Zaporizhzhya National University
www.znu.edu.ua

SemData
semdata-project.eu

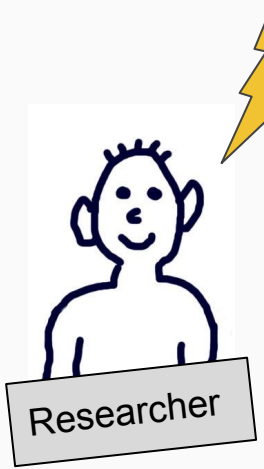
Probabilistic Topic Modelling for Controlled Snowball Sampling in Citation Network Collection

Hennadii Dobrovolskyi, Nataliya Keberle, Olga Todoriko
Department of Computer Science,
Zaporizhzhya National University, Zaporizhzhya, Ukraine,
gen.dobr@gmail.com, nkeberle@gmail.com, o-sun@rambler.ru

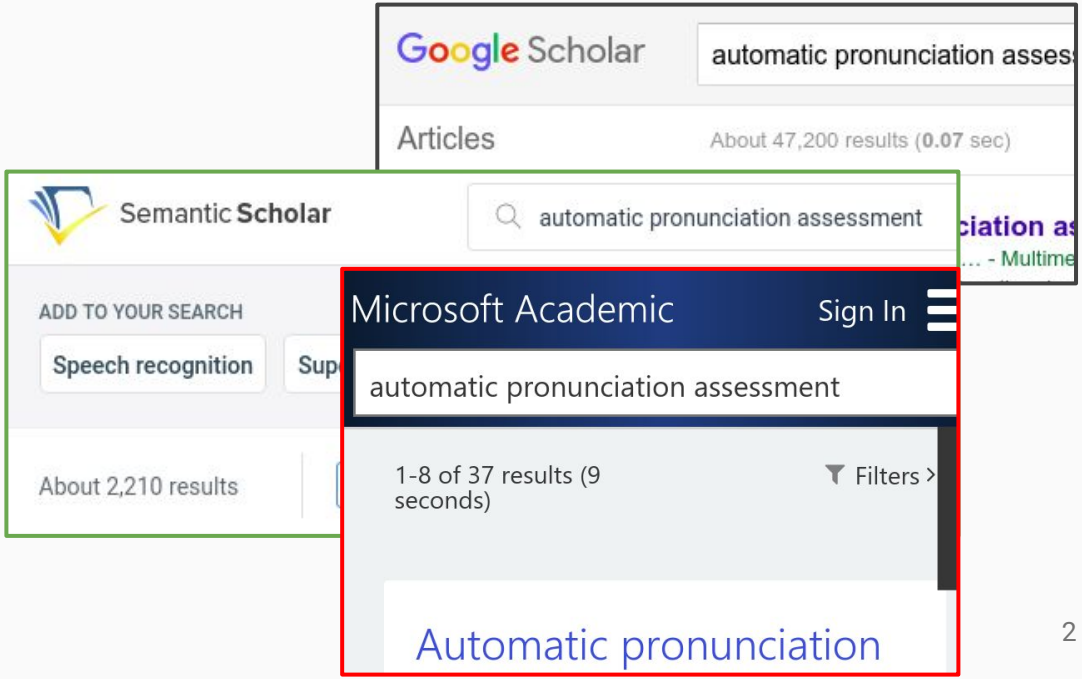
International Conference on Knowledge Engineering and Semantic Web (KESW-2017),
Szczecin, Poland, Nov 08 - Nov 10, 2017

Motivation

"Automatic pronunciation assessment"

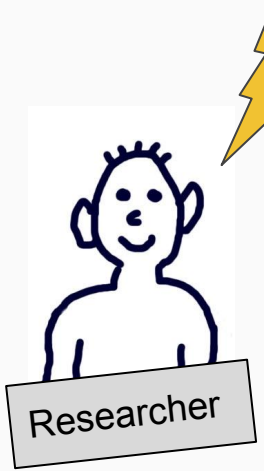


1) The Library
2) Bibliography papers
3) Keyword search: Google Scholar, Microsoft Academic, SemanticScholar



Motivation

“Automatic pronunciation assessment”



Intelligent techniques

4) Coauthor networks
5) Exploratory search
6) Ontology based (e.g. Klink-2)
7) Citation networks

Our Results

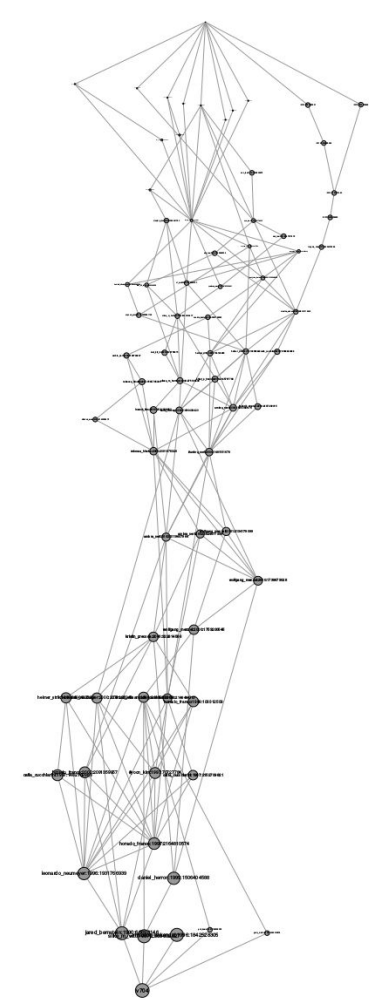
1. Software implemented to collect representative citation network.

2. Plan of Reading

https://github.com/gendobr/snowball

Main citation path for the collected citation network. Nodes are marked as (first author : year : MS_Academic_Id)

v703
nancy_f_chen:2016:2575689684
wei_li:2016:2401896499
wenping_hu:2015:2091856355
wenping_hu:2014:1965370992
joost_van_doremalen:2009:2132049498
maxine_eskenazi:2009:2016114400
helmer_strik:2007:2139565824
khiet_p_truong:2005:2145767788
khiet_truong:2006:1496430420
#ambra_neri:2004:1481151678
ambra_neri:2002:2119607964
kristin_precoda:2000:322814586
leonardo_neumeyer:2000:2070133242
#yoon_kim:1997:70727784
horacio_franco:1997:2164810574
leonardo_neumeyer:1996:1931766939
jared_bernstein:1990:66698146
v704



Top 73 nodes of the collected citation network.

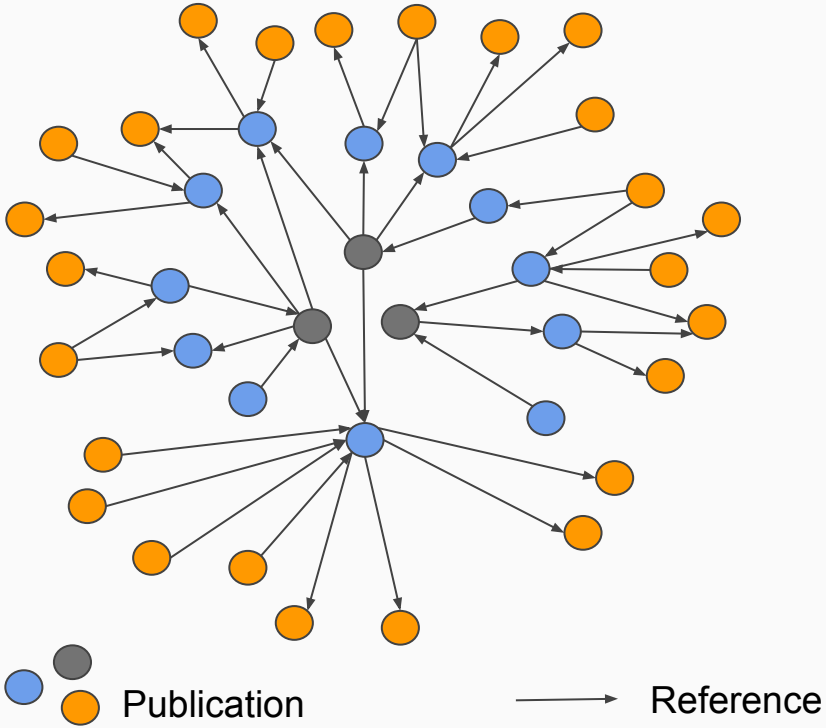
Citation networks

Directed graph

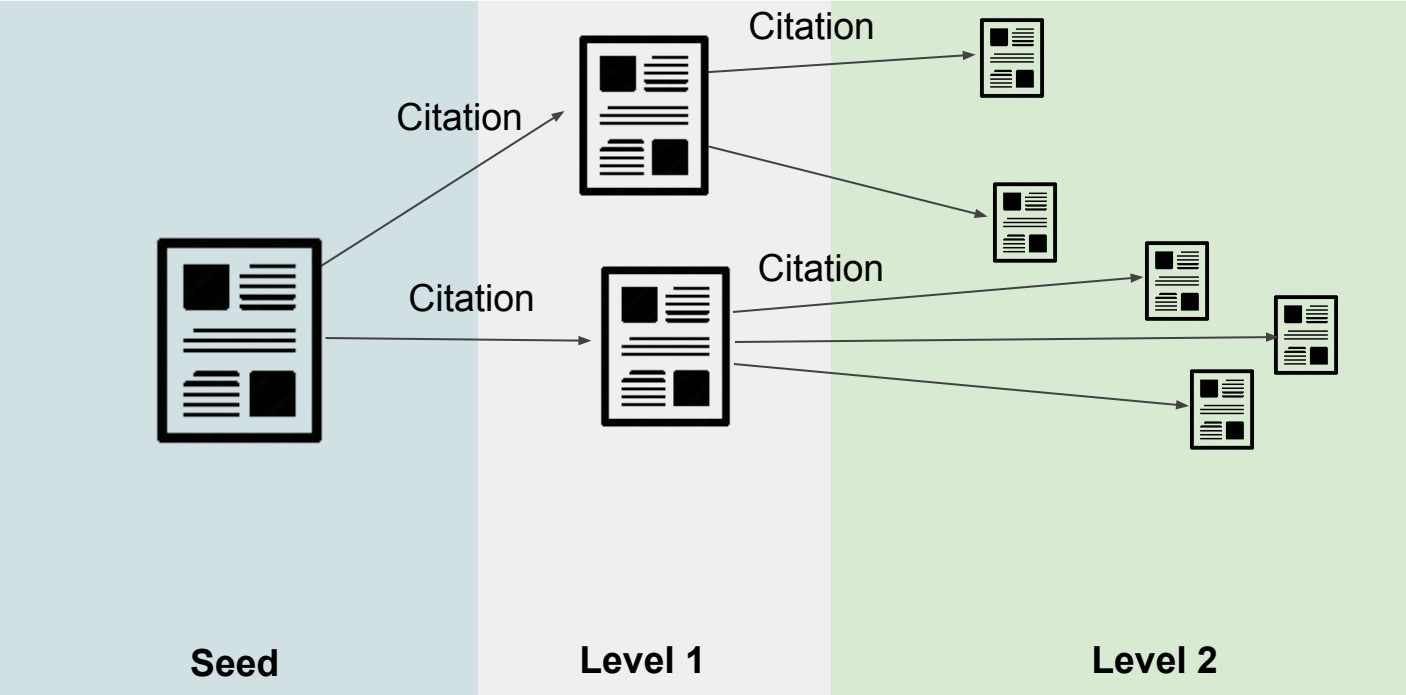
Quality: references are selected carefully by the authors

Automation: search engines can follow references

Completeness: phenomena of ”small world” - proven property of scale-free networks

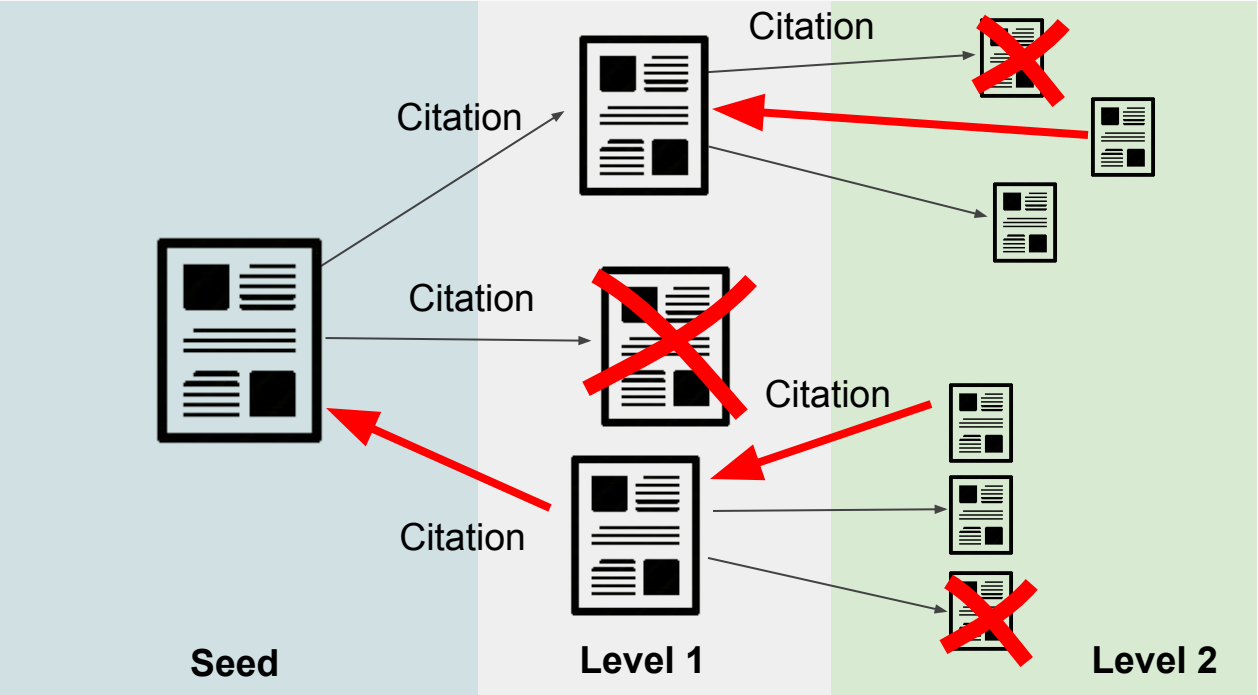


Classical Snowball Sampling



Following all the citations we collect **too many publications**

Controlled Snowball Sampling



Keep only the papers similar to the seed papers

Similarity - from **probabilistic** topic model

Joint Probability Estimate

Title + abstract

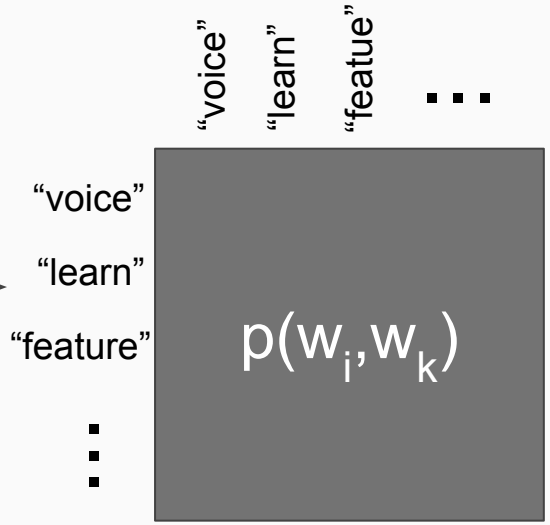
Automatic pronunciatic
assessment for Mandarin
Chinese

Jan 1st 2004, International Conference on Multimed
volume 3, pp 1979-1982, DOI: 10.1109/ICME.2004.1

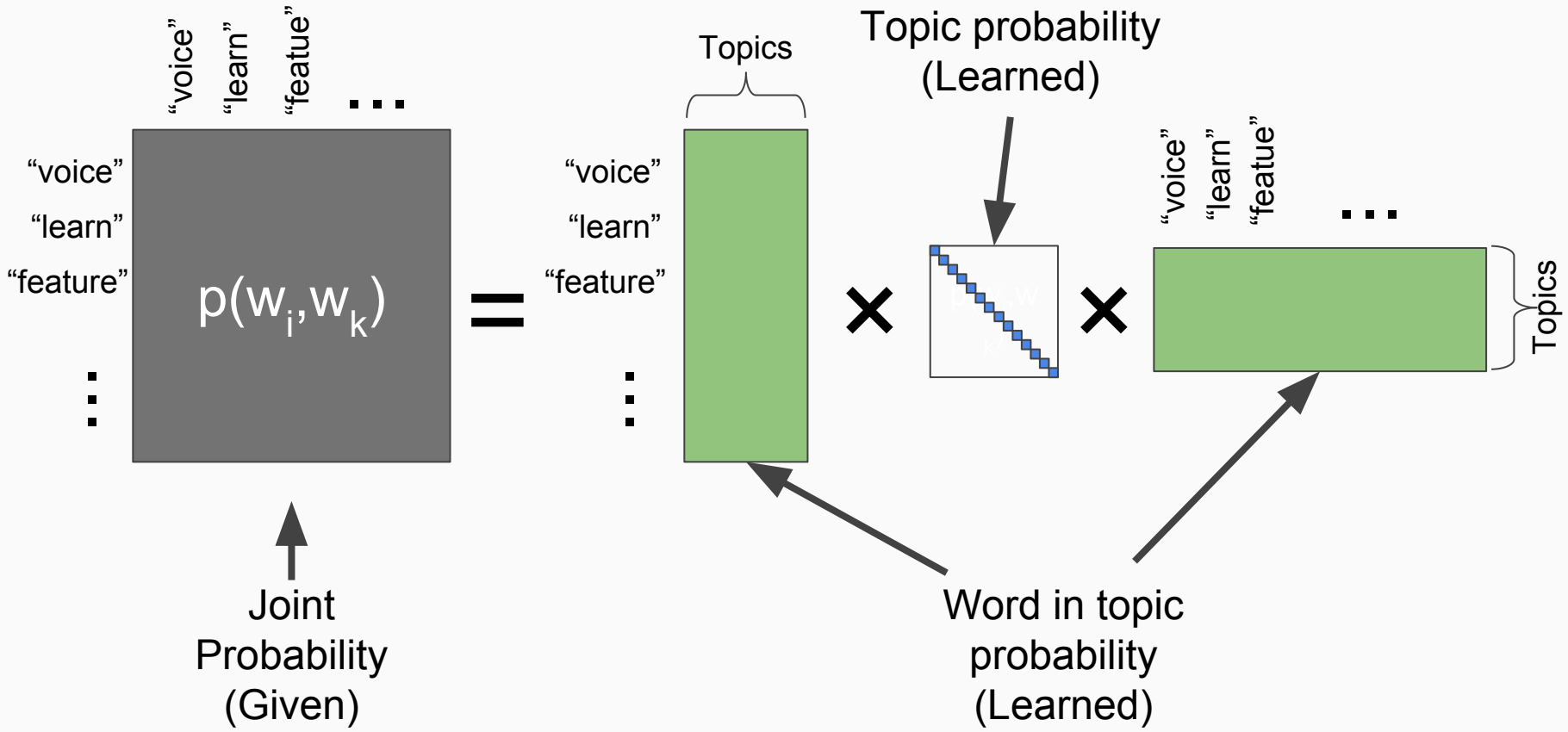
Jiang-Chun Chen (National Tsing Hua University),
Jyh-Shing Roger Jang (National Tsing Hua Universit
Juii-Yi Li (National Tsing Hua University),
Ming-Chun Wu (National Tsing Hua University)

This work describes the algorithms used in a pr

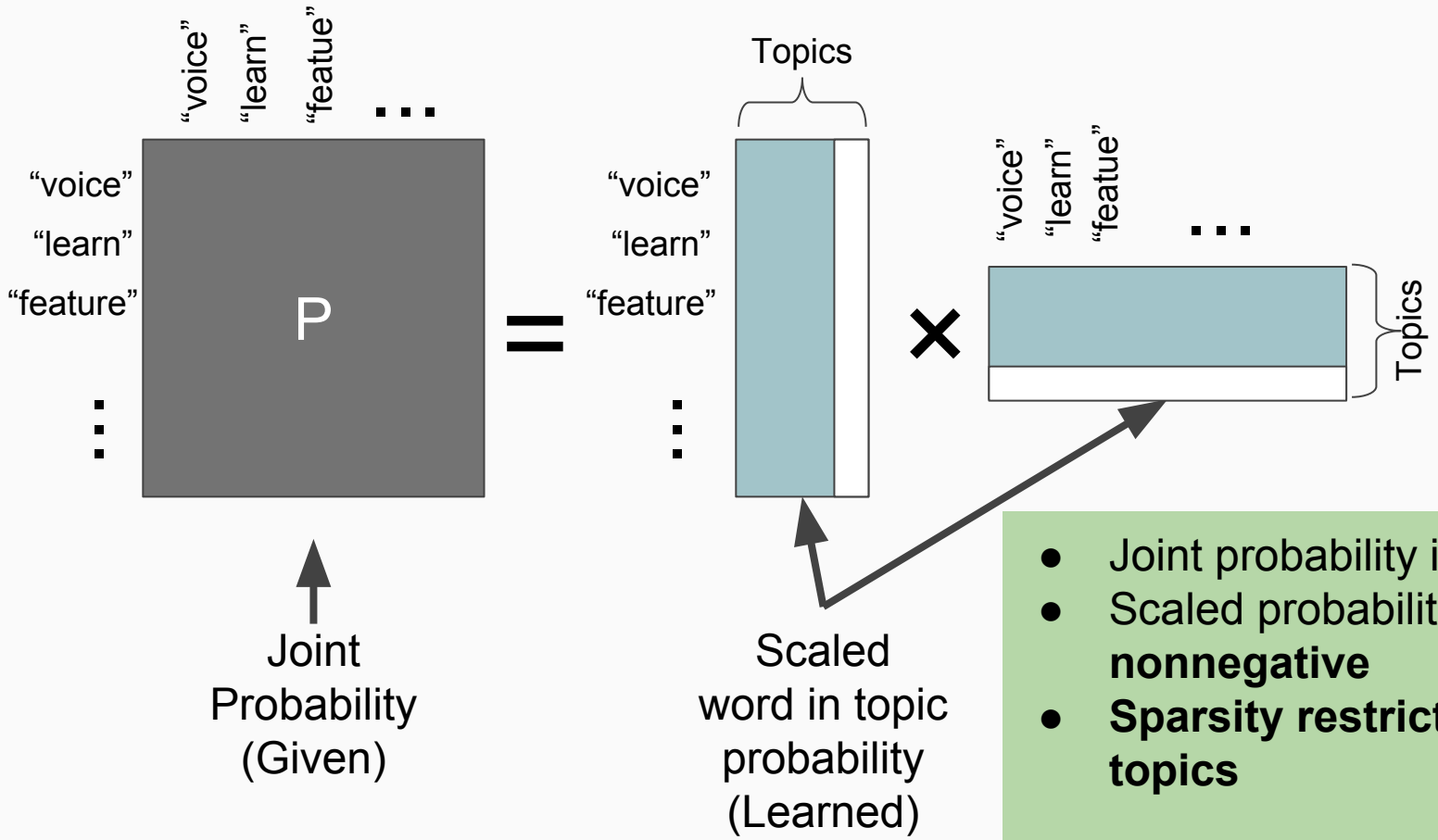
1. Tokenization
2. Stemming
3. Keep nouns and adjectives
4. Remove stop-words
5. Remove rare words
6. Count word co-occurrences



Probabilistic Topic Model from Word-Word Co-Occurrence



Sparse Symmetric Nonnegative Matrix Factorization (SSNMF)



- Joint probability is **symmetric**
- Scaled probability is **nonnegative**
- **Sparsity restricts number of topics**

Seed Papers Selection

Valid seed papers should be 5-10 years old and have to be widely cited *.

- some seminal papers of the knowledge domain pointed by experts
- papers selected by the researcher/supervisor

Best seeds are:

- reviews,
- foundational or framing articles on the topic of interest.

* Lecy, Jesse D., and Kate E. Beatty. "Representative literature reviews using constrained snowball sampling and citation network analysis." (2012).

Experiment Details

Seed papers selected for phrase:

”automatic pronunciation assessment”

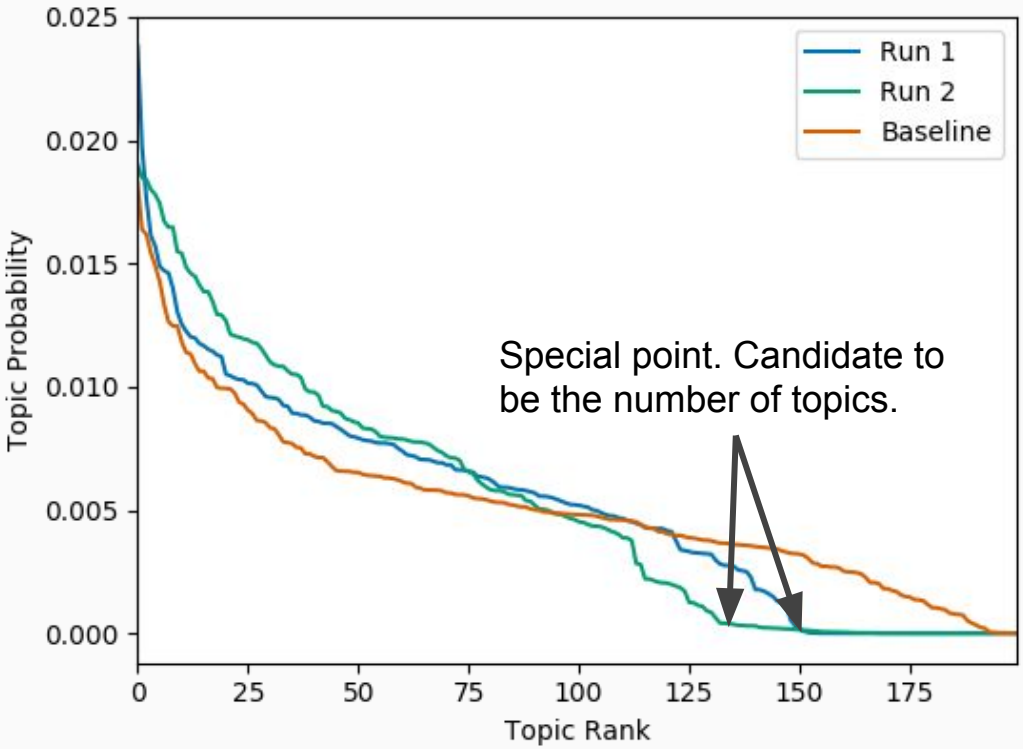
Source of seed papers: Google Scholar

Use Microsoft Academic Knowledge API to search for publications

SSNMF-based Principal Components

Like classic PCA we keep only topics that have large probabilities.

Sparsity requirement forces some topics to have tiny probabilities.



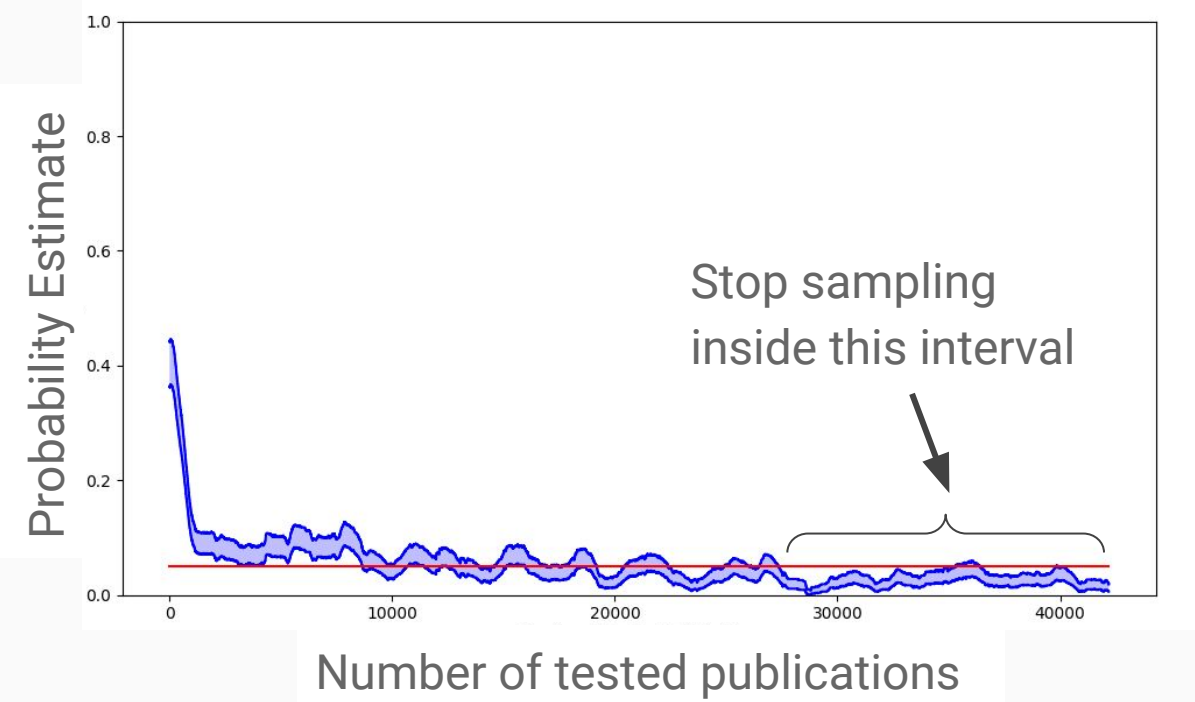
Topics ranked by value of topic probability for different SSNMF runs for sparsity parameter λ = 0.005 and **random initial states**. Baseline is topic probabilities for sparsity parameter λ = 0

Saturation of the Controlled Snowball Sampling

Observed saturation: when processing the publications sequentially we can either
(a) accept *N*th publication and add it to snowball or
(b) don't accept.

Colored strip is 0.95 confidence interval of Poisson distribution of event (a).

Red line is the acceptance probability 0.05



Acceptance probability estimate as a function of the number of already tested abstracts *N*.

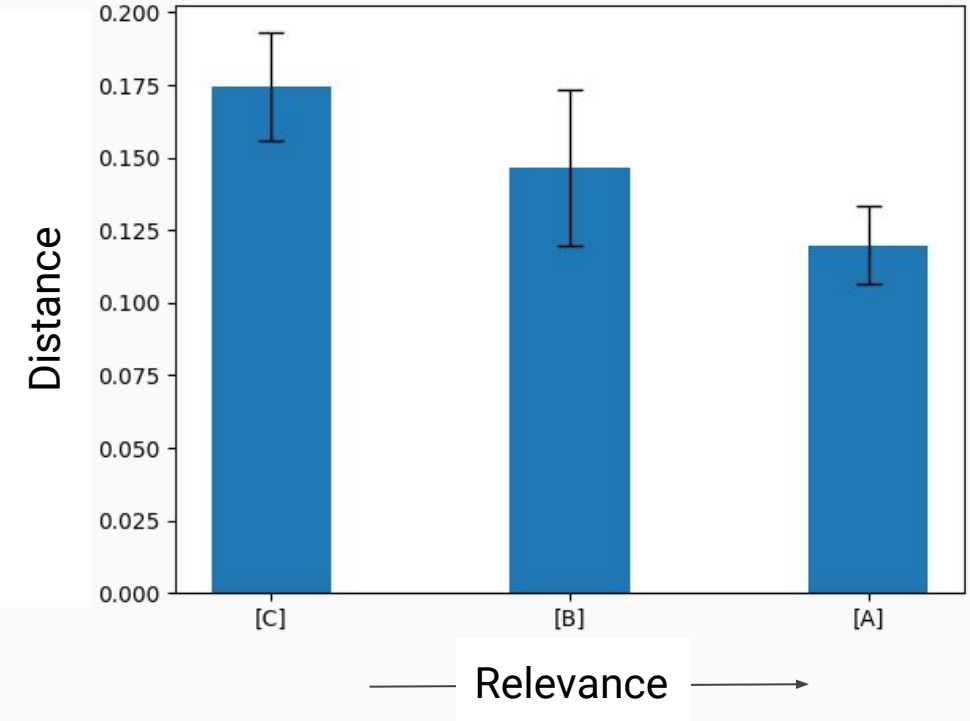
Estimate against Existing References

The biased lists of references got from [A,B,C] are estimated.

[A] and [B] concern the pronunciation training

[B] is PhD thesis also containing references to related domains

[C] concerns the related domain

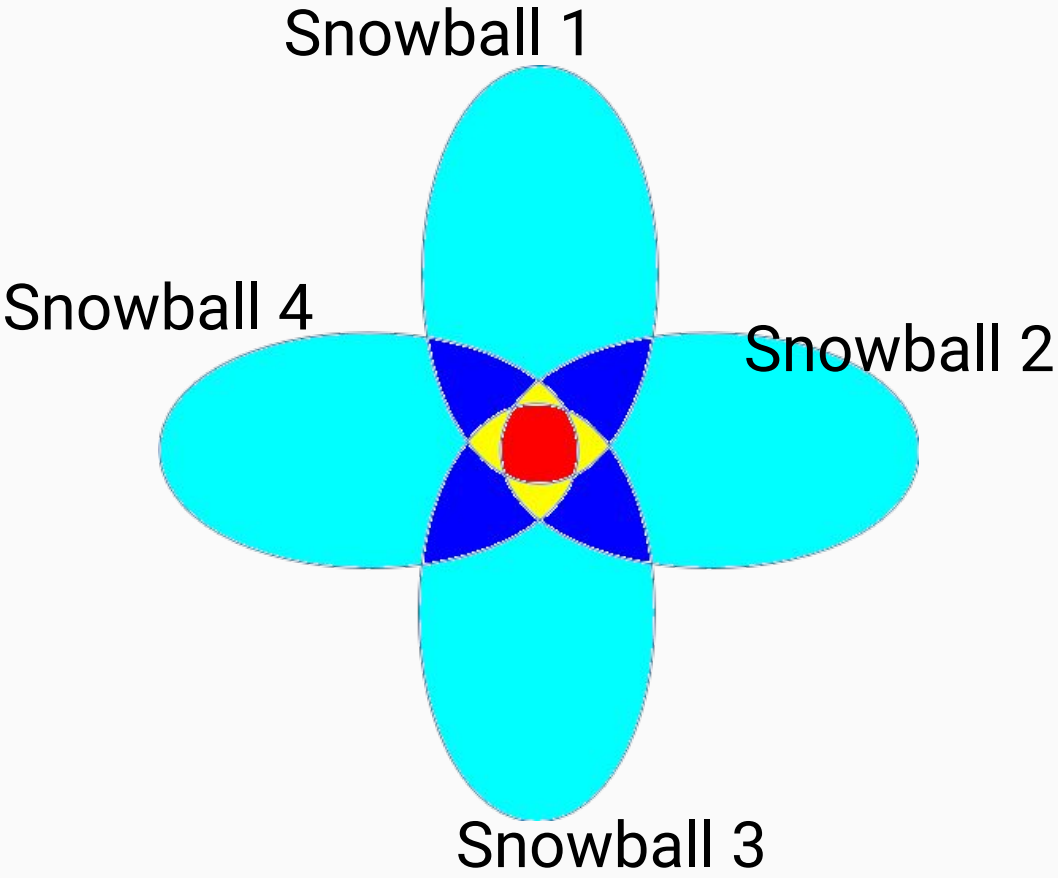


Distance from sample lists of references to seed papers

[A] Fouz-González, J.: Trends and directions in computer-assisted pronunciation training. Springer, (2015)
[B] Lee, A.: Language-independent methods for computer-assisted pronunciation training. MIT, (2016)
[C] López, et.al.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, (2013)

Robustness with Respect to Seed Papers Variation

- 66% of relevant papers are detected every time
- 14% - in 80% of runs
- 14% - in 60% of runs
- 6% - at least once

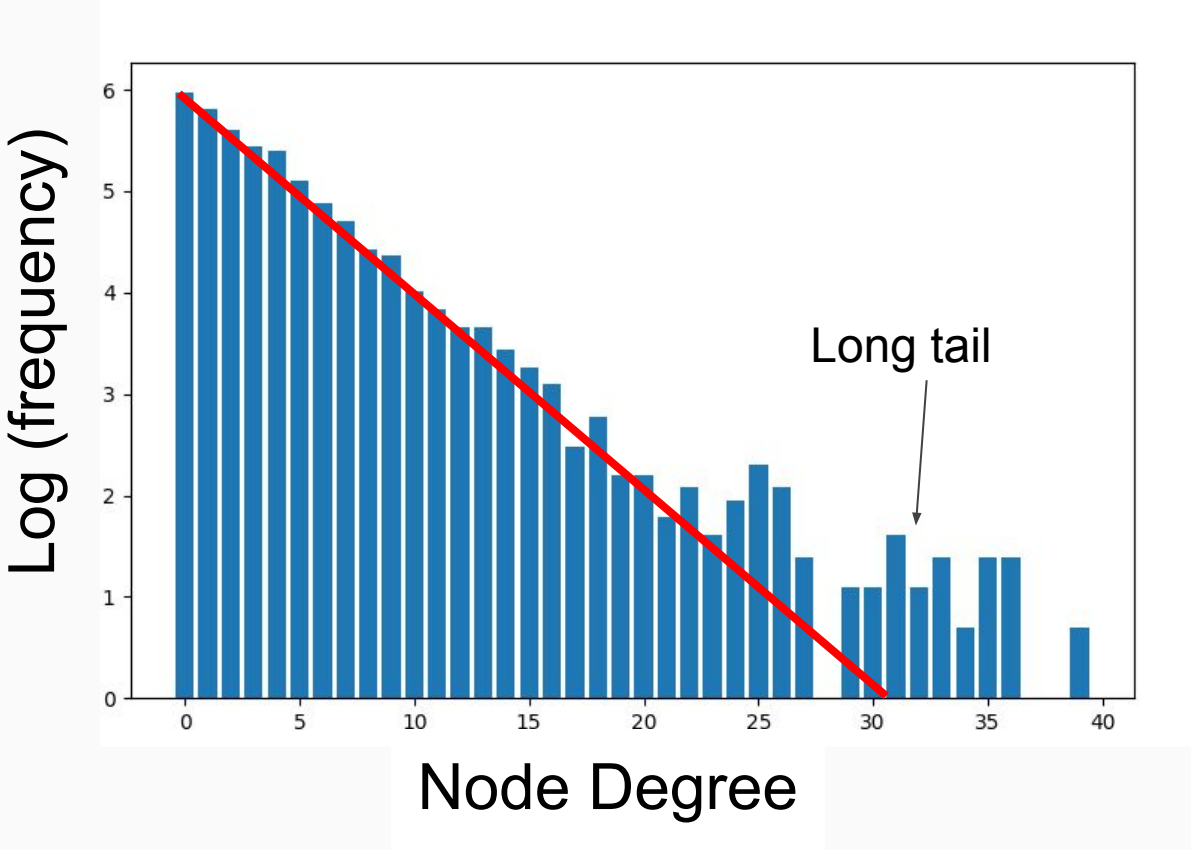


The Collected Citation Network is “Small World”

Long tail of the distribution of node degrees is the feature of “small-world” networks.

Distribution of node degrees in the collected citation network has a long tail.

Red baseline shows “big world” distribution

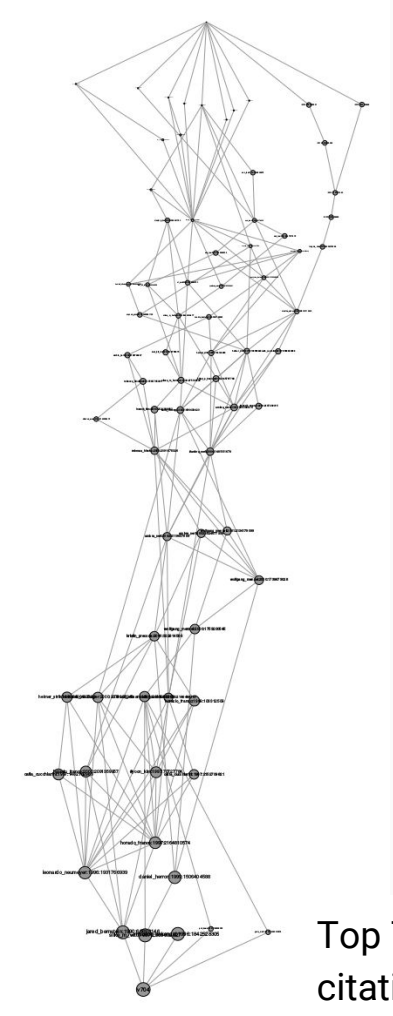


Distribution of node degrees in the collected citation network

Citation Network Analysis

Main path analysis is applied to the collected citation network.

List of the top 73 most significant publications is created.



Top 73 nodes of the collected citation network.

Main citation path for the collected citation network. Nodes are marked as (first author : year : MS_Academic_Id)

v703
nancy_f_chen:2016:2575689684
wei_li:2016:2401896499
wenping_hu:2015:2091856355
wenping_hu:2014:1965370992
joost_van_doremalen:2009:2132049498
maxine_eskenazi:2009:2016114400
helmer_strik:2007:2139565824
khiet_p_truong:2005:2145767788
khiet_truong:2006:1496430420
#ambra_neri:2004:1481151678
ambra_neri:2002:2119607964
kristin_precoda:2000:322814586
leonardo_neumeyer:2000:2070133242
#yoon_kim:1997:70727784
horacio_franco:1997:2164810574
leonardo_neumeyer:1996:1931766939
jared_bernstein:1990:66698146
v704

Conclusions

Software implemented to collect representative citation network.

Demonstrate

- Robustness
- Saturation
- Completeness

We can get list of most important publications in most of scientific domains.

https://github.com/gendobr/snowball

Thanks!

Hennadii Dobrovolskyi,
gen.dobr@gmail.com,

Nataliya Keberle,
nkeberle@gmail.com,

Olga Todoriko
o-sun@rambler.ru

Department of Computer Science
Zaporizhzhya National University
Zhukovskogo st. 66, 69600,
Zaporizhzhya, Ukraine,

https://github.com/gendobr/snowball