2018-CS512 Course Project Presentation Schedule

All the final project reports and project presentation slides are due in the system at 11:59pm May 8, 2018.

All the presentations are 12 minutes including presentation + questions/answers

**Attendance and grading requirement:** You **required** to attend your own session and the other same half-day session and evaluate all the projects presented in these two sessions except your own project. For other sessions, you are encouraged but not required to attend. Example: If your project is A1, you are required to attend and evaluate all the projects presented in Sessions A and B except A1.

**Class Presentation Classrooms:** 3405 SC Monday Morning / 3403 SC Monday Afternoon, Tuesday

**Peer-review Form.** The review form will be open for 24hrs, from 0:00am May 10, 2018 to 11:59pm May 10, 2018. So make sure to submit your peer-review form before 11:59pm May 10, 2018.

May 7 (Monday) (9:30-10:45am) (3405 SC): Session A: 5 presentations

**Presentation A1: Group:** Paul Ruales (per2), Pranav Boopathy (boopath2), Charles Shang (cshang4)

**Title:** Disease Inference via Automatically Constructed Health Knowledge Graph

**Abstract:** The demand for clinical decision support systems and more accurate online sources of health information has substantially increased in recent years (1). Heterogenous information networks contain rich information in interconnected multi-typed environments similar to healthcare. The electronic medical record (EMR) contains patient-specific health information that can be used to construct highly accurate knowledge graphs. Previous work on constructing health knowledge graphs have either involved intensive manual labor or rudimentary concept extraction techniques (1, 2, 3). This study focuses on the automatic construction of a health knowledge graph representing diseases and symptoms from an EMR. Unstructured data is mined using AutoPhrase and SetExpan to extract statistically relevant concepts and statistical models are used to estimate the edge strengths between symptoms and diseases. On the task of disease inference, this study shows that even with a very small dataset our knowledge graph performs comparably with the one built by Rotemensch et al from ER-specific data (1). Additionally, we show that combining the two knowledge graphs significantly boosts disease inference performance on ICU specific data, showing that information can be shared between hospitals without the drawback of PHI exposure.

**Link to slides:** [https://docs.google.com/presentation/d/1WLaO3Kf9jf_FINE6bCWWVlapIGxQSUxel-541BgQtQ/edit?usp=sharing](https://docs.google.com/presentation/d/1WLaO3Kf9jf_FINE6bCWWVlapIGxQSUxel-541BgQtQ/edit?usp=sharing)

**Link to final project report:**

pdf (in case link above fails)

**Presentation A2 Group:** Xikun Zhang (xikunz2), Shibi He (shibihe)

**Title:** Similarity Modeling on Heterogeneous Networks via Automatic Path Discovery

**Abstract:** Heterogeneous networks are widely used to model real-world semi-structured data. The key challenge of learning over such networks is the modeling of node similarity under both network structures and contents. To deal with network structures, existing works all assume a given or enumerable set of meta-paths and then leverage them for the computation of path-based proximities or embeddings. However, expert knowledge for given meta-paths is not always available, and as the length of considered meta-paths increases, the number of possible paths grows exponentially, which makes the path enumeration process very hard. On the other hand, while there are often rich contents around network nodes, they have never been leveraged to further improve the modeling of node similarity. In this work, to properly model node similarity in heterogeneous networks, we propose to automatically discover useful paths for pairs of nodes under both structure and content information in the networks. To this end, we combine reinforcement learning and deep embedding into a novel semi-supervised joint learning framework. Specifically, the unsupervised deep embedding component captures the contents of heterogeneous nodes, while the supervised reinforcement learning component explores the structures of network paths. Furthermore, the two components are jointly trained in a closed loop to mutually enhance each other. Extensive experiments on three real-world heterogeneous networks demonstrate the supreme advantages of our algorithm.

**Link to slides:** autopath presentation.pptx

**Link to final project report:**

**Presentation A3: Group:** aming, ni (amingni2)

**Title:** Tweet Sentiment Analysis Using Syntactic Tweets and Ensemble Neural Nets

**Abstract:** Microblogging like Twitter has become a very popular platform for Internet users, and users often generate millions of tweets a day to express their opinions about certain things and share their life experience about the world. If sentiment analysis techniques can be applied on tweets effectively with only small amount of training data, we can gain very valuable information for marketing purposes. For example, companies may want to monitor consumers’ opinions about their products on Twitter platform. Therefore, the ability to classify short tweets into sentiment categories is becoming more important for business. In this project, we have applied sentiment analysis on Twitter data. In addition, we have improved traditional sentiment analysis methods by using syntactic information and ensemble learning. First of all, we have extracted syntactic features from tweets, which have been shown to be useful in sentiment analysis. Then, we have combined these features with traditional methods of feature extraction using machine learning techniques. Finally, we have applied ensemble learning to improve the accuracy of our model. This project has shown that combining traditional methods with syntactic information can improve the performance of sentiment analysis on Twitter data.
scales with limited data is very valuable. In this paper, we propose to (1) augment small training tweets with Syntactic Examples generated by Markov Chain Model and Word Lists, (2) randomly sample sub-embeddings from full trained word embeddings using Ensemble Neural Nets (BiLSTM + CNN).

Link to slides: [http://web.engr.illinois.edu/~amingni2/cs512/slide.pdf](http://web.engr.illinois.edu/~amingni2/cs512/slide.pdf)

Link to final project report:

**Presentation A4:** Group: Xinwei He (xhe17), Naijing Zhang (nzhang31)

**Title:** Higher-Order Clustering in Heterogeneous Information Networks via Joint Non-Negative Tensor Factorization

**Abstract:** Network motifs are crucial building blocks of understanding and modeling complex networks for its capacity in characterizing higher-order interactions. Meanwhile, heterogeneous information networks (HINs) are ubiquitous in real-world applications. We are hence planning to combine the two notions and propose a HiN-Motif clustering model by Tensor factorization techniques.

Link to slides: [https://drive.google.com/file/d/1LbVP5ZXwJbh2XWldHbF5N3_7McaHE5qB/view?usp=sharing](https://drive.google.com/file/d/1LbVP5ZXwJbh2XWldHbF5N3_7McaHE5qB/view?usp=sharing)

Link to final project report:

**Presentation A5:** Group: Chiawei Chen (chiawei2), Daniel You (tjyou2)

**Title:** Enriching Multi-label Entity Typing with Subword Information

**Abstract:** This paper presents a new approach to tackle large scaled multi-label entity typing with extra subword information. We focus on PubMed abstract dataset, and observe that most words in bio-medical area consist of multiple meaningful subwords. For instance, “Antipsychotic” can be decomposed into two subwords “anti” and “psychotic”, and we can reconstruct the meaning of the original words from those subwords. Therefore, if there's an unseen mention, we can propagate information through subwords and get the understanding of this unseen mention. Plus, bio-medical words share lots of subwords with each other; thus, we can utilize this advantage to connect label prediction with subwords. Furthermore, due to sparsity of context, we add an attention mechanism to help our model focus on important words within a sentence, and find that using attention and subword features can boost each other which lead to better performance.

Link to slides: [https://docs.google.com/presentation/d/1O876T-9SYYDklHJjDSr2Pb8zugp3pjhuZ973v8B7Xavg/edit#slide=id.g39bc4f9e4d_3_11](https://docs.google.com/presentation/d/1O876T-9SYYDklHJjDSr2Pb8zugp3pjhuZ973v8B7Xavg/edit#slide=id.g39bc4f9e4d_3_11)

Link to final project report:

**May 7 (Monday) (11:00-12:15pm) (3405 SC) Session B: 5 presentations**

**Presentation B1:** Group: Zubin Pahuja (zphauja2), Yuchen He (he34), Xiaocheng Ou (xo4), Zhiyuan Zheng (zheng55)

**Title:** Crowd-Sourcing for Universal Web Data Extraction

**Abstract:** Our project attempts to do automatic label assignment for web extracted data and improve the labeling accuracy with the user credential guided True labeling algorithm. We propose a pipelined architecture that would first get the semi-structured data on any web page using crowd-sourced wrappers through a Chrome Extension. Then in the second step, perform the task of automatic label assignment to the extracted data from the web page. This is a supervised Multi-class classification task with external knowledge-base. Finally using the crowd-sourced corpus to find credible users and assign label provided by trustworthy users. The credibility of a user is based on his label accuracy and label type weight. Most prior work tackles the problem of domain-specific data extraction, and no previous studies have been conducted on human-in-the-loop data extraction with the use of crowd-sourcing and truth finding. Our result showed a great improvement in web data label with the use of crowd-sourcing and truth finding.


Link to final project report:

**Presentation B2:** Group: Sarah Christensen (sac2), Kent Quanrud (quanrud2)

**Title:** Combinatorial Approaches to Phrase Mining

**Abstract:** We revisit phrase mining from a combinatorial point of view. We first develop a discrete model, which (informally speaking) interprets phrase mining as selecting phrases whose induced bag-of-phrase representation of documents maximizes the total entropy subject to an upper bound on the total descriptive complexity of the bag-of-phrase summaries. We observe that the implicit optimization problem can be viewed as an instance of submodular maximization subject to a knapsack constraint, for which the greedy algorithm has provable guarantees on the
Link to slides: cq-combinatorial-phrase-mining-slides.pdf

Link to final project report:

Presentation B3: Group: Luning Wang (luningw2), Liuyi Shi (liuyis2)
Title: Movie box-office prediction with social media mining

Abstract: Social media provides the snapshot of the worlds status with up-to-minute information. Movies are an important aspect of social media. Box office prediction has long been the problem of movie market. With accurate prediction the cinema can schedule the arrangement of movies and movie makers can get prepared about the domestic market. In this paper we give a brief idea about how to use Twitter data to predict movie market. We use sentimental analysis to find peoples opinion on a certain movie and use the count of tweets and the ratio of positive and negative opinions as two features. Then we train our classifier on these features to predict if a certain movie will be a hit, loss or a neutral one. This method can be used for cinemas to rearrange the schedule of movies, for audience to change the movies they go to see if they do not have a specific preference and for movie producer to have a better understanding of the market.

Link to slides: https://drive.google.com/file/d/1GnsXA6SUdG5UwhupwzKX1V-ALvp9GoBj/view?usp=sharing

Link to final project report:

Presentation B4: Group: Yunyi Zhang (yzhan238), Xinwei He (xhe17), Luyu Gao (luyugao2)
Title: Taxonomy Embedding in Poincaré Space: A General Exploration

Abstract: Graph embedding is an important branch in Data Mining and Machine Learning, and most of recent studies are focused on preserving the hierarchical structure with less dimensions. One of such models, called Poincaré Embedding, achieves the goal by using Poincaré Ball model to embed hierarchical structure in hyperbolic space instead of traditionally used Euclidean space. However, Poincaré Embedding suffers from two major problems: (1) performance drops as depth of nodes increases since nodes tend to lay at the boundary; (2) the embedding model is constrained with pre-constructed structures and cannot be easily extended. In this paper, we first raise several techniques to overcome the problem of low performance for deep nodes, such as using partial structure, adding regularization, and exploring sibling relations in the structure. Then we also extend the Poincaré Embedding model by extracting information from text corpus and propose a joint embedding model with Poincaré Embedding and Word2vec.

Link to slides: https://docs.google.com/presentation/d/1Tnjw557cM-ul62htetKDwGqJAITUEk8PTdL4hh1EY/edit?usp=sharing

Link to final project report:

Presentation B5: Group: Harshit Agarwal (hagarwa3), Zach Kimberg (kimberg2)
Title: Transfer Learning from Alternative Mechanisms to Optimize Differentially Private GAN Release

Abstract: In order to gain the maximum benefits from the large amounts of data generated by modern software, it is important that the data be shared with researchers and those best able to gain and use the insights it can generate. But, this runs into the problems of privacy. Even with names removed, it is quite possible, and increasingly likely, that the data can be matched with other public data sets and de-anonymized. When the data set contains sensitive data, such as medical, this can be problematic. Differential privacy is the most common technique used now to prevent this de-anonymization and help people feel secure in sharing their information. And, GANs represent an incredibly powerful differentially private mechanism that can produce an infinitely sized synthetic dataset instead of traditionally used Euclidean space. However, GANs can be difficult to train and don't achieve good privacy bounds without additional public data to train on. We present a new mechanism to leverage the power of alternative differential privacy mechanisms in accordance with the GAN training to relieve the necessity for public data. It can be used to create a jumpstarting data set for initial GAN training and then can be refined using the original data set. Additionally, the jumpstart dataset can be used to further optimize the training process. Using this, it is possible to better utilize the privacy budget and produce a better quality output GAN.

Link to slides: https://docs.google.com/presentation/d/1a1ashHDzQOGar3bWu-TB6cdq2XV7-BCq9PbOOk5LyljM/edit?usp=sharing

Link to final project report

May 7 (Monday) (2:00-3:15pm) (3403 SC): Session C: 5 presentations

Presentation C1: Group: Xuan Wang (xwang174), Yu Zhang (yuz9), Yinyin Chen (ychen409), Liyuan Liu (ll2)
Title: Open Information Extraction in Biomedical Literature
Abstract: Biomedical open information extraction (OpenIE) is an important task to extract rich structured information from the biomedical literature. Previous work on OpenIE in general domain are restricted to subject-verb-object relationships. Moreover, they ignore the entity type information, which lead to noisy and redundant output. Recent work on meta-pattern extraction produce structured n-ary relationships, but cannot deal with long and complicated sentence structures. We propose a framework that first resolve the long and complicated sentence structures and then use meta-patterns to extract the n-ary tuples with typed entities and their relationships. Our method achieves the highest precision in comparison with the state-of-the-art baselines, and keeps the distinctiveness and simplicity of extracted tuples. This work shows great promise in automatically extracting rich structured information from the biomedical literature with high quality.

Link to slides: cs512_project_presentation

Link to final project report:

Presentation C2 Group: Yuntae(Chris) Kim (ykim157), Ramya Narayanaswamy (rpn2), Dariusz Kuc (dkuc2)

Title: Data-driven Supervised Morpheme Extraction

Abstract: With the ever-increasing availability of large unstructured data and the need to extract useful information, text segmentation is one of the predominant research topics in data mining. Some of the key applications of text segmentation are natural language processing, topic modeling, machine translation etc. Various forms of text segmentation exist, and popular segmentation techniques are focused on character sequences, words, phrases, and sentences. However, text Segmentation to extract morphemes, which are smallest meaningful unit of a language has proven to improve effectiveness of downstream data mining tasks. Most of the previous research in morpheme extraction is unsupervised with best accuracy of ~20%. We model morpheme extraction as a classification task with data-driven statistical features, resulting in accuracy of ~65%.

Link to slides: Data-Driven Supervised Morpheme Extraction

Link to final project report:

Presentation C3 Group: Yichen Feng (feng36), Shuchen Song (ssong18)

Title: What Meta-Graphs to Use and How

Abstract: Heterogeneous information networks (HIN) have drawn significant research attention recently, due to its power of modeling multi-typed multi-relational data and facilitating various downstream applications. All algorithms on HIN leverages meta-graphs or meta-paths (a special case of meta-graphs). However, they usually assume a given set of meaningful meta-graphs, and none of them attempts to study the general quality and utility of different meta-graphs. Moreover, as embedding has shown to be a fundamental solution for various network mining tasks, no existing work has studied how to simultaneously leverage multiple meta-graphs for HIN embedding in an unsupervised manner. In our work, by generalizing prolific studies on homogeneous networks, especially spectral clustering, we fitly aim to provide a theoretically sound and empirically useful metric for the quality of meta-graphs on any arbitrary HIN. It helps to understand the high-order organizations of HIN and obtain the candidate set of meaningful meta-graphs. Finally, theoretical guarantees on the utility of the combination of multiple meta-graphs w.r.t. the utilities of individual meta-graphs are derived to close the gap between our meta-graph selection and combination modules. To the best of our knowledge, this is the first research effort to provide theoretical analysis on the utility of meta-graphs and their combinations, especially regarding HIN embedding.

Link to slides: Presentation Slides

Link to final project report:

Presentation C4: Group: Yuning Mao (yuningm2), Jiaming Shen (js2), Xiaotao Gu (xiaotao2), Sha li (shal2)

Title: Meta-Path Guided Embedding for Heterogeneous Information Network Consolidation

Abstract: Most real-world data can be modeled as heterogeneous information networks (HINs) consisting of vertices of multiple types and their relationships. However, current HINs can only be constructed from structured meta-data while much richer unstructured data is not leveraged. To tackle this issue, we propose a new task, heterogeneous information networks consolidation, which aims to enhance existing HINs with attributes extracted from unstructured data. Our proposed framework models vertices as low-dimensional vectors to explore similarities embedded in the network structure. To accommodate user preferences at defining similarity semantics, it accepts user-defined meta-paths as guidance to learn vertex vectors in a user-preferred embedding space. Moreover, an efficient and parallel sampling-based optimization algorithm is developed to learn embeddings in large-scale HINs. After the embeddings of nodes in the HINs are learned, a parallel learning to match algorithm is utilized to consolidate the existing HINs with seed attributes discovered from unstructured text data.

Link to slides: 512 project presentation.pdf

Link to final project report:
Presentation C5: Group: Yu Lu (yulu3), Shidi Zhao (szhao41), Xingyu Xiang(xxiang6)

Title: Wicked Fast Latent Keyphrase Inference

Abstract: For document representation, latent keyphrase inference (LAKI) is developed to incorporate latent document keyphrases into document representation. The idea of representing document by its domain keyphrase brings LAKI very high accuracy in document classification. But however, since LAKI inferred the dependency between domain keyphrases and content units via a complex Bayesian network. It slows down the efficiency of LAKI in both training and inference process. To solve this problem, we propose WAKI (Wicked fast latent keyphrase inference) to do document keyphrase inference. In WAKI, instead of utilizing Bayesian network, we use a weight matrix to convert the bag of phrases representation into document keyphrase representation. Information on quality of phrase, semantic similarity and prior frequency of phrase are all incorporated into the weight matrix. The introduce of matrix operations improved the efficiency of LAKI by an order of magnitude in both training and inference phases while achieving higher accuracy on document classification. Moreover, since the weight matrix is sparse, the online inference time increase very slow as the number of total phrases grows. All of these make WAKI an efficient, scalable and effective architecture compared to LAKI.

Link to slides: WAKI

Link to final project report:

May 8 (Tuesday) (9:30-10:45am) (3403 SC): Session D: 5 presentations

Presentation D1: Group: Zecheng Zhang (zzhan147), Yuncheng Wu (ywu101), Zesheng Wang (zwang180)

Title: Open domain event extraction

Abstract: The spread of mobile devices and access to the Internet make text data ubiquitous and the amount of text data still increases rapidly. Many approaches have been developed to get useful and human understandable knowledge from large amount of text data. One of the most useful one is the open domain event extraction. Good quality of open domain event extraction can significantly better systems such as information retrieval system and information extraction system; with good event extraction they can mine higher quality of knowledge from large corpus or large amount of text data. The main task of open domain event extraction is to get corresponding event type, identify trigger words and roles of arguments given a raw sentence. Our method follows a pipeline which combines some current natural language processing techniques and provide good quality of trigger words clustering, event type and role of argument prediction. Our analysis shows that our method possibly is a promising approach on open domain event extraction.

Link to slides: https://docs.google.com/presentation/d/1R2OmBhAcRYYTmZmGQUNPbjsfIBOFS1tCQcdHy6iELF4/edit?usp=sharing

Link to final project report:

Presentation D2 Group: Collin Gress (gress2), Xiyao Shi (xshi27)

Title: Approximate Multi-Mapper: an Algorithm for Strided Approximate Nucleotide Sequence Mapping Under Edit Distance

Abstract: Mapping random read nucleotide sequences to reference sequences, which involves finding positions in the reference where the random reads likely came from, is of great importance to clinicians wishing to gain knowledge from DNA samples. State of the art DNA sequencing machines are fast but are very error prone, often causing error rates in their reads of up to 15%. This makes mapping these reads to reference genomes difficult since the mapping algorithm must account for the errors. Algorithms often model the errors using the edit distance metric, but, because calculating the edit distance between two sequences is quadratic in the sequence length, the algorithms must find efficient ways to avoid these expensive calculations. In our mapping algorithm, we present a method of determining mapping positions based on the distance of two sequences' CGK-embeddings in the Hamming space. Our algorithm employs several other heuristics in order to prune unnecessary calculations and compete with other state of the art mapping algorithms in terms of precision, recall, and running time.

Link to slides: https://www.dropbox.com/s/dcjx48a6naedkru/Approximate%20Multi-Mapper%20%29%28%20%29%20%20%29%20%20%29.pptx?dl=0

Link to final project report:

Presentation D3: Group: Yu Meng (yumeng5)

Title: Weakly-supervised Text Categorization from Different Sources

Abstract: Supervised and semi-supervised text categorization have been well studied in data mining and NLP applications. Many conventional approaches rely on large amount of labeled documents to train a good classifier that is used later to categorize unlabeled data. In real-world scenarios, however, labeling a large number of documents is time consuming and sometimes even impossible. Therefore, it is often desirable to use methods that leverage weak supervision sources instead of massive labeled documents. In this paper, we propose a framework for text categorization that can leverage different kinds of weak supervision sources from users, including class surface names, class keywords or few labeled documents. Our experiments on several real-world datasets show that our framework outperforms several state-of-the-art weakly-supervised and semi-supervised algorithms, and that although weak supervision sources can be of different forms, if leveraged properly, they can all contribute to accurate text categorization.
**Presentation D4:** Group: Guangyuan Wang (gwang10), Zongyi Wang (zhang195)

**Title:** Learning Content-rich Diffusion Network Embedding

**Abstract:** Information networks are ubiquitous in the real world, while embedding, as a kind of network representation, has received attention from many researchers because of its effectiveness in preserving the semantics of the network and its broad application including classification, link prediction etc. Previously, many methods have been proposed to learn network embedding from non-attributed and static networks. Networks are treated simply as nodes and links. However, information is not merely encoded in the structure, nodes itself may have their intrinsic attributes and little research have been done to incorporate this information. Furthermore, outside information will also diffuse on the network over the time and for a specific time we can have a different diffusion structure. In this paper, we will be introducing the idea of diffusion network, and we propose two models for embedding learning that aim to efficiently capture the rich content of nodes as well as that of the diffusion. We then evaluate the quality of our embedding by conducting node classification experiments, the result of which shows that our method for generating embedding outperforms other baselines.

**Link to slides:** Learning_Content-rich_Diffusion_Network_Embedding.pptx

**Link to final project report:**

**Presentation D5:** Group: Bing Liang (bingl3), Jieyu Zhang (jieyu2), Ruiyang Wang (ruiyang7)

**Title:** Semi-supervised Heterogeneous Network Consolidation via Neural Label Propagation

**Abstract:** We propose a task, heterogeneous information networks consolidation. Adding data to missing attributes based on exist attributes. Previously, people use pre-set meta-path approach to learn the network embedding to implement missing data. It's hard to transit to new dataset, and need to set meta-path carefully. Our approach abandon meta-path. We use Neural Module Networks (NMN) to learn the logic behind missing data. Our model archives a remarkable results.

**Link to slides:** HNCvNLP

**Link to final project report:**

---

**May 8 (Tuesday) (11:00-12:15pm) (3403 SC) Session E: 5 presentations**

**Presentation E1:** Group: Cheng Ding (chengd2), Yicong Yang (yyang153), Xiaolu Qi (xiaoluq2)

**Title:** On-The-Fly Knowledge Base Construction With Coreference Resolution

**Abstract:** With the spread of the Internet, there is more and more information on the Internet, people can find the useful information easily through search engine. However, there are some deficiencies in search engine: Because of too many return results, it's difficult for users to get the exact information they need, and the fundamental of search engine focus on the grammar of key words rather than the semantic. It's difficult for users to accurately express their requirement with simple query words. Besides, even the largest knowledge bases are still limited in up-to-date coverage of what happens in the real world and miss out on many relevant predicates that precisely capture the wide variety of relationships among entities. To overcome these limitations, we try to extract entities and their relations from text from massive text corpora and build on-the-fly knowledge bases. We use ReMine as our open IE tool. To improve the precision of ReMine result and for the future application in QA system, we combine with the coreference methods.

**Link to slides:** https://drive.google.com/file/d/18dVlW2KzumMVPrCYViLadLO2ehMpQrDM/view?usp=sharing

**Link to final project report:**

**Presentation E2** Group: Yundi Fei (yfei5), Yimeng Zhou (yzhou97)

**Title:** Weighted PathSim for Measuring Movie and Actor Similarity

**Abstract:** Similarity search between actors and movies is a basic operation for recommendation and clustering purpose. The primary similarity can be the similarity between their metadata. Another approach is to construct a heterogeneous network of movie and cast node, and use PathSim to calculate similarity between the same type nodes. This work extends the heterogeneous network of movie and actor with a new type of node of topic. Each topic node represents a topic mined from all movie overview data using TopMine, and is connected to movies with such topic. The edges between a topic and a movie be weighted to represent the probability of a movie belonging to a topic. Therefore, we modify PathSim to consider not only the count of meta-paths, but also the weight of the meta-paths. The modified PathSim can calculate similarity
between movies using Movie-Topic-Movie meta-path, and can calculate similarity between actors using Actor-Movie-Topic-Movie-Actor meta-
path. The similarity measures can be used to find top-k similar results just like in PathSim. Our method gives a new definition of actor and movie
similarity which can be used for further extension and application.

Link to slides: https://docs.google.com/presentation/d/1YICnR_urMiUdg5wu8uJVvakkpQmkK1LR-a0K3inVEQI/view?usp=sharing

Link to final project report:

Presentation E3: Group: Yiran Zhao (zhao97), Shengzhong Liu (sl29)

Title: Dependency and Pattern Mining in Github Sequential Events.

Abstract: Large volumes of event data are becoming increasingly available in a wide range of applications, such as healthcare analytics, smart
cities, social sensing and crowdsourcing, as well as recommendation system construction. These events are usually causally dependent on each
other, and present some frequent event patterns in the time series sequence, with respect to who perform the action and what they do. The
dependencies and patterns usually reveals the user habits and interactions among different parties. Analyzing the trends, hidden rules,
interactions and patterns behind social event data is both interesting and challenging. It possesses significant meaning in user behavior modeling
and online marketing strategy design. This paper presents the event models and sequential patterns contained in the Github dataset, which
contains millions of users, repositories and even more events generated by users. We analyze the interaction of events that are close in time by
modeling them with point process models. Time separated interactions presented as sequential patterns are also discovered.

Link to slides: https://wiki.illinois.edu/wiki/download/attachments/420151317/present512-zyr-lsz.pptx

Link to final project report:

Presentation E4: Group: Zhengzhi Lou (zlou4), Yuchen Li (li215)

Title: Counting Temporal Motifs in Heterogeneous Info Network

Abstract: Network motifs are crucial building blocks of understanding and modeling complex networks for its capacity in characterizing higher-
order interactions. Meanwhile, heterogeneous information networks (HINs) are ubiquitous in real-world applications, which often come with rich
temporal information. We are hence motivated to study temporal motifs in heterogeneous information networks. With examples from real world
datasets, we demonstrate HIN motifs can be armed with substantially more expressive power by incorporating temporal information.
Furthermore, counting temporal HIN motif instances in large-scale networks is time consuming, and we develop efficient counting algorithms for
the HIN motifs that are of the most interests in the literature.

Link to slides: https://docs.google.com/presentation/d/1zUrUaB8xz1AoNCHE63DXuuff2vAWX0kg2Gb8t1x9e18/edit?usp=sharing

Link to final project report:

Presentation E5: Group: Qianyang Peng (qp3), Renxuan Wang (renxuan2)

Title: AutophrasePy: an Integrated Python Package for AutoPhrase

Abstract: AutoPhrase is a phrase mining tool which has been widely adopted in many mainstream applications. However, it has several
limitations, such as not being user configurable and requiring independent execution. AutophrasePy is a programming project aiming at improving
AutoPhrase. In this project, we mainly achieved two goals. First, we extended the language support of AutoPhrase, adding the functionality
of Japanese and Arabic text training and segmentation. Second, we developed a Python package integrating all functions of AutoPhrase, which
supports pip installation, automatic wiki-files downloading and customized model selection. We provided user-friendly APIs, which allows users to
do phrase mining under any directory they like with just two simple Python functions.

Link to slides: https://docs.google.com/presentation/d/1i5kVqYzut0IBxAr0xMRHkaFLz8-HNH1Ox7M8eMjx9n8/edit?usp=sharing

Link to final project report:

May 8 (Tuesday) (2:00-3:15pm) (3403 SC): Session F: 5 presentations

Presentation F1: Group: Nestor Bermudez (nab6) and Gaurush Hiranandani (gaurush2)

Title: ACA: Assistant for Causal Analysis

Abstract: Analysts working in large-scale financial companies continuously observe stock trading of companies established in various sectors.
The trade volume of stocks is not only affected by news specific to the company but also by many socio-political events. In this project, we create
an assistant to such analysts, which first detects anomalies in the trade volume and then mines causes for those anomalies from the news
articles. We form a principled approach, which leverages concepts from anomaly detection, causal analysis, and text mining, in order to construct
ACA -- Assistant for Causal Analysis.

Link to slides: PDF
Presentation F2: Group: Kejia Jiang (kjiang9), Yining Wang (ywang308), Xinyan Zhou (xzhou14)

Title: Open Domain Event Extraction and Schema Induction

Abstract: Event extraction aims at identifying and typing trigger words and participants (arguments). However, in open domain, pre-defined event types and event schemas suffer from low coverage. Recent progress in Reading Comprehension and Question Answering shows a promising solution to automated information extraction from massive corpus. It would be interesting to mine valuable domain-specific event types and schemas on target domain. In this project, we will be investigating various deep learning techniques on task of event extraction and transfer knowledge from existing domain to target domain accordingly.

Link to slides: https://docs.google.com/presentation/d/1TrD8uYKnhuEQMxrhZDTx2omMJ7j6NiiW7Q7wEITYLQ/edit#slide=id.g39c9ac9d67_0_67

Link to final project report:

Presentation F3: Group: Yihui Cui (ycui11), Yisi Liu (yisil2), Haocheng Zhang (hzhang91)

Title: Yelp Sentimental Analysis with Multi-vew Network Embedding

Abstract: This project is aiming at sentiment analysis on Yelp dataset. In order to quantify the degree of the attitude, we set the star score as an indirect standard to reflect the reviewer’s attitude. Unlike previous sentiment analysis models which using text analysis or machine learning classification, we propose a new model using the meta-based-method. The meta-path-based methods are commonly used in network mining. There already have some existing mature methods, for example, metapath2vec for scalable network embedding learning in heterogeneous network. In our project, we simplify the multi-view heterogeneous network mining. The method is to extract several meta-path to form simple homogeneous network which can be treated as subgraphs and different views of the overall heterogeneous network. After extracting the network embeddings from those homogeneous networks, we simply concatenate these paths together to learn the network representations for later prediction use.

Link to slides: https://docs.google.com/presentation/d/1TUlr0GRSUQPeRjW5fG536pqcOxnJBF-NFPulhusiFbo/edit?usp=sharing

Link to final project report:

Presentation F4: Group: Saar Kuzi (skuzi2), Janina Sarol (mjsarol)

Title: Identifying Elliptical Coordinated Compound Noun Phrases using Word Embeddings

Abstract: Elliptical coordinated compound phrases (CCNP), where a word or a set of words are omitted from a compound noun phrase, are often used in technical writing in order to save space. For example, the phrase “efficient systems and efficient algorithms” can be reduced to “efficient systems and algorithms”. The meaning of the original phrase is preserved and can be easily understood by humans. However, machines may have problems interpreting such phrases. In this work we focus on the task of CCNP disambiguation. That is, given a phrase we need to decide whether it is a CCNP or not. We propose a novel supervised approach for that purpose. We use word embedding-based features to capture semantic similarity of the different phrase components. An empirical evaluation demonstrates the effectiveness of our approach as compared to several baselines.

Link to slides: ccnp.pdf

Link to final project report:

Presentation F5: Group: Xiwen Ma (xiwenma2), Junting Wang (junting3)

Title: PathSim and PTE based Top-K Similarity Measure

Abstract: Due to the limitation of PathSim that it can only select single meta path to conduct similarity measure, we integrate PathSim and PTE to leverage a new method for more meaningful similarity measure. Our interest is to find authors with both similar paper quality and paper content. By using PathSim with select meta path APCPA, we can ensure the quality measure, and by using PTE to embed paper abstract and then cluster, we can ensure the content measure. Our final similarity measure is a combination of two results from PathSim and PTE. We also conduct experiment to evaluate different cluster number in PTE and different coefficient in the final similarity measure formula.

Link to slides: https://docs.google.com/presentation/d/1tdR8tMp-8LESogbetAW4ojd7sYqfOe2H4MYF7F1bfJg/edit#slide=id.g39c8d548c1_0_683

Link to final project report:

Presentation F6: Group: Wei Chen Lin (wclin2), Xi Li (xli2), Yekai Yu (yekaiyu2)
**Title:** Text summarization based on quality phrase

**Abstract:** Automatic summarization is designed to catch the main idea of a given textual document. With massive corpus, like news, generated every day, it's necessary to have text summarization to help people mine the information. To quickly and precisely perform text summarization on unknown domain corpus, in this paper, an adopted text ranking algorithm is proposed. The algorithm employs the quality phrases and some sentence-level textual features to improve the quality of the summarization results. The summary is evaluated by using ROUGE to quantify the similarity between selected sentences and article title.

**Link to slides:** Text Summarization - Slides

**Link to final project report:**

---

**Presentation F7: Group: Heting Gao(hgao17), Yunan Zhang(yunanz2)**

**Title:** Macross: Activity Mining based on Metapath Guided Cross-Modal Embedding

**Abstract:** To mine activity of spatial-temporal tweet data for a user specified query, a metapath guided embedding approach is explored in this project. Given a dataset of tweets, we extract and aggregate location and time and construct a heterogeneous information network using the aggregated space and time, as well as stop-word filtered frequent words. Metapath2vec is used to construct vector representations for times, locations and frequent words such that co-occurrence pairs of nodes are closer in latent space. The vector representations will be used to infer related time, locations or keywords for a user query.

**Link to slides:** Metapath Based Activity Mining

**Link to final project report:**

---

**Online Students:**

**Presentation 1: Group: Marvin Biscocho (biscoch2), Lawrence Wadle (lwadl2)**

**Title:** Cyberbullying Tweets in Events

**Abstract:**

While interesting opportunities are available in the world of big data, there also exists many challenges, such as cyberbullying. With mobile devices and internet anonymity, users can hide behind a screen and torment others without regrets. A common source of this unhealthy environment can be found on social networking sites, such as Twitter, Facebook, Reddit, etc. Studies have linked toxic environments to depression. While many works have studied identifying cyberbullying, we study the correlation of cyberbullying tweets about the surrounding news topics, which stir up controversy and may provide motives for cyberbullying. For cyberbullying instances, Twitter was chosen as many large datasets are available. For current event articles, NYTimes has an API available for data collection. The period studied is 06/2009 and 8/2009. Repetitive tweets containing negative sentiment involving multiple parties are studied and linked to news articles via AutoPhrase and Doc2Vec. The techniques applied can aid in detecting cyberbullying instances in social networking sites.

**Link to slides:** https://www.dropbox.com/s/foddgr6q5bjsfp1/Topic%20Modelling%20Approach%20on%20Cyberbullying%20Instances%20with%20Relevant.pptx.pdf?dl=0

**Link to final project report:**

---

**Presentation 2: Group: Jian Du (jiandu2)**

**Title:** Crowd Sentiment And Stock Performance

**Abstract:** In this paper, I use more than 11,000 comment articles of the S&P 500 constituents stocks downloaded from a semi-professional stock commentary website Seeking Alpha (www.seekingalpha.com) to construct sentiment measures and study how sentiment responds to stock past performance and if crowd-sentiment carries any predictive power of future cross-sectional and market-level stock returns. Two approaches to measure sentiment are used and compared: (1) the first approach is based on simple negative and positive word frequency method; (2) the second approach utilizes phrase mining and word embedding methodology. The results show that the sentiment in the articles reflect the past company/stock performance, and based on simple OLS regression setup, the sentiment measure does not seem to contain extra predictive information about future stock performance.

**Link to slides:** https://www.dropbox.com/s/hvs6hp6668tvn1g/jiandu2_CrowdSentiment%26StockPerformance_slides.pdf?dl=0

**Link to final project report:**