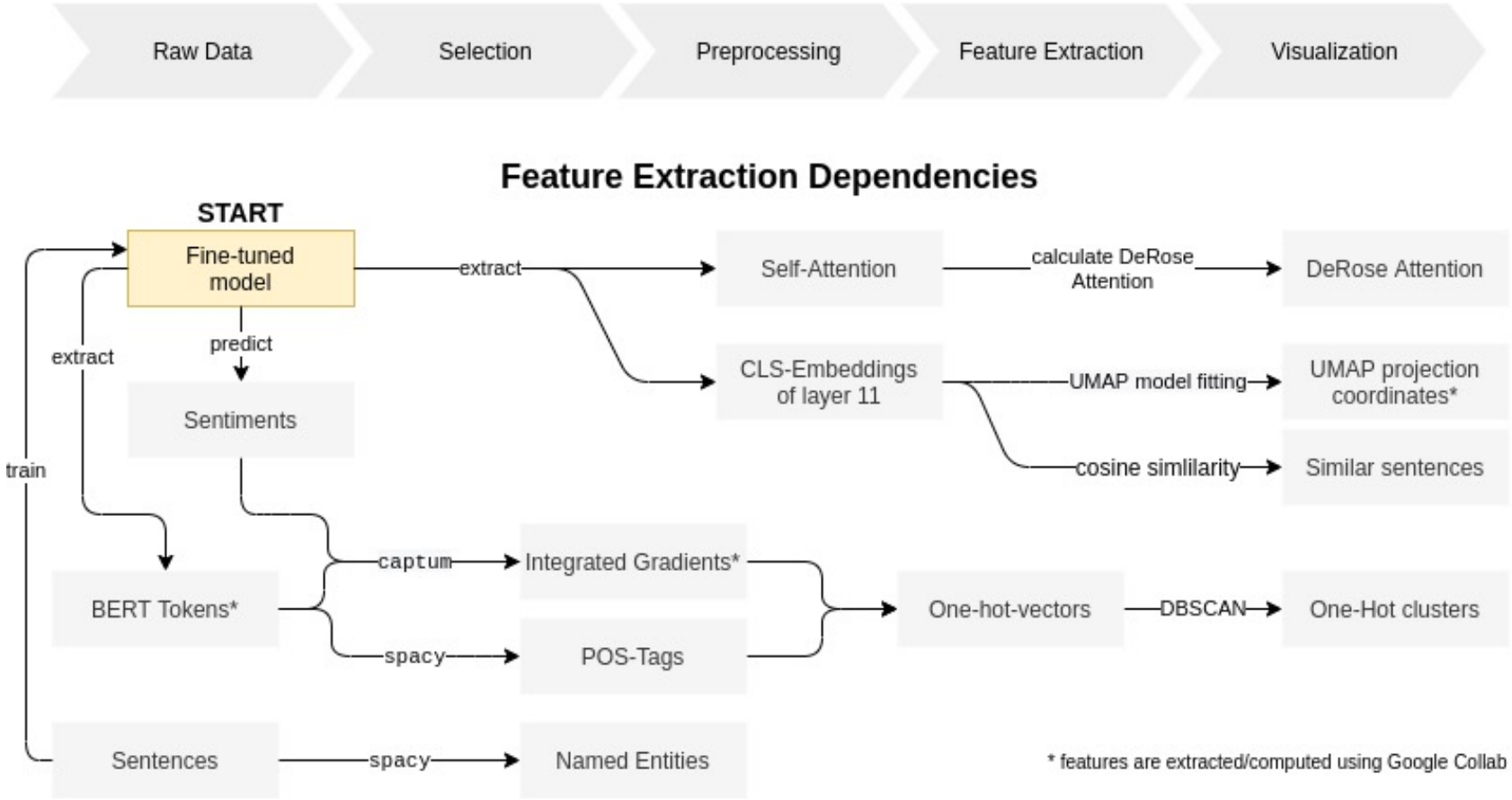
- Word2vec (Context independent model)
- Context dependent models via Language Models (esp. transformer models like BERT)

Advanced Features

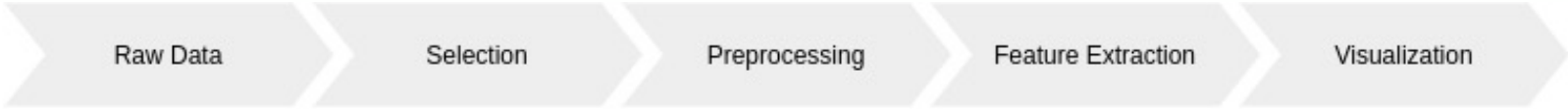
- Sentiment, Topic, Entities , …



Text Analysis Pipeline



Text Analysis Pipeline



Common Preprocessing Steps

- Basic cleaning like: new-line and whitespace removal
- Stop-word removal
- Stemming
- Sentence detection
- Tokenization
- POS-Tagging
- …

(Contextualized) Language Models - Transformer

- Recent (2017) developments: transformer and attention
  - STAR choice for NLP task over RNNs
  - strong models like BERT & GPT-3
- Statistical representation of langugage
- Somewhat understaning of Language & Context
- Applicable for all common text analysis tasks like Text Classification (e.g. Sentiment Analysis), Text Generation, Question Answering, ….

Text Visualizations in Research (XAI)

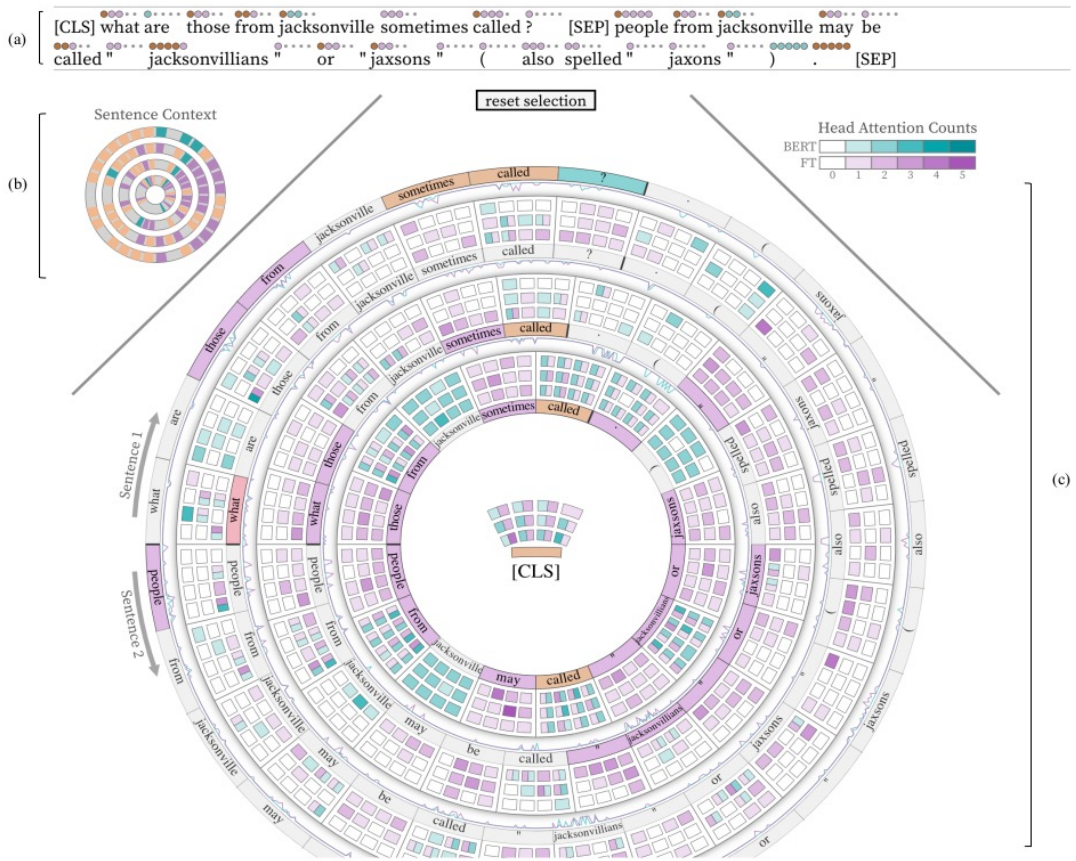


Fig 1 from: Joseph F. DeRose, Jiayao Wang, and Matthew Berger. 2020. Attention Flows: Analyzing and Comparing Attention Mechanisms in Language Models. abs/2009.07053 (2020). arXiv:2009.07053 https://arxiv.org/abs/2009.07053

Cheonbok Park, Jaegul Choo, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, and Yeonsoo Lee. 2019. SANVis: Visual Analytics for Understanding Self-Attention Networks. In 30th IEEE Visualization Conf., IEEE VIS 2019 - Short Papers, 2019. IEEE, 146–150. https://doi.org/10.1109/VISUAL.2019.8933677

Part 3

Examples

https://textvis.lnu.se/
https://lingvis.io

Refferences and further reading

- Phonetics vs. Phonology http://www.phon.ox.ac.uk/jcoleman/PHONOLOGY1.htm
- Computational Methods for Document Analysis Lecture 2019, University Konstanz, Prof. Dr. Daniel A. Keim
- Natural Language Processing Library, https://spacy.io/
- Various Deep Learning Language Models: Huggingface, https://huggingface.co/
- D. A. Keim and D. Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis," 2007 IEEE Symposium on Visual Analytics Science and Technology, 2007, pp. 115-122, doi: 10.1109/VAST.2007.4389004.
- https://lingvis.io
  - For related publications see: https://www.researchgate.net/project/LingVISio
- https://textvis.lnu.se/
  - Kucher, Kostiantyn; Kerren, Andreas (2015). [IEEE 2015 IEEE Pacific Visualization Symposium (PacificVis) - Hangzhou, China (2015.4.14-2015.4.17)] 2015 IEEE Pacific Visualization Symposium (PacificVis) - Text visualization techniques: Taxonomy, visual survey, and community insights. , (), 117–121.
- https://www.sbert.net/
  - Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.,
  - Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
- Stanford CS224N: NLP with Deep Learning | Winter 2021
  - https://www.youtube.com/watch?v=rmVRLeJRkl4&list=PLoROMvodv4rOSH4v6133s9LFPRHjEmbmJ
- Joseph F. DeRose, Jiayao Wang, and Matthew Berger. 2020. Attention Flows: Analyzing and Comparing Attention Mechanisms in Language Models. abs/2009.07053 (2020). arXiv:2009.07053 https://arxiv.org/abs/2009.07053
- Cheonbok Park, Jaegul Choo, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, and Yeonsoo Lee. 2019. SANVis: Visual Analytics for Understanding Self-Attention Networks. In 30th IEEE Visualization Conf., IEEE VIS 2019 - Short Papers, 2019. IEEE, 146–150. https://doi.org/10.1109/VISUAL.2019.8933677