# 3. Graph-Based Algorithms For Natural Language Processing And Information Retrieval

**Rada Mihalcea, University of North Texas, and Dragomir Radev, University of Michigan**

Graph theory is a well studied discipline, and so are the fields of natural language processing and information retrieval. However, most of the times, they are perceived as different disciplines, with different algorithms, different applications, and different potential end-users.

The goal of this tutorial is to provide an overview of methods and applications in natural language processing and information retrieval that rely on graph-based algorithms. This will include techniques for graph traversal, minimum path length, min-cut algorithms, minimum spanning trees, random walks, etc. and their application to information retrieval and Web search, text understanding (word sense disambiguation and semantic classes), parsing, text summarization, keyword extraction, text clustering, and others.

## 3.1 Tutorial Outline

1. Graph-based Algorithms Basics
   * Vectors, matrices, graphs
   * Graph representations and notations
   – Traversal, min-cut/max-flow, matching
   * Algorithms for graph traversal
   * Minimum path length
   * Minimum spanning trees
   * Min-cut/max-flow algorithms
   * Graph-matching algorithms
   – Ranking, clustering, learning
   * Eigenvector analysis
   * Node-ranking algorithms
   * Graph-based centrality
   * Graph-based clustering
   * Machine learning on graphs
2. Information Retrieval applications
   * Web-page ranking
   * Text classification and clustering
3. Natural language processing applications
   – Semantics
   * Word sense disambiguation
   * Semantic classes
   * Textual entailment
   * Sentiment classification
   – Syntax, Summarization
   * Dependency parsing
   * Prepositional attachment
   * Keyword extraction
   * Text summarization

## 3.2 Target Audience

This tutorial is intended for researchers and practitioners who seek a general understanding of the application of graph-theoretical representations and algorithms to natural language processing and information

retrieval. It is introductory in nature, no special knowledge or background is required.

Rada Mihalcea is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, graph-based algorithms for natural language processing and information retrieval, minimally supervised natural language learning, and multilingual natural language processing. She has published more than 80 articles in books, journals, and proceedings, in these and related areas. She is the president of the ACL Special Group on the Lexicon (SIGLEX), and a board member for the ACL Special Group on Natural Language Learning (SIGNLL). She serves on the editorial board of the journal of Computational Linguistics, the journal of Language Resources and Evaluations, and the recently established journal of Interesting Negative Results in Natural Language Processing and Machine Learning. Her research is supported by NSF, Google, and the state of Texas.

Dragomir Radev is an Associate Professor of Information, of Computer Science and Engineering, and of Linguistics at the University of Michigan. He has a PhD in Computer Science from Columbia University. He has held numerous posts within NAACL and ACL. He is on the editorial boards of Information Retrieval and the Journal of Artificial Intelligence Research and was recently nominated to the board of the Journal of Natural Language Engineering. He has co-chaired 5 ACL/NAACL workshops and given 6 tutorials at venues like SIGIR, AAAI, and RANLP. Dragomir's current interests are in text summarization, information extraction, information retrieval, graph models, semi-supervised learning, and machine translation. He has more than 50 peer-reviewed papers as well as more than 50 talks at various universities and other venues. Dragomir's work has been funded by NSF, NIH, and ONR.