PUBLICATION RECOMMENDATION SYSTEM

LUKAS VLCEK

PUBLICATION RECOMMENDATIONS

- Rapid increase in (especially technical) literature
- Search for related literature is slow, especially for non-experts
- Need for automated categorization and recommendation systems
- Essential ability for meta-research in science and industry

APPROACH

- Build a model of semantic relations between publications based on categorization available in Wikipedia
- Recommendation rankings based on
  - semantic closeness (shortest path between the original and related publication)
  - publication impact (degree of centrality within the area of interest)
- Tools
  - Pandas for basic dataset manipulation
  - NetworkX library for graph analysis
  - BeautifulSoup, requests for web scraping

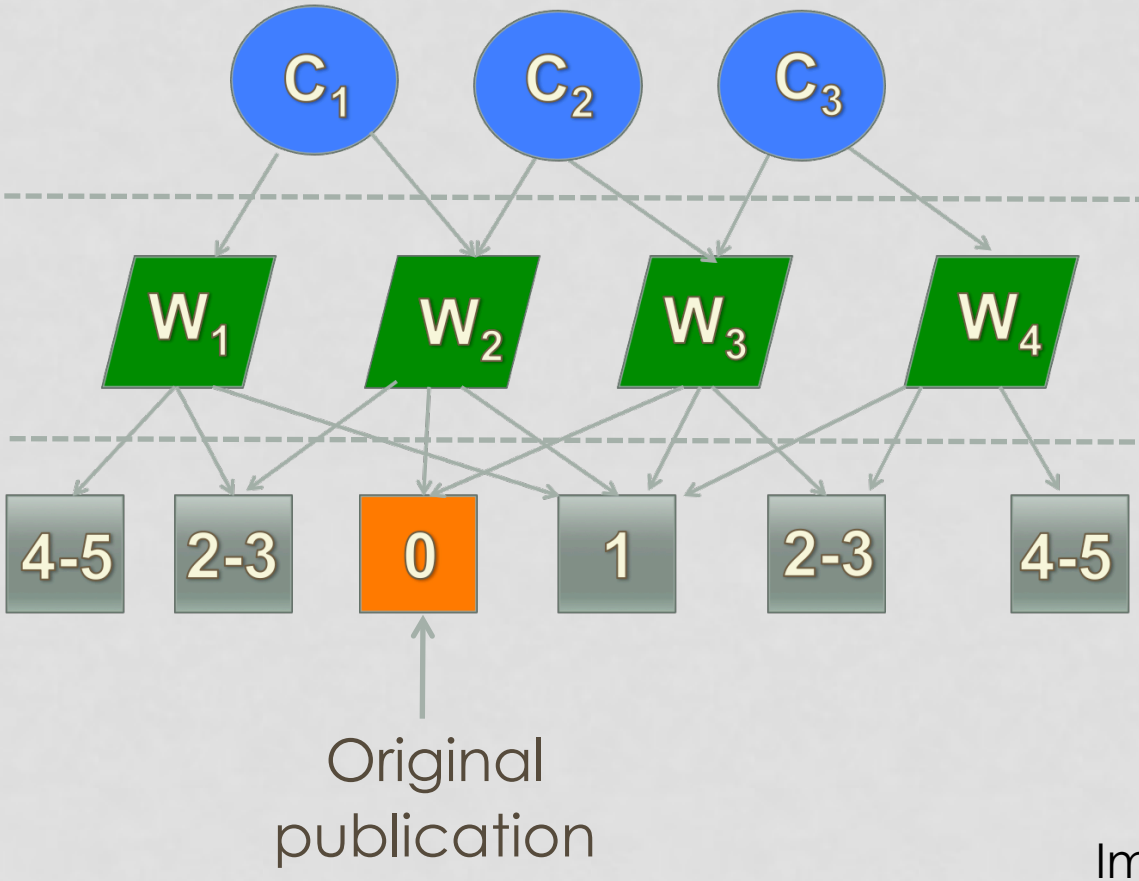
DATASETS: SOURCES AND AGGREGATION

1. Preprocessed dataset of publication references on Wikipedia pages (TSV format)
   - Features: Wiki page ID and title, publication ID
   - 3.8 M records
   - Minor data cleaning (convert NaN to 'NaN')

2. Web scraping from Wikipedia and Google search results
   - Topic-category relationship
   - Publication information

MODEL STRUCTURE

Multipartite graph of semantic relationships and citations



Categories

Wikipedia web pages

Publications with rankings

Implemented using NetworkX

TEST 1
"DESIGNING GREAT BEERS"

1. (6 cite) **The Oxford Companion to Beer**

2. (2 cite) **The Encyclopedia of Beer: The Beer Lover's Bible**

3. (2 cite) **The Best Breweries and Brewpubs of Illinois**

4. (2 cite) **Fermenting Revolution: How to Drink Beer and Save the World**

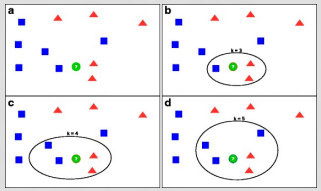
5. (1 cite) **Pale Ale: History, Brewing Techniques, Recipes**

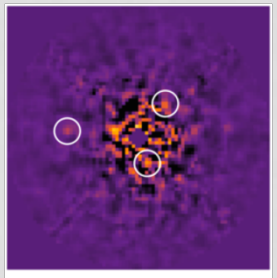
6. (1 cite) **The World Guide to Beer: The Brewing Styles, the Brands, the ...**

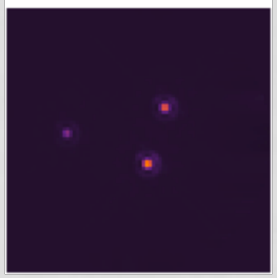
7. (1 cite) **IPA: Brewing Techniques, Recipes and the Evolution of India ...**

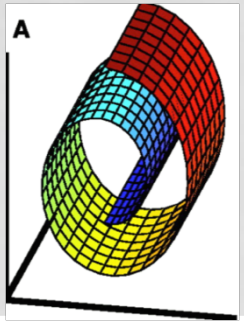
TEST 2

"LEARNING THE PARTS OF OBJECTS BY NON-NEGATIVE MATRIX FACTORIZATION"









1. (10 cite) **Ten quick tips for machine learning in computational biology.**

2. (8 cite) **Detection and Characterization of Exoplanets and Disks using Projections on Karhunen-Loeve Eigenimages**

3. (6 cite) **K-Corrections and Filter Transformations in the Ultraviolet, Optical, and Near-Infrared**

4. (6 cite) **Non-negative Matrix Factorization: Robust Extraction of Extended Structures**

5. (4 cite) **Deep Learning in Neural Networks: An Overview**

6. (4 cite) **Nonlinear dimensionality reduction by locally linear embedding**

CONCLUSIONS

- Built a publication recommendation system based on semantic closeness of publications and their impact (citations)
- Two subsystems
  - Directed graph of relations between publications
  - Data acquisition system (web scraping)
- The system recommends relevant publications ordered roughly in their perceived importance. The recommendations have sped up literature research in the fields of interest (beer, machine learning)
- The model can be further improved by extending the current publication citation dataset, and searching further in the related publications by including additional level of parent categories.