Human-Centered AI
Course

LV 706.046 AK HCI 2020
Mini-Projects from Explainable AI

Professor: Andreas Holzinger
Supervisor: Bernd Malle

Human-Centered AI (Holzinger Group)
Institute for Interactive Systems and Data Science, TU Graz
and
Institute for Medical Informatics, Statistics & Documentation, Medical University Graz

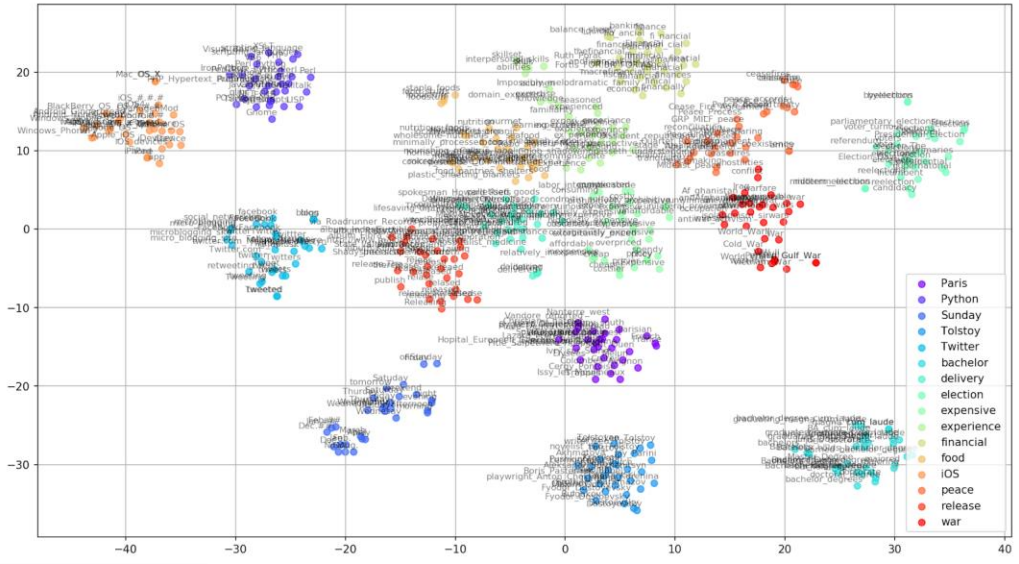
Online: https://human-centered.ai/lv-706-046-ak-hci-2020-explainable-ai

General

- Both mini-projects offer a lot of freedom as far as details are concerned – I will provide some general goals, but you are welcome to go on your own forays!
- Underlying data for both tasks are e-commerce datasets, since those are readily available & unproblematic from a privacy standpoint
- Communication will happen online for the start, since I will only be back in Austria early May
- Task details, materials and code will follow within the next 2 weeks – I am still collecting all the parts & assembling the pipelines ;-)

1. Visualization & Analysis of Word Vector Embeddings

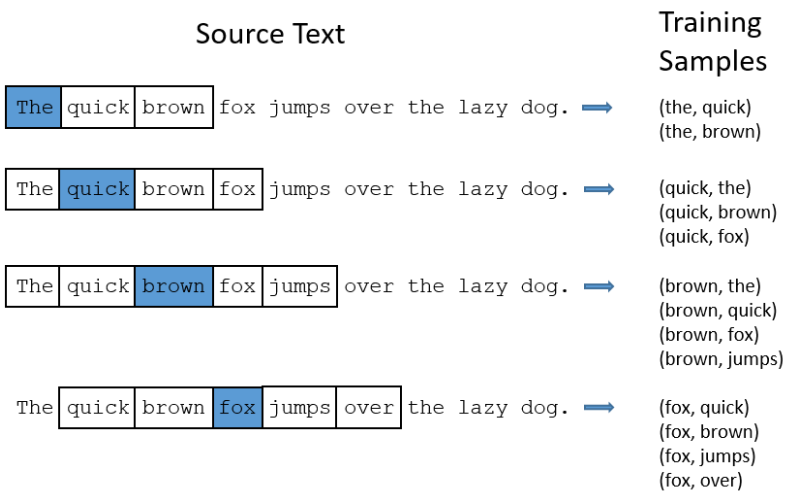
- of different models
- via different techniques
- describing (explaining?) how they discriminate / cluster

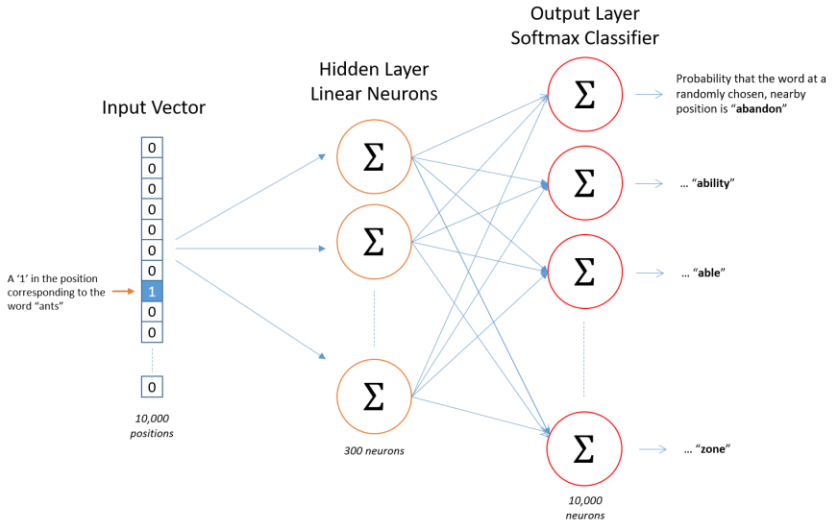


https://towardsdatascience.com/google-news-and-leo-tolstoy-visualizing-word2vec-word-embeddings-with-t-sne-11558d8bd4d

https://www.depends-on-the-definition.com/guide-to-word-vectors-with-gensim-and-keras/

*"**Word embeddings** are mathematical models that encode **word** rel ations within a **vector** space. They are created by an unsupervised training process based on **cooccurrence** information between wordsi n a large corpus"*





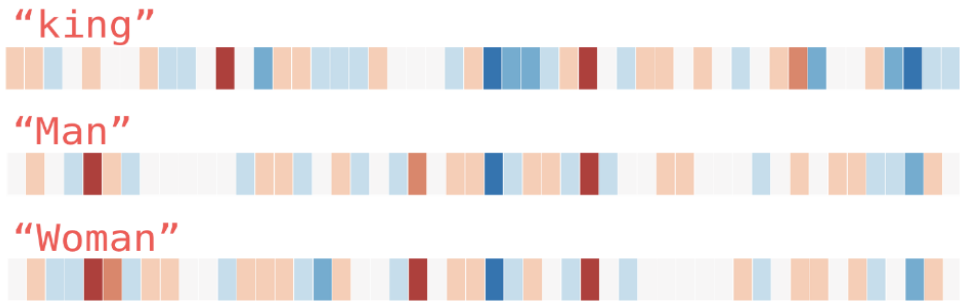
"define a fake task for the NN"
- context prediction in the case of skip-gram

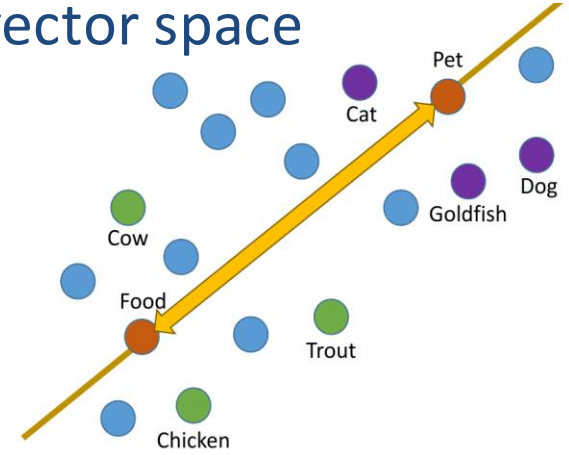
"build a shallow architecture, train & throw away the outputs"
- we only need the embeddings

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Tasks:

- learn how to interpret similarity in the vector space



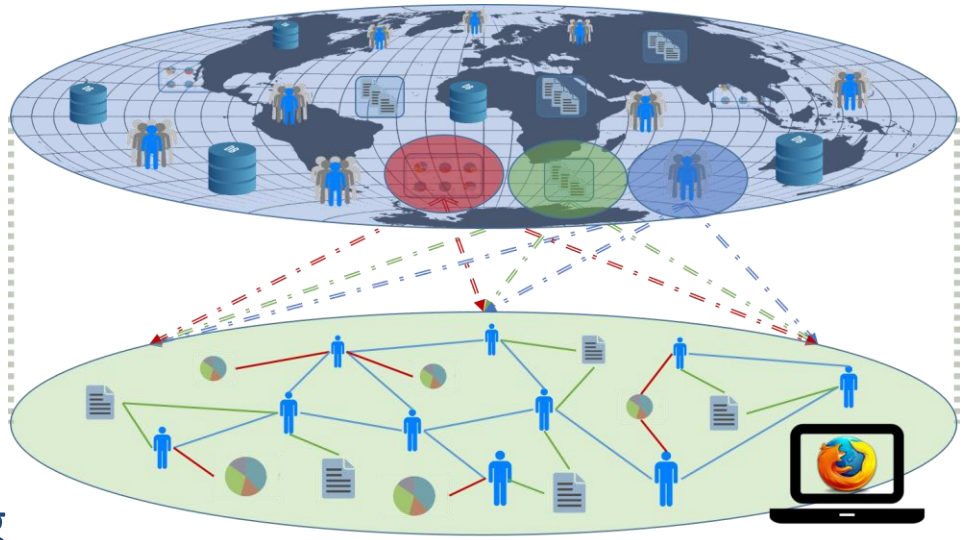


- understand the influence of source corpora on the embeddings
- analyze neighborhoods – what is (un)expected?
- predict & test consequences of changes in parameter settings / pre-processing of input data on the resulting model
- *=> "develop an intuition for embeddings as a basis for future explanations"*

Heimerl, F., & Gleicher, M. (2018). Interactive Analysis of Word Vector Embeddings. *Computer Graphics Forum*, *37*(3). doi:10.1111/cgf.13417

http://jalammar.github.io/illustrated-word2vec/

Visualization of personal recommender graphs (*"Local Sphere"*) & their change over time

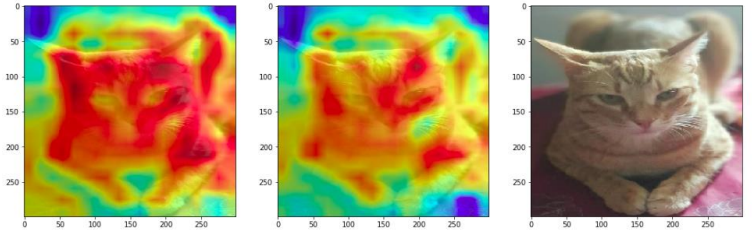


Drawing from globally available resources (e.g. a webshop database enriched similarities) each user derives her own local sub-graph representing her context / potential interests. As she interacts with the system (explores & follows / ignores visual clues), this context is refined & the graph should respond accordingly.

Bernd Malle, Nicola Giuliani, Peter Kieseberg, and Andreas Holzinger. The More the Merrier - Federated Learning from Local Sphere Recommendations. In Machine Learning and Knowledge Extraction, IFIP CD-MAKE, Lecture Notes in Computer Science LNCS 10410, pages 367–374. Springer, Cham, 2017. doi: 10.1007/978-3-319-66808-6 24.

Project 2 – Background

- Explainability of deep learning systems is a must, but still in the early stages (except for cat pics ;-)



- non-visual and / or higher dimensional data are not intuitive to the human brain – decisions made in those spaces aren't either

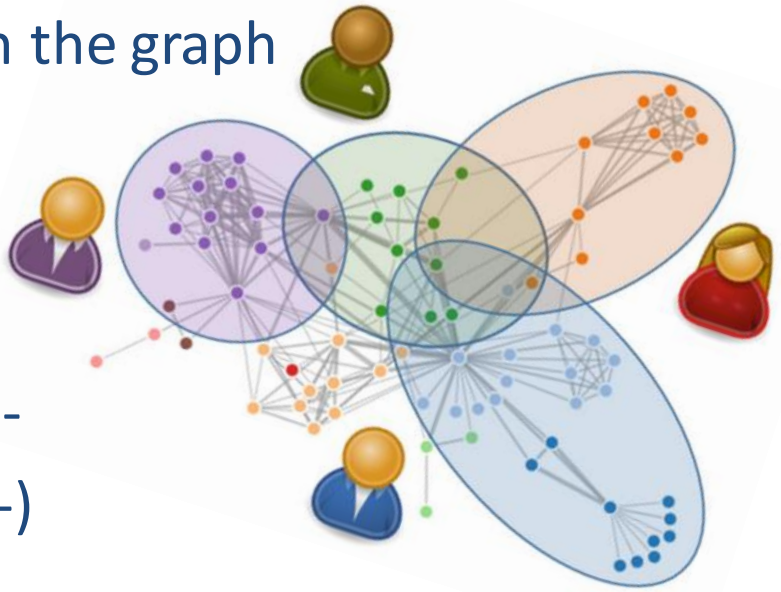
- graphs are a convenient way to break-down high-dimensional information by reducing their complexity to concepts like similarity, connection, and influence.

- understanding which graph metrics will change with user interactions will help develop an intuition about what factors in the original high-dimensional space are relevant for decisions!

https://medium.com/google-developer-experts/interpreting-deep-learning-models-for-computer-vision-f95683e23c1d

Tasks:

- Research graph visualization algorithms pertinent to recommenders (node & edge types, cluster)
- Either extend our existing (Graphinius) VIS library or decide on a different one (but make sure it's properly extensible)
- Highlight recommendations and influence factors (if available)
- Visualize continuous changes in the graph due to user interaction (vids)
- If time permits, visualize several local spheres together
- Graphs, recommender & event-stream will be provided by us ;-)



Thank you!