INTERNSHIP TOPICS - 2017

DATA SCIENCE AND MINING (DASCIM) TEAM@ LIX, ECOLE POLYTECHNIQUE

http://www.lix.polytechnique.fr/dascim/

February 28, 2017

Instructions

Thank you for your interest in the internship topics of our team. Please read carefully the instructions below.

- Please fill in the following form, specifying your details, which topic(s) you are interested in, and a link to a motivation letter and your CV: http://tinyurl.com/hbtr2bm
- There will be a technical interview (programming algorithms and fundamental ML in a language of your choice for a fixed time interval - i.e. 2 hours)
- After the technical interview, shortlisted people will be interviewed by team members to ensure best match of choices and skills to internships.

If you have further questions please contact us at: mvazirg at lix.polytechnique.fr

1 Heterogeneous Text-Graph Node Embeddings and Applications

1.1 Description

Many graphs have text associated with their nodes or edges. However, state-of-the-art graph node embeddings [1, 2] make only use of the network structure, and are general in that they don't try to optimize performance on a specific downstream task. Embedding the nodes of the network while taking all this information into account is still an area of active research [3, 4].

The internship will aim at designing new techniques that:

- Exploit the rich information in heterogeneous networks (specifically text-enhanced networks).
- Learn task-specific graph node embeddings.
- Have applications in graph and text mining tasks.

1.2 Desired skills

Python, deep learning, NLP, word embeddings, graph node embeddings, graph theory, linear algebra.

1.3 Supervisors

The intern will work under the guidance of Antoine Tixier (DaSciM) and Fragkiskos Malliaros (UC San Diego).

References

[1] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In *KDD*, 2014.

[2] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." In *KDD*, 2016.

[3] Chen, Ting, and Yizhou Sun. "Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification." arXiv preprint arXiv:1612.02814 (2016).

[4] Tang, Jian, Qu, Meng, and Mei, Qiaozhu. "PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks". In *KDD*, 2015.

2 Weighted K-truss Decomposition and Applications

2.1 Description

The K-truss decomposition algorithm[1, 2] is a triangle-based extension of the k-core algorithm [3] that was originally developed to study cohesiveness in social networks. It finds many powerful applications in graph and data mining. Currently, however, K-truss works only for unweighted and undirected graphs, which is a major limitation. Proposing an extension of the algorithm to richer representations would make for an impactful contribution to the field of graph theory. The objective of the internship will be to investigate:

- different ways of incorporating weights and edge direction into triangles,
- modifications that would need to be done at the algorithm level,
- impact on complexity,
- applications in tasks including community evaluation in social networks and keyword extraction for text summarization.

2.2 Desired skills

Python, graph theory, linear algebra, algorithms.

2.3 Supervisors

The intern will work under the guidance of Antoine Tixier (DaSciM), Fragkiskos Malliaros (UC San Diego) and Apostolos Papadopoulos (DaSciM and Aristotle University of Thessaloniki).

References

[1] Cohen, J. (2008). Trusses: Cohesive subgraphs for social network analysis. National Security Agency Technical Report, 16.

[2] Wang, J and Cheng, J. Truss decomposition in massive networks. In *VLDB*, 2012.

[3] Seidman, S. B. (1983). Network structure and minimum degree. Social networks, 5(3), 269-287.

3 Updatable core decompositions

In graph analysis, the k-core decomposition is one of the most fundamental tools [1, 2, 3]. One of the main advantage of this structure is that it is very efficient to compute [4]. As a network grows, there is a need to recompute the decomposition in order to update properties of the graph and its' nodes. On small graphs, one may re-apply a k-core decomposition algorithm on the entire graph but in big graphs -with millions of nodes and edges- even the most efficient implementation consumes a lot of resources needlessly.

For this reason, various models have been developed to compute the k-core decomposition under different cenarios such as streaming [5, 6] and in incrementally updating existing decompositions while the graph grows [7].

While the k-core is a very popular structure, it does not capture semantics of the network that are usually modeled with direction or weight. The DaSciM team has developed and analyzed models for core decomposition on directed, weighted and other types of graphs [8, 9].

In this intership, the main goal is to create updateable decomposition algorithms for directed and weighted graphs:

1. Identify the most efficien sollutions for k-core decomposition under an environmet where large graphs continuously grow and new edges and/or nodes are added.
2. Extend these algorithms into a directed and weighted scenarios
3. Provide implementations of these algorithms for large scale deployment on real graphs.

Supervisors: Giatsidis Christos, Michalis Vazirgiannis

3.1 Desired skills

Python (and Java) programming.

References

[1] Shai and Havlin, Shlomo and Kirkpatrick, Scott and Shavitt, Yuval and Shir, Eran.A model of Internet topology using k-shell decompositionCarmi.Proceedings of the National Academy of Sciences.104 11150–11154. 2007.

[2] Goltsev, Alexander V., Sergey N. Dorogovtsev, and Jose Ferreira F. Mendes. "k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects." Physical Review E 73.5 (2006): 056101.

[3] Miorandi, Daniele, and Francesco De Pellegrini. "K-shell decomposition for dynamic complex networks." Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on. IEEE, 2010.

[4] Batagelj, Vladimir, Andrej Mrvar, and Matja Zavernik. "Partitioning approach to visualization of large graphs." International Symposium on Graph Drawing. Springer Berlin Heidelberg, 1999.

[5] Saryce, Ahmet Erdem, et al. "Streaming algorithms for k-core decomposition." Proceedings of the VLDB Endowment 6.6 (2013): 433-444.

[6] Saryce, Ahmet Erdem, et al. "Incremental k-core decomposition: algorithms and evaluation." The VLDB Journal 25.3 (2016): 425-447.

[7] Li, Rong-Hua, Jeffrey Xu Yu, and Rui Mao. "Efficient core maintenance in large dynamic graphs." IEEE Transactions on Knowledge and Data Engineering 26.10 (2014): 2453-2465.

[8] Giatsidis, Christos, Dimitrios M. Thilikos, and Michalis Vazirgiannis. "D-cores: Measuring collaboration of directed graphs based on degeneracy." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

[9] Giatsidis, Christos, Dimitrios M. Thilikos, and Michalis Vazirgiannis. "Evaluating cooperation in communities with the k-core structure." Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011.

4 Graph Mining for Multi-label Classification

Multi-label classification involves building models that may assign multiple labels to each data instances, as opposed to traditional multi-class (single-label) classification. The main challenge is to model dependence among labels and exploit this dependence efficiently. If labels are viewed as nodes on a graph, dependence can be viewed as edges connecting these nodes. Relatively little work has been carried out so far to study possible connections between graph mining and multi-label learning. This internship will develop some of these connections, for example, using graph kernel methods and methods for community detection or influence maximization on models of multi-label label dependence and thus create new methods or improve the performance of existing ones.

4.1 Desired Skills

Good programming ability. Knowledge about supervised classification (about multi-label classification is a bonus) and familiarity with graph mining concepts.

4.2 Supervisors

Jesse Read and Michalis Varzigiannis

References

[1] K. Dembczyski, W. Waegeman, W. Cheng, E. Hllermeier (2012). On label dependence and loss minimization in multi-label classification. MLJ. 88:1, pp 5–45.

[2] H. Su and J. Rousu (2015). Multilabel Classification through Random Graph Ensembles. MLJ, 99(231).

[3] M. Vazirgiannis, C. Giatsidis, F. D. Malliaros (2013) Graph Mining Tools for Community Detection and Evaluation in Social Networks and the Web. WWW Tutorial. `http://www.lix.polytechnique.fr/~mvazirg/WWW2013_tutorial`

5 Recurrent Neural Networks for Prediction in Data-Streams with Temporal Dependence

Machine learning in data streams is an increasingly active and relevant research topic. In the age of big data, learning must often be done rapidly in an online fashion, as data becomes available in a continuous stream. For example, predicting electricity demand, tracking and modelling in sensor networks, intrusion detection, and modelling the failure of hardware (predictive maintenance). There are many online algorithms that can be applied to data streams, but these typically do not take into account temporal dependence. Recurrent Neural Networks can be suitable for this application, due to their memory structure.

This internship will investigate different representations for data stream and time series data, and the development and application of recurrent neural networks in this scenario.

5.1 Desired Skills

Python. Knowledge of machine learning methods, in particular neural networks. Familiarity with data streams and/or time-series data.

5.2 Supervisors

Jesse Read

References

[1] J. Gama (2010). A survey on learning from data streams: current and future trends. Prog Artif. Intell. 45–55.

[2] R. Jzefowicz, W. Zaremba, I. Sutskever (2015). An Empirical Exploration of Recurrent Network Architectures. ICML 2015: 2342–2350.

6 Sequential and Structured Output with Multi-label Models

Multi-output (a.k.a. multi-target) learning involves building a model that can produce multiple outputs per data instance. Such models have generally been used in a multi-label context: assigning multiple labels/categories/tags to text documents, images, and so forth. However, such models can be employed much more widely, in diverse problems such as sequence and graph prediction, recommender systems, and multi-dimensional anomaly detection. Challenges involve dealing with large datasets with many output variables, complex dependency structures, and missing data (i.e., semi-supervised). with many output variables, complex dependency structures, and missing data (i.e., semi-supervised). This internship will choose one or several of such problems, and extend/develop and evaluate multi-output models suited for application.

6.1 Desired Skills

Java (familiarity with WEKA would be a plus) *or* Python (familiarity with SCIKIT-LEARN would be a plus). An understanding of machine learning, in particular knowledge of supervised learning algorithms. Knowledge of multi-label/multi-output learning is a plus.

6.2 Supervisors

Jesse Read

References

[1] K. Dembczyski, W. Waegeman, W. Cheng, E. Hllermeier (2012). On label dependence and loss minimization in multi-label classification. MLJ. 88(1), pp 5–45.

[2] M. Zhang and Z. Zhou (2014). A Review on Multi-Label Learning Algorithms. IEEE TKDE 26(8), pp 1819–1837.

7 Analysis of the Weibo Social Graph

Social network analysis (SNA), also referred to as "structural analysis", is of central interest in our era as it constitutes the strategy of investigating social structures through the use of network and graph theories. Actually, it is the mapping and measuring of relationships (flows) between people, groups and other connected entities. A social network can be typically represented by a graph, whose nodes represents the users, and the edges the interpersonal ties (relationships or interactions) among them. Since the origins of social network analysis, there has been interest in identifying the most relevant actors of a social network. Several metrics based on connections, distributions and segmenation can be used in order to give us insight into the various roles and groupings in a network – who are the connectors, mavens, leaders, bridges, isolates, where are the clusters and who is in them, who is in the core of the network, and who is on the periphery. Apart from the aforementioned metrics, it is critical for us to know which is the probability a node to influence its neighbors. More specifically, we are interesting to know how a node or a small set of seed nodes can propagate some ideas (rumours, trends, etc) to other distant in the network. The specific phenomenon is known as *spread of influence* and has long been studied in the fields of sociology, communication, marketing, etc.

In the context of this internship, we will focus on the in-depth analysis of a real dynamic social network, called as Sina Weibo. Sina Weibo, known as the "Chinese Twitter", is one of the most popular media platforms in China. By the third quarter of 2015, it has had more than 222 million subscribers and 100 million daily users. It also shares a lot of features with Twitter. For instance, a user may post with a 140-character limit, mention or talk to other people using "@UserName" formatting, add hashtags with "#HashName#" formatting, follow other users to make their posts appear in one's own timeline, re-post with "//@UserName" similar to Twitter's retweet function "RT @UserName", select posts for one's favorites list, and verify the account if the user is a celebrity.

7.1 Desired Skills

Excellent programming skills (Python). Familiarity with graph mining concepts.

7.2 Supervisors

Nikolaos Tziortziotis, Fragkiskos Malliaros, and Michalis Vazirgiannis.

References

[1] Sina Weibo revenue and statistics - App Industry Insights. `http://www.businessofapps.com/sina-weibo-revenue-and-statistics/`

[2] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*, 2010.

[3] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John Hopcroft. The Lifecycle and Cascade of Social Messaging Groups. In *WWW*, 2016.

[4] E. Otte, and R. Rousseau (2002). Social network analysis: a powerful strategy, also for the information sciences.

[5] A. Mislove, M. Marcon,K. P. Gummadi, P. Druschel, and B. Bhattacharjee (2007). Measurement and Analysis of Online Social Networks.

[6] F. Riquelme, P. Gonzlez-Cantergiani (2016). Measuring user influence on Twitter: A survey.

[7] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi (2010). Measuring User Influence in Twitter: The Million Follower Fallacy.

8 Node Embedding Algorithms for Community Detection

In principle, supervised representation learning methods and node embeddings, have mainly been used in supervised learning tasks in graphs, including node classification and link prediction. The main goal of this internship topic is to examine the capabilities of node embeddings in the task of community detection in graphs. In particular, we will empirically evaluate the performance of recently proposed node embedding methods, with respect to both functional and structural properties [5] of the extracted communities. Furthermore, we aim to compare the performance of such methods to other traditional node embeddings techniques (e.g., spectral clustering) as well as to widely used community detection algorithms.

8.1 Desired skills

Good knowledge of graph theory, linear algebra and probabilities. Familiarity with node embedding methods will be a plus. We will also use the Python programming language.

8.2 Supervisors

Fragkiskos Malliaros (UC San Diego), Apostolos N. Papadopoulos (Aristotle University of Thessaloniki) and Michalis Vazirgiannis (DaSciM).

References

[1] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online Learning of Social Representations. In *KDD*, 2014.

[2] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei. LINE: Large-scale Information Network Embedding. In *WWW*, 2015.

[3] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *KDD*, 2016.

[4] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. Learning Deep Representations for Graph Clustering. In *AAAI*, 2014.

[5] Marc Mitri, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. Sensitivity of Community Structure to Network Uncertainty. In *SDM*, 2017

9 Online Learning of Influence Probabilities on Social Graphs

Nowadays, there has been tremendous interest in the phenomenon of influence propagation in social networks. Actually, a social network plays can be seen as a medium for the spread of information, ideas, and influence among its members. For instance, a number of free samples of a product can be given to a few influential social network users (known as seed nodes), with the hope that they will influence their friends to buy it. The most of the works in this area assume that they have as input to their problems a social graph with edges labeled with probabilities of influence between users. Nevertheless, in real life the probability information may not be available in advance or to be incomplete. In this internship, we will study (empirically and theoretically) the IM problem in the absence of complete information on influence probabilities.

9.1 Desired skills

Strong background in graph theory, probabilities, statistics, multi-armed bandits, and good programming skills (C++, Python).

9.2 Supervisors

Nikolaos Tziortziotis, and Michalis Vazirgiannis.

References

[1] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the Spread of Influence through a Social Network. In SIGKDD, 2003.

[2] Siyu Lei, Silviu Maniu, Luyi Mo, Reynold Cheng, and Pierre Senellart. Online Influence Maximization. In KDD, 2015.

[3] Wei Chen, Yajun Wang, Yang Yuan, Qinshi Wang. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms. In JMLR, 2016.

[4] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning Influence Probabilities In Social Networks. In WSDM, 2010.

[5] Yixin Bao, Xiaoke Wang, Zhi Wang, Chuan Wu, Francis C.M. Lau. Online Influence Maximization in Non-Stationary Social Networks. In IEEE/ACM IWQoS, 2016.

10 Learning Node Embeddings in Signed Graphs

The goal of this internship topic is to study the problem of learning node embeddings in signed networks, i.e., low dimensional vector representations of the nodes, and their applications to learning tasks over signed social networks including the ones of link predictions and node classification. In particular, we will examine how recent graph representation learning methods (e.g., DeepWalk and node2vec) can be extended to signed graphs.

10.1 Desired skills

Good knowledge of graph theory, supervised learning, linear algebra and probabilities. Familiarity with node embedding methods will be a plus. We will also use the Python programming language.

10.2 Supervisors

Fragkiskos Malliaros (UC San Diego)

References

[1] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. A Survey of Signed Network Mining in Social Media. ACM Computing Surveys (CSUR), 49(3), 2016.

[2] Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S. Dhillon. Prediction and Clustering in Signed Networks: A Local to Global Perspective. J. Mach. Learn. Res., 15(4), 2014

[3] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online Learning of Social Representations. In *KDD*, 2014.

[4] Jerome Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jrgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization. In *SDM*, 2010.

11 Benchmarking Influence Maximization in Complex Networks

Spreading processes have been gaining great interest in the research community. This is justified mainly by the fact that they occur in a plethora of applications ranging from the spread of news and ideas to the diffusion of influence and social movements and from the outbreak of a disease to the promotion of commercial products. Of crucial importance towards understanding and being able to control such processes in complex networks, is to identify those entities that can act as influential spreaders. Identification of such nodes can ensure propagation of information to a great part of the network, optimization of the available resources or even control the spreading.

There exist various models that can simulate a diffusion process. Famous are the epidemic models, namely SIR (i.e, Susceptible Infected Recovered) and SIS (i.e, Susceptible Infected Susceptible) as well as the Rumor Dynamics model (Maki and Thomson model - MT). Those are until now used in order to locate single spreaders in networks. On the other hand, models such as the Linear Threshold (LT), Independent Cascade (IC) and Heat Diffusion Models have been used in the case were the task is to locate a group of privileged nodes that by acting all together can maximize the total spread of influence, usually called the Influence Maximization problem. For those different subgroups of problems, a plethora of algorithms and heuristics have been proposed in order to identify those entities in a graph that can provide the best possible information diffusion.

Nevertheless all those algorithms are tested on different datasets and the results are evaluated using different metrics.The goal of this internship will be to create a benchmark were all those algorithms will be implemented for a well-known set of datasets that can be found online and with common metrics in order to have a fair comparison for the performance of the algorithms in terms of efficiency and time complexity. The benchmark will be a tool for anyone wanting to find which algorithm suits their needs in order to locate superspreaders in the network in question.

11.1 Desired skills

Good knowledge of graph theory, sound mathematical background. Programming skills (Python or C++).

11.2 Supervisors

Maria Rossi and Michalis Vazirgiannis

References

[1] David Kempe, Jon Kleinberg, and va Tardos. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[2] Maksim Kitsak et al. Identification of influential spreaders in complex networks. Nature physics 6.11 (2010): 888-893.

[3] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. 2010 IEEE International Conference on Data Mining. IEEE, 2010.

[4] Hao Ma, et al. Mining social networks using heat diffusion processes for marketing candidates selection. Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

12 Analysis of a Real Spreading Process

Information and influence spread have been attracting a lot of attention due to the important role they play to numerous applications such as viral marketing, adoption of innovations and more generally spread of behavior and social norms. Various previous studies have been focused on understanding the patterns during a spreading process. Most studies have been focusing on individual-based diffusion data and on inferring the diffusion network.

In this internship the goal is to study real information spreading processes using data from Weibo (a chinese microblogging website). We will report results after analyzing temporal patterns of user-to-user influence. We will explain the observed time-varying dynamics of user activities during the spreading of a specific topic. We will finally investigate the topological characteristics of individuals that are influenced and that participate in a diffusion process and present the patterns that are detected.

12.1 Desired skills

Good knowledge of graph theory, sound mathematical backgroun. Programming skills (Python or C++).

12.2 Supervisors

Maria Rossi and Michalis Vazirgiannis

References

[1] David Kempe, Jon Kleinberg, and va Tardos. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[2] Maksim Kitsak et al. Identification of influential spreaders in complex networks. Nature physics 6.11 (2010): 888-893.

[3] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. 2010 IEEE International Conference on Data Mining. IEEE, 2010.

[4] De Domenico, M., et al. "The Anatomy of a Scientific Rumor." Scientific Reports 3 (2013): 2980.

[5] Borge-Holthoefer, Javier, et al. "The dynamics of information-driven coordination phenomena: A transfer entropy analysis." Science advances 2.4 (2016): e1501158.

13 Analysis of the World Wealth and Income Database

The World Wealth and Income Database (World WID) is one of the most known resources for research in economic data. The effort is led by the Paris School of Economics (group of Thomas Piketty). Our team completed the design and development of a Bigdata solution for the WID database enabling efficient time series data management. Currently, only simple analytics are offered on these data, such as the distribution of variables across several attributes and percentiles with location and temporal criteria. See the details here: http://wid.world/

The goal of this internship is to perform advanced data analysis on the aforementioned data by studying deeper correlations among the multiple dimensions concerning income, location, time, and other features present in the data. Also machine learning methods will be employed to learn models that will be able to predict income distribution across the globe for the future.

13.1 Desired skills

- Analytical skills
- Programming: Python or Matlab
- Good background in statistical learning
- Experience on time series analysis and Apache Spark will be highly valued

13.2 Supervisors

Nikos Tziortziotis, Michalis Vazirgiannis

References

[1] Matsubara, Yasuko, Yasushi Sakurai, and Christos Faloutsos. "Autoplait: Automatic mining of co-evolving time sequences." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

[2] Ahmed, Nesreen K., et al. "An empirical comparison of machine learning models for time series forecasting." Econometric Reviews 29.5-6 (2010): 594-621.

14 Text Classification using CNNs with Graph of Words based input

Neural networks with convolutional and pooling layers (CNNs) have been proven very effective for classification tasks. Given a large structure, CNNs identify the sets of features of the structure that are most informative for the prediction task, and generate a new representation of the structure based on these sets of features. For example, in text categorization, small sequences of words may be good indicators of the topic of a document. CNNs learn to identify such local indicators, regardless of their position in the document, and use these indicators to generate a new representation for each document. Graph-of-words (GOW) is a graph-based document representation that is used as an alternative to bag-of-words. GOW represents each textual document as a graph whose vertices correspond to unique terms of the document and whose edges represent co-occurrences between the terms within a fixed-size sliding window. The GOW representation was found to be very effective in several tasks.

The goal of this internship is to design and implement a CNN for text categorization which in contrast to existing CNNs takes as input the GOW representation of documents. The CNN should be capable of identifying areas of the graph (i.e. subgraphs) that are good indicators of the class to which the document belongs. Since there is no canonical ordering for the nodes of a graph and since graphs cannot be easily represented as fixed-size vectors, the main challenge of this internship is how to feed the input graphs to the CNN.

14.1 Desired skills

- Analytical skills
- Programming: Python
- Experience on neural network design will be highly valued

14.2 Supervisors

Antoine Tixier, Polykarpos Meladianos, Giannis Nikolentzos, Michalis Vazirgiannis

References

[1] Kim Yoon. "Convolutional Neural Networks for Sentence Classification" Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

[2] Johnson Rie and Zhang Tong. "Effective Use of Word Order for Text Categorization with Convolutional Neural Networks" Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

[3] Tixier Antoine, Skianis Konstantinos and Vazirgiannis Michalis. "GoWvis: a web application for Graph-of-Words-based text visualization and summarization", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, 2016.